involve

a journal of mathematics

msp

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

On the minuscule representation of type B_n

William J. Cook and Noah A. Hughes

(Communicated by Ravi Vakil)

We study the action of the Weyl group of type B_n acting as permutations on the set of weights of the minuscule representation of type B_n (also known as the spin representation). Motivated by a previous work, we seek to determine when cycle structures alone reveal the irreducibility of these minuscule representations. After deriving formulas for the simple reflections viewed as permutations, we perform a series of computer-aided calculations in GAP. We are then able to establish that, for certain ranks, the irreducibility of the minuscule representation cannot be detected by cycle structures alone.

1. Introduction

The original motivation for this project was to extend results found in [Cook et al. 2005]. In that paper the authors present a constructive method for solving the inverse problem in differential Galois theory. This problem seeks to determine if certain groups can appear as differential Galois groups of systems of linear differential equations and, if so, given that group, determine such a system of equations.

In [Cook et al. 2005] the authors present a construction which relies on the existence of minuscule modules whose irreducibility can be detected by examining the cycle structures of the corresponding Weyl group viewed as permutations of weights. While each simple Lie algebra has infinitely many isomorphism classes of finite-dimensional irreducible representations, not every simple Lie algebra possesses a minuscule representation. Those which do, have only a handful.

Minuscule representations have the interesting property that all of their weights lie in a single Weyl group orbit. This then implies that all of the weight spaces are 1-dimensional. The irreducibility of such a module is guaranteed by the transitive action of the Weyl group. We set out to find when this transitivity (and thus irreducibility) can be seen from the cycle structures of the Weyl group elements (viewed as permutations) alone.

The authors in [Cook et al. 2005] were able to show that each algebra of type A_n ($n \geq 1$), C_n ($n \geq 3$), D_n ($n \geq 4$), E_6, or E_7 possesses a minuscule representation having the desired property. Since E_8, F_4, and G_2 have no minuscule representations at all, these cases must be discarded. This leaves type B_n as the final case to be considered. Using calculations performed in Maple (a computer algebra system), the authors were able to show that B_2, B_3, B_5, and B_7 have a conforming minuscule representation. They also showed that B_4's irreducibility cannot be seen from cycle structures alone. The status of the other type-B_n cases were left open.

In this paper, we focus on simple Lie algebras of type B_n. Such algebras have only one minuscule representation which is also known as the spin representation. After some introductory material, we explicitly determine the action of the Weyl group of type B_n on the weights of its minuscule representation. We then produce results obtained from calculations performed in [GAP 2017]; our code can be found in the online supplement. We are able to show that the irreducibility of the minuscule representation of type B_n can be detected by cycle structures alone when $n = 1, 2, 3, 5$, and 7 and that irreducibility cannot be detected when $n = 4, 6, 8, 9, \ldots, 14$. We conjecture that this continues to be true for all higher ranks as well.

2. Simple Lie algebras

We give a brief account of the background needed to discuss minuscule representations. We recommend [Erdmann and Wildon 2006] for a gentle introduction to this material or the texts [Humphreys 1972] and [Carter 2005] for more complete discussions.

A *Lie algebra* is a vector space $\mathfrak{g}$ (over $\mathbb{C}$) equipped with a bilinear multiplication $[\,\cdot\,,\,\cdot\,] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, called the *Lie bracket*, which is alternating, i.e., $[x, x] = 0$ for all $x \in \mathfrak{g}$, and satisfies the Jacobi identity $[[x, y], z] + [[y, z], x] + [[z, x], y] = 0$ for all $x, y, z \in \mathfrak{g}$. For each $g \in \mathfrak{g}$ we define $\mathrm{ad}(g) : \mathfrak{g} \to \mathfrak{g}$ to be left multiplication by g: $\mathrm{ad}(g)(x) = [g, x]$. A *subalgebra* $\mathfrak{h}$ of $\mathfrak{g}$ is a subspace of $\mathfrak{g}$ which is closed under the Lie bracket; i.e., $\mathfrak{h} \subseteq \mathfrak{g}$ such that for all $x, y \in \mathfrak{h}$ we have $[x, y] \in \mathfrak{h}$. An *ideal* $\mathfrak{i}$ of $\mathfrak{g}$ is a subspace of $\mathfrak{g}$ which absorbs multiplication by elements of $\mathfrak{g}$; i.e., $\mathfrak{i} \subseteq \mathfrak{g}$ such that for all $x \in \mathfrak{i}$ and $g \in \mathfrak{g}$ we have $[g, x] \in \mathfrak{i}$. We call $\mathfrak{g}$ *abelian* if $[x, y] = 0$ for all $x, y \in \mathfrak{g}$. A nonabelian Lie algebra with no proper nontrivial ideals is called *simple*. This means that $\mathfrak{g}$ is simple if $[\mathfrak{g}, \mathfrak{g}] \neq \mathbf{0}$ and if $\mathfrak{i}$ is an ideal of $\mathfrak{g}$, then $\mathfrak{i} = \mathbf{0}$ or $\mathfrak{g}$.

As an example, $\mathbb{R}^3$ equipped with the familiar cross product is a 3-dimensional simple Lie algebra (over the field of real numbers $\mathbb{R}$). If we let $\mathfrak{gl}_n$ denote the $n \times n$ complex matrices, then $\mathfrak{gl}_n$ becomes the *general linear* Lie algebra when given the *commutator bracket* $[A, B] = AB - BA$. The set of all trace-zero $n \times n$ complex matrices is called the *special linear* Lie algebra $\mathfrak{sl}_n$. It is a subalgebra of $\mathfrak{gl}_n$ and turns out to be simple when $n \geq 2$.

Let $\varphi : \mathfrak{g}_1 \to \mathfrak{g}_2$ be a linear map between two Lie algebras. We call φ a *homomorphism* if $\varphi([x, y]) = [\varphi(x), \varphi(y)]$ for all $x, y \in \mathfrak{g}_1$. Of course, a bijective homomorphism is an *isomorphism*.

One of the early triumphs of Lie theory was Killing and Cartan's classification of all finite-dimensional simple Lie algebras (over $\mathbb{C}$). Killing and Cartan were able to show that each finite-dimensional simple Lie algebra was isomorphic to one of the algebras on their list:

$$A_n \ (n \geq 1), \quad B_n \ (n \geq 2), \quad C_n \ (n \geq 3), \quad D_n \ (n \geq 4),$$
$$E_6, \ E_7, \ E_8, \quad F_4, \quad \text{and} \quad G_2.$$

Algebras of types A through D are called *classical algebras*. Those of types E, F, and G are called *exceptional algebras*. We refer the reader to [Erdmann and Wildon 2006] for an accessible introduction to this classification.

A *Cartan subalgebra* $\mathfrak{h}$ of a simple Lie algebra $\mathfrak{g}$ is a subalgebra which is nilpotent, i.e.,

$$\underbrace{[[\cdots[[\mathfrak{h}, \mathfrak{h}], \mathfrak{h}], \ldots], \mathfrak{h}]}_{k-\text{times}} = \mathbf{0}$$

for some integer $k > 0$, and self-normalizing, i.e., if $x \in \mathfrak{g}$, $y \in \mathfrak{h}$, and $[x, y] \in \mathfrak{h}$ then $x \in \mathfrak{h}$. Equivalently, a Cartan subalgebra is a maximal toral subalgebra (a *toral* subalgebra is a subalgebra $\mathfrak{h}$ such that for all $h \in \mathfrak{h}$, the linear endomorphism $\mathrm{ad}(h) : \mathfrak{g} \to \mathfrak{g}$ is diagonalizable). Every Cartan subalgebra of a finite-dimensional simple Lie algebra $\mathfrak{g}$ has the same dimension. This dimension is called the *rank* of the simple Lie algebra.

Since all toral subalgebras $\mathfrak{h}$ are abelian, we have that for all $x, y \in \mathfrak{h}$, the maps $\mathrm{ad}(x)$ and $\mathrm{ad}(y)$ commute and so the space of endomorphisms $\mathrm{ad}(\mathfrak{h})$ can be simultaneously diagonalized. Thus $\mathfrak{g}$ decomposes into a collection of simultaneous eigenspaces for $\mathrm{ad}(\mathfrak{h})$ for any toral subalgebra $\mathfrak{h}$. By choosing $\mathfrak{h}$ to be maximal toral, our eigenspaces are in some sense maximally refined.

For what follows, let $\mathfrak{g}$ be a simple Lie algebra and let $\mathfrak{h}$ be a Cartan subalgebra of $\mathfrak{g}$. Let $n = \dim(\mathfrak{h})$ be the rank of $\mathfrak{g}$. Since $\mathrm{ad}(\mathfrak{h})$ is simultaneously diagonalizable, $\mathfrak{g} = \prod_{\alpha \in \mathfrak{h}^*} \mathfrak{g}_\alpha$, where $\mathfrak{h}^* = \{f : \mathfrak{g} \to \mathbb{C} \mid f \text{ is linear}\}$ is the dual space of $\mathfrak{h}$ and $\mathfrak{g}_\alpha = \{g \in \mathfrak{g} \mid [h, g] = \alpha(h)g \text{ for all } h \in \mathfrak{h}\}$ when $\alpha \in \mathfrak{h}^*$. When nontrivial, $\mathfrak{g}_\alpha$ is a simultaneous eigenspace corresponding to the eigenvalue $\alpha(h)$ for each $h \in \mathfrak{h}$. Since $\mathfrak{h}$ is abelian and self-normalizing, $\mathfrak{g}_0 = \mathfrak{h}$. If $\mathbf{0} \neq \alpha \in \mathfrak{h}^*$ and $\mathfrak{g}_\alpha \neq \mathbf{0}$, we call α a *root* and $\mathfrak{g}_\alpha$ a *root space* of $\mathfrak{g}$. Let $\Delta \subset \mathfrak{h}^*$ be the set of roots of $\mathfrak{g}$.

Given a set of roots Δ, there exists a subset $\Pi \subseteq \Delta$ such that each root can be expressed as a nonpositive or nonnegative integral linear combination of elements of Π. In this case we call the elements of Π *simple roots*. A root system may have many equivalent collections of simple roots. The cardinality of a set of simple roots

is exactly the rank of $\mathfrak{g}$ (i.e., the dimension of $\mathfrak{h}$). Let us fix such a set of simple roots $\Pi = \{\alpha_1, \ldots, \alpha_n\} \subseteq \Delta$. So for each $\alpha \in \Delta$ there exists $c_1, \ldots, c_n \in \mathbb{Z}$ such that $\alpha = c_1\alpha_1 + \cdots + c_\ell\alpha_n$ with either all $c_i \geq 0$ (for a *positive root*) or all $c_i \leq 0$ (for a *negative root*).

3. The Weyl group and irreducible modules

The simple roots, $\Pi = \{\alpha_1, \ldots, \alpha_n\}$, form a basis for $\mathfrak{h}^*$. The *fundamental weights* $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ form another important basis for $\mathfrak{h}^*$. The root and weight bases are related by the *Cartan matrix* of $\mathfrak{g}$. In particular, if $A = (a_{ij})_{1 \leq i, j \leq n}$ is the Cartan matrix, then $\alpha_i = a_{i1}\lambda_1 + a_{i2}\lambda_2 + \cdots + a_{in}\lambda_n$ for $1 \leq i \leq n$.

For each $1 \leq i \leq n$, we define $\sigma_i : \mathfrak{h}^* \to \mathfrak{h}^*$ by $\sigma_i(\lambda_j) = \lambda_j - \delta_{ij}\alpha_i$ and extend linearly (where δ_{ij} is the Kronecker delta). The map σ_i is called the *simple reflection* associated with the simple root α_i. Let $\mathfrak{W}(\mathfrak{g}) = \langle \sigma_1, \sigma_2, \ldots, \sigma_n \rangle$ be the group generated by the simple reflections (generated as a subgroup of, for example, $GL(\mathfrak{h}^*)$). This is called the *Weyl group* of $\mathfrak{g}$.

A (finite-dimensional) vector space M (over $\mathbb{C}$) equipped with a bilinear $\mathfrak{g}$-action $(g, \boldsymbol{v}) \mapsto g \cdot \boldsymbol{v}$ is a $\mathfrak{g}$-*module* if $[x, y] \cdot \boldsymbol{v} = x \cdot (y \cdot \boldsymbol{v}) - y \cdot (x \cdot \boldsymbol{v})$ for all $x, y \in \mathfrak{g}$ and $\boldsymbol{v} \in M$. A homomorphism $\varphi : \mathfrak{g} \to \mathfrak{gl}(M)$ (where $\mathfrak{gl}(M)$ is equipped with the commutator bracket) is called a *representation*. It is not hard to show that every module gives rise to a representation and vice versa. Specifically, given a module action or representation, one can define the other structure as $x \cdot \boldsymbol{v} = (\varphi(x))(\boldsymbol{v})$. For what follows, we will treat the words "module" and "representation" as synonyms.

Let $\varphi : M_1 \to M_2$ be a linear map between two $\mathfrak{g}$-modules. If $\varphi(g \cdot \boldsymbol{v}) = g \cdot \varphi(\boldsymbol{v})$ for all $g \in \mathfrak{g}$ and $\boldsymbol{v} \in M_1$, then φ is a $\mathfrak{g}$-module map. A bijective module map is called a ($\mathfrak{g}$-module) isomorphism.

A subspace closed under the action of $\mathfrak{g}$ is called a *submodule*. A nontrivial module ($M \neq \boldsymbol{0}$) which has no nontrivial proper submodules (if N is a submodule, then $N = \boldsymbol{0}$ or $N = M$) is called an *irreducible* module. If M is a $\mathfrak{g}$-module and $\lambda \in \mathfrak{h}^*$, we define $M_\lambda = \{\boldsymbol{v} \in M \mid h \cdot \boldsymbol{v} = \lambda(h)\boldsymbol{v} \text{ for all } h \in \mathfrak{h}\}$. If $M_\lambda \neq \boldsymbol{0}$, we say that M_λ is a *weight space* (whose elements are *weight vectors*) with *weight* λ. Just as $\mathfrak{g}$ is a direct sum of root spaces, $\mathfrak{g}$-modules are direct sums of weight spaces: $M = \prod_{\lambda \in \mathfrak{h}^*} M_\lambda$.

Let M be an irreducible $\mathfrak{g}$-module. There exists a (unique) weight $\lambda \in \mathfrak{h}^*$ of M such that given any other weight $\mu \in \mathfrak{h}^*$ we have $\mu = \lambda - \sum_{i=1}^n b_i\alpha_i$, where $b_i \in \mathbb{Z}$ and $b_i \geq 0$. So every other weight is obtained by subtracting certain collections of positive roots from this weight. Such a weight, λ, is unique and is called the *highest weight* of M. If $\lambda \in \mathfrak{h}^*$ and there exists $c_i \in \mathbb{Z}$, $c_i \geq 0$ such that $\lambda = \sum_{i=1}^n c_i\lambda_i$ (the λ_i's are the fundamental weights), then λ is a *dominant integral weight*.

Highest weights of finite-dimensional irreducible modules are dominant integral. Conversely, each dominant integral weight is the highest weight of some

finite-dimensional irreducible module. Two irreducible modules with the same highest weight are isomorphic, so we have a bijection between the set of dominant integral weights and the isomorphism classes of finite-dimensional irreducible modules.

Let λ be a dominant integral weight for some simple Lie algebra of type X_n. We denote the irreducible highest-weight X_n-module with highest weight λ by $L(X_n, \lambda)$ or just $L(\lambda)$ when the algebra is understood.

4. Minuscule representations

There are many equivalent ways of defining minuscule weights. In fact, six equivalent conditions are given in [Bourbaki 2005, Chapter VIII, Section 7.3]. The following definition best fits our purposes:

Definition 4.1. Suppose $L(\lambda)$ is an irreducible finite-dimensional $\mathfrak{g}$-module with nonzero highest weight $\lambda \in \mathfrak{h}^*$. Then λ is a *minuscule weight* and $L(\lambda)$ is a *minuscule module* if the Weyl group $\mathfrak{W}(\mathfrak{g})$ acts transitively on the set of weights of $L(\lambda)$, i.e., $\mathfrak{W}(\mathfrak{g}) \cdot \lambda$ is the set of all weights of $L(\lambda)$.

Given a $\mathfrak{g}$-module M, we know M decomposes into weight spaces M_λ for $\lambda \in \mathfrak{h}^*$. The dimension of a weight space M_λ is called the *multiplicity* of the weight λ.

If $\mu = w \cdot \lambda$ for $\mu, \lambda \in \mathfrak{h}^*$ and $w \in \mathfrak{W}(\mathfrak{g})$, then M_μ and M_λ have the same dimension. Therefore, weights lying in an orbit of the Weyl group all have the same multiplicity. Thus since the weights of a minuscule module all lie in a single Weyl group orbit, the weight spaces in a minuscule module must all have the same multiplicity as the highest weight. But the highest-weight space for an irreducible module is always 1-dimensional. Therefore, all the weight spaces in a minuscule module are 1-dimensional and the dimension of a minuscule module is the same as the number of its weights.

Both [Humphreys 1972, Section 13, p. 72, Exercise 13] and [Bourbaki 2005, Chapter VIII, Section 7.3, p. 132] give the following table of minuscule weights for finite-dimensional simple Lie algebras:

type	A_n	B_n	C_n	D_n	E_6	E_7
minuscule weights	$\lambda_1, \ldots, \lambda_n$	λ_n	λ_1	$\lambda_1, \lambda_{n-1}, \lambda_n$	λ_1, λ_6	λ_7

Note that algebras of types F_4, E_8, and G_2 have no minuscule representations.

For further information about minuscule representations we direct the reader to either [Bourbaki 2005, Chapter VII, Section 7.3] or the book [Green 2013], which is entirely devoted to the study of minuscule representations and contains a wealth of information about them.

5. Strictly transitive sets

Recall that the original motivation for this project was to extend results found in [Cook et al. 2005]. Following that paper, let us denote the conjugacy class of a permutation σ by $\bar{\sigma}$. We say a collection of conjugacy classes, $\{C_1, \ldots, C_\ell\}$ of the symmetric group S_m is *strictly transitive* if for any choice of $\tau_i \in C_i$ ($i = 1, \ldots, \ell$) the subgroup generated by $\tau_1, \ldots, \tau_\ell$ acts transitively. Lemma 3.7 in [Cook et al. 2005] states that $\{C_1, \ldots, C_\ell\}$ is strictly transitive if and only if for some (and therefore any) set of representatives $\{\tau_1, \ldots, \tau_\ell\}$ (with $\tau_i \in C_i$) and for any $1 \leq j \leq m - 1$, there is an element τ_k leaving no set of cardinality j invariant.

As an example, working in S_4, $\{\overline{(1234)}\}$ is strictly transitive by itself (leaving only the empty set and $\{1, 2, 3, 4\}$ invariant). Also, $\{\overline{(123)}, \overline{(12)(34)}\}$ is strictly transitive since an element from $\overline{(123)}$ only allows invariant sets of cardinalities 0, 1, 3, and 4 whereas elements in $\overline{(12)(34)}$ only allow invariant sets of sizes 0, 2, and 4. So putting these two criteria together, cardinalities 1, 2, and 3 are ruled out. On the other hand, $\{\overline{(1)}, \overline{(12)}, \overline{(12)(34)}\}$ is not strictly transitive since selecting the permutations (1), (12), and $(12)(34)$ allows the set $\{1, 2\}$ (of cardinality 2) to remain invariant.

Recall that the Weyl group permutes the weights of a representation. Thus if $\mathfrak{g}$ is a simple Lie algebra and M is a $\mathfrak{g}$-module with $\dim(M) = m$, then $\mathfrak{W}(\mathfrak{g})$ can be viewed as a subgroup of the symmetric group S_m, say

$$\mathfrak{W}(\mathfrak{g}) \cong W \subseteq S_m.$$

For the construction in [Cook et al. 2005] to work for a Lie group with corresponding Lie algebra $\mathfrak{g}$, the authors needed an irreducible representation where the conjugacy classes of the corresponding permutation representation of the Weyl group form a strictly transitive set.

To have any hope of W having a strictly transitive set of conjugacy classes we must have that the weights of M lie in a single orbit of $\mathfrak{W}(\mathfrak{g}) \cong W$. This means that the construction cannot go through unless M is a minuscule representation. This in turn implies that the construction cannot work for algebras of type E_8, F_4, or G_2 (where there are not minuscule representations).

Now let M (with $\dim(M) = m$) be a minuscule $\mathfrak{g}$-module with corresponding Weyl group W (viewed as permutations of the weights of M). The conjugacy classes of W form a strictly transitive set if and only if the cycle structures in W do not allow invariant sets of cardinality j for $1 \leq j \leq m - 1$. Essentially this means that the conjugacy classes of W form a strictly transitive set only if the irreducibility of M is visible directly from the cycle structures of W. So for the construction in [Cook et al. 2005] to go through we need a representation whose irreducibility can be established by examining the cycle structures of the Weyl group elements acting as permutations on the weights of this representation.

6. Seeing irreducibility from cycle structures

The problem of identifying a minuscule representation with corresponding Weyl group action possessing a strictly transitive set of conjugacy classes was solved in [Cook et al. 2005] for a simple Lie algebra of type A_n, C_n, D_n, E_6, or E_7. Again, algebras of types F_4, E_8, and G_2 have no minuscule representations so there are no strictly transitive sets associated with representations there. We will briefly review the results found in [Cook et al. 2005]. For more detail we refer the reader to Section 4 of that paper.

Recall that $L(A_n, \lambda_i)$ (where $n = 1, 2, \dots$) is minuscule for all $i = 1, \dots, n$. Focusing on $i = 1$, the minuscule module $L(A_n, \lambda_1)$ (where $n = 1, 2, \dots$) is $(n+1)$-dimensional. It turns out that the Coxeter element (i.e., the product of all of the simple reflections) of the Weyl group is represented by an $(n+1)$-cycle, since such a cycle leaves only sets of cardinalities 0 and $n + 1$ invariant. Thus we have a strictly transitive set, and so the irreducibility of $L(A_n, \lambda_1)$ is visible from cycle structures alone.

For type C_n (where $n = 3, 4, \dots$), the only minuscule module is the $(2n)$-dimensional representation $L(C_n, \lambda_1)$. As with type A_n, it turns out that the Coxeter element is represented by a $(2n)$-cycle. This means that the irreducibility of $L(C_n, \lambda_1)$ is visible from cycle structures alone.

Each algebra of type D_n (where $n = 4, 5, \dots$) possesses three minuscule modules: $L(D_n, \lambda_1)$, $L(D_n, \lambda_{n-1})$, and $L(D_n, \lambda_n)$. The first of these, $L(D_n, \lambda_1)$, is $(2n)$-dimensional. If the weights are suitably labeled by $1, 2, \dots, 2n$, it turns out that the product of the first $n - 1$ simple reflections yields the permutation $\tau_1 = (1, 2, \dots, n)(n + 1, \dots, 2n)$ and the Coxeter element is $\tau_2 = (1, \dots, n - 1, n+1, \dots, 2n-1)(n, 2n)$. Representatives from the class $\bar{\tau}_1$ leave sets of cardinalities $0, n$, and $2n$ invariant whereas representatives from $\bar{\tau}_2$ leave sets of cardinalities $0, 2, 2n - 2$, and $2n$ invariant. Since $n \geq 4$, intersecting these two criteria leaves just 0 and $2n$. Therefore, $\{\bar{\tau}_1, \bar{\tau}_2\}$ is a strictly transitive set and so the irreducibility of $L(D_n, \lambda_1)$ is visible from cycle structures alone.

The algebra of type E_6 possess two minuscule modules: $L(E_6, \lambda_1)$ and $L(E_6, \lambda_6)$. These are both 27-dimensional. The corresponding permutation representations of the Weyl group possess elements τ_1 and τ_2 with respective cycle structures $12 + 12 + 3$ (two 12-cycles and a 3-cycle) and $9 + 9 + 9$ (three 9-cycles). This means that elements from $\bar{\tau}_2$ only allow invariant sets of cardinality 0, 9, 18, and 27. Notice that cardinalities 9 and 18 are not allowed by elements of $\bar{\tau}_1$. Therefore, $\{\bar{\tau}_1, \bar{\tau}_2\}$ is a strictly transitive set.

The only minuscule module of E_7 is the 56-dimensional representation $L(E_7, \lambda_7)$. The corresponding permutation representation of the Weyl group possesses elements τ_1 and τ_2 with respective cycle structures $18 + 18 + 18 + 2$ (three 18-cycles and a

transposition) and $14+14+14+14$ (four 14-cycles). This means that elements from $\bar{\tau}_2$ only allow invariant sets of cardinality 0, 14, 28, 42 and 56. Notice that cardinalities 14, 28 and 42 are not allowed by elements of $\bar{\tau}_1$. Therefore, $\{\bar{\tau}_1, \bar{\tau}_2\}$ is a strictly transitive set.

Finally, algebras of type B_n (where $n = 2, 3, \dots$) only have one minuscule representation: $L(B_n, \lambda_n)$. This is a 2^n-dimensional representation and the focus of this project. In [Cook et al. 2005], it is stated that when $n = 2, 3, 5$, and 7 the Weyl group corresponding to the minuscule module $L(B_n, \lambda_n)$ possesses a strictly transitive set. However, the Weyl group in the case $n = 4$ does not. For other ranks the problem is left open.

7. The action of $\mathfrak{W}(B_n)$ on the minuscule representation

We now focus on simple Lie algebras of type B_n (where $n = 2, 3, \dots$). Algebras of type B_n can be realized as the *special orthogonal* Lie algebras $\mathfrak{so}_{2n+1}$. Specifically, letting I_n denote the $n \times n$ identity matrix, we have that the special orthogonal Lie algebra is the following set of $(2n+1) \times (2n+1)$ complex matrices:

$$\mathfrak{so}_{2n+1} = \left\{ X \in \mathfrak{gl}_{2n+1} \ \middle| \ X^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & I_n \\ 0 & -I_n & 0 \end{bmatrix} = - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & I_n \\ 0 & -I_n & 0 \end{bmatrix} X \right\}.$$

This is a $(2n^2+n)$-dimensional simple Lie algebra of rank n. Let us fix a collection of simple roots $\Pi = \{\alpha_1, \dots, \alpha_n\}$ and corresponding fundamental weights $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ for this algebra. We have that the Cartan matrix (the change of basis matrix from Λ to Π) is

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -2 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{bmatrix}$$

with corresponding Dynkin diagram
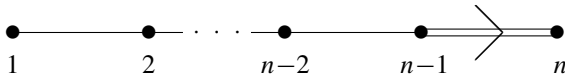
Explicitly we have the following relationships between our fundamental weights and simple roots:

$$\alpha_1 = 2\lambda_1 - \lambda_2, \quad \alpha_2 = -\lambda_1 + 2\lambda_2 - \lambda_3, \quad \dots,$$

$$\alpha_{n-2} = -\lambda_{n-3} + 2\lambda_{n-2} - \lambda_{n-1}, \quad \alpha_{n-1} = -\lambda_{n-2} + 2\lambda_{n-1} - 2\lambda_n, \quad \alpha_n = -\lambda_{n-1} + 2\lambda_n.$$

Let $\epsilon_1, \ldots, \epsilon_n$ be the standard basis for $\mathbb{R}^n$. In addition, consider $\alpha_i = 4(\epsilon_i - \epsilon_{i+1})$ for $i = 1, \ldots, n-1$ and $\alpha_n = 4\epsilon_n$. By Lemma 5.1 in [Green 2008], $\Pi = \{\alpha_1, \ldots, \alpha_n\}$ is a set of simple roots for a root system of type B_n.

Recall, see [Humphreys 1972, Section 13.2, Table 1, p. 69], that for type B_n,

$$\lambda_i = \alpha_1 + 2\alpha_2 + \cdots + (i-1)\alpha_{i-1} + i(\alpha_i + \cdots + \alpha_{n-1} + \alpha_n) \quad \text{for } i = 1, \ldots, n-1,$$
$$\lambda_n = \tfrac{1}{2}(\alpha_1 + 2\alpha_2 + \cdots + n\alpha_n).$$

In terms of the standard basis we have that $\lambda_i = 4(\epsilon_1 + \cdots + \epsilon_i)$ for $i = 1, \ldots, n-1$ and $\lambda_n = 2(\epsilon_1 + \cdots + \epsilon_n)$. This in turn implies $\epsilon_1 = \tfrac{1}{4}\lambda_1$, $\epsilon_j = \tfrac{1}{4}\lambda_j - \tfrac{1}{4}\lambda_{j-1}$ (where $j = 2, \ldots, n-1$), and $\epsilon_n = \tfrac{1}{2}\lambda_n - \tfrac{1}{4}\lambda_{n-1}$.

Recall that the Weyl group is generated by the simple reflections: $\sigma_i(\lambda_j) = \lambda_j - \delta_{ij}\alpha_i$ $(i = 1, \ldots, n)$. Notice that ϵ_j only involves λ_{j-1} and λ_j for $j = 2, \ldots, n$ and ϵ_1 only involves λ_1. Therefore, since $\sigma_i(\lambda_k) = \lambda_k$ for $k \neq i$, we have $\sigma_i(\epsilon_j) = \epsilon_j$ if $j \neq i$ or $i + 1$.

For $1 < i < n$,

$$\sigma_i(\epsilon_i) = \sigma_i\left(\tfrac{1}{4}\lambda_i - \tfrac{1}{4}\lambda_{i-1}\right) = \tfrac{1}{4}\sigma_i(\lambda_i) - \tfrac{1}{4}\sigma_i(\lambda_{i-1})$$
$$= \tfrac{1}{4}\lambda_i - \tfrac{1}{4}\alpha_i - \tfrac{1}{4}\lambda_{i-1} = \epsilon_i - \tfrac{1}{4}\alpha_i = \epsilon_i - (\epsilon_i - \epsilon_{i+1}) = \epsilon_{i+1}.$$

Likewise, $\sigma_i(\epsilon_{i+1}) = \epsilon_i$. Therefore, for $i = 2, \ldots, n-1$, we see σ_i switches ϵ_i and ϵ_{i+1} and leaves the other ϵ_j fixed. A similar calculation shows that σ_1 switches ϵ_1 and ϵ_2 leaving the other basis vectors fixed.

Notice $\sigma_n(\epsilon_j) = \epsilon_j$ for $j = 1, \ldots, n-1$. Finally, consider

$$\sigma_n(\epsilon_n) = \sigma_n\left(\tfrac{1}{2}\lambda_n - \tfrac{1}{4}\lambda_{n-1}\right) = \tfrac{1}{2}\sigma_n(\lambda_n) - \tfrac{1}{4}\sigma_n(\lambda_{n-1})$$
$$= \tfrac{1}{2}\lambda_n - \tfrac{1}{2}\alpha_n - \tfrac{1}{4}\lambda_{n-1} = \epsilon_n - \tfrac{1}{2}\alpha_n = \epsilon_n - 2\epsilon_n = -\epsilon_n.$$

Thus σ_n leaves all but the last basis vector fixed and switches the sign of the final basis vector.

If we label $\epsilon_1, \ldots, \epsilon_n$ by $1, \ldots, n$, then we have that the Weyl group is acting as signed permutations on $\{\pm 1, \ldots, \pm n\}$. In fact, the permutation representation of the Weyl group $\mathfrak{W}(C_n)$ acting on the weights of the minuscule $L(C_n, \lambda_1)$ can be realized in this way. This is part of the reason it was relatively easy for the authors of [Cook et al. 2005] to resolve the type C_n case.

Even though types B_n and C_n have isomorphic Weyl groups (both groups are isomorphic to the group of signed permutations on $\{1, \ldots, n\}$), the permutation representation of $\mathfrak{W}(B_n)$ acting on the weights of the minuscule representation $L(B_n, \lambda_n)$ is much more complicated than $\mathfrak{W}(C_n)$ acting on the weights of $L(C_n, \lambda_1)$.

Let Ψ be the set of 2^n vectors of the form $(\pm 2, \ldots, \pm 2)$. By Proposition 5.2 in [Green 2008], Ψ is a set of roots for $L(B_n, \lambda_n)$. Notice that

$$\lambda_n = 2(\epsilon_1 + \cdots + \epsilon_n) = (2, \ldots, 2)$$

is the highest weight. We know that $\mathfrak{W}(B_n)$ permutes the elements of Ψ. Consider the signs of the coordinates of an element of Ψ. We can treat these like reversed binary digits (interpret $+$ as 0 and $-$ as 1) then add 1 to this number. For example: $(-2, +2, +2)$ is interpreted as $001_2 + 1 = 2$ and $(+2, -2, -2)$ is interpreted as $110_2 + 1 = 7$.

Then σ_i for $i = 1, \ldots, n-1$ has the effect (after adjusting for the addition of 1) of switching the j and $(j+1)$-th digits of the reversed binary number and σ_n has the effect of flipping the final digit of the reversed binary number. This gives us the following:

Theorem 7.1. *The simple reflections of the Weyl group $\mathfrak{W}(B_n)$ acting on the weights of the minuscule representation $L(B_n, \lambda_n)$ can be represented by the permutations*

$$\sigma_j = \prod_{p=0}^{2^{(n-j-1)}-1} \prod_{k=1}^{2^{j-1}} (p2^{j+1} + 2^{j-1} + k, \, p2^{j+1} + 2^j + k), \quad 1 \le j \le n-1,$$

$$\sigma_n = \prod_{k=1}^{2^{n-1}} (k, \, 2^{n-1} + k).$$

8. Experimental results for type B_n

Using [Theorem 7.1](#) and [GAP 2017], for $n \le 14$, we were able to find complete lists of cycle structures for the elements in $\mathfrak{W}(B_n)$ viewed as permutations of weights of the minuscule module. (Our GAP code can be found in the online supplement.) These lists allowed us to conclude that the cycle structures for types B_n when $n = 1, 2, 3, 5,$ and 7 yield strictly transitive sets. Thus the irreducibility of $L(B_n, \lambda_n)$ can be seen from cycle structure alone when $n = 1, 2, 3, 5,$ and 7.

The same cannot be concluded for other values of n. Below we elaborate on our method for determining irreducibility from cycle structures by examining the cycle structures of B_n for the ranks $n = 1, 2, 3, 4,$ and 5.

Note that, viewed as permutations, $\mathfrak{W}(B_1) = \{(1), (12)\}$. For our purposes we describe the cycle structures in this group by $1 + 1$ for the identity (two 1-cycles) and 2 for the transposition (12) (a single 2-cycle). This identification allows us to read off the possible dimensions of invariant subspaces allowed by each cycle structure. If we can find a cycle structure (or a collection of cycle structures) that only allows for dimensions of 0 and 2^n we know we can conclude irreducibility from the cycle structures alone. In this case, the 2-cycle structure guarantees the irreducibility of our minuscule representation. We will understand why after the following examples.

When $n = 2$, we have $\mathfrak{W}(B_2) = \langle (23), (13)(24) \rangle$ with cycle structures

$$1 + 1 + 1 + 1 \;=\; 1 + 1 + 2 \;=\; 2 + 2 \;=\; 4.$$

So every element in $\mathfrak{W}(B_2)$ viewed as a permutation is of the form four 1-cycles, two 1-cycles and a 2-cycle, two 2-cycles or a 4-cycle. Any partial sum of a type of cycle structure is a possible dimension for an invariant subspace of our minuscule representation allowed by that cycle structure. So the cycle structure $1 + 1 + 2$ allows for possible dimensions of 0, 1, 2, $3 = 1 + 2$ and $4 = 1 + 1 + 2$. However, the pair of cycles $2 + 2$ only allows dimensions 0, 2, and $4 = 2 + 2$. Critically, we also have that the cycle structure 4 (a 4-cycle) allows for dimensions of only 0 and 4. Hence, we conclude that any invariant subspace of our minuscule representation must be of dimension 0 or 4. So irreducibility of our minuscule representation is visible from examining cycle structures alone.

Next $\mathfrak{W}(B_3) = \langle (23)(67),\ (35)(46),\ (15)(26)(37)(48) \rangle$ and has cycle structures

$$\begin{aligned}
1 + 1 + \cdots + 1 &= 1 + 1 + 1 + 1 + 2 + 2 = 1 + 1 + 3 + 3 \\
= 2 + 2 + 2 + 2 &= 2 + 6 = 4 + 4.
\end{aligned}$$

In this case there is no structure of the form $2^3 = 8$ to guarantee irreducibility. Instead we may consider the structures $2 + 6$ and $4 + 4$ simultaneously: $2 + 6$ allows for the possible dimensions 0, 2, 6, and 8, while $4 + 4$ allows for 0, 4, and 8. These lists of possible dimensions of invariant subspaces intersect at just 0 and 8. Hence, irreducibility follows from cycle structures.

The first case in which this method fails is that of $n = 4$:

$$\mathfrak{W}(B_4) = \big\langle (2, 3)(6, 7)(10, 11)(14, 15),\ (3, 5)(4, 6)(11, 13)(12, 14),$$
$$(5, 9)(6, 10)(7, 11)(8, 12),\ (1, 9)(2, 10) \cdots (8, 16) \big\rangle.$$

In this realization of $\mathfrak{W}(B_4)$ we find the cycle structures

$$\begin{aligned}
1 + 1 + \cdots + 1 &= 1 + 1 + \cdots + 1 + 2 + 2 + 2 + 2 \\
= 1 + 1 + 2 + 4 + 4 + 4 &= 1 + 1 + 1 + 1 + 3 + 3 + 3 + 3 \\
= 2 + 2 + \cdots + 2 &= 1 + 1 + 1 + 1 + 2 + 2 + \cdots + 2 \\
= 2 + 2 + 6 + 6 &= 4 + 4 + 4 + 4 = 8 + 8.
\end{aligned}$$

Each of these cycle structures allows for an invariant subspace of dimension 8. So even though B_4's minuscule module is irreducible, cycle structures alone will not reveal this to us.

For B_5, we have that $\mathfrak{W}(B_5)$ has cycles structures of the forms $8 + 8 + 8 + 8$ and $2 + 10 + 10 + 10$. The form $8 + 8 + 8 + 8$ only allows for submodules of dimensions 0, 8, 16, 24, and 32, whereas $2 + 10 + 10 + 10$ only allows for submodules of dimensions 0, 2, 10, 12, 20, 22, 30, and 32. Thus, only 0 and 32 are allowed, so irreducibility follows.

sums up the results for ranks $6 \leq n \leq 12$. We see that the cycle structures for B_7 imply the irreducibility of its minuscule representation.

rank	invariant subspace dimensions allowed by cycle structures
6	0, 24, 40, 64
7	0, 128
8	0, 16, 32, 112, 128, 144, 224, 240, 256
9	0, 144, 224, 288, 368, 512
10	0, 64, 144, 224, 240, 320, 400, 464, 480, 544, 560, 624, 704, 784, 800, 880, 960, 1024
11	0, 288, 464, 528, 640, 704, 1344, 1408, 1520, 1584, 1760, 2048
12	0, 48, 112, 176, 224, 288, 352, 400, 464, 528, 576, 640, 704, 752, 816, 880, 928, 992, 1056, 1104, 1168, 1232, 1280, 1344, 1408, 1456, 1520, 1584, 1632, 1696, 1760, 1808, 1872, 1936, 1984, 2048, 2112, 2160, 2224, 2288, 2336, 2400, 2464, 2512, 2576, 2640, 2688, 2752, 2816, 2864, 2928, 2992, 3040, 3104, 3168, 3216, 3280, 3344, 3392, 3456, 3520, 3568, 3632, 3696, 3744, 3808, 3872, 3920, 3984, 4048, 4096
13	0, 624, 704, 1328, 1456, 2160, 2288, 2912, 2992, 3616, 3744, 4448, 4576, 5280, 5904, 6032, 6736, 6864, 7488, 7568, 8192
14	0, 368, 704, 1456, 2160, 2912, 3616, 3696, 4368, 5072, 5152, 5824, 6528, 5200, 6608, 6864, 8064, 8320, 9520, 9776, 9856, 10560, 11232, 11312, 12016, 12688, 12768, 13472, 14224, 14928, 15680, 16016, 16384

Table 1. Summary of results for B_n, where $6 \le n \le 12$.

We were not able to get GAP to complete calculations for any higher-rank cases. The problem is that Weyl groups grow very fast as rank is increased. In fact $\mathfrak{W}(B_n)$ is isomorphic to a semidirect product of S_n and $(\mathbb{Z}_2)^n$, so $|\mathfrak{W}(B_n)| = 2^n \cdot n!$. Even at rank 14 we have a group of order $2^{14} \cdot 14!$ acting on a set of $2^{14} = 16384$ weights! However, by randomly sampling $\mathfrak{W}(B_n)$ for ranks of up to $n = 23$, we obtained strong evidence that the number of allowed invariant subspace dimensions blows up as rank is increased. We conjecture that the irreducibility of the minuscule representation cannot be seen from cycle structures alone after rank 7. We found this quite surprising given the nature of the minuscule representations for the other types of algebras.

References

[Bourbaki 2005] N. Bourbaki, *Lie groups and Lie algebras, Chapters 7–9*, Springer, 2005. MR Zbl

[Carter 2005] R. W. Carter, *Lie algebras of finite and affine type*, Cambridge Studies in Advanced Mathematics **96**, Cambridge University Press, 2005. MR Zbl

[Cook et al. 2005] W. J. Cook, C. Mitschi, and M. F. Singer, "On the constructive inverse problem in differential Galois theory", *Comm. Algebra* **33**:10 (2005), 3639–3665. MR Zbl

[Erdmann and Wildon 2006] K. Erdmann and M. J. Wildon, *Introduction to Lie algebras*, Springer, 2006. MR Zbl

[GAP 2017] "GAP – Groups, Algorithms, and Programming", version 4.8.7, 2017, available at http://www.gap-system.org.

[Green 2008] R. M. Green, "Representations of Lie algebras arising from polytopes", *Int. Electron. J. Algebra* **4** (2008), 27–52. MR Zbl

[Green 2013] R. M. Green, *Combinatorics of minuscule representations*, Cambridge Tracts in Mathematics **199**, Cambridge University Press, 2013. Zbl

[Humphreys 1972] J. E. Humphreys, *Introduction to Lie algebras and representation theory*, Graduate Texts in Mathematics **9**, Springer, 1972. MR Zbl

cookwj@appstate.edu   Department of Mathematical Sciences,
                             Appalachian State University, Boone, NC, United States

noah.hughes@uconn.edu   Department of Mathematics, University of Connecticut,
                             Storrs, CT, United States

Pythagorean orthogonality of compact sets

Pallavi Aggarwal, Steven Schlicker and Ryan Swartzentruber

(Communicated by Kenneth S. Berenhaut)

The Hausdorff metric h is used to define the distance between two elements of $\mathcal{H}(\mathbb{R}^n)$, the hyperspace of all nonempty compact subsets of $\mathbb{R}^n$. The geometry this metric imposes on $\mathcal{H}(\mathbb{R}^n)$ is an interesting one — it is filled with unexpected results and fascinating connections to number theory and graph theory. Circles and lines are defined in this geometry to make it an extension of the standard Euclidean geometry. However, the behavior of lines and segments in this extended geometry is much different from that of lines and segments in Euclidean geometry. This paper presents surprising results about rays in the geometry of $\mathcal{H}(\mathbb{R}^n)$, with a focus on attempting to find well-defined notions of angle and angle measure in $\mathcal{H}(\mathbb{R}^n)$.

1. Background

In this section we provide the definition of the Hausdorff metric and some known results about lines and segments of compact sets. The Hausdorff metric h was introduced by Felix Hausdorff in the early twentieth century as a way to measure the distance between compact sets. The space $\mathbb{R}^n$ will be our underlying space, and we will denote by $\mathcal{H}(\mathbb{R}^n)$ the hyperspace of all nonempty compact subsets of $\mathbb{R}^n$. The standard Euclidean metric on $\mathbb{R}^n$ will be denoted by d_E. The Hausdorff metric on $\mathcal{H}(\mathbb{R}^n)$ is defined as follows.

Definition 1.1. The *Hausdorff distance $h(A, B)$* between sets A and B in $\mathcal{H}(\mathbb{R}^n)$ is

$$h(A, B) = \max\{d(A, B), d(B, A)\},$$

where

$$d(a, B) = \min_{b \in B}\{d_E(a, b)\}$$

for $a \in A$ and

$$d(A, B) = \max_{a \in A}\{d_E(a, B)\}.$$

Example 1.2. Let A be the closed interval $[0, 2]$ and B be the closed interval $[3, 4]$ in $\mathcal{H}(\mathbb{R})$. Then $d(A, B) = |0 - 3| = 3$ and $d(B, A) = |4 - 2| = 2$. This example

illustrates that it is possible to have $d(A, B) \neq d(B, A)$, which necessitates using the maximum of $d(A, B)$ and $d(B, A)$ as the Hausdorff distance between A and B. Thus we have that $h(A, B) = d(A, B) = 3$.

The proof that h is a metric can be found in many topology texts; see [Barnsley 1993; Edgar 1990] for example.

Line segments, lines, and rays in $\mathcal{H}(\mathbb{R}^n)$ can be defined in a way that makes them analogous to segments, lines, and rays in $\mathbb{R}^n$. In $\mathbb{R}^n$ we can think of the line segment $\overline{ab}$ as the set of all points c that lie between a and b, that is, the points c that satisfy $d_E(a, b) = d_E(a, c) + d_E(c, b)$. We follow the convention in [Blumenthal 1953] and write acb to indicate that c lies between a and b. Similarly, the ray $\overrightarrow{ab}$ can be thought of as all points c such that acb or abc, and the line $\overleftrightarrow{ab}$ is the set of all points c such that cab, acb, or abc. We can naturally extend these notions to define segments, lines, and rays in $\mathcal{H}(\mathbb{R}^n)$.

Definition 1.3. The set $C \in \mathcal{H}(\mathbb{R}^n)$ lies *between* the sets A and B in $\mathcal{H}(\mathbb{R}^n)$ if

$$h(A, C) + h(C, B) = h(A, B).$$

We write ACB to signify that C lies between A and B.

The definition of betweenness in $\mathcal{H}(\mathbb{R}^n)$ then allows us to define segments, lines, and rays in $\mathcal{H}(\mathbb{R}^n)$.

Definition 1.4. Let A and B be distinct sets in $\mathcal{H}(\mathbb{R}^n)$:

(1) The *segment* $\overline{AB}$ is the collection of all sets $C \in \mathcal{H}(\mathbb{R}^n)$ that satisfy ACB.

(2) The *ray* $\overrightarrow{AB}$ is the collection of all sets $C \in \mathcal{H}(\mathbb{R}^n)$ that satisfy ACB or ABC.

(3) The *line* $\overleftrightarrow{AB}$ is the collection of all sets $C \in \mathcal{H}(\mathbb{R}^n)$ that satisfy CAB, ACB, or ABC.

Example 1.5. Let A be the circle of radius 1 and B be the circle of radius 2 in $\mathcal{H}(\mathbb{R}^2)$, both centered at the origin. Then $h(A, B) = d(A, B) = d(B, A) = 1$. Figure 1 (left image) illustrates a set C satisfying CAB (the shaded set), a set C
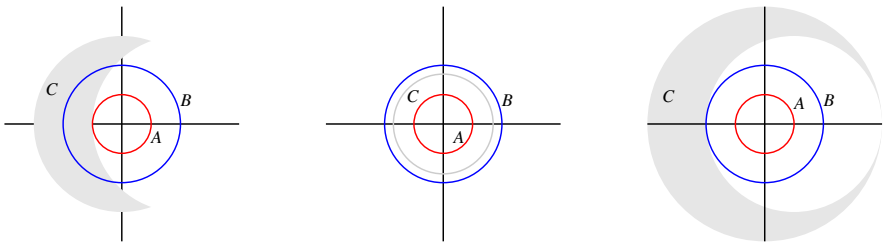
Figure 1. Sets C satisfying CAB, ACB, and ABC.
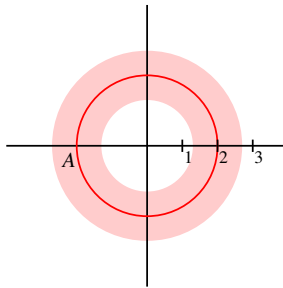
Figure 2. A dilation of a circle.

(middle image) satisfying ACB (the circle C centered at the origin of radius $1 + s$ for any $0 < s < 1$), and a set C (right image) satisfying ABC (the shaded set).

It is reasonable to ask how one goes about finding a set C that satisfies CAB, ACB, or ABC. The key lies in the dilation of a set.

Definition 1.6. Given $A \in \mathcal{H}(\mathbb{R}^n)$ and $s \in \mathbb{R}$ with $s \geq 0$, the *s-dilation* of A is the set

$$(A)_s = \{x \in \mathbb{R}^n \mid d(x, A) \leq s\}.$$

As an example, the 0.7-dilation of the circle of radius 2 in $\mathcal{H}(\mathbb{R}^2)$ is the shaded region shown in Figure 2.

Dilations are useful mainly because $h(A, (A)_s) = s$ and any set C that satisfies $h(A, C) = s$ is a subset of $(A)_s$ [Braun et al. 2005, Theorem 4]. Thus when we want to find a set C that satisfies ACB with $h(A, C) = s$, for example, we can restrict our search to subsets of $(A)_s \cap (B)_{h(A,B)-s}$. In fact, the set $X = (A)_s \cap (B)_{h(A,B)-s}$ itself satisfies AXB with $h(A, X) = s$ for any $0 < s < h(A, B)$, as the following lemma attests.

Lemma 1.7 [Bogdewicz 2000, Lemma 3.6]. *Let $A, C \in \mathcal{H}(\mathbb{R}^n)$, $h(A, C) = q$ and let*

$$W = (A)_s \cap (C)_{q-s}$$

for each $s \in [0, q]$. Then $h(A, W) = s$ and $h(W, C) = q - s$.

If we restrict ourselves to the subspace of single point sets, then the Hausdorff metric is just the Euclidean metric. In this way, the standard Euclidean geometry can be embedded in the geometry of $\mathcal{H}(\mathbb{R}^n)$. In general, though, lines and segments behave quite differently in $\mathcal{H}(\mathbb{R}^n)$ than they do in Euclidean geometry. For example, in Euclidean geometry, given two points a and b, for any $s \geq 0$ there is exactly one point c on $\overrightarrow{ab}$ (and $\overrightarrow{ba}$) with $d_E(a, c) = s$. It is demonstrated in [Bay et al. 2005] that, under certain circumstances, there are no sets C that satisfy BAC with $h(A, C) = s$ for all s larger than some real number s_0. Hence some lines in $\mathcal{H}(\mathbb{R}^n)$ are actually

just halflines. On the other hand, if there exists $a \in A$ such that $d(a, B) \neq h(A, B)$ or $b \in B$ such that $d(b, A) \neq h(A, B)$, then there are infinitely many sets C that satisfy ACB and $h(A, C) = s$ for any $0 < s < h(A, B)$ [Blackburn et al. 2009, Lemma 2.3]. This prompts the following definition:

Definition 1.8. Let $A \neq B \in \mathcal{H}(\mathbb{R}^n)$. The elements $C, C' \in \mathcal{H}(\mathbb{R}^n)$ are *at the same location* on $\overleftrightarrow{AB}$ if C and C' satisfy

- ACB and $AC'B$,
- CAB and $C'AB$, or
- ABC and ABC',

with $h(A, C) = h(A, C') = s$ for some s.

2. Some results about rays in $\mathcal{H}(\mathbb{R}^n)$

The discussion in the previous section indicates that the geometry of the Hausdorff metric is often surprising and counterintuitive. To develop the geometry more fully, we are interested in defining and measuring angles in $\mathcal{H}(\mathbb{R}^n)$. As in Euclidean geometry, we can consider an angle as being formed by two rays with a common endpoint. In Euclidean geometry, however, if the point c is on the ray $\overrightarrow{ab}$, then the rays $\overrightarrow{ac}$ and $\overrightarrow{ab}$ are the same. There is no guarantee that the same result is true in $\mathcal{H}(\mathbb{R}^n)$. In this section we examine some behavior of rays in $\mathcal{H}(\mathbb{R}^n)$ that will be pertinent if we want to define angle measure.

One result that we will use comes from [Montague 2008].

Lemma 2.1. *Let $A, B \in \mathcal{H}(\mathbb{R}^n)$. Fix $s, t \in \mathbb{R}$ with $s, t > 0$, and $s + t = h(A, B)$. Then, for $C \in \mathcal{H}(\mathbb{R}^n)$, the following statements are equivalent*:

(1) *The set C is a subset of $(A)_s \cap (B)_t$, $A \subseteq (C)_s$, and $B \subseteq (C)_t$.*

(2) *The set C is between A and B, and $h(A, C) = s$.*

Proof. (1) $\Rightarrow$ (2). Suppose $C \subseteq (A)_s \cap (B)_t$, $A \subseteq (C)_s$, and $B \subseteq (C)_t$. Since $C \subseteq (A)_s$, we know $d(C, A) \leq s$, and since $A \subseteq (C)_s$, we know $d(A, C) \leq s$. Similarly, $d(B, C) \leq t$ and $d(C, B) \leq t$. Thus $h(A, C) \leq s$ and $h(C, B) \leq t$. Since $s + t = h(A, B) \leq h(A, C) + h(C, B)$, and $h(A, C) \leq s$ and $h(C, B) \leq t$, we conclude that $h(A, C) = s$ and $h(C, B) = t$. Therefore, C is between A and B and $h(A, C) = s$.

(2) $\Rightarrow$ (1). Suppose C is between A and B, and $h(A, C) = s$. Since C is a distance s from A, by [Braun et al. 2005, Theorem 2], we know that $C \subseteq (A)_s$. Given ACB and $h(A, C) = s$, we also have that $h(C, B) = t = h(A, B) - s$. Then, we know that $C \subseteq (B)_t$, by [Braun et al. 2005, Theorem 2]. Thus, $C \subseteq (A)_s \cap (B)_t$. Also, we know $d(a, C) \leq d(A, C) \leq h(A, C) = s$ for all $a \in A$, so $A \subseteq (C)_s$. Likewise, $d(b, C) \leq h(B, C) = t$ for all $b \in B$, so $B \subseteq (C)_t$. $\square$
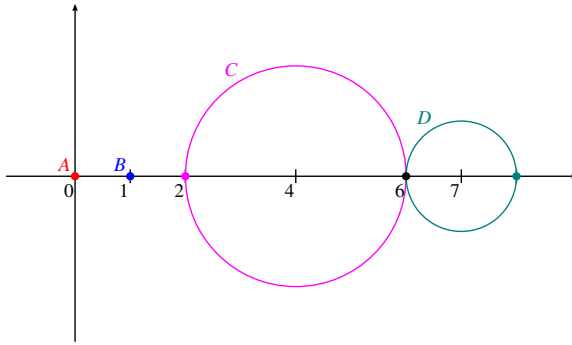
Figure 3. D in $\overrightarrow{AC}$.

Our first result addresses the question of whether rays $\overrightarrow{AC}$ and $\overrightarrow{AD}$ must be the same if C and D both lie on ray $\overrightarrow{AB}$.

Proposition 2.2. *For $A, B, C, D \in \mathcal{H}(\mathbb{R}^n)$, if $C \in \overrightarrow{AB}$ with ABC, and $D \in \overrightarrow{AC}$ with ACD, then $D \in \overrightarrow{AB}$ with ABD and BCD.*

Proof. It is not difficult to show that ABD is equivalent to BCD, so we focus on the latter. Let $s = h(A, B)$ and $t = h(B, C)$. Then ABC implies $h(A, C) = s + t$. Lemma 2.1 tells us that

$$B \subseteq (A)_s \cap (C)_t, \quad A \subseteq (B)_s, \quad \text{and} \quad C \subseteq (B)_t.$$

Similarly, since we have ACD, for some fixed $x, y \in \mathbb{R}$ with $x, y > 0$, $x + y = h(A, D)$, and $h(A, C) = x$, it follows that

$$C \subseteq (A)_x \cap (D)_y, \quad A \subseteq (C)_x, \quad \text{and} \quad D \subseteq (C)_y.$$

Thus C must be a subset of both $(B)_t$ and $(A)_x \cap (D)_y$, from which it follows that

$$C \subseteq (B)_t \cap (A)_x \cap (D)_y.$$

Hence we have $C \subseteq (B)_t \cap (D)_y$, $B \subseteq (C)_t$, and $D \subseteq (C)_y$. We conclude that C is between B and D by Lemma 2.1. $\qquad \square$

One consequence of Proposition 2.2 is the following corollary, whose proof is left to the reader.

Corollary 2.3. *For $A, B, C, D \in \mathcal{H}(\mathbb{R}^n)$, if $C \in \overrightarrow{AB}$ with ACB, and $D \in \overrightarrow{AC}$ with ADC, then $D \in \overrightarrow{AB}$.*

It is not always true that $\overrightarrow{AB}$ contains the same sets as $\overrightarrow{AC}$. The following example demonstrates that $C \in \overrightarrow{AB}$ with ABC and $D \in \overrightarrow{AB}$ does not necessarily imply that $D \in \overrightarrow{AC}$.

Example 2.4. Let $A = \{(0, 0)\}$, $B = \{(1, 0)\}$, C be the circle of radius 2 centered at $(4, 0)$, and D be the unit circle centered at $(7, 0)$ in $\mathcal{H}(\mathbb{R}^2)$ as illustrated in Figure 3.

Note that $h(A, B) = 1$, $h(B, C) = 5$, and $h(A, C) = 6$, so $C \in \overrightarrow{AB}$. Similarly, $h(B, D) = 7$, and $h(A, D) = 8$, so $D \in \overrightarrow{AB}$. However, $h(C, D) = 4$, so $D \notin \overrightarrow{AC}$. Therefore, $D \in \overrightarrow{AB}$ does not necessarily imply that $D \in \overrightarrow{AC}$.

3. Pythagorean triples in $\mathcal{H}(\mathbb{R}^n)$

The results in the previous section about rays provide some caution about the idea of defining angle measure. We can define an angle to be a union of two rays that emanate from a given point, but we might expect the measure of an angle, if possible, to be more complicated than in Euclidean geometry.

To consider angle measure in $\mathcal{H}(\mathbb{R}^n)$, we are motivated by an approach used by Wildberger [2005], who presents an alternative to classical trigonometry that does not rely on the general notion of angle. The concept of *spread*, or proportion of distance, utilizes the idea of orthogonality. In order to use this approach in $\mathcal{H}(\mathbb{R}^n)$, we will need to define and understand orthogonality. We define orthogonality in $\mathcal{H}(\mathbb{R}^n)$ as Pythagorean orthogonality.

Definition 3.1. The sets A, B, and Q in $\mathcal{H}(\mathbb{R}^n)$ form a *Pythagorean triple* (or right triangle with $\overline{AQ}$ as hypotenuse) if

$$h(A, B)^2 + h(B, Q)^2 = h(A, Q)^2.$$

Example 3.2. Let $a, b, q > 0$ with $q^2 = a^2 + b^2$, and let $A = \{(a, 0)\}$, $B = \{(0, 0)\}$, and $Q = \{(0, b)\}$. Note that $h(A, B) = a$, $h(B, Q) = b$, and $h(A, Q) = q$. In this case the sets A, B, and Q form a Pythagorean triple, and we can see that the idea of Pythagorean triples in $\mathcal{H}(\mathbb{R}^n)$ is really a generalization of the concept of Pythagorean orthogonality in $\mathbb{R}^n$. An example of infinite sets that form a Pythagorean triple is the collection A, B, and Q, where A is the circle centered at the origin of radius 1, B is the circle centered at the origin of radius 4, and Q is the disk centered at the origin of radius 6. In this case we have $h(A, B) = 3$, $h(B, Q) = d(Q, B) = 4$, and $h(A, Q) = d(Q, A) = 5$.

The question we want to address now is, given sets $A, B \in \mathcal{H}(\mathbb{R}^n)$, must there exist a set (or sets) Q that forms a Pythagorean triple with $\overline{AB}$ as hypotenuse such that $h(A, Q) = s$? In fact, we will prove that there are infinitely many such Q at a fixed location s from A. It is important to note that any such set Q must lie within the intersection $(A)_s \cap (B)_{\sqrt{r^2 - s^2}}$. For the remainder of this section we set the conditions that

- A and B are in $\mathcal{H}(\mathbb{R}^n)$ with $r = h(A, B) = d(A, B) > 0$;
- s and t are positive numbers with $r^2 = s^2 + t^2$ (note that this implies $r > s$, $r > t$, and $s + t > r$); and
- $Q_s = (A)_s \cap (B)_t$.

We will refer to these conditions as our *Pythagorean conditions*. We present several lemmas that we will use to establish our first result about Pythagorean triples in $\mathcal{H}(\mathbb{R}^n)$. By $N_\epsilon(q)$ we mean the open ϵ-neighborhood $\{x \in \mathbb{R}^n : d_E(x, q) < \epsilon\}$ centered at point q. We denote the closure of a set S as $\bar{S}$ and the boundary of S as ∂S.

We state the first lemma, which is [Blackburn et al. 2009, Lemma 2.3].

Lemma 3.3. *Let A and B be elements of $\mathcal{H}(\mathbb{R}^n)$. If $d(B, A) > 0$, then there exist $b \in B$ and $a \in \partial A$ such that $d_E(b, a) = d(b, A) = d(B, A)$.*

Lemma 3.4. *Let $B \in \mathcal{H}(\mathbb{R}^n)$ and let $t > 0$. If $y \in \partial(B)_t$, then $d(y, B) = t$.*

The proof of Lemma 3.4 is straightforward and is left to the reader.

We know that if $h(A, B) = r$ and $u + v = r$, then $h(A, (A)_u \cap (B)_v) = u$ and $h(B, (A)_u \cap (B)_v) = v$. However, in the case where $r^2 = s^2 + t^2$ (so that $s + t \neq r$), we cannot conclude that $h(A, Q_s) = s$ and $h(B, Q_s) = t$, but we do have the inequalities.

Lemma 3.5. *Given the Pythagorean conditions,*

$$h(A, Q_s) \leq s \quad and \quad h(B, Q_s) \leq t.$$

Proof. We will demonstrate that $h(A, Q_s) \leq s$. The argument that $h(B, Q_s) \leq t$ is similar and is left to the reader. We first show that $d(A, Q_s) \leq s$. Let $a \in A$, and let $b \in B$ such that $d_E(a, b) = d(a, B) \leq d(A, B) = r$. Let $x \in \overrightarrow{ab}$ with $d_E(a, x) = s$. If $d_E(a, x) \geq d_E(a, b)$, then $b \in (A)_s$ and so $b \in Q_s$. If $d_E(a, x) < d_E(a, b)$, then $x \in \overline{ab}$. So $d_E(b, x) = d_E(a, b) - s \leq r - s < t$ and $x \in (A)_s \cap (B)_t$. In either case we have $d(a, Q_s) \leq s$, which demonstrates that $d(A, Q_s) \leq s$.

The fact that $Q_s \subseteq (A)_s$ implies that $d(Q_s, A) \leq s$. Therefore, we have $h(A, Q_s) \leq s$. □

To demonstrate that there are infinitely many sets Q such that A, B, and Q form a Pythagorean triple with $\overline{AB}$ as hypotenuse, we will next show that we can remove a small neighborhood from Q_s without affecting the inequalities in Lemma 3.5.

Lemma 3.6. *Given the Pythagorean conditions, let $a \in A$ and $b \in B$ such that $d_E(a, b) = r$. Then $N_\epsilon(q)$ is in the interior of Q_s, where $\epsilon = \frac{1}{2}(s + t - r)$ and $q \in \overline{ab}$ with $d_E(q, b) = t - \epsilon$.*

Proof. It is not difficult to show that $r > t - \epsilon > 0$ and so q exists. Let $x \in N_\epsilon(q)$. Then

$$d_E(x, b) \leq d_E(x, q) + d_E(q, b) < \epsilon + (t - \epsilon) = t.$$

Now

$$d_E(a, q) = d_E(a, b) - d_E(b, q) = s - \epsilon,$$

so it follows that $d_E(x, a) < s$. □
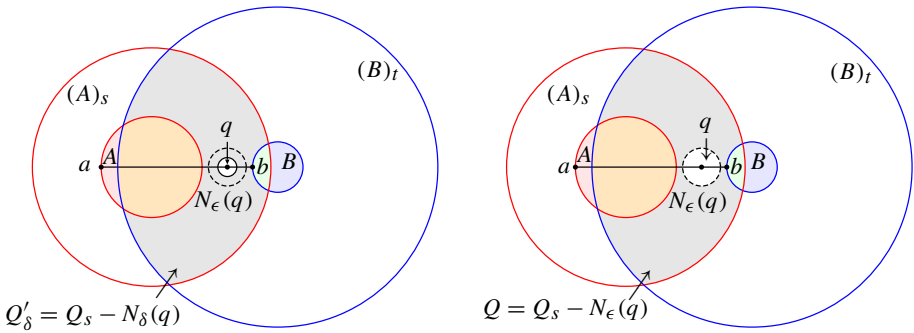
Figure 4. Left: $Q'_\delta = Q_s \setminus N_\delta(q)$. Right: $Q = Q_s \setminus N_\epsilon(q)$.

Lemma 3.7. *Given the Pythagorean conditions, let $a \in A$ and $b \in B$ such that $d_E(a, b) = r$, let $\epsilon = \frac{1}{2}(s + t - r)$, and let $q \in \overline{ab}$ with $d_E(q, b) = t$. Let $0 \le \delta < \epsilon$ and let $Q'_\delta = Q_s \setminus N_\delta(q)$. Then $h(A, Q'_\delta) \le s$ and $h(B, Q'_\delta) \le t$.*

Proof. A picture illustrating the theorem is shown in Figure 4, left (where A is the disk centered at $(-3, 0)$ of radius 2 and B is the disk centered at $(2, 0)$ with radius 1 in $\mathbb{R}^2$). First note that $\epsilon > 0$. Also note that Q'_δ is closed and a subset of Q_s (that is, Q'_δ is just Q_s with a neighborhood around a point in its interior removed), so it is an element in $\mathcal{H}(\mathbb{R}^n)$. We know that $h(A, Q_s) \le s$ by Lemma 3.5. Since $Q'_\delta \subseteq Q_s$ and $d(x, A) \le s$ for every $x \in Q_s$, we have $d(x, A) \le s$ for every $x \in Q'_\delta$. So $d(Q'_\delta, A) \le d(Q_s, A) \le s$. Now we show that $d(A, Q'_\delta) \le s$.

Let $x \in A$. We consider two cases. First, suppose that $x \in N_\delta(q)$. Let q' be the point on the boundary of $\overline{N_\delta(q)}$ closest to x. Since Q'_δ is closed, it follows that $\partial \overline{N_\delta(q)} \subseteq Q'_\delta$. So $q' \in Q'_\delta$ and $d_E(x, q') \le \delta \le s$ (note that $t < r$ and so $s > \frac{1}{2}(s + t - r) = \epsilon$). Thus, $d(x, Q'_\delta) \le s$. For the second case, assume that $x \notin N_\delta(q)$. Let $q_x \in Q_s$ such that $d_E(x, q_x) = d(x, Q_s) \le h(A, Q_s) \le s$. If $q_x \notin N_\delta(q)$, then $q_x \in Q'_\delta$ and $d(x, Q'_\delta) \le s$. If $q_x \in N_\delta(q)$, let q' be the point on $\partial N_\delta(q) \cap \overline{xq_x}$. Then $d_E(x, q') < d_E(x, q_x) \le s$ and $d(x, Q'_\delta) \le s$. Therefore, $d(x, Q'_\delta) \le s$ for every $x \in A$ and $d(A, Q'_\delta) \le s$. A similar argument shows $h(B, Q'_\delta) \le t$. □

Theorem 3.8 will demonstrate that, under certain conditions, we have infinitely many Pythagorean triples with $\overline{AB}$ as hypotenuse.

Theorem 3.8. *Given the Pythagorean conditions, let $a \in A$ and $b \in B$ such that $d_E(a, b) = r$, let $\epsilon = \frac{1}{2}(s + t - r)$, and let $q \in \overline{ab}$ with $d_E(q, b) = t$. Let $0 \le \delta < \epsilon$ and let $Q'_\delta = Q_s \setminus N_\delta(q)$. If $Q_s \cap \partial(A)_s \ne \varnothing$ and $Q_s \cap \partial(B)_t \ne \varnothing$, then A, B, and Q'_δ form a Pythagorean triple with $h(A, Q'_\delta) = s$.*

Proof. An illustration of Theorem 3.8 is shown in Figure 4. Lemma 3.7 shows that $h(A, Q'_\delta) \le s$. To verify that $h(A, Q'_\delta) = s$ we use the hypothesis that Q_s

contains a point $z \in \partial(A)_s$. Lemma 3.4 shows that $d(z, A) = s$. The proof of Lemma 3.7 demonstrated that if $x \in N_\delta(q)$, then $d(x, A) < s$. Therefore, $z \in Q'_\delta$ and $d(Q'_\delta, A) = s$. A similar argument shows that $h(Q'_\delta, B) = t$. $\square$

Theorem 3.8 shows that under certain conditions, there are infinitely many sets Q'_δ such that A, B, and Q'_δ form a Pythagorean triple with $h(A, Q'_\delta)$ the same for every δ. In other words, there can be infinitely many different Pythagorean triples with a fixed $\overline{AB}$ as hypotenuse. The next question we address is if this is always the case. In other words, can we find Pythagorean triples with $\overline{AB}$ as hypotenuse if Q_s does not contain boundary points of $(A)_s$ or $(B)_t$?

Lemma 3.9 shows that we cannot have $Q_s \cap \partial(B)_t = \varnothing$, and helps us understand when $Q_s \cap \partial(A)_s \neq \varnothing$.

Lemma 3.9. *Assume the Pythagorean conditions*:

(1) *If $0 < s < d(B, A)$, then $Q_s \cap \partial(A)_s \neq \varnothing$.*

(2) *$Q_S \cap \partial(B)_t \neq \varnothing$.*

Proof. Assume $0 < s < d(B, A)$. Let $b \in B$ and $a \in A$ such that $d_E(b, a) = d(b, A) = d(B, A) > s$. Let $x \in \overline{ba}$ such that $d_E(a, x) = s$. Thus, $x \in (A)_s$ and $x \in (B)_t$, and $x \in Q_s$. Now let $z \in \overline{bx}$ with $d_E(z, x) > 0$. We will show that $z \notin (A)_s$. Suppose to the contrary that $z \in (A)_s$. Then there is an $a_z \in A$ with $d_E(z, a_z) \le s$. Applying the triangle inequality along with the fact that $d_E(b, x) = d_E(b, a) - s = d(b, A) - s$ shows that

$$d_E(b, a_z) < d(b, A) - s + s = d(b, A),$$

which is impossible. We conclude that $z \notin (A)_s$ and so every neighborhood around x contains a point in $(A)_s$ and a point not in $(A)_s$. Therefore, $x \in \partial(A)_s$ and $Q_s \cap \partial(A)_s \neq \varnothing$.

The proof of the second assertion follows the same argument as part (1), noting that t is always between 0 and $r = d(A, B)$. $\square$

Note that we cannot draw any conclusions about $Q_s \cap \partial(A)_s$ if $s \ge d(B, A)$, as the next examples illustrate.

Example 3.10. Let A be an annulus centered at the origin with inner radius r_A and outer radius R_A and let B be the single point set consisting of the point at the origin, as shown in Figure 5, left. Then $0 < d(B, A) = r_A < d(A, B) = R_A$. In this case, if $s \ge d(B, A)$, then $Q_s = (B)_t$ and $Q_s \cap \partial(A)_s = \varnothing$.

Example 3.11. Let $A = \{(0, 0), (1, 0)\}$ and $B = \{(-1, 0)\}$ in $\mathbb{R}^2$, as shown in Figure 5, right. Then $0 < d(B, A) = 1 < d(A, B) = 2$. In this case, $s \ge d(B, A)$ does not imply $Q_s \cap \partial(A)_s = \varnothing$. In fact, here we will have $Q_s \cap \partial(A)_s = \varnothing$ only when $1 + s > t$.
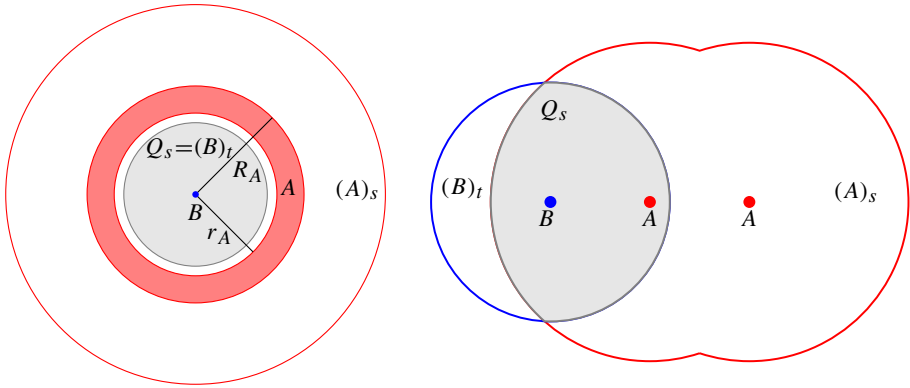
Figure 5. Left: $Q_s \cap \partial(A)_s = \varnothing$. Right: $Q_s \cap \partial(A)_s \neq \varnothing$.

As a consequence of Theorem 3.8 and Lemma 3.9, to determine if we can always find a Pythagorean triple with $\overline{AB}$ as hypotenuse we only have to consider the remaining case when $s \geq d(B, A)$ and $Q_s \cap \partial(A)_s = \varnothing$. The next theorem shows what happens when Q_s does not contain a boundary point of $(A)_s$.

Theorem 3.12. *Assume the Pythagorean conditions. If $s \geq d(B, A)$ and Q_s does not contain a boundary point of $(A)_s$, then $(B)_t \subseteq (A)_s$.*

Proof. Assume that $s \geq d(A, B)$ and that Q_s does not contain a boundary point of $(A)_s$. Now suppose to the contrary that $(B)_t \nsubseteq (A)_s$. Then there is a point $y \in (B)_t - (A)_s$. Since $y \in (B)_t$, there exists $b \in B$ such that $d_E(y, b) = d(y, B) \leq t$. Also, $b \in B$ implies there is a point $a \in A$ such that $d_E(b, a) = d(b, A) \leq d(B, A)$. Since $y \notin (A)_s$, we know that $d_E(y, a) > s$. Let $x \in \overline{ay}$ with $d_E(a, x) = s$.

Claim. $\overline{xy} \subseteq (\{b\})_{d_E(b, y)}$.

Proof of the claim. Let $z \in \overline{xy}$ and construct the triangles $\triangle ayb$ and $\triangle azb$ as shown in Figure 6. Let θ be the angle formed at point a. The law of cosines shows that

$$d_E(b, y)^2 = d_E(b, a)^2 + d_E(a, y)^2 - 2d_E(b, a)d_E(a, y)\cos\theta.$$

Substituting $d_E(a, z) + d_E(z, y)$ for $d_E(a, y)$ and using the law of cosines again yields

$$d_E(b, y)^2 = d_E(b, z)^2 + d_E(z, y)\big(2d_E(a, z) + d_E(z, y) - 2d_E(b, a)\cos\theta\big).$$

Now $d_E(a, z) \geq s \geq d(B, A) \geq d_E(b, a)$, so

$$2d_E(a, z) + d_E(z, y) - 2d_E(b, a)\cos\theta \geq 2s + d_E(z, y) - 2d_E(b, a)\cos\theta$$

$$\geq 2d_E(b, a) + d_E(z, y) - 2d_E(b, a)\cos\theta$$

$$= d_E(z, y) + 2d_E(b, a)(1 - \cos\theta) > 0.$$
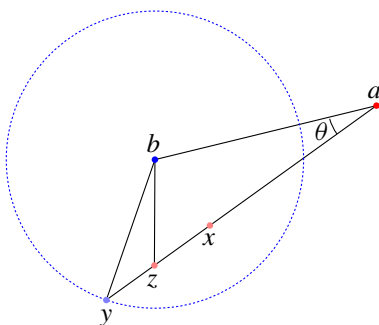
Figure 6. Showing $d_E(b, x) \leq t$.

Therefore,

$$d_E(b, y)^2 > d_E(b, z)^2$$

and $d_E(b, z) < d_E(b, y)$ as desired, completing the proof of the claim. □

We proceed with the proof of Theorem 3.12. The set $G = (A)_s \cap \overline{xy}$ is the intersection of two compact sets and so is compact. Moreover, since $x \in (A)_s$, the set G is nonempty. Now let $w \in G$ such that $d_E(a, w) = \max_{g \in G}\{d_E(a, g)\}$. In other words, w is the point on $\overline{xy}$ in $(A)_s$ closest to y. Since $y \notin (A)_s$ we know that $w \neq y$. Therefore, no point in $\overline{yw}$ other than w can be in $(A)_s$. Thus, we have that $w \in \partial(A)_s$. Since $w \in (\{b\})_{d_E(b, y)}$ as well, we have found a point in $Q_s \cap \partial(A)_s$, a contradiction. We conclude that $(B)_t \subseteq (A)_s$. □

Theorem 3.8, Lemma 3.9, and Theorem 3.12 combine to leave one remaining case to consider to determine if we can always find a Pythagorean triple with $\overline{AB}$ as hypotenuse: when Q_s does not contain a boundary point of $(A)_s$, or $(B)_t \subseteq (A)_s$.

Theorem 3.13. *Assume the Pythagorean conditions. Let $a \in A$ and $b \in B$ such that $d_E(a, b) = d(a, B) = d(A, B) = r$. Let $Q = (B)_t \setminus N_s(a)$. If $(B)_t \subseteq (A)_s$, then $h(A, Q) = s$ and $h(B, Q) = t$ and the sets A, B, and Q form a Pythagorean triple with $\overline{AB}$ as hypotenuse.*

Proof. To show that A, B, and Q form a Pythagorean triple we need to know that Q is not empty, $h(A, Q) = s$, and $h(B, Q) = t$. First we show that Q is not empty.

Since $r > s$ there exists a $c \in \overline{ab}$ such that $d_E(a, c) = s$. Then $c \notin N_s(a)$. We also have

$$d_E(b, c) = r - s < t,$$

and $c \in (B)_t$. Therefore, $Q \neq \varnothing$.

Next we prove that $h(A, Q) = s$. The fact that $Q \subseteq (B)_t \subseteq (A)_s$ implies that $d(Q, A) \leq s$.

Now we demonstrate that $d(A, Q) \leq s$. Let $a' \in A$. There exists $b' \in B$ such that $d_E(a', b') = d(a', B) \leq \max_{x \in A}\{d(x, B)\} = d(a, B) = r$. Note that $d_E(a', b') \leq r = d(a, B) \leq d_E(a, b')$. Since $d(a, B) = r$ and $N_s(a) \subseteq N_r(a)$, we know that $b' \notin N_s(a)$ and $b' \in Q$. If $d_E(a', b') \leq s$, then $d(a', Q) \leq s$. Now assume that $d_E(a', b') > s$. Let $c' \in \overline{a'b'}$ such that $d_E(a', c') = s$. Then

$$d_E(a, b') \leq d_E(a, c') + d_E(c', b')$$

so

$$d_E(a, c') \geq d_E(a, b') - d_E(c', b') \geq d_E(a', b') - d_E(c', b') = d_E(a', c') = s.$$

Thus, $c' \notin N_s(a)$. Also,

$$d_E(b', c') = d_E(a', b') - d_E(a', c') \leq r - s < t$$

and so $c' \in Q$. Therefore, $d(a', Q) \leq s$. We conclude that $d(A, Q) \leq s$.

Now, we show that $d(a, Q) = s$. Since $Q = (B)_t \setminus N_s(a)$, we see that $d(a, Q) \geq s$. Recall $c \in Q$ with $d_E(a, c) = s$. So $d(a, Q) = d_E(a, c) = s$. Therefore $d(A, Q) = s$ and $h(A, Q) = s$.

Next we prove that $h(B, Q) = t$. Recall that $Q \subseteq (B)_t \subseteq (A)_s$, so it follows that $d(Q, B) \leq t$.

Now we demonstrate that $d(B, Q) \leq t$. Let $b^* \in B$. There exists $a^* \in A$ such that $d_E(b^*, a^*) = d(b^*, A) \leq r$. Note that $d_E(b^*, a^*) \leq r = d_E(a, b) \leq d_E(a, b^*)$. Since $N_s(a) \subseteq N_r(a)$, we know that $b^* \notin N_s(a)$ and $b^* \in Q$. So $d(b^*, Q) = 0$ and $d(B, Q) = 0 < t$.

Finally, we show that $d(Q, B) = t$. Let $W = \overrightarrow{ab} \cap (B)_t$ and let $w \in W$ such that $d_E(a, w)$ is a maximum. By definition, $w \in (B)_t$. Let $x \in \overrightarrow{bw}$ such that $d_E(b, x) = t$. (Note that $d_E(b, x) \leq d_E(b, w)$.) Then $x \in W$ and so $d_E(a, w) \geq d_E(a, b) + d_E(b, x) = r + t > s$. Thus, $w \notin N_s(a)$ and $w \in Q$. If $w' \in \overrightarrow{ab}$ with $d_E(a, w') > d_E(a, w)$, then $w' \notin W$ and so $w' \notin (B)_t$. Thus, $w \in \partial(B)_t$. Lemma 3.4 shows that $d(w, B) = t$ and so $d(Q, B) = t$. $\square$

Corollary 3.14. *Assume the Pythagorean conditions. If $(B)_t \subseteq (A)_s$, then there are infinitely many sets $Q \in \mathcal{H}(\mathbb{R}^n)$ that form a Pythagorean triple with A and B with $\overline{AB}$ as hypotenuse and $h(A, Q) = s$.*

Proof. Figure 7 will be a useful reference for this proof. Let $a \in A$ and $b \in B$ such that $d_E(a, b) = d(A, B) = h(A, B) = r$. Let $C = (B)_t \setminus N_s(a)$. Let $\mu = \min\{\frac{1}{2}(s + t - r), r - s\}$ and $q \in \overline{ab}$ such that $d_E(a, q) = s + \mu$. Thus, $q \notin N_s(a)$. Since $d_E(a, q) + d_E(q, b) = d_E(a, b) = r$ and $d_E(a, q) = s + \mu$, we have that

$$d_E(b, q) = d_E(a, b) - d_E(a, q) = r - (s + \mu) < t - \mu < t.$$

We conclude that $q \in N_t(b) \subset (B)_t$ and $q \notin N_s(a)$. It follows $q \in (B)_t \setminus N_s(a) = C$.
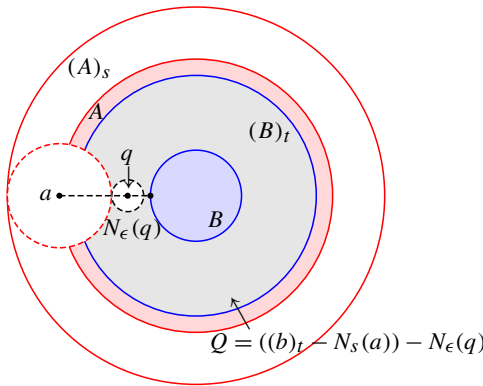
Figure 7. The shaded set is $Q = C - N_\epsilon(q)$.

Now we demonstrate that $N_\mu(q) \subseteq N_t(b) \subseteq (B)_t$. Let $z \in N_\mu(q)$. Two applications of the triangle inequality show that $d_E(a, z) > s$ and $d_E(b, z) < t$. It follows that $z \notin N_s(a)$ and $N_\mu(q) \cap N_s(a) = \varnothing$, and that $z \in N_t(b)$ and $N_\mu(q) \subseteq N_t(b) \subset (B)_t$. Now let $0 < \epsilon < \mu$ and let $Q = C - N_\epsilon(q)$. We will demonstrate that A, B, and Q form a Pythagorean triple.

Because Q is a closed subset of C, it is an element of $\mathcal{H}(\mathbb{R}^n)$. Theorem 3.13 shows that $h(A, C) = s$ and $h(B, C) = t$. Let $x \in C$. Since $Q \subseteq C$ and $d(x, A) \leq s$, we have that $d(Q, A) \leq d(C, A) \leq s$. Likewise, $d(x, B) \leq t$, so $d(Q, B) \leq d(C, B) \leq t$.

Let $x \in A$. As in the proof of Lemma 3.7, we can show that $d(x, Q) \leq s$ and $d(A, Q) \leq s$. We proceed to prove that $h(A, Q) = s$. Theorem 3.13 shows that for $c \in \overline{ab}$ with $d_E(a, c) = s$, we have $d_E(a, c) = d(a, C) = s$. Now

$$d_E(c, q) = d_E(a, q) - d_E(a, c) = (s + \mu) - s = \mu > \epsilon,$$

so $c \in Q$ and $d(A, Q) = s$.

The proof that $h(B, Q) = t$ is similar and is left to the reader. □

In summary, in this section we have demonstrated that for any distinct sets A and B in $\mathcal{H}(\mathbb{R}^n)$ and any $0 < s < h(A, B) = d(A, B)$, there are infinitely many sets Q such that A, B, and Q form a Pythagorean triple with $\overline{AB}$ as hypotenuse and $h(A, Q) = s$.

4. Projections

To continue our attempt to develop angle measure in $\mathcal{H}(\mathbb{R}^n)$, we return to Wildberger's approach. To measure *spread*, we will need to determine if, given two rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$ in $\mathcal{H}(\mathbb{R}^n)$, we can find a set P on the ray $\overrightarrow{AC}$ that creates a Pythagorean triple with $\overline{AB}$ as hypotenuse. We will call such a set P a *projection* of B onto $\overleftrightarrow{AC}$.

The previous section shows that there are infinitely many Pythagorean triples to consider, but we need to know if there is a specific one that can be found on a given ray.

Unfortunately, Example 4.1 will demonstrate that our quest for projections will not always be successful.

Example 4.1. Let $A = \{a\}$, $B = \{b\}$ with $h(a, b) = d_E(a, b) = r > 0$, and let c be a point such that $\overline{ab} \perp \overline{ac}$. Let $q = d_E(a, c)$. Choose c_1 so that bac_1 and $\sqrt{r^2 + q^2} - r < d_E(a, c_1) < q$ (since $q > \sqrt{r^2 + q^2} - r$ we can find many such c_1), and let $C = \{c, c_1\}$. We will show that there is no set P that lies on $\overleftrightarrow{AC}$ and forms a Pythagorean triple with A and B.

Observe that $h(A, B) = d_E(a, b) = r$ and that $h(A, C) = d(C, A) = d_E(c, a) = q$. Note that

$$h(B, C) = d_E(c_1, b) = r + d_E(a, c_1) > \sqrt{r^2 + q^2}$$

so A, B, and C themselves do not form a Pythagorean triple. Suppose to the contrary that there is a set P such that $P \in \overleftrightarrow{AC}$ and P forms a Pythagorean triple with A and B with hypotenuse $\overline{AB}$. Let $h(A, P) = s$, where $0 < s < r$. Then since P forms a Pythagorean triple with A and B, $h(B, P) = t = \sqrt{r^2 - s^2}$. First, we consider the case when P is between A and C, implying that $h(P, C) = q - s$ and $q > s$. We note that $P \subseteq (A)_s \cap (B)_t \cap (C)_{q-s}$. It follows that $P \subseteq (A)_s \cap (C)_{q-s}$. Let $W = (A)_s \cap (C)_{q-s}$. Lemma 1.7 shows that W is between A and C. Let $w \in \overline{ac}$ such that $d_E(a, w) = s$. Then $d_E(w, c) = d_E(a, c) - d_E(a, w) = q - s$ and $w \in W$.

Now we will show that $w \in P$. Since $(A)_s$ is a disk of radius s, the only point of $(A)_s$ that lies on $\overline{ac}$ that is a distance s from a is w. Let $a_s \in (A)_s$ such that $a_s \neq w$. Then $d_E(a_s, a) \leq s$ and the triangle inequality shows that

$$d_E(c, a_s) \geq d_E(a, c) - d_E(a_s, a) \geq q - s.$$

Since $a_s \neq w$, we must have $d_E(c, a_s) > q - s$. Now $d(C, P) \geq d(c, P)$ and $d(c, P)$ is achieved at some point $p \in P \subseteq (A)_s$. If $p \neq w$, then

$$q - s = h(C, P) \geq d(C, P) = d_E(c, p) > q - s.$$

We conclude that $p = w$ and $w \in P$.

Since $w \in P$, we must have that $w \in (B)_t = (\{b\})_t$ as well. The points a, b, and w form a right triangle as in Figure 8, with $d_E(b, w) = \sqrt{r^2 + s^2}$. Since $w \in (B)_t$ it follows that $d_E(b, w) \leq t$. Recall that $t = \sqrt{r^2 - s^2}$, so $d_E(b, w) \leq \sqrt{r^2 - s^2}$. But this makes $\sqrt{r^2 - s^2} \geq \sqrt{r^2 + s^2}$, which is impossible since $s > 0$. Thus, no set P exists such that P forms a Pythagorean triple with A and B, and P is between A and C.

For the second case, suppose that such a P exists such that A is between P and C and $h(A, P) = s$. In this case we have $h(C, P) = q + s$ and $P \subseteq (A)_s \cap (B)_t \cap (C)_{q+s}$. Again, we begin by examining the characteristics of $(A)_s \cap (C)_{q+s}$.
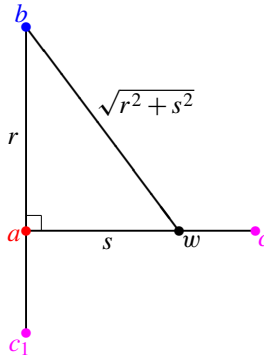
Figure 8. A right triangle with vertices a, b, and w.

Let $a_s \in (A)_s$. Then $d_E(a, a_s) \leq s$. Since $d_E(c, a) = q$, by the triangle inequality we have

$$d_E(c, a_s) \leq d_E(c, a) + d_E(a, a_s) \leq q + s.$$

It follows that $a_s \in (C)_{q+s}$ and $(A)_s \subseteq (C)_{q+s}$, so $P \subseteq (A)_s \cap (B)_t$.

Let $w \in \overleftrightarrow{ac}$ with wac and $d_E(a, w) = s$. Now we will show that $w \in P$. Note that $d(A, C) = d_E(a, c_1) < d_E(a, c) = q$. Let $w' \in P$. Since $P \subseteq (A)_s$ we know that $w' \in (A)_s$ and we must have $d_E(w', a) \leq s$. By the triangle inequality,

$$d_E(w', c_1) \leq d_E(a, c_1) + d_E(a, w') < q + s.$$

Thus, $d(w', C) < q + s$ and $d(P, C) < q + s$. Now let us turn to $d(C, P)$.

The above argument shows that $d_E(c_1, w') < q + s$, from which it follows that $d(c_1, P) < q + s$. From the triangle inequality we can see that

$$d_E(c, w') \leq d_E(c, a) + d_E(a, w') \leq q + s.$$

Note that if $d_E(c, w') = q + s$ then $w' \in \overleftrightarrow{ac}$, $d_E(a, w') = s$, and $w'ac$. This forces $w' = w$. So in order to have $h(P, C) = q + s$ we must have $w \in P$.

Since $w \in P$ and $P \subseteq (B)_t$, we have that $w \in (B)_t$ as well. The points a, b, w form a right triangle as in [Figure 8](#). We can see that $d_E(b, w) = \sqrt{r^2 + s^2}$, but $d_E(b, w) \leq t = \sqrt{r^2 - s^2}$. This forces $\sqrt{r^2 - s^2} \geq \sqrt{r^2 + s^2}$, which is impossible since $s > 0$. Therefore, there does not exist a set P that forms a Pythagorean triple with A and B such that A is between P and C.

The remaining case where P satisfies ACP is similar and is left to the reader.

Therefore, there exists no set $P \in \overleftrightarrow{AC}$ that forms a Pythagorean triple with A and B such that $\overline{AB}$ is a hypotenuse.

[Example 4.1](#) illustrates the extreme case that there is no set P that lies on $\overleftrightarrow{AB}$ that makes a Pythagorean triple with A and B with $\overline{AB}$ as hypotenuse. We now
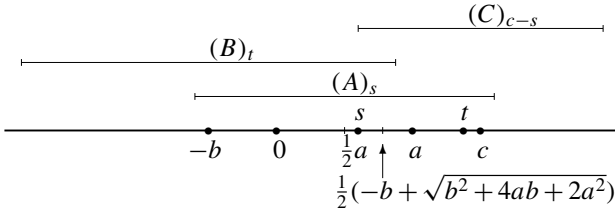
Figure 9. Infinitely many sets P.

show that the other extreme is possible — that there can be infinitely many sets P that satisfy both the betweenness and Pythagorean triple conditions.

Example 4.2. Let $A = \{0, a\}$, $B = \{-b\}$, and $C = \{c\}$ in $\mathcal{H}(\mathbb{R})$, where $a, b, c > 0$ and $c > a$. First we note that $h(A, B) = d(A, B) = a + b$ and $h(A, C) = d(A, C) = c$. Since $a, b > 0$, a little algebra shows that

$$\tfrac{1}{2}a < \tfrac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2}).$$

So for any $a, b > 0$ the interval between $\tfrac{1}{2}a$ and $\tfrac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2})$ is not empty. Let

$$\tfrac{1}{2}a \le s \le \tfrac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2})$$

and assume in addition that $c \ge s + \tfrac{1}{2}a$. Let $t = \sqrt{h(A, B)^2 - s^2} = \sqrt{(a + b)^2 - s^2}$. It follows that

$$(A)_s = [-s, a + s], \quad (B)_t = [-b - t, t - b], \quad \text{and} \quad (C)_{c-s} = [s, 2c - s].$$

See Figure 9 for an illustration. Now let us determine $(A_s) \cap (C)_{c-s}$. The fact that $c \ge s + \tfrac{1}{2}a$ implies that $2c - s \ge s + a$, and so

$$(A)_s \cap (C)_{c-s} = [s, a + s].$$

Now let $P_s = (A)_s \cap (B)_t \cap (C)_{c-s}$. We will show that $s \le t - b \le a + s$, which means that $P_s = [s, t - b]$.

Let $y = 2s^2 + 2bs - (2ab + a^2)$. Using the quadratic formula and the fact that $\tfrac{1}{2}(-b - \sqrt{b^2 + 4ab + 2a^2}) < 0 < \tfrac{1}{2}a$, we have $\tfrac{1}{2}a \le s \le \tfrac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2})$ and

$$s \le t - b \tag{1}$$

when $\tfrac{1}{2}a \le s \le \tfrac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2})$. Since $s \le t - b$ it follows that $[s, t - b]$ is not empty.

The fact that $t = \sqrt{(a + b)^2 - s^2} \le a + b$ implies that

$$t - b \le a, \tag{2}$$

and since $s > 0$, we have $t - b < a + s$. Thus we can conclude that $P_s = [s, t - b]$ when $\frac{1}{2}a \leq s \leq \frac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2})$.

Now we check that the set P_s satisfies $h(A, B)^2 = h(B, P_s)^2 + h(A, P_s)^2$ and $h(A, P_s) + h(P_s, C) = h(A, C)$.

First, we calculate $h(A, P_s)$. We know $P_s \subseteq (A)_s$, which implies $d(P_s, A) \leq s$. Inequalities (1) and (2) show that $d(A, P_s) = \max\{d_E(0, s), d_E(a, t - b)\}$. We know $d_E(0, s) = s$ and $d_E(a, t - b) = a - (t - b)$. Inequality (1) implies

$$s \geq a - s \geq a - (t - b),$$

and so $d_E(a, t - b) \leq s = d_E(0, s)$. Thus, $h(A, P_s) = s$.

Now we determine $h(P_s, B)$. Note that since $t - b \geq s$ we have $d(B, P_s) = d_E(-b, s) \leq d(P_s, B) = d_E(t - b, -b) = t$. Therefore $h(B, P_s) = t$.

Next, we find $h(C, P_s)$. Here we have $d(C, P_s) = d_E(c, t - b) \leq d(P_s, C) = d_E(s, c) = c - s$. Therefore, $h(C, P_s) = c - s$.

Because $h(A, P_s) + h(C, P_s) = s + (c - s) = c = h(A, C)$, we have AP_sC. Additionally, notice that $h(A, P_s)^2 + h(B, P_s)^2 = s^2 + t^2 = s^2 + (a + b)^2 - s^2 = (a+b)^2 = h(A, B)^2$. Therefore, $P_s \subseteq (A)_s \cap (B)_t \cap (C)_{c-s}$ lies between A and C and forms a Pythagorean triple with A and B for $\frac{1}{2}a \leq s \leq \frac{1}{2}(-b + \sqrt{b^2 + 4ab + 2a^2})$.

Example 4.2 shows us that there are infinitely many sets A, B, C such that there exists an infinite number of sets P_s that lie between A and C and form a Pythagorean triple with A and B. It turns out that each of the sets P_s from Example 4.2 lies on the same ray $\overrightarrow{AC}$ (the proof is left to the reader).

The previous examples demonstrate that the behavior of Pythagorean triples in $\mathcal{H}(\mathbb{R}^n)$ is quite different than from those in $\mathbb{R}^n$. The situation in $\mathcal{H}(\mathbb{R}^n)$ is even stranger than we have already seen, as our final example illustrates.

Example 4.3. Let $A = \{0\}$, $B = \{-b\}$, $C = \{c\}$, for some $b, c > 0$. It is straightforward to see that BAC. For $0 < s < b$, let $t = \sqrt{b^2 - s^2}$ such that $(A)_s = [-s, s]$, $(B)_t = [-b - t, -b + t]$, and $(C)_{c+s} = [-s, 2c + s]$. Calculations similar to those in Example 4.2 show that $P_s = (A)_s \cap (B)_t \cap (C)_{c+s} = [-s, -b + t]$, that P_s satisfies $P_s AC$, and that P_s forms a Pythagorean triple with A and B. Unlike in Example 4.2 where P_s was defined only in a restricted subinterval of $(0, h(A, B))$, in this situation we can use any value for s from 0 up to $h(A, B)$. Thus not only can we form a projection from a line to itself, we can do so for every value in the interval $[0, h(A, B)]$.

5. Conclusions

While Euclidean geometry is embedded in the Hausdorff metric geometry as single point sets, the Hausdorff metric geometry is quite different. As we have seen, there can be infinitely many different sets at the same location that form a Pythagorean

triple with given sets A and B and hypotenuse $\overline{AB}$. Unfortunately, our attempt to measure angles in $\mathcal{H}(\mathbb{R}^n)$ using Pythagorean orthogonality to determine spread does not work in general since we cannot always project a given set onto a given ray (even though we can do this in infinitely many different ways in other cases). It may be that a different notion of orthogonality will allow us to proceed. For example, in Euclidean geometry the segment $\overline{ab}$ is orthogonal to the line ℓ that contains b if the distance from a to any point on ℓ is a minimum. Defining orthogonality in $\mathcal{H}(\mathbb{R}^n)$ in terms of minimum distances might provide different results.

References

[Barnsley 1993] M. F. Barnsley, *Fractals everywhere*, 2nd ed., Academic Press Professional, Boston, 1993. MR Zbl

[Bay et al. 2005] C. Bay, A. Lembcke, and S. Schlicker, "When lines go bad in hyperspace", *Demonstratio Math.* **38**:3 (2005), 689–701. MR Zbl

[Blackburn et al. 2009] C. C. Blackburn, K. Lund, S. Schlicker, P. Sigmon, and A. Zupan, "A missing prime configuration in the Hausdorff metric geometry", *J. Geom.* **92**:1-2 (2009), 28–59. MR Zbl

[Blumenthal 1953] L. M. Blumenthal, *Theory and applications of distance geometry*, Oxford Univeristy Press, 1953. MR Zbl

[Bogdewicz 2000] A. Bogdewicz, "Some metric properties of hyperspaces", *Demonstratio Math.* **33**:1 (2000), 135–149. MR Zbl

[Braun et al. 2005] D. Braun, J. Mayberry, A. Powers, and S. Schlicker, "A singular introduction to the Hausdorff metric geometry", *Pi Mu Epsilon Journal* **12**:3 (2005), 129–138.

[Edgar 1990] G. A. Edgar, *Measure, topology, and fractal geometry*, Springer, 1990. MR Zbl

[Montague 2008] D. Montague, "Finite betweenness under the Hausdorff metric", REU report, 2008.

[Wildberger 2005] N. J. Wildberger, *Divine proportions: rational trigonometry to universal geometry*, Wild Egg, Sydney, 2005. MR Zbl

pallavi054@yahoo.com   *California Institute of Technology, Pasadena, CA, United States*

schlicks@gvsu.edu   *Department of Mathematics, Grand Valley State University, Allendale, MI, United States*

rdswartzentruber@gmail.com   *Eastern Mennonite University, Harrisonburg, VA, United States*

Different definitions of conic sections in hyperbolic geometry

Patrick Chao and Jonathan Rosenberg

(Communicated by Józef H. Przytycki)

In classical Euclidean geometry, there are several equivalent definitions of conic sections. We show that in the hyperbolic plane, the analogues of these same definitions still make sense, but are no longer equivalent, and we discuss the relationships among them.

1. Introduction

Throughout this paper, $\mathbb{E}^n$ will denote Euclidean n-space and $\mathbb{H}^n$ will denote hyperbolic n-space. Recall that (up to isometry) these are the unique complete simply connected Riemannian n-manifolds with constant curvature 0 and -1, respectively. We will use $d(x, y)$ for the Riemannian distance between points x and y in either of these geometries. We will sometimes identify $\mathbb{E}^n$ with affine n-space $\mathbb{A}^n(\mathbb{R})$ over the reals, which can then be embedded as usual in projective n-space $\mathbb{P}^n(\mathbb{R})$ (the set of lines through the origin in $\mathbb{A}^{n+1}(\mathbb{R})$). While this paper is about 2-dimensional geometry, we will sometimes need to consider the case $n = 3$ as well as $n = 2$.

1.1. *Hyperbolic geometry.* *Hyperbolic geometry* is a form of non-Euclidean geometry, which modifies Euclid's fifth axiom, the parallel postulate. The parallel postulate has an equivalent statement, known as *Playfair's axiom*.

Definition 1 (Playfair's axiom). Given a line l and a point p not on l, there exists only one line through p parallel to l.

In hyperbolic geometry, this is modified by allowing an infinite number of lines through p parallel to l. This has interesting effects, resulting in the angles in a triangle adding up to less than π radians, and a relation between the area of the triangle and the *angular defect*, the difference between π and the sum of the angles.

Hyperbolic geometry may also be considered to be the Riemannian geometry of a surface of constant negative curvature. When this curvature is normalized to -1, there are especially nice formulas, such as the fact that the area of a triangle is equal to the angular defect.

However, since a surface of negative curvature cannot be embedded in a surface of zero curvature, hyperbolic geometry requires "models" to represent hyperbolic space on a flat sheet of paper. There are many such models, including the Poincaré disk and the Poincaré upper half-plane model. These all have varying metrics and methods of representing lines (i.e., geodesics) and shapes.

In the *Poincaré disk model* of $\mathbb{H}^2$, $\{z \in \mathbb{R}^2 : |z| < 1\}$, geodesics are either circular arcs orthogonal to the unit circle or else lines through the origin. The metric is defined as

$$(ds)^2 = \frac{(dx)^2 + (dy)^2}{(1 - x^2 - y^2)^2}.$$

In the *upper half-plane model* of $\mathbb{H}^2$, $\{(x, y) \in \mathbb{R}^2 : y > 0\}$, geodesics are either lines orthogonal to the x-axis or else circular arcs orthogonal to the x-axis. The metric is defined as

$$(ds)^2 = \frac{(dx)^2 + (dy)^2}{y^2}.$$

In the *Klein disk model* or *Beltrami–Klein model* of $\mathbb{H}^2$, the points in the model are the points of the open unit disk in the Euclidean plane, and the geodesics are the intersections with the open disk of chords joining two points on the unit circle. The formula for the metric in this model is rather complicated:

$$(ds)^2 = \frac{(dx)^2 + (dy)^2}{1 - x^2 - y^2} + \frac{(x\,dx + y\,dy)^2}{(1 - x^2 - y^2)^2}.$$

1.2. *Conics.*

Definition 2. One of the oldest notions in geometry, going all the way back to Apollonius, is that of *conic sections* in $\mathbb{E}^2$. There are at least four equivalent definitions of a conic section C:

(1) A smooth irreducible algebraic curve in $\mathbb{A}^2(\mathbb{R})$ of degree 2.

(2) The intersection of a right circular cone in $\mathbb{E}^3$ (with vertex at the origin, say) with a plane not passing through the origin, this plane in turn identified with $\mathbb{E}^2$.

(3) The *two focus definition*: Fix two points $a_1, a_2 \in \mathbb{E}^2$. An *ellipse C* is the locus of points $x \in \mathbb{E}^2$ such that $d(x, a_1) + d(x, a_2) = c$, where $c > 0$ is a fixed constant. Similarly, a *hyperbola C* is the locus of points $x \in \mathbb{E}^2$ such that $|d(x, a_1) - d(x, a_2)| = c$, where $c > 0$ is a fixed constant. The points a_1 and a_2 are called the *foci* of the conic, and the line joining them (assuming $a_1 \neq a_2$) is called the *major axis*. A *circle* is

the special case of an ellipse where $a_1 = a_2$. A *parabola* is the limiting case of a one-parameter family of ellipses $C_t(a_1, a_2, c_t)$, where a_1 is fixed and we let a_2 run off to infinity along the major axis keeping $c_t - d(a_1, a_2)$ fixed.

(4) The *focus/directrix definition*: Fix a point $a_1 \in \mathbb{E}^2$, called the *focus*, and a line ℓ not passing through a_1, called the *directrix*. A *conic* C is the locus of points with $d(x, a_1) = \varepsilon d(x, \ell)$, where $\varepsilon > 0$ is a constant called the *eccentricity*. If $\varepsilon < 1$ the conic is called an *ellipse*; if $\varepsilon = 1$ the conic is called a *parabola*; if $\varepsilon > 1$ the conic is called a *hyperbola*. A *circle* is the limiting case of an ellipse obtained by fixing a_1 and sending $\varepsilon \to 0$ and $d(a_1, \ell) \to \infty$ while keeping $r = \varepsilon d(a_1, \ell)$ fixed.

Note that these definitions come from totally different realms. Definition 2(1) is from algebraic geometry. Definition 2(2) uses a totally geodesic embedding of $\mathbb{E}^2$ into $\mathbb{E}^3$. Definitions 2(3) and 2(4) use only the metric geometry of $\mathbb{E}^2$.

Since Definition 2(1) is phrased in terms of algebraic geometry, it naturally leads to a definition of a *conic in* $\mathbb{P}^2(\mathbb{R})$ as a smooth irreducible algebraic curve of degree 2. Such a curve must be given (in homogeneous coordinates) by a homogeneous quadratic equation $Q(x) = 0$, where Q is a nondegenerate indefinite quadratic form on $\mathbb{R}^3$. This is the equation of a cone, and intersecting the cone with an affine plane not passing through the origin (the vertex of the cone) gives us back Definition 2(2).

1.3. *Contents of this paper.* The topic of this paper is studying what happens to Definitions 2(1)–(4) when we replace $\mathbb{E}^2$ by $\mathbb{H}^2$. This is an old problem, and is discussed for example in [Story 1882; Coxeter 1998; Fladt 1958; 1964; Molnár 1978]. However, as we will demonstrate, the analogues of Definitions 2(1)–(4) are no longer equivalent in $\mathbb{H}^2$. Thus there is some confusion in the literature, and those who talk about conic sections in $\mathbb{H}^2$ (as recently as [Csima and Szirmai 2014; 2015]) do not always all mean the same thing. Our main results are Theorems 11, 13, 14, and 15 in Section 3, which clarify the relationships among these definitions (especially the two-focus and focus-directrix definitions) in $\mathbb{H}^2$. The following summarizes our results:

- Circles: Definition 3 $\Leftrightarrow$ Definition 7. Definitions 3 and 7 are included in Definitions 4 and 6. Definitions 3 and 7 $\not\Leftrightarrow$ Definition 8.
- Horocycles (paracycles), hypercycles: These are included in Definitions 5 and 6, and not included in Definitions 7 and 8.
- Ellipses: Definition 7 $\not\Leftrightarrow$ Definition 8. But when Definition 8 gives a closed curve, it is included in Definition 7.
- Hyperbolas: Definition 7 $\subsetneqq$ Definition 8.
- Parabolas: Definition 7 $\not\Leftrightarrow$ Definition 8. Neither kind of parabola is ever closed.

2. Other Axiomatizations

Before discussing conics in $\mathbb{H}^2$, we first explain still another definition of conic sections in $\mathbb{P}^2(\mathbb{R})$, which is the definition found in [Coxeter 1998, Chapter III] and with a slight variation in [Story 1882]. This definition uses the notion of a *polarity* p in $\mathbb{P}^2(\mathbb{R})$. This is a particular type of mapping of points to lines and lines to points preserving the incidence relations of projective geometry (or in the language of [Coxeter 1998, §3.1], a *correlation*). It can be explained in terms of algebraic geometry as follows. If Q is a nondegenerate quadratic form on $\mathbb{R}^3$, then there is an associated nondegenerate symmetric bilinear form defined by $B(x, y) = \frac{1}{2}(Q(x + y) - Q(x) - Q(y))$, and if V is a linear subspace of $\mathbb{R}^3$ of dimension $d = 1$ or 2, then the orthogonal complement $V^{\perp, B}$ of V with respect to B is a linear subspace of dimension $3 - d$. Thus the process p of taking orthogonal complements with respect to B sends points in $\mathbb{P}^2(\mathbb{R})$, which are 1-dimensional linear subspaces of $\mathbb{R}^3$, to lines (copies of $\mathbb{P}^1(\mathbb{R})$), which are 2-dimensional linear subspaces of $\mathbb{R}^3$, and *vice versa*. Given a polarity p, the associated *conic* is the set C of points $x \in \mathbb{P}^2(\mathbb{R})$ such that x lies on the line $p(x)$, i.e., the set of 1-dimensional linear subspaces V of $\mathbb{R}^3$ for which $V \subset V^{\perp, B}$, or in other words, for which V is B-isotropic. Thus if we identify the point $x \in \mathbb{P}^2(\mathbb{R})$ with its homogeneous coordinates, or with a basis vector for V up to rescaling, this becomes the condition $B(x, x) = 0$, or $Q(x) = 0$, which is just Definition 2(1). (Note that if Q is definite, the conic is empty, so we are forced to take Q to be indefinite in order to get anything interesting.) Conversely, it is well known [Coxeter 1998, §4.72] that every polarity arises from a nonsingular symmetric matrix or equivalently from a nondegenerate quadratic form Q, so the polarity definition of conics in [Coxeter 1998, Chapter III] is equivalent to Definition 2(1).

We now introduce several possible definitions of conic sections in $\mathbb{H}^2$.

Definition 3 (a metric circle). A *circle C* in $\mathbb{H}^2$ is the locus of points a fixed distance $r > 0$ from a *center* $x_1 \in \mathbb{H}^2$; i.e., $C = \{x \in \mathbb{H}^2 : d(x, x_1) = r\}$.

Definition 4 (analogue of Definition 2(2)). A *right circular cone* in $\mathbb{H}^3$ is defined as follows. Fix a point $x_0 \in \mathbb{H}^3$ (say the origin, if we are using the standard unit ball in $\mathbb{R}^3$ as our model of $\mathbb{H}^3$) and fix a plane P in $\mathbb{H}^3$ (a totally geodesic copy of $\mathbb{H}^2$) not passing through x_0. There is a unique ray starting at x_0 and intersecting P perpendicularly. Let x_1 be the intersection point (the closest point on P to x_0), and fix a radius $r > 0$. We then have the circle C in P centered at x_1 with radius r. The cone $c(x_0, C)$ through x_0 and C is then the union of the lines (geodesics) through x_0 passing through a point of C. The point x_0 is called the *vertex* of the cone. A *conic section* (in the literal sense!) in $\mathbb{H}^2$ is then the intersection of a plane P' in $\mathbb{H}^3$ (not passing through x_0) with $c(x_0, C)$.

Since we can take $P' = P$ in the above definition, it is obvious that a circle (as in Definition 3) is a special case of a conic section in the sense of Definition 4.

In the Poincaré ball model of $\mathbb{H}^3$ with x_0 the origin, geodesics through x_0 are just straight lines for the Euclidean metric, so it's easy to see that a right circular cone with vertex x_0 is also a right circular cone in the Euclidean sense in $\mathbb{R}^3$. On the other hand, planes in $\mathbb{H}^3$ not passing through x_0 correspond to Euclidean spheres perpendicular to the unit sphere (the boundary of the model of $\mathbb{H}^3$). Thus a conic section in the sense of Definition 4 is the intersection of a right circular cone with a sphere, and is thus (in terms of the algebraic geometry of $\mathbb{A}^3(\mathbb{R})$) an algebraic curve of degree ≤ 4. To view this conic in the usual Poincaré disk model of $\mathbb{H}^2$, we apply an isometry (stereographic projection) from P to the unit disk in $\mathbb{C}$. Since this is a rational map, we see that any conic section in the sense of Definition 4 is an algebraic curve (in fact of degree ≤ 4) when viewed in the disk model of $\mathbb{H}^2$. Alternatively, if we use the Klein ball model of $\mathbb{H}^3$ with x_0 the origin, then a right circular cone with vertex x_0 will again look like a Euclidean right circular cone, while a 2-plane in $\mathbb{H}^3$ will be the intersection of the ball with a Euclidean 2-plane, and any conic section in the sense of Definition 4 will also be a conic section in the Euclidean sense of Definition 2. Thus Definition 4 is equivalent to the following:

Definition 5 (analogue of Definition 2(1)). A *conic in $\mathbb{H}^2$ in the algebraic sense* is the intersection of a smooth irreducible algebraic curve of degree 2 in $\mathbb{A}^2(\mathbb{R})$ with the open unit disk, viewed as the Klein disk model for $\mathbb{H}^2$. (This is a nonconformal model in which points of $\mathbb{H}^2$ are points of the open unit disk, and the straight lines are intersections with the open disk of straight lines in the plane.) Such a conic is closed (compact) if and only if it is a circle or ellipse not intersecting the unit circle (the *absolute* in the terminology of [Story 1882] and [Coxeter 1998]).

Definition 5 is the definition of conics used in [Story 1882; Coxeter 1998].

Still another approach to defining conics may be found in [Molnár 1978], based on the axiom system for $\mathbb{E}^2$ and $\mathbb{H}^2$ developed in [Bachmann 1973]. First we need to discuss Bachmann's approach to metric geometry. Bachmann observes that in either $\mathbb{E}^2$ and $\mathbb{H}^2$, there is a unique isometry which is reflection in a given line a or around a given point A. Thus we can identify lines and points with certain distinguished involutory elements $\mathcal{S}$ (the reflections in lines) and $\mathcal{P}$ (the reflections around points) of the isometry group G. More is true: every element of G is a product of at most three elements of $\mathcal{S}$. Elements of $\mathcal{S}$ are orientation-reversing; elements of $\mathcal{P}$ are orientation-preserving. The product of two elements a, $b \in \mathcal{S}$ is a nontrivial involution if and only if $a \neq b$ and $ab = ba$; in this case, the lines associated to a and b are perpendicular (we write $a \perp b$) and $ab \in \mathcal{P}$ is the reflection around the unique intersection point of a and b. Furthermore, every element of G of order 2 belongs to $\mathcal{S}$ or to $\mathcal{P}$, but not to both. A point $A \in \mathcal{P}$ lies on a line $a \in \mathcal{S}$ exactly

when there exists $b \in S$ commuting with a such that $A = ab$. Thus a *metric plane* $\mathcal{M}$ can be identified with a group G together with a distinguished generating set S consisting of involutions and the set $\mathcal{P}$ of nontrivial products of commuting elements of S, satisfying certain axioms. We won't need the axioms here, since they will be evident in the cases we are interested in. In the case of $\mathbb{E}^2$, $G = \mathbb{R}^2 \rtimes O(2)$, the usual Euclidean motion group, and in the case of $\mathbb{H}^2$, $G = \mathrm{PGL}(2, \mathbb{R}) \cong O^+(2, 1)$. If $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ has determinant $+1$, then it operates on the upper half-plane by linear fractional transformations, which are orientation-preserving, and if it has determinant -1, then it operates on the upper half-plane by

$$z \mapsto \frac{a\bar{z} + b}{c\bar{z} + d},$$

and this conjugate-linear map is an orientation-reversing isometry of $\mathbb{H}^2$.

Bachmann also points out that the metric plane $(\mathcal{M}, S, \mathcal{P})$ corresponding to $\mathbb{E}^2$ or $\mathbb{H}^2$ can be embedded naturally in a *projective-metric plane* $(\mathcal{PM}, S', \mathcal{P}')$, in such a way that $S \subseteq S'$ and $\mathcal{P} \subseteq \mathcal{P}'$. In the case of $\mathbb{E}^2$, this is just the usual embedding of $\mathbb{A}^2(\mathbb{R})$ in $\mathbb{P}^2(\mathbb{R})$ by adjoining a copy of $\mathbb{P}^1(\mathbb{R})$ at ∞, and the associated group is $\mathrm{PGL}(3, \mathbb{R})$. In the case of $\mathbb{H}^2$, $\mathcal{PM}$ is again a copy of $\mathbb{P}^2(\mathbb{R})$, but its points and lines consist of *ideal points* and *ideal lines* of $\mathbb{H}^2$. A simple way to visualize the embedding of $\mathbb{H}^2$ in $\mathbb{P}^2(\mathbb{R})$ is to use the (nonconformal) Klein model of $\mathbb{H}^2$, in which points are points in the interior of the unit disk in $\mathbb{R}^2$, and lines are the intersections of ordinary straight lines in $\mathbb{A}^2(\mathbb{R})$ with the unit disk. Then each point or line of $\mathbb{H}^2$ obviously corresponds to a unique point or line of $\mathbb{P}^2(\mathbb{R})$. When viewed as $\mathcal{PM}$ in this way, $\mathbb{P}^2(\mathbb{R})$ carries a canonical polarity, namely the one associated to the unit circle in $\mathbb{A}^2(\mathbb{R})$, viewed as a conic in the sense of the polarity definition at the beginning of this section. When we embed $\mathbb{A}^2(\mathbb{R})$ in $\mathbb{P}^2(\mathbb{R})$ as usual via $(x, y) \mapsto [x, y, 1]$ (homogeneous coordinates denoted by square brackets), this polarity is associated to the quadratic form $Q: (x, y, z) \mapsto x^2 + y^2 - z^2$, since $Q(\cos\theta, \sin\theta, 1) = 0$ for any real angle θ.

Definition 6 [Molnár 1978, Definition 4.1]. A *conic C in the sense of Molnár*, with foci $A, B \in \mathbb{P}^2(\mathbb{R})$, is defined by choosing a line x_1 in $\mathbb{P}^2(\mathbb{R})$ which is not a boundary line (i.e., x_1 is not tangent to the unit circle) and not passing through either A or B, and with A and B not each other's reflections across x_1. Then C consists of points X_{11} and X chosen as follows. X_{11} is the intersection of the lines a_{11} through A and B^{x_1} (the reflection of B across x_1) and b_{11} through B and A^{x_1}. (The line x_1 is chosen so that a_{11} and b_{11} are not boundary lines.) The other points X are defined by fixing a point Y on x_1 and taking the lines a through Y and A and b though Y and B, and then if neither a nor b is a boundary line, letting X be the intersection of a_{11}^a and b_{11}^b (the reflections of a_{11} and b_{11} across a and b, respectively). Appropriate modifications are made if a or b is a boundary line.
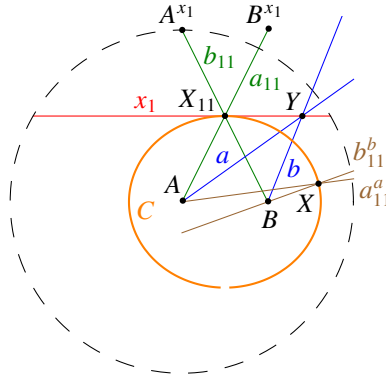
Figure 1. Molnár's construction of a conic.

As is quite evident, Molnár's definition is quite complicated but results in a conic section in $\mathbb{H}^2$ being the intersection of a conic in $\mathbb{P}^2(\mathbb{R})$ with the unit disk (in the Klein model). We will not consider this definition further, but it's closely related to Definition 5. A picture of the construction with $A = (0, 0)$, $B = (0.5, 0)$, $x_1 = \{y = 0.5\}$ is shown in Figure 1.

3. Main results

Definition 7 (analogue of Definition 2(3)). The definition of *two focus conics* in Definition 2(3) immediately goes over to $\mathbb{H}^2$, simply by replacing the Euclidean distance by the hyperbolic distance. Note that the case of a circle was already mentioned in Definition 3.

The last definition is the only one that is not immediately obvious. However, if we were to carry Definition 2(4) over to $\mathbb{H}^2$ without change, then since in the upper half-plane or disk models of $\mathbb{H}^2$, the distance function is the log of an algebraic expression, in the case of irrational eccentricity ε we would effectively get the equation

$$\text{(algebraic expression)} = \text{(algebraic expression)}^\varepsilon,$$

which is a transcendental equation, and could not possibly agree with the other definitions of conic sections. This explains the modification made in [Story 1882]. The use of the hyperbolic sine comes from its role in hyperbolic geometry via the solution of the Jacobi equation.

Definition 8 (analogue of Definition 2(4)). Fix a point $a_1 \in \mathbb{H}^2$, called the *focus*, and a line (geodesic) ℓ not passing through a_1, called the *directrix*. A *conic* C is the locus of points $x \in \mathbb{H}^2$ with $\sinh d(x, a_1) = \varepsilon \sinh d(x, \ell)$, where $\varepsilon > 0$ is a constant called the *eccentricity*. If $\varepsilon < 1$ the conic is called an *ellipse*; if $\varepsilon = 1$ the conic is

called a *parabola*; if $\varepsilon > 1$ the conic is called a *hyperbola*. (Note: in the case of the parabola, but only in this case, the hyperbolic sines cancel and can be removed from the definition.) A *circle* is the limiting case of an ellipse obtained by fixing a_1 and sending $\varepsilon \to 0$ and $d(a_1, \ell) \to \infty$ while keeping $r = \varepsilon \sinh d(a_1, \ell)$ fixed.

3.1. Circles. We begin now to compare the various definitions. We start with the circle, which is the most straightforward. Definition 3 clearly coincides with Definition 4, in the sense that if we intersect a right circular cone with a plane perpendicular to the axis, the result is a circle in the sense of Definition 3. We also have the following.

Proposition 9. *Definition 3 coincides with the case of circles in Definition 5, but with the Klein model replaced by the Poincaré model. In other words, an ordinary circle in $\mathbb{A}^2(\mathbb{R})$, contained in the open unit disk, when viewed as a curve in the Poincaré disk model of $\mathbb{H}^2$ is a metric circle in $\mathbb{H}^2$, and vice versa. Similarly, an ordinary circle contained in the upper half-plane, when viewed as a curve in the Poincaré upper half-plane model of $\mathbb{H}^2$, is a metric circle in $\mathbb{H}^2$, and vice versa.*

Proof. First consider the disk model. If the center is the origin, this is clear since the hyperbolic distance from 0 to z in $\{z : |z| < 1\}$ in $\mathbb{C}$ is a (nonlinear) function

$$\tanh^{-1}(|z|) = \tfrac{1}{2} \log\left(\frac{1 + |z|}{1 - |z|}\right)$$

of the Euclidean distance $|z|$ from 0 to z, so that each Euclidean circle centered at 0 is also a hyperbolic circle (of a different radius), and *vice versa*. However, any circle in $\mathbb{H}^2$ can be mapped to a circle centered at 0 via an isometry of $\mathbb{H}^2$, and since linear fractional transformations send circles to circles [Ahlfors 1978, Chapter 3, §3.2, Theorem 14], the general case follows. The case of the half-plane model also follows since there is a linear fractional transformation relating this model to the disk model. $\square$

Remark 10. However, one should note that the center of a circle in the unit disk or the upper half-plane may differ, depending on whether one considers it as a Euclidean circle or a metric circle in $\mathbb{H}^2$. For example, the metric circle in $\mathbb{H}^2$ (in the upper half-plane model) around the point i with hyperbolic metric radius $\log 2$ has Euclidean equation

$$\frac{|z-i|}{|z+i|} = \tanh\left(\tfrac{1}{2} \log 2\right) = \tfrac{1}{3} \quad \text{or} \quad \left|z - \tfrac{5}{4}i\right| = \tfrac{3}{4},$$

so its center as a Euclidean circle is $\tfrac{5}{4}i$.

Metric circles in $\mathbb{H}^2$, when drawn in the Klein disk model, only appear to be circles when centered at the origin. Otherwise, they are ellipses.

However, the focus/directrix definition of circles is quite different.

Theorem 11. *The definition of circle in Definition 8 does **not** agree with the definition of circle in Definitions 3, 4, 5, and 7.*

Proof. Consider a circle in the sense of Definition 8. Without loss of generality, we work in the upper half-plane model of $\mathbb{H}^2$ and set $a_1 = i$, $\ell = \{z \in \mathbb{H}^2 : |z| = R\}$, where we let $R \to +\infty$. In this case $d(a_1, \ell) = \log R$ and we want to keep $r = \varepsilon \sinh(\log R) = \varepsilon(R^2 - 1)/(2R)$ constant, so we take $\varepsilon = 2rR/(R^2 - 1)$. For $z \in \mathbb{H}^2$,

$$d(z, \ell) = \tfrac{1}{2}d(z, w), \quad \text{where } w = R^2/\bar{z} = \text{reflection of } z \text{ across } \ell.$$

Then the equation $\sinh d(z, a_1) = \varepsilon \sinh d(z, \ell)$ becomes

$$\sinh\left(2\tanh^{-1}\left|\frac{z-i}{z+i}\right|\right) = \frac{2rR}{R^2 - 1}\sinh\left(\tanh^{-1}\left|\frac{z - R^2/\bar{z}}{z - R^2/z}\right|\right).$$

The left-hand side simplifies to

$$\frac{2|z+i|\,|z-i|}{|z+i|^2 - |z-i|^2} = \frac{|z^2 + 1|}{2\,\mathrm{Im}\,z}.$$

On the right-hand side,

$$\left|\frac{z - R^2/\bar{z}}{z - R^2/z}\right| = \frac{R^2 - |z|^2}{\left|R^2 - |z|^2 z/\bar{z}\right|} = \frac{R^2 - a}{\sqrt{R^4 + a^2 - 2R^2 a \cos\theta}},$$

where $a = |z|^2$ and $\theta = 2\arg z$. Then

$$\lim_{R \to \infty} \frac{2rR}{R^2 - 1}\sinh\left(\tanh^{-1}\left(\frac{R^2 - a}{\sqrt{R^4 + a^2 - 2R^2\cos\theta}}\right)\right) = \frac{\sqrt{2}\,r}{|z|\sqrt{1 - \cos\theta}}.$$

Thus Definition 8 gives for our circle the equation

$$\frac{|z^2 + 1|}{2\,\mathrm{Im}\,z} = \frac{\sqrt{2}\,r}{|z|\sqrt{1 - \cos\theta}} = \frac{r}{|z|}\csc(\tfrac{1}{2}\theta) = \frac{r}{|z|}\frac{|z|}{\mathrm{Im}\,z} = \frac{r}{\mathrm{Im}\,z}$$

or

$$|z^2 + 1| = 2r \tag{1}$$

in the upper half-plane. This is an algebraic curve but not a metric circle. Figure 2 shows the case of $r = 0.25$ (in solid color) as drawn with Mathematica. This curve passes through the points $i\sqrt{3/2}$, $i\sqrt{1/2}$, and $i \pm \sqrt{(\sqrt{17} - 4)/2}$; the circle centered on the imaginary axis tangent to it at $i\sqrt{3/2}$ and $i\sqrt{1/2}$ is shown with a dashed line in the same figure. The curves are close but do not coincide.   □

Aside from circles, there are various other circle-like curves that play a role in hyperbolic geometry. These may be considered to be conics according to certain definitions. Note also that they are distinct from the circles of Definition 8.
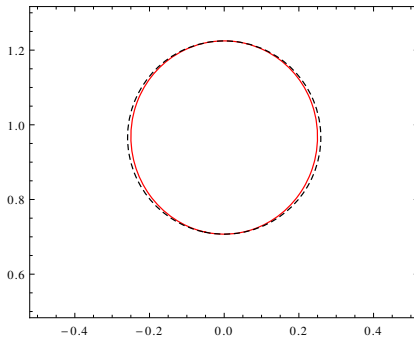
Figure 2. A "circle" with focus i and $r = 0.25$ (solid color) and a tangent metric circle (dashed) in the upper half-plane.
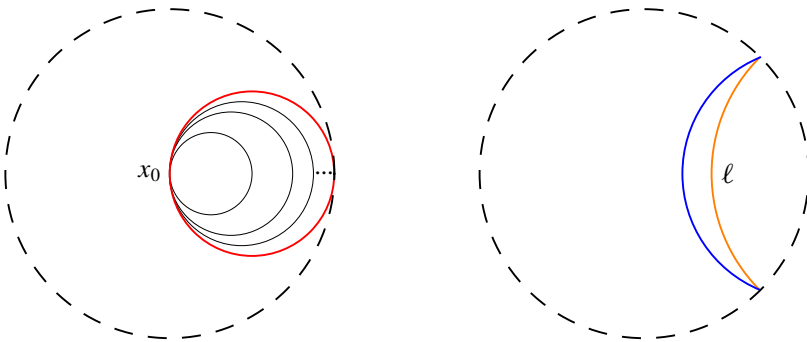
Figure 3. Left: a horocycle (solid red) as a limit of circles (black) through x_0 with radii going to infinity. Right: a hypercycle (solid blue) and a straight line ℓ (orange) with the same ideal limits at infinity.

Definition 12. A *horocycle* (occasionally called a *paracycle*) in the Poincaré disk model of $\mathbb{H}^2$ is the intersection of the disk with a circle tangent to the unit circle (and lying inside the circle). A *hypercycle* in the Poincaré disk model of $\mathbb{H}^2$ is the intersection of the disk with a circle meeting the unit circle in exactly two points. These have well-known intrinsic definitions. A horocycle is the limit of a sequence of circles C_n (in the sense of Definition 3) all passing through a fixed point x_0, with centers x_n all lying on a fixed ray through x_0 and with radii $d(x_n, x_0) = r_n \to \infty$. See Figure 3, left. A hypercycle is a curve on one side of a given line ℓ whose points all have the same orthogonal distance from ℓ. See Figure 3, right. Note that horocycles and hypercycles are clearly conics in the sense of Definition 5. But they are not covered by Definitions 7 and 8. Molnár [1978] observed that metric circles (Definition 3), horocycles, and hypercycles are all special cases of Definition 6 when the two foci coincide.
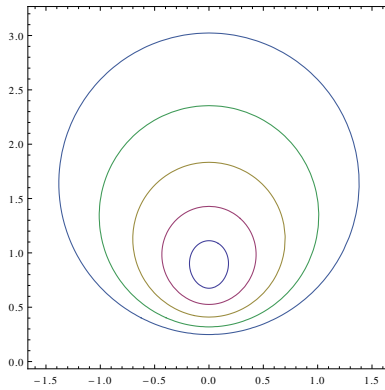
Figure 4. Two-focus ellipses in the upper half-plane with foci at i and $\frac{3}{4}i$, as drawn with Mathematica.

3.2. Ellipses. Next, we consider the case of the (noncircular) ellipse. There are two main competing definitions: Definition 7 and Definition 8.

Theorem 13. *The definition of ellipse in Definition 8 does **not always** agree with the definition of ellipse in Definition 7. However, there are cases where they coincide. More precisely, when Definition 8 gives a closed curve in $\mathbb{H}^2$, this curve is also a two-focus ellipse.*

Proof. We will work in the upper half-plane model of $\mathbb{H}^2$ and, without loss of generality, put one focus at i and let the imaginary axis be an axis of the ellipse. For an ellipse with the "two-focus definition" and foci at i and bi, $b > 0$, the equation is

$$2\tanh^{-1}\left(\frac{|z-i|}{|z+i|}\right) + 2\tanh^{-1}\left(\frac{|z-bi|}{|z+bi|}\right) = c,$$

which can be rewritten as the algebraic equation

$$(x^2 + y^2 + 1 + \sqrt{(x^2 - y^2 + 1)^2 + 4x^2 y^2})$$
$$\times (x^2 + y^2 + b^2 + \sqrt{(x^2 - y^2 + b^2)^2 + 4x^2 y^2}) = 4be^c y^2 \quad (2)$$

with $c > 0$. Plots of this equation for $b = \frac{3}{4}$ and for various values of c are shown in Figure 4. The minimal value of c to have the foci inside the ellipse is the hyperbolic distance between the foci, or $|\log b|$. As c increases, the curves get bigger and bigger and look more like circles. Now that since (2) implies that $d(z, i) \leq c$, any ellipse in the sense of Definition 7 is automatically compact (closed) in $\mathbb{H}^2$.

Now consider the focus-directrix definition for an ellipse in the upper half-plane, with a focus at i and directrix $|z| = r$, $r > 1$ (this choice makes the imaginary axis an axis of the ellipse). The distance from z to the directrix is half the distance to
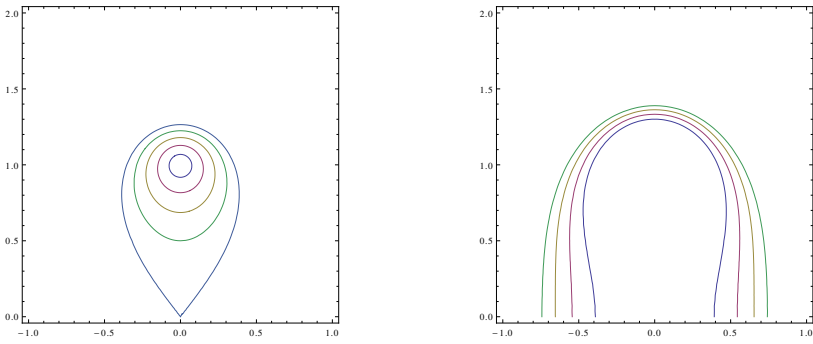
Figure 5. Focus/directrix ellipses in the upper half-plane with focus at i and directrix $|z| = 2$, as drawn with Mathematica. On the left, cases with $\varepsilon \leq 0.5$. The case $\varepsilon = 0.5$ is a lemniscate. On the right, cases with ε from 0.6 to 0.9.

the reflection of z across the directrix, which is $r^2/\bar{z}$. Thus the equation becomes

$$\sinh\left(2\tanh^{-1}\left(\frac{|z-i|}{|z+i|}\right)\right) = \varepsilon \sinh\left(\tanh^{-1}\left(\frac{|z-r^2/\bar{z}|}{|z-r^2/z|}\right)\right),$$

which simplifies (after squaring both sides) to

$$r^2\left(1+x^4+y^4+2y^2(-1+\varepsilon^2)+2x^2(1+y^2+\varepsilon^2)\right) = \varepsilon^2(r^4+(x^2+y^2)^2). \quad (3)$$

This is a relatively simple quartic equation in x and y, basically the Cassini oval equation, and has some interesting features. For example, if one sets $\varepsilon = 1/r$, this reduces to a lemniscate passing through $(0, 0)$ (an ideal boundary point of $\mathbb{H}^2$). When $\varepsilon > 1/r$, the curve (viewed in $\mathbb{H}^2$) is not closed and approaches two distinct ideal boundary points. Pictures of this behavior appear in Figure 5. As a check that having two distinct ideal boundary points is not just an artifact of the calculation, one can check that upon substituting $r = 3$ and $\varepsilon = \frac{1}{2}$ into (3), one gets two points with $y = 0$, namely $x = \pm\sqrt{3/7}$.

To illustrate another difference between the two definitions, consider the case of the two-focus definition when the foci coincide, i.e., $b = 1$ in (2). Then (2) reduces to

$$x^2 + y^2 + 1 + \sqrt{(x^2-y^2+1)^2 + 4x^2y^2} = 2e^{c/2}y$$

or

$$(x^2-y^2+1)^2 + 4x^2y^2 - (2e^{c/2}y - x^2 - y^2 - 1)^2 = 0,$$

which simplifies to the equation of a circle:

$$x^2 + \left(y - \cosh\left(\tfrac{1}{2}c\right)\right)^2 = \sinh^2\left(\tfrac{1}{2}c\right). \quad (4)$$

However, the focus/directrix equation (3) never reduces to a circle.

However, perhaps rather surprisingly, focus/directrix ellipses with $\varepsilon < 1/r$ (this is the case where the curve is closed) turn out to be special cases of two-focus ellipses. A rather horrendous calculation with Mathematica or MuPAD shows for example that (2) with $b = 2$ and $c = \log(\frac{5}{2})$ is equivalent to (3) with

$$\varepsilon = \frac{\sqrt{209}}{21}, \quad r = \sqrt{\frac{11}{19}}.$$

To see this, rewrite (3) in the form

$$x^2 + y^2 + 1 + \sqrt{(x^2 - y^2 + 1)^2 + 4x^2 y^2} = \frac{20y^2}{x^2 + y^2 + 4 + \sqrt{(x^2 - y^2 + 4)^2 + 4x^2 y^2}}$$

$$= \frac{20y^2(x^2 + y^2 + 4 - \sqrt{(x^2 - y^2 + 4)^2 + 4x^2 y^2})}{(x^2 + y^2 + 4)^2 - ((x^2 - y^2 + 4)^2 + 4x^2 y^2)},$$

simplify, and rewrite in the form $E + \sqrt{B} = F\sqrt{D}$, where

$$B = (x^2 - y^2 + 1)^2 + 4x^2 y^2 \quad \text{and} \quad D = (x^2 - y^2 + 4)^2 + 4x^2 y^2.$$

Square both sides, again simplify and regroup to get the term with $\sqrt{B}$ by itself, and finally square again. After factoring out y^2, one finally ends up with the equation

$$20x^4 + 40x^2 y^2 + 325x^2 + 20y^4 - 116y^2 + 80 = 0,$$

which is equivalent to (3) for the given parameters. Other values of r and ε (with $r\varepsilon < 1$) can be handled similarly; one just needs to solve for the values of b and c giving the same y-intercepts. $\qquad \square$

3.3. *Parabolas.* Next, we consider the case of the parabola. Here the result is rather simple.

Theorem 14. *The definitions of parabolas in* Definition 8 *and in* Definition 7 *never agree. In all cases, however, a parabola in* $\mathbb{H}^2$ *is not closed.*

Proof. Without loss of generality, we can again use the Poincaré upper half-plane model of $\mathbb{H}^2$ and put one focus at i and take the axis of the parabola to be the imaginary axis. The two-focus definition of Definition 7 is the limiting case of (2) as we keep $be^c = \frac{1}{2}C$ fixed and let $b \to 0_+$. (This is because $d(i, ib) = |\log b| = -\log b$ for $0 < b < 1$ and we want $c - d(i, ib) = c + \log b$ to be held constant.) Then (2) reduces to

$$(x^2 + y^2 + 1 + \sqrt{(x^2 - y^2 + 1)^2 + 4x^2 y^2})(x^2 + y^2) = Cy^2, \qquad (5)$$

or equivalently (after regrouping and squaring to get rid of the radical, then factoring out a y^2)

$$2(C - 2)(x^2 + y^2)^2 + 2C(x^2 + y^2) - C^2 y^2 = 0. \qquad (6)$$
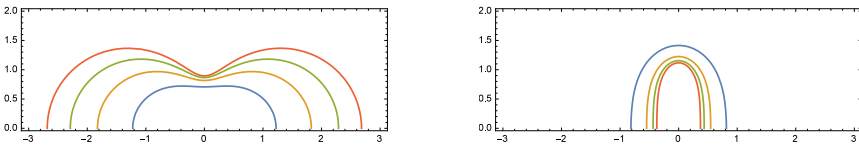
Figure 6. Focus/directrix parabolas in the upper half-plane with focus at i and directrix $|z| = r$, as drawn with Mathematica. On the left, cases with $r < 1$. These are Cassini ovals. On the right, cases with $r > 1$. Of course, if one were wearing "hyperbolic glasses", all would look roughly the same.
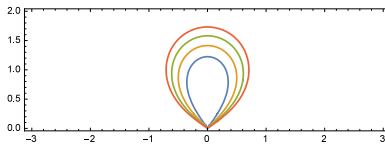

Figure 7. Two-focus parabolas in the upper half-plane with focus at i, as drawn with Mathematica. Note the lemniscate shape.

This is the equation of a lemniscate through the origin. (Remember that 0, however, is only an ideal boundary point of $\mathbb{H}^2$.) Definition 8 simply gives (3) with $\varepsilon = 1$, which reduces to

$$1 - r^2 + 4x^2 + (1 - 1/r^2)(x^2 + y^2)^2 = 0, \tag{7}$$

which is a Cassini oval equation. Note that (6) and (7) never agree, since for $r \neq 1$ (we don't want the directrix of the parabola to pass through the focus), the curve given by (7) doesn't pass through the origin. Pictures of the various kinds of parabolas, plotted by Mathematica, are shown in Figures 6 and 7.   □

3.4. *Hyperbolas.*   Finally, we consider the case of the hyperbola.

Theorem 15. *The definition of hyperbola in Definition 8 does **not always** agree with the definition of hyperbola in Definition 7. However, the two-focus hyperbola from Definition 7 is a special case of the focus-directrix hyperbola of Definition 8.*

Proof. Consider the two-focus hyperbola. Fix $c > 0$. (When $c = 0$, the definition degenerates to the bisector of the line segment joining the two foci, which is a straight line i.e., a geodesic.) We will work in the upper half-plane model of $\mathbb{H}^2$ and, without loss of generality, put one focus at i and the other focus at ib, $b > 1$. The equation of the two-focus hyperbola is then

$$2\tanh^{-1}\left(\frac{|z-i|}{|z+i|}\right) - 2\tanh^{-1}\left(\frac{|z-bi|}{|z+bi|}\right) = \pm c,$$
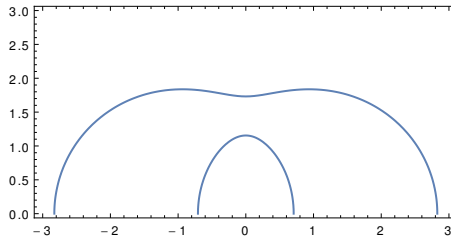
Figure 8. A hyperbola in the upper half-plane with foci at i and $2i$, $c = \log\left(\frac{3}{2}\right)$, as drawn with Mathematica.

which can be rewritten as the algebraic equation

$$b^2 + x^2 + y^2 + \sqrt{b^4 + 2b^2(x^2 - y^2) + (x^2 + y^2)^2}$$
$$= be^{\pm c}(1 + x^2 + y^2 + \sqrt{1 + 2(x^2 - y^2) + (x^2 + y^2)^2}) \quad (8)$$

with $c > 0$. Note that the hyperbola should intersect its axis (here the imaginary axis) at two points of the form iy, $1 < y < b$, so we want $0 < c < \log b$, and the two y-intercepts are at $i\sqrt{be^{\pm c}}$. Comparing this with the y-intercepts for the focus-directrix hyperbola (3) (the equation is the same as for the ellipse — the only difference is the value of the eccentricity ε), we see that this agrees with a focus-directrix hyperbola with parameters satisfying

$$\frac{\sqrt{r + r^2\varepsilon}}{\sqrt{r + \varepsilon}} = \frac{\sqrt{b}}{e^{c/2}}, \quad \frac{\sqrt{r - r^2\varepsilon}}{\sqrt{r - \varepsilon}} = \sqrt{b}e^{c/2}$$

or

$$r = \sqrt{\frac{-b + 2b^2e^c - be^{2c}}{b - 2e^c + be^{2c}}}, \quad \varepsilon = \frac{\sqrt{b(-1 + 2be^c - e^{2c})(b - 2e^c + be^{2c})}}{b(e^{2c} - 1)}. \quad (9)$$

Note that since $c < \log b$, the value of ε is > 1. Just as an example, if $b = 2$ and $c = \log\left(\frac{3}{2}\right)$, after removing some superfluous factors, equation (8) reduces to $24 + 6x^4 - 26y^2 + 6y^4 + 3x^2(-17 + 4y^2) = 0$, which agrees with the focus-directrix hyperbola with focus i, directrix $|z| = \sqrt{11/7}$, and eccentricity $\varepsilon = \sqrt{77}/5$. A graph of this hyperbola, drawn with Mathematica, appears in Figure 8.

So this analysis shows that every two-focus hyperbola is also a focus-directrix hyperbola. The converse fails, however. Indeed, one can see from (3) that the focus-directrix hyperbola with $r = \varepsilon > 1$ degenerates to the equation

$$(r^2 + 1)x^2 + (r^2 - 1)y^2 = \frac{r^4 - 1}{2},$$

which, surprisingly, is an *ellipse* in Cartesian coordinates. This has only one y-intercept in the upper half-plane, at the point $i\sqrt{(r^2 + 1)/2}$. So this "hyperbola" has only one vertex, the other vertex having gone to $+\infty i$, and this cannot be written as a two-focus hyperbola. □

References

[Ahlfors 1978] L. V. Ahlfors, *Complex analysis: an introduction to the theory of analytic functions of one complex variable*, 3rd ed., McGraw-Hill, New York, 1978. MR

[Bachmann 1973] F. Bachmann, *Aufbau der Geometrie aus dem Spiegelungsbegriff*, 2nd ed., Springer, 1973. MR Zbl

[Coxeter 1998] H. S. M. Coxeter, *Non-Euclidean geometry*, 6th ed., Mathematical Association of America, Washington, DC, 1998. MR Zbl

[Csima and Szirmai 2014] G. Csima and J. Szirmai, "Isoptic curves of conic sections in constant curvature geometries", *Math. Commun.* **19**:2 (2014), 277–290. MR Zbl

[Csima and Szirmai 2015] G. Csima and J. Szirmai, "Isoptic curves of generalized conic sections in the hyperbolic plane", preprint, 2015. arXiv

[Fladt 1958] K. Fladt, "Die allgemeine Kegelschnittsgleichung in der ebenen hyperbolischen Geometrie, II", *J. Reine Angew. Math.* **199** (1958), 203–207. MR Zbl

[Fladt 1964] K. Fladt, "Elementare Bestimmung der Kegelschnitte in der hyperbolischen Geometrie", *Acta Math. Acad. Sci. Hungar.* **15** (1964), 247–257. MR Zbl

[Molnár 1978] E. Molnár, "Kegelschnitte auf der metrischen Ebene", *Acta Math. Acad. Sci. Hungar.* **31**:3-4 (1978), 317–343. MR Zbl

[Story 1882] W. E. Story, "On non-Euclidean properties of conics", *Amer. J. Math.* **5**:1-4 (1882), 358–381. MR

prc@berkeley.edu   *Montgomery Blair High School, Silver Spring, MD, United States*

Current address: *University of California, Berkeley, CA*

jmr@math.umd.edu   *Department of Mathematics, University of Maryland, College Park, MD, United States*

The Fibonacci sequence under a modulus: computing all moduli that produce a given period

Alex Dishong and Marc S. Renault

(Communicated by Kenneth S. Berenhaut)

The Fibonacci sequence $F = 0, 1, 1, 2, 3, 5, 8, 13, \ldots$, when reduced modulo m is periodic. For example, $F \bmod 4 = 0, 1, 1, 2, 3, 1, 0, 1, 1, 2, \ldots$. The period of $F \bmod m$ is denoted by $\pi(m)$, so $\pi(4) = 6$. In this paper we present an algorithm that, given a period k, produces all m such that $\pi(m) = k$. For efficiency, the algorithm employs key ideas from a 1963 paper by John Vinson on the period of the Fibonacci sequence. We present output from the algorithm and discuss the results.

1. The problem

Consider the usual Fibonacci sequence $F = 0, 1, 1, 2, 3, 5, 8, \ldots$, with $F_0 = 0$, $F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}$. When reduced modulo m, the Fibonacci sequence is periodic. For example, $F \bmod 4 = 0, 1, 1, 2, 3, 1, 0, 1, 1, \ldots$. The period of $F \bmod m$ is denoted by $\pi(m)$, so we see that $\pi(4) = 6$. The properties of $\pi(m)$ have been studied extensively; see, e.g., [Gupta et al. 2012; Robinson 1963; Vinson 1963; Wall 1960]. One might ask, of course, if there are any other values of m such that $\pi(m) = 6$. The answer is no (you can verify this by hand), but it turns out that there are 10 different moduli m such that $\pi(m) = 24$ (namely, 6, 9, 12, 16, 18, 24, 36, 48, 72, 144). Our goal is to construct an efficient algorithm that, given a period k, produces all m such that $\pi(m) = k$.

It is instructive to first consider how one might solve the problem by brute force. If $\pi(m) = k$, then $F_k \equiv 0 \pmod{m}$ and $F_{k+1} \equiv 1 \pmod{m}$. That is, m divides both F_k and $F_{k+1} - 1$. For brute force, we fix k, find all common divisors of F_k and $F_{k+1} - 1$, and then apply the π function to these divisors to see which ones produce the desired value of k. Computing $\pi(m)$ is not difficult but it requires factoring m as a product of primes, then factoring $p \pm 1$ for each prime p that divides m. See [Wall 1960] for theorems on $\pi(m)$ and [Flanagan et al. 2015] for an algorithm for $\pi(m)$ (as well as many other facts about the Fibonacci sequence under a modulus).

By employing key ideas from a 1963 paper by John Vinson on the period of the Fibonacci sequence, we were able to produce an algorithm that does not require computing $\pi(m)$. Instead, the moduli we seek can be produced with simple divisibility tests.

2. The algorithm

In this section we present Theorem 2.1 on which our algorithm is based, pseudo-code for the algorithm, and some output. In the next section we provide a proof of Theorem 2.1.

First, we note that $\pi(2) = 3$ but it is known that for $m > 2$, $\pi(m)$ must be even. By inspecting a few small cases, it is easy to see that no moduli produce a period of 4, and the smallest even period is 6. Let $L = 2, 1, 3, 4, 7, \ldots$ denote the Lucas sequence: $L_0 = 2$, $L_1 = 1$, and $L_n = L_{n-1} + L_{n-2}$. It is well-known that $L_n = F_{2n}/F_n = F_{n-1} + F_{n+1}$.

Theorem 2.1. *Given any even $k \geq 6$:*

(1) *If $k \equiv 2 \pmod 4$, then $\pi(m) = k$ if and only if $m \mid L_{k/2}$, and $m \nmid F_q$ for all q such that $q \mid k$ and $q \neq k$.*

(2) *If $k \equiv 4 \pmod 8$, then $\pi(m) = k$ if and only if $m \mid F_{k/2}$, and $m \nmid L_{k/4}$, and $m \nmid F_q$ for all q such that $q \mid \frac{k}{2}$ and $q \neq \frac{k}{2}$ or $\frac{k}{4}$.*

(3) *If $k \equiv 0 \pmod 8$, then $\pi(m) = k$ if and only if $m \mid F_{k/2}$, and $m \nmid F_q$ for all q such that $q \mid \frac{k}{2}$ and $q \neq \frac{k}{2}$.*

The algorithm follows immediately from the theorem.

Algorithm 2.2. Given an integer $k \geq 2$, to produce the set of all m such that $\pi(m) = k$:

```
Input:  an integer k ≥ 2
If k = 3, then return {2}.
If k ∈ {2, 4} or if k is odd, then return {}.
If k mod 4 = 2:
   Let M = {m : m | L_{k/2}}.
   Let F = {F_q : q | k and q ≠ k}.
If k mod 8 = 4:
   Let M = {m : m | F_{k/2} and m ∤ L_{k/4}}.
   Let F = {F_q : q | k/2 and q ≠ k/2 and q ≠ k/4}.
If k mod 8 = 0:
   Let M = {m : m | F_{k/2}}.
   Let F = {F_q : q | k/2 and q ≠ k/2}.
Return {m ∈ M : m ∤ f for all f ∈ F}
```
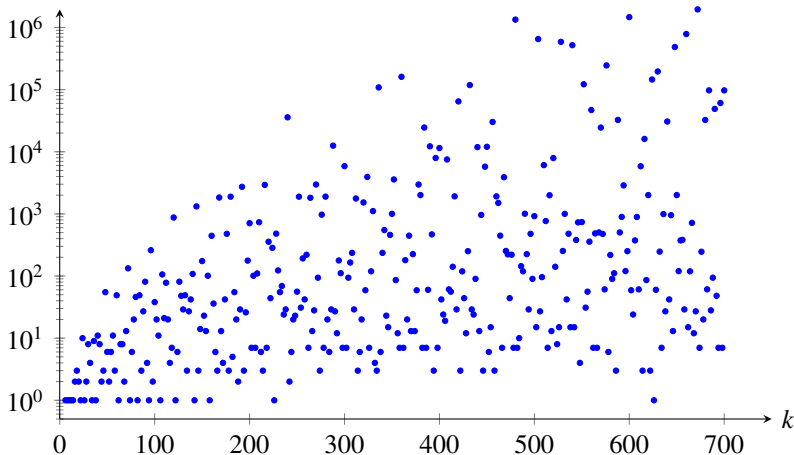
Figure 1. The number of m such that $\pi(m) = k$ for a given k.

Figure 1 shows the results when the algorithm is run on all even k from 6 to 700 and the size of the output set is calculated. The value of k appears on the horizontal axis, and the number of moduli m such that $\pi(m) = k$ is expressed on the vertical axis.

What surprised us most in this study was the incredible number of moduli that can produce a given period. For example, $\pi(m) = 600$ for 1,466,812 different values of m.

Moreover, the algorithm above has much greater speed than simple brute force. When we computed the moduli for all even periods k from 6 to 300, the brute force algorithm took 180.28 seconds, whereas Algorithm 2.2 completed the task in 0.62 seconds. We used the online Sage computer algebra system for our computations [Stein et al. 2016].

3. Proof of Theorem 2.1

The zeros in F mod m are evenly spaced. For example, consider F mod 5:

$$F \bmod 5 = 0, 1, 1, 2, 3, 0, 3, 3, 1, 4, 0, 4, 4, 3, 2, 0, 2, 2, 4, 1, 0, 1, 1, \ldots.$$

(To see why the zeros are evenly spaced, we can use the identities

$$F_{s+t} = F_{s-1}F_t + F_s F_{t+1},$$
$$F_{s-t} = (-1)^t (F_s F_{t+1} - F_{s+1}F_t).$$

If $F_s \equiv F_t \equiv 0$, then $F_{s+t} \equiv 0$ and $F_{s-t} \equiv 0$.)

The *rank* of F mod m, denoted by $\alpha(m)$, is the least index $i > 0$ such that $F_i \equiv 0 \pmod{m}$. We can deduce, for example, that if $m \mid F_i$, then $\alpha(m) \mid i$. The

order of F mod m, denoted by $\omega(m)$, is $\pi(m)/\alpha(m)$ (which is an integer since the zeros are evenly spaced). We see above that $\pi(5) = 20$, $\alpha(5) = 5$, and $\omega(5) = 4$.

It turns out that $\pi(2) = 3$, but for all $m > 2$, $\pi(m)$ must be even. As we see in the mod 5 example, $\alpha(m)$ need not be even. It is a remarkable fact that for any m, $\omega(m) = 1, 2$, or 4; this is proven in [Vinson 1963]. In that paper, Vinson studies the relationship between the period, rank, and order. Based on the Vinson paper, Renault was able find several other consequences, and the following theorem is a direct result of Theorem 3.35 and Corollary 3.38 in [Renault 1996].

Theorem 3.1. *For any modulus $m > 2$:*

(1) $\pi(m) \equiv 2 \pmod 4$ *if and only if* $\omega(m) = 1$. *In this case,* $\alpha(m) \equiv 2 \pmod 4$.

(2) *If* $\pi(m) \equiv 4 \pmod 8$, *then* $\omega(m) = 2$ *or* 4. *In this case,* $\alpha(m) \equiv 2 \pmod 4$ *or* $\alpha(m)$ *is odd, respectively.*

(3) *If* $\pi(m) \equiv 0 \pmod 8$, *then* $\omega(m) = 2$. *In this case,* $\alpha(m) \equiv 0 \pmod 4$.

Since $\pi(m)$ is even for $m > 2$, the above theorem describes all possible cases for $\pi(m)$. Also, even though the "in this case" portions follow obviously from their preceding statements, we can use them to draw conclusions. For example, we can see from the theorem that $\alpha(m) \equiv 0 \pmod 4$ if *and only if* $\pi(m) \equiv 0 \pmod 8$. We proceed now to the proof of Theorem 2.1.

Proof of Theorem 2.1(1). $(\Rightarrow)$ Assume $k \equiv 2 \pmod 4$ and $\pi(m) = k$. Since $k \equiv 2 \pmod 4$, Theorem 3.1 tells us that $\omega(m) = 1$. Thus, $m \nmid F_q$ for all q such that $1 \le q < k$. In particular, $m \nmid F_q$ for any q such that $q \mid k$ and $q \ne k$.

It remains to show that $m \mid L_{k/2}$. By the fact that $\pi(m) = k$ and the identity $F_{-n} = (-1)^{n+1} F_n$, we see that $F_{k-n} \equiv F_{-n} \equiv (-1)^{n+1} F_n \pmod m$. Then, since $\frac{k}{2}$ is odd,

$$F_{k/2-1} = F_{k-(k/2+1)} \equiv -F_{k/2+1} \pmod m.$$

Consequently, $m \mid F_{k/2-1} + F_{k/2+1}$. But by the identity $L_n = F_{n-1} + F_{n+1}$, this is exactly $m \mid L_{k/2}$, as required.

$(\Leftarrow)$ Assume $k \equiv 2 \pmod 4$ and (a) $m \mid L_{k/2}$ and (b) $m \nmid F_q$ for any q such that $q \mid k$ and $q \ne k$. We must show that $\pi(m) = k$.

By (a), $m \mid F_k$, so $\alpha(m) \mid k$. By (b) we find that in fact, $\alpha(m) = k$. Thus, $\pi(m) = k$, $2k$, or $4k$.

If $\pi(m) = 4k$, then $\omega(m) = 4$ and by Theorem 3.1, $\alpha(m)$ must be odd. However, $\alpha(m) \equiv 2 \pmod 4$, so this can't be the case.

If $\pi(m) = 2k$, then $\pi(m) \equiv 4 \pmod 8$, and so Theorem 2.1(2)$(\Rightarrow)$ implies $m \nmid L_{\pi(m)/4}$; that is, $m \nmid L_{k/2}$. But this contradicts our hypothesis (a) that $m \mid L_{k/2}$, and so $\pi(m) \ne 2k$.

We must conclude that $\pi(m) = k$ and the proof is complete. $\square$

Proof of Theorem 2.1 (2). ($\Rightarrow$) Assume $k \equiv 4$ (mod 8) and $\pi(m) = k$. Since $\pi(m) \equiv 4$ (mod 8), by Theorem 3.1 we know that $\omega(m) = 2$ or 4. In either case, $m \mid F_{k/2}$ and $m \nmid F_q$ where $q \mid \frac{k}{2}$ and $q \neq \frac{k}{2}, \frac{k}{4}$. Thus, it only remains to prove that $m \nmid L_{k/4}$.

For ease of notation, let $s = F_{k/2+1}$, let $a = F_{k/4+1}$, and observe that $s \not\equiv 1$ (mod m).

Claim 1. $F_{k/4-1} \equiv -sa$ (mod m).

Proof of Claim 1. Modulo m, the Fibonacci sequence starting at $F_{k/2}$ is 0, s, s, $2s$, $3s$, $5s$, ..., and in general, $F_{k/2+n} \equiv sF_n$ (mod m). In particular, $F_{(3k)/4+1} \equiv sa$. The identity $F_{-n} = (-1)^{n+1} F_n$ implies $F_{k-n} \equiv F_{-n} \equiv (-1)^{n+1} F_n$ (mod m). Since $\frac{k}{4}$ is odd, we find,

$$F_{k/4-1} \equiv F_{k-((3k)/4+1)} \equiv -F_{(3k)/4+1} \equiv -sa \pmod{m}.$$

Claim 2. $(a, m) = 1$.

Proof of Claim 2. We have $(F_{k/4-1}, F_{k/4+1}) = F_{(k/4-1,k/4+1)} = F_2 = 1$. So, there exist integers u and v such that $F_{k/4-1}u + F_{k/4+1}v = 1$. Thus, $-sau + av \equiv 1$ (mod m), and so $a(-su + v) \equiv 1$ (mod m) and we find that a is invertible mod m. That is, $(a, m) = 1$.

Consider the identity $L_n = F_{n-1} + F_{n+1}$. For contradiction,

$$m \mid L_{k/4} \implies m \mid F_{k/4-1} + F_{k/4+1} \implies -sa + a \equiv 0 \pmod{m}$$
$$\implies a(1-s) \equiv 0 \pmod{m} \implies s \equiv 1 \pmod{m}.$$

The last implication is due to the fact that $(a, m) = 1$, and we've arrived at a contradiction since $s \not\equiv 1$ (mod m). We conclude $m \nmid L_{k/4}$, as needed.

($\Leftarrow$) Assume $k \equiv 4$ (mod 8), (a) $m \mid F_{k/2}$, (b) $m \nmid L_{k/4}$, and (c) $m \nmid F_q$ for all $q \mid \frac{k}{2}$ where $q \neq \frac{k}{2}$ or $\frac{k}{4}$. We must prove that $\pi(m) = k$. By (a) and (c), $\alpha(m) = \frac{k}{4}$ or $\frac{k}{2}$. We know that the only possible values for $\omega(m)$ are 1, 2, or 4.

<u>Case 1:</u> $\alpha(m) = \frac{k}{4}$.

If $\omega(m) = 2$, then $\pi(m) = \frac{k}{2} \equiv 2$ (mod 4). However this contradicts Theorem 3.1 since $\pi(m) \equiv 2$ (mod 4) if and only if $\omega(m) = 1$.

If $\omega(m) = 1$, then $\pi(m) = \frac{k}{4} \equiv 1$ (mod 2). Again, this contradicts Theorem 3.1 since $\omega(m) = 1$ if and only if $\pi(m) \equiv 2$ (mod 4).

Thus, in Case 1 we find that $\omega(m) = 4$ and we conclude $\pi(m) = k$.

<u>Case 2:</u> $\alpha(m) = \frac{k}{2}$.

If $\omega(m) = 4$, then $\pi(m) = 2k \equiv 0$ (mod 8). But by Theorem 3.1, if $\pi(m) \equiv 0$ (mod 8), then $\omega(m) = 2$, a contradiction.

If $\omega(m) = 1$, then $\pi(m) = \frac{k}{2} \equiv 2$ (mod 4). We can now apply Theorem 2.1(1)($\Rightarrow$), and we find $m \mid L_{\pi(m)/2} = L_{k/4}$. However, this contradicts our hypothesis (b).

Thus, in Case 2 we find $\omega(m) = 2$ and we conclude $\pi(m) = k$. □

Proof of Theorem 2.1(3). ($\Rightarrow$) Assume $k \equiv 0$ (mod 8) and $\pi(m) = k$. Since $\pi(m) \equiv 0$ (mod 8), Theorem 3.1 tells us that $\omega(m) = 2$, and so $\alpha(m) = \frac{k}{2}$. Thus, $m \mid F_{k/2}$ and $m \nmid F_q$ for any q such that $1 \leq q < \frac{k}{2}$. In particular, $m \nmid F_q$ for all q such that $q \mid \frac{k}{2}$ and $q \neq \frac{k}{2}$, and this direction of the proof is complete.

($\Leftarrow$) Assume $k \equiv 0$ (mod 8), and (a) $m \mid F_{k/2}$, and (b) $m \nmid F_q$ for all q such that $q \mid \frac{k}{2}$ and $q \neq \frac{k}{2}$. We must prove that $\pi(m) = k$. By (a), we see $\alpha(m) \mid \frac{k}{2}$, and by (b), we deduce that in fact $\alpha(m) = \frac{k}{2}$. Thus $\alpha(m) \equiv 0$ (mod 4). By Theorem 3.1, this can only happen when $\omega(m) = 2$. Thus $\pi(m) = k$. □

References

[Flanagan et al. 2015] P. Flanagan, M. S. Renault, and J. Updike, "Symmetries of Fibonacci points, mod m", *Fibonacci Quart.* **53**:1 (2015), 34–41. MR

[Gupta et al. 2012] S. Gupta, P. Rockstroh, and F. E. Su, "Splitting fields and periods of Fibonacci sequences modulo primes", *Math. Mag.* **85**:2 (2012), 130–135. MR Zbl

[Renault 1996] M. Renault, *The Fibonacci sequence under various moduli*, master's thesis, Wake Forest University, 1996, available at http://webspace.ship.edu/msrenault/fibonacci/FibThesis.pdf.

[Robinson 1963] D. W. Robinson, "The Fibonacci matrix modulo m", *Fibonacci Quart* **1**:2 (1963), 29–36. MR Zbl

[Stein et al. 2016] W. A. Stein et al., *Sage mathematics software*, Version 6.7, Sage Development Team, 2016, available at http://www.sagemath.org.

[Vinson 1963] J. Vinson, "The relation of the period modulo to the rank of apparition of m in the Fibonacci sequence", *Fibonacci Quart* **1**:2 (1963), 37–45. MR Zbl

[Wall 1960] D. D. Wall, "Fibonacci series modulo m", *Amer. Math. Monthly* **67** (1960), 525–532. MR Zbl

ajdish@udel.edu   Department of Mathematical Sciences,
                                 University of Delaware, Newark, DE, United States

msrenault@ship.edu   Mathematics Department, Shippensburg University,
                                 Shippensburg, PA, United States

On the faithfulness of the representation of GL(n) on the space of curvature tensors

Corey Dunn, Darien Elderfield and Rory Martin-Hagemeyer

(Communicated by Kenneth S. Berenhaut)

We prove that the standard representation of GL(n) on the space of algebraic curvature tensors is almost faithful by showing that the kernel of this representation contains only the identity map and its negative. We additionally show that the standard representation of GL(n) on the space of algebraic covariant derivative curvature tensors is faithful.

1. Introduction

Let V be a finite-dimensional real vector space. An *algebraic curvature tensor* on V (or ACT for short) is a multilinear function

$$R : V \times V \times V \times V \to \mathbb{R}$$

that satisfies the following for all $x, y, z, w \in V$:

$$R(x, y, z, w) = -R(y, x, z, w), \quad R(x, y, z, w) = R(z, w, x, y),$$
$$0 = R(x, y, z, w) + R(z, x, y, w) + R(y, z, x, w). \tag{1-a}$$

The last of these is called the *first Bianchi identity*. Let $\mathcal{A}(V)$ be the set of all algebraic curvature tensors on V. As a set of real-valued functions, it is easy to check that $\mathcal{A}(V)$ is a vector space under the usual operations of summing the functions and scaling by real numbers; see [Gilkey 2001, p. 23].

There is another multilinear function on V that we study here. An *algebraic covariant derivative curvature tensor* on V (or ACDCT for short) is a multilinear function

$$R_1 : V \times V \times V \times V \times V \to \mathbb{R}$$

that satisfies the following for all $x, y, z, w, v \in V$:

$$
\begin{aligned}
R_1(x, y, z, w; v) &= -R_1(y, x, z, w; v), \\
R_1(x, y, z, w; v) &= R_1(z, w, x, y; v), \\
0 = R_1(x, y, z, w; v) &+ R_1(z, x, y, w; v) + R_1(y, z, x, w; v), \\
0 = R_1(x, y, z, w; v) &+ R_1(x, y, v, z; w) + R_1(x, y, w, v; z).
\end{aligned}
\tag{1-b}
$$

The first three properties of R_1 are similar to those of R, while the last property is referred to as the *second Bianchi identity*. Let $\mathcal{A}_1(V)$ be the set of ACDCT on V. $\mathcal{A}_1(V)$ is similar to $\mathcal{A}(V)$ in that it is a vector space as well; see [Gilkey 2001, p. 26].

These multilinear objects play a central role in the area of differential geometry. If g is a pseudo-Riemannian metric on a manifold M, then the curvature tensor R^g and its covariant derivative ∇R^g have the same symmetries of R and R_1, respectively, upon restriction to a point of the manifold (when one uses the Levi-Civita connection to construct them).

Let the *general linear group*, denoted $\mathrm{GL}(n)$, be the set of all invertible linear transformations $A : V \to V$. There is a natural action of $\mathrm{GL}(n)$ on both $\mathcal{A}(V)$ and $\mathcal{A}_1(V)$ that defines representations ρ and ρ_1 of $\mathrm{GL}(n)$ on $\mathcal{A}(V)$ and $\mathcal{A}_1(V)$, respectively. Define

$$
\begin{aligned}
\rho(A)(R)(x, y, z, w) &= R(A^{-1}x, A^{-1}y, A^{-1}z, A^{-1}w), \\
\rho_1(A)(R_1)(x, y, z, w; v) &= R_1(A^{-1}x, A^{-1}y, A^{-1}z, A^{-1}w; A^{-1}v).
\end{aligned}
\tag{1-c}
$$

For convenience, we simply express these actions of precomposition by the inverse of A by $\rho(A)(R) = A^*R$, and $\rho_1(A)(R_1) = A^*R_1$.

These representations have been studied by previous authors. The representation of the orthogonal group on $\mathcal{A}(V)$ decomposes into eight irreducible subspaces, see [Gilkey 2007; Blažić et al. 2006], with geometric significance. For example, one of these irreducible subspaces is the space of Weyl conformal curvature tensors. The action of $\mathrm{GL}(n)$ on the space $\mathcal{A}_1(V)$ was studied in [Strichartz 1988].

By definition, if G is a group, W is a vector space, and τ is a representation of G on W (that is, τ is a homomorphism from G to the endomorphisms of W), then τ is a *faithful* representation if $\ker(\tau)$ is trivial. In addition, τ is *almost faithful* if $\ker(\tau)$ is a discrete subgroup of G (in the event G is a Lie group, this is equivalent to $\ker(\tau)$ being a zero-dimensional subgroup of G).

It is our goal to investigate the faithfulness of the representations ρ and ρ_1 described above in (1-c). After establishing some supporting lemmas in Section 2, we establish the following theorem in Section 3:

Theorem 1.1. *The representation ρ in* (1-c) *is almost faithful. In fact,* $\ker(\rho) = \{\pm I\}$.

We go on to prove the following result concerning ρ_1.

Theorem 1.2. *The representation ρ_1 in* (1-c) *is faithful.*

We describe an immediate corollary and application to these main results concerning groups of symmetries of curvature tensors. Following [Dunn et al. 2015], we define the *structure group G_T* of an ACT or ACDCT T to be the following subgroup of $\mathrm{GL}(n)$:

$$G_T = \{A \in \mathrm{GL}(n) \mid A^*T = T\}.$$

One is interested in any data concerning structure groups for a variety of reasons, although one main purpose would be in constructing invariants — these invariants are then used to study the manifolds that these objects are derived from. See [Dunn 2009; Gilkey 2007] for more on the development of invariants from structure groups.

Corollary 1.3. *Let G_R be the structure group of the ACT R, and G_{R_1} be the structure group of the ACDCT R_1. If $I : V \to V$ is the identity map, then*

$$\bigcap_{R \in \mathcal{A}(V)} G_R = \{\pm I\} \quad \text{and} \quad \bigcap_{R_1 \in \mathcal{A}_1(V)} G_{R_1} = \{I\}.$$

Put differently, Theorems 1.1 and 1.2 demonstrate in this corollary that with exception to $\pm I$ (and only in the ACT case), there is no subgroup of $\mathrm{GL}(n)$ that preserves every ACT or every ACDCT.

2. Preliminary results

There are three preliminary results we shall need to establish our main results. The first two (Lemmas 2.1 and 2.3) concern a construction of ACTs and ACDCTs. The final preliminary result (Theorem 2.7) and a needed corollary (Corollary 2.8) concern the Jordan decomposition of a matrix.

Tensor constructions. Let $S^k(V)$ be the (vector) space of k-multilinear functions

$$\varphi : \times^k V \to \mathbb{R}$$

that are symmetric in every slot. For example, $S^2(V)$ is the set of symmetric bilinear forms, and $S^3(V)$ is the set of totally symmetric trilinear forms. If $\varphi \in S^2(V)$ and $\psi \in S^3(V)$, define

$$R_\varphi(x, y, z, w) = \varphi(x, w)\varphi(y, z) - \varphi(x, z)\varphi(y, w),$$

$$(R_1)_{\varphi,\psi}(x, y, z, w; v) = \varphi(x, w)\psi(y, z, v) + \varphi(y, z)\psi(x, w, v) \tag{2-a}$$
$$- \varphi(x, z)\psi(y, w, v) - \varphi(y, w)\psi(x, z, v).$$

The R_φ and $(R_1)_{\varphi,\psi}$ described in (2-a) are referred to as *canonical* ACTs or ACDCTs. It can be shown that $R_\varphi \in \mathcal{A}(V)$ and $(R_1)_{\varphi,\psi} \in \mathcal{A}_1(V)$; see [Gilkey

2007]. In fact, it is known [Gilkey 2007, p. 47] that

$$\mathcal{A}(V) = \operatorname{span}\{R_\varphi \mid \varphi \in S^2(V)\}, \quad \mathcal{A}_1(V) = \operatorname{span}\{(R_1)_{\varphi,\psi} \mid \varphi \in S^2(V), \psi \in S^3(V)\}.$$

Moreover, these canonical ACTs and ACDCTs have geometric significance since they arise as the curvature tensor and its covariant derivative of a hypersurface embedding [Gilkey 2007].

We use the construction found in (2-a) to produce certain ACTs and CDACTs that will be of use to us.

Lemma 2.1. *Let $\{e_1, \ldots, e_n\}$ be a basis for V. Let i, j and k be given distinct indices*:

(1) *There exists $R \in \mathcal{A}(V)$ such that, up to the symmetries listed in (1-a), the only nonzero term is $R(e_i, e_j, e_j, e_i) = 1$.*

(2) *There exists $R \in \mathcal{A}(V)$ such that, up to the symmetries listed in (1-a), the only nonzero term is $R(e_i, e_j, e_k, e_i) = 1$.*

(3) *Given constants $c_{i,j}$ and $c_{i,j,k}$, there exists $R \in \mathcal{A}(V)$ such that, up to the symmetries listed in (1-a), the only nonzero terms of R are*

$$R(e_i, e_j, e_j, e_i) = c_{ij} \quad and \quad R(e_i, e_j, e_k, e_i) = c_{ijk}.$$

Proof. We prove these results by using the construction in (2-a). To prove the first assertion, define $\varphi \in S^2(V)$ by setting $\varphi(e_i, e_i) = \varphi(e_j, e_j) = 1$ and all other entries equal to zero. It is a now a routine check that $R_\varphi(e_i, e_j, e_j, e_i) = 1$ and all other curvature entries up to the symmetries listed in (1-a) are zero.

To prove the second assertion, define φ_1 to have the nonzero entries

$$\varphi_1(e_i, e_i) = \varphi_1(e_j, e_k) = \varphi_1(e_k, e_j) = 1.$$

We now have the following nonzero entries of R_{φ_1} up to the symmetries listed in (1-a):

$$R_{\varphi_1}(e_i, e_j, e_k, e_i) = 1 \quad and \quad R_{\varphi_1}(e_j, e_k, e_k, e_j) = -1.$$

By the first assertion, there exists an ACT $\tilde{R}$ such that the only nonzero entry up to the symmetries listed in (1-a) is $\tilde{R}(e_j, e_k, e_k, e_j) = 1$. We now complete the second assertion by defining $R = R_{\varphi_1} + \tilde{R}$.

To prove the final assertion, let the constants c_{ij} and c_{ijk} be given, and for every i, j, and k, using the previous assertions define the ACTs $R_{ij}, R_{ijk} \in \mathcal{A}(V)$ such that up to the symmetries listed in (1-a), the only nonzero entries of these ACTs are

$$R_{ij}(e_i, e_j, e_j, e_i) = 1 \quad and \quad R_{ijk}(e_i, e_j, e_k, e_i) = 1.$$

We can now define $R = \sum_{i,j} c_{ij} R_{ij} + \sum_{i,j,k} c_{ijk} R_{ijk}.$ $\qquad \square$

Remark 2.2. The notation R_{ij} and R_{ijk} will be used in the proof of Theorem 1.1.

We can prove a similar result concerning the construction of an ACDCT that has certain prescribed entries.

Lemma 2.3. *Let $\{e_1, \ldots, e_n\}$ be a basis for V. Let i, j be given distinct indices*:

(1) *There exists $R_1 \in \mathcal{A}_1(V)$ such that, up to the symmetries listed in (1-a), the only nonzero term is $R_1(e_i, e_j, e_j, e_i; e_j) = 1$.*

(2) *Given constants c_1 and c_2, there exists an $R_1 \in \mathcal{A}_1(V)$ such that*

$$R_1(e_1, e_2, e_2, e_1; e_1) = c_1 \quad and \quad R_1(e_1, e_2, e_2, e_1; e_2) = c_2.$$

Proof. We use the construction in (2-a). To prove the first assertion, define $\varphi \in S^2(V)$ and $\psi \in S^3(V)$ by having the nonzero values

$$\varphi(e_i, e_i) = 1, \quad \psi(e_j, e_j, e_j) = 1.$$

It is now a routine check that $(R_1)_{\varphi, \psi}(e_i, e_j, e_j, e_i; e_j) = 1$ is the only nonzero entry up to the symmetries listed in (1-a):

To prove the second assertion, let ${}^1R_1, {}^2R_1 \in \mathcal{A}(V)$ be given such that the only nonzero entries up to the symmetries listed in (1-a) are

$${}^1R_1(e_1, e_2, e_2, e_1; e_1) = 1 \quad and \quad {}^2R_1(e_1, e_2, e_2, e_1; e_2) = 1.$$

We then define $R_1 = c_1({}^1R_1) + c_2({}^2R_1)$, which satisfies the given conditions. □

Remark 2.4. According to the symmetries in (1-b), we have

$$R_1(e_i, e_j, e_j, e_i; e_j) = R_1(e_j, e_i, e_i, e_j; e_j).$$

So the ACDCT guaranteed to exist from Lemma 2.3 can be chosen to have the final index match the fourth index, or chosen to match the third, provided the first four indices are of the form (i, j, j, i) — or any of the other dependent forms derivable from the symmetries in (1-b).

Remark 2.5. The notation ${}^1R_1$ and ${}^2R_1$ will be used in the proof of Theorem 1.2.

Jordan normal form. We recall a familiar result from linear algebra concerning the Jordan normal form of a matrix; see [Adkins and Weintraub 1992] for details. To properly state this result, we make the following definitions.

Definition 2.6. Let $\lambda \in \mathbb{R}$, and let $a + b\sqrt{-1} \in \mathbb{C}$ with $a, b \in \mathbb{R}$ and $b > 0$. The *real Jordan block of size k corresponding to λ* is the $k \times k$ matrix $J(\lambda, k)$ of real numbers

$$J(\lambda, k) = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 & \cdots \\ 0 & 0 & \lambda & 1 \\ & \vdots & & & \ddots \end{bmatrix}.$$

The *complex Jordan block of size k corresponding to* $a + b\sqrt{-1}$ is the $2k \times 2k$ matrix $J(a, b, k)$ of real numbers

$$
J(a, b, k) = \begin{bmatrix} a & b & 1 & 0 & 0 & 0 \\ -b & a & 0 & 1 & 0 & 0 \cdots \\ 0 & 0 & a & b & 1 & 0 \\ 0 & 0 & -b & a & 0 & 1 \\ & \vdots & & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.
$$

We briefly recall the direct sum operation on matrices before we state Theorem 2.7. If A and B are matrices of any size, then one may create the new matrix $A \oplus B$ by defining

$$
A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix},
$$

where the 0 entries above denote the 0-matrix of appropriate size.

We can now state the famous result concerning the Jordan decomposition of a linear transformation.

Theorem 2.7 (Jordan normal form of a linear transformation). *Let $A : V \to V$ be any linear transformation. There exists a basis $\mathcal{B}$ for V such that the matrix representation $[A]_{\mathcal{B}}$ of A on $\mathcal{B}$ is the direct sum of Jordan blocks corresponding to the eigenvalues of A. Furthermore, the unordered collection of Jordan blocks is uniquely determined by A.*

Note that the expression of A into a direct sum of Jordan blocks is referred to as "expressing A in its Jordan normal form". We shall need the following corollary in Section 4.

Corollary 2.8. *Let $A : V \to V$ be any linear transformation. There exists a basis $\{e_1, \ldots, e_n\}$ for V such that* $\mathrm{span}\{e_1, e_2\}$ *is A-invariant.*

Proof. Find a basis for V that expresses A in its Jordan normal form, which is possible by Theorem 2.7. If A has any complex eigenvalues, rearrange this basis so that the corresponding complex Jordan block appears first. The result is now true in this case, since $Ae_1 = ae_1 - be_2$ and $Ae_2 = be_1 + ae_2$.

If A has no complex eigenvalues, then the Jordan normal form of A is a direct sum of real Jordan blocks of various sizes, and these real Jordan blocks are all upper triangular. Hence, their direct sum is upper triangular. Now we have

$$
Ae_i \in \mathrm{span}\{e_1, \ldots, e_i\}
$$

for every $i \geq 1$. In particular, this holds for $i = 2$. ☐

3. The representation on $\mathcal{A}(V)$ is almost faithful

Here, we establish Theorem 1.1. The general method of proof of Theorem 1.1 (and with minor adjustments, Theorem 1.2 in the next section) is as follows. If $A \in \mathrm{GL}(n)$ is given and $A \neq \pm I$, then we produce an ACT R such that $A^*R \neq R$. By expressing A^{-1} in its Jordan normal form, this comes down to a number of cases, and in each case we use Lemma 2.1 (or Lemma 2.3 in the next section) to produce the needed ACT (or ACDCT).

Proof of Theorem 1.1. Since any ACT R inputs four entries and is multilinear, we have $(\pm I)^*R = R$ for every R. In the language of representations, $\rho(\pm I)$ is the identity map on $\mathcal{A}(V)$; hence $\pm I \in \ker \rho$. We prove that if $A \neq \pm I$, then $A \notin \ker(\rho)$ by finding an ACT R for which $\rho(A)(R) \neq R$, demonstrating that $\rho(A)$ is not the identity on $\mathcal{A}(V)$.

Note that since $A \neq \pm I$ and each of these is self-inverse, $A^{-1} \neq \pm I$. We decompose A^{-1} into its Jordan normal form and proceed by cases depending on the first Jordan block in this form. Recall that since $A \in \mathrm{GL}(n)$, none of its eigenvalues are equal to 0.

(1) **The first Jordan block of A^{-1} is $J(\lambda, 1)$.** We break this case into several subcases.[1] Note that in what follows, $\lambda \in \mathbb{R}$.

(a) *The second Jordan block of A^{-1} is $J(\eta, 1)$.* Note that $\eta \in \mathbb{R}$. There are now three possibilities for the third Jordan block of A^{-1}:

(i) *All other Jordan blocks of A^{-1} are real and of size 1.* Since $A^{-1} \neq \pm I$, not all of the eigenvalues are 1 and not all of the eigenvalues are -1. Thus there is at least one real eigenvalue γ of A^{-1} that differs from either λ or η. Without loss of generality, suppose $\gamma \neq \lambda$, and $J(\gamma, 1)$ is the third Jordan block of A^{-1}. Then for an arbitrary ACT R, we have

$$A^*R(e_1, e_2, e_2, e_1) = \lambda^2\eta^2 R(e_1, e_2, e_2, e_1),$$
$$A^*R(e_1, e_3, e_3, e_1) = \lambda^2\gamma^2 R(e_1, e_3, e_3, e_1),$$
$$A^*R(e_2, e_3, e_3, e_2) = \eta^2\gamma^2 R(e_2, e_3, e_3, e_2),$$
$$A^*R(e_2, e_1, e_3, e_2) = \eta^2\lambda\gamma R(e_2, e_1, e_3, e_2).$$

Using the notation of Lemma 2.1, choose $R = R_{12} + R_{13} + R_{23} + R_{213}$. Then if $A^*R = R$, the above equations results in the system of equations

$$\lambda^2\eta^2 = \lambda^2\gamma^2 = \eta^2\gamma^2 = \eta^2\lambda\gamma = 1.$$

---

[1]It is not surprising that this is the most complicated case: $J(\lambda, 1)$ is a Jordan block of $\pm I$ when $\lambda = \pm 1$ and so further distinguishing features of A^{-1} are needed.

Since these eigenvalues are nonzero, we see from the first three of these that $\lambda^2 = \eta^2 = \gamma^2$, and so $\lambda^4 = \eta^4 = \gamma^4 = 1$. Thus, each of these is either ± 1. But $\lambda \neq \gamma$, so they must be of opposite signs. In this case, $\lambda\gamma = -1$, but then in the final expression above $\eta^2\lambda\gamma = -1$, a contradiction.

(ii) *In the remaining Jordan blocks, there exists a real Jordan block of size greater than or equal to* 2. Suppose the next Jordan block is $J(\gamma, k)$ for $k \geq 1$. Notice that for any ACT R we would then have

$$A^*R(e_1, e_3, e_4, e_1) = \lambda^2\gamma R(e_1, e_3, e_3, e_1) + \lambda^2\gamma^2 R(e_1, e_3, e_4, e_1).$$

So, if $R = R_{13}$, then $R_{13}(e_1, e_3, e_4, e_1) = 0$, while

$$A^*R_{13}(e_1, e_3, e_4, e_1) = \lambda^2\gamma \neq 0,$$

a contradiction if $A^*R = R$.

(iii) *The remaining Jordan blocks of A^{-1} are complex.* Suppose the next Jordan block of A^{-1} is $J(a + b\sqrt{-1}, k)$ for $k \geq 1$. If R is an arbitrary ACT, then

$A^*R(e_1, e_3, e_3, e_1)$
$$= \lambda^2 a^2 R(e_1, e_3, e_3, e_1) - 2\lambda^2 ab R(e_1, e_3, e_4, e_1) + \lambda^2 b^2 R(e_1, e_4, e_4, e_1).$$

Then we recall that $b \neq 0$ and notice that if $R = R_{13} + a/(2b)R_{134}$, then $A^*R = R$ implies the left side of this equation is 1, while the right side is 0, a contradiction.

(b) *The second Jordan block is real and of size at least* 2. Suppose the second Jordan block is $J(\eta, k)$ for $k \geq 2$. It will be helpful in comparison to Case (2a) later to note here that our assumptions have

$$A^{-1}(e_1) = \lambda e_1, \quad A^{-1}(e_2) = \eta e_2, \quad A^{-1}(e_3) = \eta e_3 + e_2. \quad\quad (3\text{-a})$$

Now if R is an arbitrary ACT, we have

$A^*R(e_1, e_3, e_3, e_1)$
$$= \lambda^2 R(e_1, e_2, e_2, e_1) + \lambda^2\eta^2 R(e_1, e_3, e_3, e_1) + 2\lambda^2\eta R(e_1, e_2, e_3, e_1).$$

So if $R = -\eta^2 R_{12} + R_{13}$ and $A^*R = R$, then the left side of the equation is 1, while the right side is 0, another contradiction.

(c) *The second Jordan block is complex.* Suppose the second Jordan block is $J(a + b\sqrt{-1}, k)$ for $k \geq 1$. Then for an arbitrary ACT R, we have

$A^*R(e_1, e_2, e_2, e_1)$
$$= \lambda^2 a^2 R(e_1, e_2, e_2, e_1) + \lambda^2 b^2 R(e_1, e_3, e_3, e_1) + 2\lambda^2 ab R(e_1, e_2, e_3, e_1).$$

Recalling that $b \neq 0$, if $R = R_{12} + a/(2b)R_{123}$ and $A^*R = R$, then the left side of the equation is 1, while the right side is 0, another contradiction.

(2) **The first Jordan block of A^{-1} is $J(\lambda, 2)$.** There are two cases to consider concerning the second Jordan block:

(a) *The remaining Jordan blocks of A^{-1} are all real.* If the second Jordan block is $J(\eta, k)$ for $k \geq 1$, then we have

$$A^{-1}(e_1) = \lambda e_1, \quad A^{-1}(e_2) = \lambda e_2 + e_1, \quad A^{-1}(e_3) = \eta e_3. \qquad \text{(3-b)}$$

Comparing (3-b) to (3-a), one sees that under a permutation of the basis vectors, one reproduces the Case (1b) above.

(b) *There exists a complex Jordan block in A^{-1}.* Suppose the second Jordan block is $J(a + b\sqrt{-1}, k)$ for some k. Then for an arbitrary ACT R, we have

$A^*R(e_1, e_3, e_3, e_1)$
$$= \lambda^2 b^2 R(e_1, e_4, e_4, e_1) + \lambda^2 a^2 R(e_1, e_3, e_3, e_1) + 2\lambda^2 ab R(e_1, e_3, e_4, e_1).$$

Recalling that $b \neq 0$, if $R = R_{13} + a/(2b)R_{134}$ and $A^*R = R$, then the left side of the equation is 1, while the right side is 0, another contradiction.

(3) **The first Jordan block of A^{-1} is $J(\lambda, m)$ for $m \geq 3$.** For an arbitrary ACT R, we have

$$A^*R(e_1, e_3, e_3, e_1) = \lambda^2 R(e_1, e_2, e_2, e_1) + 2\lambda^3 R(e_1, e_2, e_3, e_1) + \lambda^4 R(e_1, e_3, e_3, e_1).$$

If $R = -\lambda^2 R_{12} + R_{13}$ and $A^*R = R$, then the left side of the equation is 1, while the right side is 0, another contradiction.

(4) **The first Jordan block of A^{-1} is $J(a, b, 1)$.** There are two cases to consider:

(a) *All remaining Jordan blocks of A^{-1} are real.* Permuting basis vectors only reorders the Jordan blocks in A^{-1}. Thus, if there are other real Jordan blocks after one complex Jordan block, one may reorder the basis vectors to have the real Jordan blocks appear first. Thus, we reproduce one of the previous cases.

(b) *There exists another complex Jordan block in A^{-1}.* For an arbitrary ACT R, we have

$A^*R(e_1, e_2, e_3, e_1) = ac(a^2 + b^2)R(e_1, e_2, e_3, e_1) + bc(a^2 + b^2)R(e_2, e_1, e_3, e_2)$
$$- ad(a^2 + b^2)R(e_1, e_2, e_4, e_1) - bd(a^2 + b^2)R(e_2, e_1, e_4, e_2).$$

Recall that b and d are nonzero. If $R = ac/(bd)R_{214} + R_{123}$ and $A^*R = R$, then the left side of the equation is 1, while the right side is 0, another contradiction.

(5) The first Jordan block of A^{-1} is $J(a, b, m)$ for $m \geq 2$. For an arbitrary ACT R, we have

$$A^* R(e_1, e_2, e_3, e_1) = b(a^2 + b^2) R(e_1, e_2, e_2, e_1) + a^2(a^2 + b^2) R(e_1, e_2, e_3, e_1)$$
$$+ ab(a^2 + b^2) R(e_2, e_1, e_3, e_2) - ab(a^2 + b^2) R(e_1, e_2, e_4, e_1)$$
$$- b^2(a^2 + b^2) R(e_2, e_1, e_4, e_2).$$

If $R = (a^2/b^2) R_{214} + R_{123}$ and $A^* R = R$, then the left side of the equation is 1, while the right side is 0, another contradiction.

To summarize, when given any Jordan decomposition of A^{-1} and $A \neq \pm I$, there exists an ACT R for which $A^* R \neq R$. Since $(\pm I)^* R = R$, the only ACT for which $A^* R = R$ for all R is when $A = \pm I$. As a result, $\rho(A)$ is the identity endomorphism on the space of algebraic curvature tensors precisely when $A = \pm I$.   □

4. The representation on $\mathcal{A}_1(V)$ is faithful

We conclude the paper by establishing Theorem 1.2.

Proof of Theorem 1.2. Unlike the proof of Theorem 1.1, by Corollary 2.8 we only need to consider three possible Jordan forms that occupy the upper 2×2 part of the matrix A^{-1}.

(1) There is a complex Jordan block in A^{-1}. Suppose the first Jordan block of A^{-1} is $J(a, b, k)$. For an arbitrary ACDCT R_1, we have

$$A^* R_1(e_1, e_2, e_2, e_1; e_1) = (a^2 + b^2)^2 \big(a R_1(e_1, e_2, e_2, e_1; e_1) - b R_1(e_1, e_2, e_2, e_1; e_2) \big).$$

Recall that $b \neq 0$. Using the notation of Lemma 2.3, if $R_1 = ({}^1 R_1) + a/b({}^2 R_1)$ and $A^* R_1 = R_1$, then the left side of the equation is 1, while the right side is 0, a contradiction.

(2) There are only real Jordan blocks, and there is at least one of size 2 or more. Suppose the first Jordan block of A^{-1} is $J(\lambda, k)$ for $k \geq 2$. For an arbitrary ACDCT R_1, we have

$$A^* R_1(e_1, e_2, e_2, e_1; e_2) = \lambda^4 (R_1(e_1, e_2, e_2, e_1; e_1)) + \lambda^5 (R_1(e_1, e_2, e_2, e_1; e_2)).$$

If $R_1 = -\lambda({}^1 R_1) + ({}^2 R_1)$ and $A^* R_1 = R_1$, then the left side of the equation is 1, while the right side is 0, another contradiction.

(3) There are only real Jordan blocks, all of which have size 1. Suppose without loss of generality that the first Jordan block of A^{-1} is $J(\lambda, 1)$. The next Jordan block $J(\eta, 1)$, by assumption, is a real one of size 1, and hence we have the relations

$A^{-1}e_1 = \lambda e_1$, and $A^{-1}e_2 = \eta e_2$. For an arbitrary CDACT R_1, we have

$$A^*R_1(e_1, e_2, e_2, e_1; e_1) = \lambda^3 \eta_2 R_1(e_1, e_2, e_2, e_1; e_1),$$

$$A^*R_1(e_1, e_2, e_2, e_1; e_2) = \lambda^2 \eta_3(\lambda R_1(e_1, e_2, e_2, e_1; e_2)).$$

If $R_1 = (^1R_1)$ and $A^*R_1 = R_1$, we conclude $\lambda^3\eta^2 = 1$. If $R_1 = {}^2R_1$, then we conclude $\lambda^2\eta^3 = 1$. Since both must happen simultaneously, we have $\lambda = \eta$, and $\lambda^5 = 1$, so $\lambda = \eta = 1$. We have shown that for *any* other real Jordan block $J(\eta, 1)$, for any k, $\eta = \lambda = 1$. Thus since there are only real Jordan blocks of size 1, the only way $A^*R_1 = R_1$ for all $R_1 \in \mathcal{A}(V)$ is if all Jordan blocks of A^{-1} are $J(1, 1)$ and as a result $A^{-1} = A = I$. $\qquad\square$

Acknowledgments

References

[Adkins and Weintraub 1992] W. A. Adkins and S. H. Weintraub, *Algebra — an approach via module theory*, Graduate Texts in Mathematics **136**, Springer, 1992. MR Zbl

[Blažić et al. 2006] N. Blažić, P. Gilkey, S. Nikčević, and U. Simon, "Algebraic theory of affine curvature tensors", *Arch. Math.* (*Brno*) **42** (2006), 147–168. MR Zbl

[Dunn 2009] C. Dunn, "A new family of curvature homogeneous pseudo-Riemannian manifolds", *Rocky Mountain J. Math.* **39**:5 (2009), 1443–1465. MR Zbl

[Dunn et al. 2015] C. Dunn, C. Franks, and J. Palmer, "On the structure group of a decomposable model space", *Beitr. Algebra Geom.* **56**:1 (2015), 199–216. MR Zbl

[Gilkey 2001] P. B. Gilkey, *Geometric properties of natural operators defined by the Riemann curvature tensor*, World Scientific, River Edge, NJ, 2001. MR Zbl

[Gilkey 2007] P. B. Gilkey, *The geometry of curvature homogeneous pseudo-Riemannian manifolds*, ICP Advanced Texts in Mathematics **2**, Imperial College Press, London, 2007. MR Zbl

[Strichartz 1988] R. S. Strichartz, "Linear algebra of curvature tensors and their covariant derivatives", *Canad. J. Math.* **40**:5 (1988), 1105–1143. MR Zbl

cmdunn@csusb.edu   *Mathematics Department, California State University at San Bernardino, San Bernardino, CA, United States*

dcelderf@ncsu.edu   *Mathematics Department, North Carolina State University, Raleigh, NC, United States*

rory.martinhageme001@umb.edu   *Mathematics Department, Rutgers University, Piscataway, NJ, United States*

Quasipositive curvature on a biquotient of Sp(3)

Jason DeVito and Wesley Martin

(Communicated by Kenneth S. Berenhaut)

Suppose $\phi_3 : \mathrm{Sp}(1) \to \mathrm{Sp}(2)$ denotes the unique irreducible complex 4-dimensional representation of $\mathrm{Sp}(1) = \mathrm{SU}(2)$, and consider the two subgroups $H_i \subseteq \mathrm{Sp}(3)$ with $H_1 = \{\mathrm{diag}(\phi_3(q_1), q_1) : q_1 \in \mathrm{Sp}(1)\}$ and $H_2 = \{\mathrm{diag}(\phi_3(q_2), 1) : q_2 \in \mathrm{Sp}(1)\}$. We show that the biquotient $H_1 \backslash \mathrm{Sp}(3) / H_2$ admits a quasipositively curved Riemannian metric.

1. Introduction

Manifolds of positive sectional curvature have been studied extensively. Despite this, there are very few known examples of positively curved manifolds. In fact, other than spheres and projective spaces, every known compact simply connected manifold admitting a metric of positive curvature is diffeomorphic to an Eschenburg space [Eschenburg 1982; Aloff and Wallach 1975], Eschenburg's inhomogeneous flag manifold, the projectivized tangent bundle of $\mathbb{K}P^2$ with $\mathbb{K} \in \{\mathbb{C}, \mathbb{H}, \mathbb{O}\}$ [Wallach 1972], a Bazaikin space [Barden 1965], or the Berger space [1961], or a certain cohomogeneity one manifold which is homeomorphic, but not diffeomorphic, to $T^1 S^4$ [Dearricott 2011; Grove et al. 2011].

Because of the difficulty in constructing new examples, attention has turned to the easier problem of finding examples with quasi- or almost positive curvature. Recall that a Riemannian manifold is said to be quasipositively curved if it admits a nonnegatively curved metric with a point p for which the sectional curvatures of all 2-planes at p are positive. A Riemannian manifold is called almost positively curved if the set of points for which all 2-planes are positively curved is dense. Examples of manifolds falling into either of these cases are more abundant. See [DeVito et al. 2014; Dickinson 2004; Eschenburg and Kerin 2008; Gromoll and Meyer 1974; Kerin 2011; 2012; Kerr and Tapp 2014; Petersen and Wilhelm 1999; Tapp 2003; Wilhelm 2001; Wilking 2002].

In [DeVito et al. 2014], the first author, together with DeYeso, Ruddy, and Wesner, proves that there are precisely 15 biquotients of the form $\mathrm{Sp}(3) /\!\!/ \mathrm{Sp}(1)^2$ and show

that eight of them admit quasipositively curved metrics. We show that their methods can be adapted to work on a ninth example, called N_9 in [DeVito et al. 2014]. That is, we show N_9 admits a metric of quasipositive curvature as well.

To describe this example, we first set up notation. Let $\phi_3 : \mathrm{Sp}(1) = \mathrm{SU}(2) \to \mathrm{Sp}(2)$ denote the unique irreducible complex 4-dimensional representation of $\mathrm{Sp}(1)$. Further, let $G = \mathrm{Sp}(3)$, and let $H_1 = \{\mathrm{diag}(\phi_3(q_1), q_1) \in G : q_1 \in \mathrm{Sp}(1)$ and $H_2 = \{\mathrm{diag}(\phi_3(q_2), 1) \in G : q_2 \in \mathrm{Sp}(1)\}$. Finally, set $H = H_1 \times H_2 \subseteq G \times G$.

Theorem 1.1. *The biquotient $H_1 \backslash G / H_2$ admits a metric of quasipositive curvature.*

In fact, we show the metric constructed on G in [DeVito et al. 2014] is H invariant and the induced metric on N_9 is quasipositively curved.

Finally, we point out that one of the first steps in the proof, Proposition 2.3, does not hold for any of the remaining inhomogeneous biquotients of the form $\mathrm{Sp}(3) /\!/ \mathrm{Sp}(1)^2$. In particular, a new approach is needed to determine whether these other biquotients admit metrics of quasipositive curvature.

The outline of this paper is as follows. Section 2 will cover the necessary background, leading to a system of equations parameterized by $p \in G$, which govern the existence of a zero curvature plane at $[p^{-1}] \in G /\!/ H$. In Section 3, we find a particular point $p \in G$ for which there are no nontrivial solutions to the system of equations, establishing Theorem 1.1.

2. Background

We will use the setup of [DeVito et al. 2014]. As the calculations will be done on the Lie algebra level, we now describe all the relevant Lie algebras.

We recall the Lie algebra $\mathfrak{sp}(n)$ consists of all $n \times n$ quaternionic skew-Hermitian matrices with Lie bracket given by the commutator. That is, $\mathfrak{sp}(n) = \{A \in M_n(\mathbb{H}) : A + \bar{A}^t = 0\}$, where $\mathbb{H}$ denotes the skew-field of quaternions, and the Lie bracket is given by $[A, B] = AB - BA$. When $n = 1$, this Lie algebra is simply $\mathrm{Im}\,\mathbb{H}$.

Then the Lie algebra of $G = \mathrm{Sp}(3)$, denoted $\mathfrak{g} = \mathfrak{sp}(3)$, consists of the 3×3 skew-Hermitian matrices over $\mathbb{H}$. Further, we set $K = \mathrm{Sp}(2) \times \mathrm{Sp}(1)$, block diagonally embedded into G via $(A, q) \mapsto \mathrm{diag}(A, q) \in G$. Then one easily sees that $\mathfrak{k} = \mathfrak{sp}(2) \oplus \mathfrak{sp}(1)$ is embedded into $\mathfrak{g}$ via $(B, r) \mapsto \mathrm{diag}(B, r)$.

We also use the description of ϕ_3 on the Lie algebra level given by [DeVito et al. 2014, Proposition 4.5].

Proposition 2.1. *For $t = t_i + t_j + t_k \in \mathrm{Im}\,\mathbb{H} = \mathfrak{sp}(1)$,*

$$\phi_3(t) = \begin{bmatrix} 3t_i & \sqrt{3}(t_j + t_k) \\ \sqrt{3}(t_j + t_k) & 2(t_k - t_j) - t_i \end{bmatrix}$$

defines the unique irreducible 4-dimensional representation of $\mathfrak{sp}(1) = \mathfrak{su}(2)$.

It follows that, for $H_1 = \{\text{diag}(\phi_3(q_1), q_1) : q_1 \in \text{Sp}(1)\} \subseteq \text{Sp}(3)$,

$$\mathfrak{h}_1 = \left\{ \begin{bmatrix} 3t_i & \sqrt{3}(t_j + t_k) & \\ \sqrt{3}(t_j + t_k) & 2(t_k - t_j) - t_i & \\ & & t \end{bmatrix} : t \in \text{Im}\,\mathbb{H} \right\}.$$

Likewise, for $H_2 = \{\text{diag}(\phi_3(q_2), 1) : q_2 \in \text{Sp}(1)\} \subseteq G$, we have

$$\mathfrak{h}_2 = \left\{ \begin{bmatrix} 3s_i & \sqrt{3}(s_j + s_k) & \\ \sqrt{3}(s_j + s_k) & 2(s_k - s_j) - s_i & \\ & & 0 \end{bmatrix} : s \in \text{Im}\,\mathbb{H} \right\}.$$

The metric we will use is constructed in [DeVito et al. 2014] via a combination of Cheeger deformations [1973] and Wilking's doubling trick [2002]. More specifically, we let g_0 denote the bi-invariant metric on G with $g_0(X, Y) = -\text{Re}\,\text{Tr}(XY)$ for $X, Y \in \mathfrak{g}$. We let g_1 denote the left G-invariant, right K-invariant metric obtained by Cheeger deforming g_0 in the direction of K. That is, g_1 is the metric induced on G by declaring the canonical submersion $(G \times K, g_0 + g_0|_K) \to G$ with $(p, k) \mapsto pk^{-1}$ to be a Riemannian submersion.

We now equip $G \times G$ with the metric $g_1 + g_1$ and consider the isometric action of $G \times H_1 \times H_2$ on $G \times G$ given by $(p, h_1, h_2) * (p_1, p_2) = (pp_1 h_1^{-1}, pp_2 h_2^{-1})$. This action is free and induces a metric on the orbit space $\Delta G \backslash (G \times G)/(H_1 \times H_2)$.

Following Eschenburg [1984], the orbit space $\Delta G \backslash (G \times G)/(H_1 \times H_2)$ is canonically diffeomorphic to the biquotient $H_1 \backslash G / H_2$, which is called N_9 in [DeVito et al. 2014]. To see this, one verifies that the map $G \times G \to G$, sending (p_1, p_2) to $p_1^{-1} p_2$, descends to a diffeomorphism of the orbit spaces. We use this diffeomorphism to transport the submersion metric on $\Delta G \backslash (G \times G)/(H_1 \times H_2)$ to $H_1 \backslash G / H_2$ and let g_2 denote this metric on $H_1 \backslash G / H_2$.

We note that since g_0 is bi-invariant, it is nonnegatively curved. It follows from O'Neill's formula [1966] that g_1 and g_2 are nonnegatively curved as well.

We now describe the points having 0-curvature planes in $(H_1 \backslash G / H_2, g_2)$. To do this, we let

$$\mathfrak{p} = \left\{ \begin{bmatrix} 0 & 0 & z_1 \\ 0 & 0 & z_2 \\ -\bar{z}_1 & -\bar{z}_2 & 0 \end{bmatrix} : z_1, z_2 \in \mathbb{H} \right\} \subseteq \mathfrak{g}$$

denote the g_0-orthogonal complement of $\mathfrak{k}$: $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$. Then, for $X \in \mathfrak{g}$ we can write it as $X = X_\mathfrak{k} + X_\mathfrak{p}$, where $X_\mathfrak{k}$ is the projection of X onto $\mathfrak{k}$, and similarly for $X_\mathfrak{p}$. We also let $\text{Ad}_\mathfrak{p} : \mathfrak{g} \to \mathfrak{g}$ denote the adjoint map $\text{Ad}_\mathfrak{p}(X) = pXp^{-1}$. Then, as shown in [DeVito et al. 2014, Corollary 2.8], we have the following description of points $[p^{-1}] \in H_1 \backslash G / H_2$ containing 0-curvature planes.

Theorem 2.2. *There is a 0-curvature plane at $[p^{-1}] \in (H_1 \backslash G / H_2, g_2)$ if and only if there are linearly independent vectors $X, Y \in \mathfrak{g}$ satisfying the following equations*:

(A) $g_0(X, \mathrm{Ad}_p \, \mathfrak{h}_1) = g_0(X, \mathfrak{h}_2) = g_0(Y, \mathrm{Ad}_p \, \mathfrak{h}_1) = g_0(Y, \mathfrak{h}_2) = 0$,

(B) $[X, Y] = [X_{\mathfrak{k}}, Y_{\mathfrak{k}}] = [X_{\mathfrak{p}}, Y_{\mathfrak{p}}] = 0$,

(C) $[(\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{k}}, (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{k}}] = [(\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{p}}, (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{p}}] = 0$.

It is clear from inspecting these equations that if $\mathrm{span}\{X, Y\} = \mathrm{span}\{X', Y'\}$, then X and Y satisfy all three conditions if and only if X' and Y' do.

We also note that there is some redundancy in these equations because (G, K) is a symmetric pair. Specifically, assuming $[X, Y] = 0$, it follows that $[X_{\mathfrak{k}}, Y_{\mathfrak{k}}] = 0$ if and only if $[X_{\mathfrak{p}}, Y_{\mathfrak{p}}] = 0$ and also that $[(\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{k}}, (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{k}}] = 0$ if and only if $[(\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{p}}, (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{p}}] = 0$. To see this, we first note that $[\mathfrak{p}, \mathfrak{p}] \subseteq \mathfrak{k}$ for a symmetric pair (G, K). Using the fact that $[\mathfrak{k}, \mathfrak{p}] \subseteq \mathfrak{p}$, we see that $[X, Y]_{\mathfrak{k}} = [X_{\mathfrak{k}}, Y_{\mathfrak{k}}] + [X_{\mathfrak{p}}, Y_{\mathfrak{p}}]$. Since condition (B) forces $[X, Y]_{\mathfrak{k}} = 0$, we see that $[X_{\mathfrak{k}}, Y_{\mathfrak{k}}] = 0$ if and only if $[X_{\mathfrak{p}}, Y_{\mathfrak{p}}] = 0$. To get the result for the vectors $\mathrm{Ad}_{p^{-1}} X$ and $\mathrm{Ad}_{p^{-1}} Y$, we note that $\mathrm{Ad}_{p^{-1}} : \mathfrak{g} \to \mathfrak{g}$ is a Lie algebra isomorphism, so $[X, Y] = 0$ if and only if $[\mathrm{Ad}_{p^{-1}} X, \mathrm{Ad}_{p^{-1}} Y] = 0$.

We now show that for many $p \in \mathrm{Sp}(3)$, if X and Y satisfy conditions (A) and (B) of Theorem 2.2, then we may replace X and Y with X', Y' having a nice form.

Proposition 2.3. *Let $\rho : \mathfrak{g} \to \mathrm{Im} \, \mathbb{H}$ with $\rho(Z) = Z_{33}$, the entry of Z in the last row and last column. Suppose $[p^{-1}] \in G /\!/ H$ is a point for which $\rho|_{\mathrm{Ad}_p \, \mathfrak{h}_1}$ is surjective. If $X, Y \in \mathfrak{g}$ satisfy conditions (A) and (B) of Theorem 2.2 at the point $[p^{-1}]$, then there are vectors $X', Y' \in \mathfrak{g}$ with $\mathrm{span}\{X, Y\} = \mathrm{span}\{X', Y'\}$ and $X'_{\mathfrak{p}} = Y'_{\mathfrak{sp}(2)} = 0$, where $Y'_{\mathfrak{sp}(2)}$ denotes the projection of Y' to $\mathfrak{sp}(2) \oplus 0 \subseteq \mathfrak{k} \subseteq \mathfrak{g}$.*

Proof. We start with the equation $[X_{\mathfrak{p}}, Y_{\mathfrak{p}}] = 0$ from condition (B). Since we can identify $\mathfrak{p}$ with $T_{[eK]} G/K$, where $G/K = \mathbb{H}P^2$ has positive sectional curvature, it follows that $[X_{\mathfrak{p}}, Y_{\mathfrak{p}}] = 0$ if and only if $X_{\mathfrak{p}}$ and $Y_{\mathfrak{p}}$ are dependent. Thus, either $X_{\mathfrak{p}} = 0$ and $X = X'$ or $X_{\mathfrak{p}} = \lambda Y_{\mathfrak{p}}$ for some real number λ. Then $X' = \lambda X - Y$ has no $\mathfrak{p}$ part. We may thus assume without loss of generality that X has no $\mathfrak{p}$ part.

Since $\mathrm{Sp}(2) \times \{I\}$ is an ideal in $K = \mathrm{Sp}(2) \times \mathrm{Sp}(1)$, the condition $[X_{\mathfrak{k}}, Y_{\mathfrak{k}}] = 0$ implies $[X_{\mathfrak{sp}(2)}, Y_{\mathfrak{sp}(2)}] = 0$. By condition (A), we know $g_0(X, \mathfrak{h}_2) = g_0(Y, \mathfrak{h}_2) = 0$, so we may interpret $X_{\mathfrak{sp}(2)}$ and $Y_{\mathfrak{sp}(2)}$ as tangent vectors on $\mathrm{Sp}(2)/\phi_3(\mathrm{Sp}(1))$. But, $\mathrm{Sp}(2)/\phi_3(\mathrm{Sp}(1))$ is the Berger space [1961] and is known to admit a normal homogeneous metric of positive curvature. So we see that $[X_{\mathfrak{sp}(2)}, Y_{\mathfrak{sp}(2)}] = 0$ if and only if $X_{\mathfrak{sp}(2)}$ and $Y_{\mathfrak{sp}(2)}$ are linearly dependent.

If $X_{\mathfrak{sp}(2)} = 0$, then the only nonvanishing entry of X is X_{33}. Since, by assumption, $\rho|_{\mathrm{Ad}_p \, \mathfrak{h}_1}$ is surjective, the condition $g_0(X, \mathrm{Ad}_p \, \mathfrak{h}_1) = 0$ forces $X = 0$, contradicting the fact that $\{X, Y\}$ is linearly independent. Hence, we may assume $X_{\mathfrak{sp}(2)} \neq 0$.

Then, we may subtract an appropriate multiple of X from Y to obtain a new vector Y' with $Y'_{\mathfrak{sp}(2)} = 0$. $\qquad\square$

We now work out conditions (A), (B), and (C) of Theorem 2.2 more explicitly at a point of the form

$$p = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}.$$

We will always assume $\theta \in \left(0, \frac{1}{4}\pi\right)$. Also, we will often identify $\mathfrak{p}$, consisting of matrices of the form

$$\begin{bmatrix} 0 & 0 & z_1 \\ 0 & 0 & z_2 \\ -\bar{z}_1 & -\bar{z}_2 & 0 \end{bmatrix},$$

with $\mathbb{H}^2$ via the canonical $\mathbb{R}$-linear isomorphism mapping such a matrix to $\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$.

We note that for points of this form, $\rho|_{\mathrm{Ad}_p \mathfrak{h}_1}$ has an image consisting of all elements of $\mathrm{Im}\,\mathbb{H}$ of the form $3\sin^2\theta\, t_i + \cos^2\theta\, t$ for $t = t_i + t_j + t_k \in \mathrm{Im}\,\mathbb{H}$. Since $\cos^2\theta \neq 0$ because $\theta \in \left(0, \frac{1}{4}\pi\right)$, this map has no kernel, so it is surjective. In particular, the conditions of Proposition 2.3 are verified at all such p, and thus, we may assume

$$X = \begin{bmatrix} x_1 & x_2 & 0 \\ -\bar{x}_2 & x_3 & 0 \\ 0 & 0 & x_4 \end{bmatrix}$$

with $x_1, x_3, x_4 \in \mathrm{Im}\,\mathbb{H}$ and $x_2 \in \mathbb{H}$. Similarly, we may assume

$$Y = \begin{bmatrix} 0 & 0 & y_1 \\ 0 & 0 & y_2 \\ -\bar{y}_1 & -\bar{y}_2 & y_3 \end{bmatrix}$$

with $y_1, y_2 \in \mathbb{H}$ and $y_3 \in \mathrm{Im}\,\mathbb{H}$

Lemma 2.4. *For a point p of the above form and $X, Y \in \mathfrak{g}$, conditions (A), (B), and (C) of Theorem 2.2 are equivalent to the following list of conditions:*

$$x_1 y_1 + x_2 y_2 - y_1 x_4 = 0, \tag{1}$$
$$-\bar{x}_2 y_1 + x_3 y_2 - y_2 x_4 = 0, \tag{2}$$
$$\{x_4, y_3\} \text{ is linearly dependent over } \mathbb{R}. \tag{3}$$

For

$$v = \begin{bmatrix} \cos\theta \sin\theta\,(x_1 - x_4) \\ -\sin\theta\,\bar{x}_2 \end{bmatrix} \in \mathbb{H}^2 \cong \mathfrak{p}$$

and

$$w = \begin{bmatrix} \mathrm{Re}(y_1) + (\cos^2\theta - \sin^2\theta)\,\mathrm{Im}(y_1) - \sin\theta\cos\theta\,y_3 \\ \cos\theta\,y_2 \end{bmatrix} \in \mathbb{H}^2 \cong \mathfrak{p},$$

the following hold:

$$\text{the set } \{v, w\} \text{ is linearly dependent over } \mathbb{R}, \tag{4}$$

$$3(x_1)_i - (x_3)_i = 0, \tag{5_i}$$

$$\sqrt{3}(x_2)_j - (x_3)_j = 0, \tag{5_j}$$

$$\sqrt{3}(x_2)_k + (x_3)_k = 0, \tag{5_k}$$

$$(x_1)_i(-2\sin^2\theta) + (x_4)_i(1 + 2\sin^2\theta) = 0, \tag{6_i}$$

$$(x_2)_j(\cos\theta - 1)2\sqrt{3} + (x_1)_j\sin^2\theta + (x_4)_j\cos^2\theta = 0, \tag{6_j}$$

$$(x_2)_k(\cos\theta - 1)2\sqrt{3} + (x_1)_k\sin^2\theta + (x_4)_k\cos^2\theta = 0, \tag{6_k}$$

$$-4\sin\theta\cos\theta\,(y_1)_i + (2\sin^2\theta + 1)(y_3)_i = 0, \tag{7_i}$$

$$2\sin\theta\cos\theta\,(y_1)_j - 2\sqrt{3}\sin\theta\,(y_2)_j + \cos^2\theta\,(y_3)_j = 0, \tag{7_j}$$

$$2\sin\theta\cos\theta\,(y_1)_k - 2\sqrt{3}\sin\theta\,(y_2)_k + \cos^2\theta\,(y_3)_k = 0. \tag{7_k}$$

Proof. We first claim that condition (A) is equivalent to (5_i) through (7_k). To begin with, we note that since $Y_{\mathfrak{sp}(2)} = 0$ and $\mathfrak{h}_2 \subseteq \mathfrak{sp}(2) \oplus 0 \subseteq \mathfrak{k}$, the equation $g_0(Y, \mathfrak{h}_2) = 0$ is automatically satisfied.

Now, a calculation shows that for $s = s_i + s_j + s_k \in \mathrm{Im}\,\mathbb{H}$,

$$0 = g_0(X, \mathfrak{h}_2) = 3s_i x_1 + 2\sqrt{3}(s_j + s_k)\,\mathrm{Im}(x_2) + (2(s_k - s_j) - s_i)x_3.$$

Then, using each of $s = i$, $s = j$, and $s = k$ respectively gives (5_i), (5_j), (5_k) which, using linearity, are therefore equivalent to the condition that $g_0(X, \mathfrak{h}_2) = 0$.

Further, with $t = t_i + t_j + t_k \in \mathrm{Im}\,\mathbb{H}$, we compute

$$\mathrm{Ad}_p\,\mathfrak{h}_1 = \left\{ \begin{bmatrix} 3\cos^2\theta\,t_i + \sin^2\theta\,t & \sqrt{3}\cos\theta\,(t_j + t_k) & \cos\theta\sin\theta\,(t - 3t_i) \\ \sqrt{3}\cos\theta\,(t_j + t_k) & 2(t_k - t_j) - t_i & -\sqrt{3}\sin\theta\,(t_j + t_k) \\ \cos\theta\sin\theta\,(t - 3t_i) & -\sqrt{3}\sin\theta\,(t_j + t_k) & 3\sin^2\theta\,t_i + \cos^2\theta\,t \end{bmatrix} \right\}.$$

A calculation now shows that the expression $g_0(X, \mathrm{Ad}_p\,\mathfrak{h}_1)$ is given by the expression

$$(3\cos^2\theta\,t_i + \sin^2\theta\,t)x_1 + 2\sqrt{3}\cos\theta\,(t_j + t_k)\,\mathrm{Im}(x_2)$$
$$+ (2(t_k - t_j) - t_i)x_3 + (3\sin^2\theta\,t_i + \cos^2\theta\,t)x_4.$$

Substituting each of $t = i$, $t = j$, and $t = k$ and using (5_i), (5_j), and (5_k) to eliminate x_3 respectively gives (6_i), (6_j), and (6_k) after using $\sin^2\theta + \cos^2\theta = 1$.

Likewise, the equation $g_0(Y, \mathrm{Ad}_p\, \mathfrak{h}_1) = 0$ is equivalent to the vanishing of the expression

$$2\cos\theta\sin\theta(-3t_i+t)\,\mathrm{Im}(y_1)-2\sqrt{3}\sin\theta(t_j+t_k)\,\mathrm{Im}(y_2)+(3\sin^2\theta\, t_i+\cos^2\theta\, t)y_3.$$

Substituting each of $t = i$, $t = j$, and $t = k$ respectively gives (7_i), (7_j), and (7_k).

We next claim that (1), (2), and (3) are equivalent to condition (B) of Theorem 2.2. Computing, we see $[X, Y] = 0$ if and only if (1) and (2) are satisfied and $[x_4, y_3] = 0$. But this latter condition is equivalent to (3) since $\mathrm{Sp}(1) = S^3$ has positive sectional curvature. Further, $X_{\mathfrak{p}} = 0$, so $[X_{\mathfrak{p}}, Y_{\mathfrak{p}}] = 0$ and since $Y_{\mathfrak{sp}(2)} = 0$, condition (3) is satisfied if and only if $[X_{\mathfrak{k}}, Y_{\mathfrak{k}}] = 0$.

Lastly, we claim that (4) is equivalent to condition (C) of Theorem 2.2. To see this, first recall that it was shown directly following Theorem 2.2 that the conditions $[(\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{k}}, (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{k}}] = 0$ and $[(\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{p}}, (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{p}}] = 0$ are equivalent, so we may focus on only one of these.

A direct calculation shows that $v = (\mathrm{Ad}_{p^{-1}} X)_{\mathfrak{p}}$ and $w = (\mathrm{Ad}_{p^{-1}} Y)_{\mathfrak{p}}$, so we need only argue that $[v, w] = 0$ if and only if v and w are dependent over $\mathbb{R}$. But we may interpret v, w as elements of $T_{[eK]}G/K$ where $G/K = \mathbb{H}P^2$ has a normal bi-invariant metric of positive sectional curvature. It follows that the bracket of v and w vanishes if and only if v and w are linearly dependent. $\qquad\square$

3. Quasipositive curvature

In this section, we prove $N_9 = H_1 \backslash \mathrm{Sp}(3)/H_2$ is quasipositively curved with the metric g_2 constructed in Section 2. As mentioned above, the metric g_2 is nonnegatively curved, so it is sufficient to find a single point for which all 2-planes have nonzero curvature. In fact, we will show the following theorem.

Theorem 3.1. *With respect to the metric g_2, N_9 is positively curved at points of the form $[p^{-1}] \in H_1\backslash G/H_2 \cong N_9$, where*

$$p = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}$$

with $\theta \in \left(0, \frac{1}{6}\pi\right)$.

We will always work with points p of the above form.

Assume $[p^{-1}] \in H_1\backslash G/H_2$ is a point for which there is a 0-curvature plane. Then, using Theorem 2.2 and Proposition 2.3, it follows that there are linearly independent $X, Y \in \mathfrak{g} = \mathfrak{sp}(3)$ with

$$X = \begin{bmatrix} x_1 & x_2 & 0 \\ -\bar{x}_2 & x_3 & 0 \\ 0 & 0 & x_4 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 0 & 0 & y_1 \\ 0 & 0 & y_2 \\ -\bar{y}_1 & -\bar{y}_2 & y_3 \end{bmatrix}$$

and which satisfy all of the conditions given by Lemma 2.4. By repeatedly applying the conditions of Lemma 2.4, we will constrain the forms of X and Y until we finally find that no such X and Y exist. This contradiction will establish that there are no zero curvature planes at $[p]^{-1}$, and hence, that N_9 is positively curved at these points.

Proposition 3.2. *If $\theta \in \left(0, \frac{1}{6}\pi\right)$, the two vectors*

$$v = \begin{bmatrix} \cos\theta \sin\theta \, (x_1 - x_4) \\ -\sin\theta \, \bar{x}_2 \end{bmatrix}, \ w = \begin{bmatrix} \mathrm{Re}(y_1) + (\cos^2\theta - \sin^2\theta) \, \mathrm{Im}(y_1) - \sin\theta \cos\theta \, y_3 \\ \cos\theta \, y_2 \end{bmatrix}$$

are both nonzero.

Proof. Suppose for a contradiction that $v = 0$. Since $0 < \theta < \frac{1}{6}\pi$, we have that $v = 0$ implies $x_2 = 0$ and $x_1 = x_4$. Then (6_i), (6_j), and (6_k) imply $x_1 = x_4 = 0$. Then (5_i), (5_j), and (5_k) imply that x_3 vanishes as well. Thus, in this case, $X = 0$, contradicting the fact that X and Y are linearly independent. Thus, $v \neq 0$.

Now, suppose $w = 0$, so $y_2 = 0$, $\mathrm{Re}(y_1) = 0$ and

$$\mathrm{Im}(y_1) = y_1 = \frac{\sin\theta \cos\theta}{\cos^2\theta - \sin^2\theta} y_3 = \tfrac{1}{2}\tan(2\theta)y_3. \tag{8}$$

This equation implies that the i, j, and k components of y_1 and y_3 are positive multiples of each other. However, (7_j) and (7_k) imply that the j and k components of y_1 and y_3 are negative multiples of each other. Thus, we must have $(y_1)_j = (y_1)_k = (y_3)_j = (y_3)_k = 0$.

Solving (7_i) for $y_1 = (y_1)_i$ and combining with (8), we see that either $y_1 = y_3 = 0$, or θ must satisfy the equation

$$\frac{\sin\theta \cos\theta}{\cos^2\theta - \sin^2\theta} = \frac{2\sin^2\theta + 1}{4\sin\theta \cos\theta}.$$

Clearing denominators and simplifying gives $2\sin^2\theta \cos^2\theta + 2\sin^4\theta + \sin^2\theta = \cos^2\theta$. Factoring $\sin^2\theta$ out of the expression $2\sin^2\theta \cos^2\theta + 2\sin^4\theta$, we see this expression simplifies to $2\sin^2\theta$. Substituting this back in gives the equation $3\sin^2\theta = \cos^2\theta$, which has no solutions in $\left(0, \frac{1}{6}\pi\right)$.

Thus, for $\theta \in \left(0, \frac{1}{6}\pi\right)$, we conclude $y_1 = y_3 = 0$, which implies $Y = 0$, again contradicting the fact that X and Y are linearly independent. □

Using (4), it follows that by rescaling X, we may thus assume $v = w$. Further, the first component of v is purely imaginary, and hence $\mathrm{Re}(y_1) = 0$, that is, $y_1 = \mathrm{Im}\, y_1$. Thus, (4) is equivalent to the following two equations:

$$\cos\theta \sin\theta (x_1 - x_4) = (\cos^2\theta - \sin^2\theta)y_1 - \sin\theta \cos\theta \, y_3, \tag{9}$$

$$y_2 = -\tan\theta \, \bar{x}_2. \tag{10}$$

Proposition 3.3. *For any $\theta \in \left(0, \frac{1}{6}\pi\right)$, x_2, y_1, and y_2 are all nonzero.*

Proof. Assume for a contradiction that $y_2 = 0$. Note that, because all the coefficients in (7_i), (7_j), (7_k) are nonzero, it follows that $y_1 = 0$ if and only if $y_3 = 0$. Because $Y \neq 0$, it follows that $y_1 \neq 0$.

Rearranging (1) gives $x_1 y_1 = y_1 x_4$. Taking lengths, we see that $|x_1| = |x_4|$. We now compare the i, j, and k components of x_1 and x_4.

For the i component, we rearrange (6_i) to obtain

$$(x_1)_i = \frac{1 + 2\sin^2\theta}{2\sin^2\theta}(x_4)_i = \left(1 + \frac{1}{2\sin^2\theta}\right)(x_4)_i.$$

Since the sum in the parentheses is positive, we conclude that $|(x_1)_i| \geq |(x_4)_i|$, with equality if and only if $(x_1)_i = (x_4)_i = 0$.

For the j component, we first remark that (10) shows that $x_2 = 0$ because $y_2 = 0$. Then, rearranging (6_j) gives

$$(x_1)_j = -\frac{\cos^2\theta}{\sin^2\theta}(x_4)_j.$$

Thus, since $0 < \theta < \frac{1}{6}\pi$, we conclude that $|(x_1)_j| \geq |(x_4)_j|$ with equality if and only if $(x_1)_j = (x_4)_j = 0$. The same argument shows $|(x_1)_k| \geq |(x_4)_k|$ with equality if and only if $(x_1)_k = (x_4)_k = 0$.

Thus, each component of x_1 is at least as large, in magnitude, as the corresponding component of x_4. Hence, since $|x_1| = |x_4|$, it follows that each of these inequalities must be equalities, so $x_1 = x_4 = 0$. Since we have already shown $x_2 = 0$, equations (5_i), (5_j), and (5_k) force $x_3 = 0$ as well. That is, $X = 0$, a contradiction. Thus, $y_2 \neq 0$.

Finally, it follows from (10) that $x_2 \neq 0$. From (1) and the fact that $x_2 y_2 \neq 0$, we see that $y_1 \neq 0$. $\qquad\square$

Proposition 3.4. *For every $\theta \in \left(0, \frac{1}{6}\pi\right)$, $x_1 \neq x_4$.*

Proof. Suppose for a contradiction that $x_1 = x_4$. Then (1) takes the form

$$0 = x_1 y_1 - y_1 x_1 - \tan\theta\,|x_2|^2 = [x_1, y_1] - \tan\theta\,|x_2|^2.$$

Since $x_1, y_1 \in \mathrm{Im}\,\mathbb{H}$, we know $[x_1, y_1] \in \mathrm{Im}\,\mathbb{H}$ as well, so we conclude that $\tan\theta\,|x_2|^2 = 0$. Since $0 < \theta < \frac{1}{6}\pi$, it follows that $x_2 = 0$, a contradiction. $\qquad\square$

Our next goal is to demonstrate the following proposition.

Proposition 3.5. *For every $\theta \in \left(0, \frac{1}{6}\pi\right)$, $\dim_{\mathbb{R}} \mathrm{span}_{\mathbb{R}}\{x_1, x_4, y_1, y_3\} = 1$.*

Proof. Since, by Proposition 3.3, $y_1 \neq 0$, the dimension of this span is at least 1, so we need only show it is at most one.

We deal first with the case $x_4 = 0$. Then (1) takes the form $x_1 y_1 - \tan\theta\,|x_2|^2 = 0$. In particular, $x_1 y_1 \in \mathbb{R}$. Since x_1 and y_1 are purely imaginary, this implies $\{x_1, y_1\}$

is linearly dependent over $\mathbb{R}$. Now (10) implies that $y_3 = -x_1 + 2y_1/\tan(2\theta)$, so $\{x_1, y_1, y_3\}$ is linearly dependent. Since x_4 vanishes, $\mathrm{span}_{\mathbb{R}}\{x_1, x_4, y_1, y_3\}$ is 1-dimensional.

We now investigate the case where $x_4 \neq 0$. By (3), we may write $y_3 = \lambda x_4$ for some real number λ. Solving (9) for y_1 and substituting into (1) gives

$$0 = \tfrac{1}{2}\tan(2\theta)\big(x_1(x_1 + (\lambda - 1)x_4) - (x_1 + (\lambda - 1)x_4)x_4\big) - \tan\theta\,|x_2|^2. \qquad (11)$$

Recalling that the square of a purely imaginary number is real, the imaginary part of (11) simplifies to

$$0 = \tfrac{1}{2}\tan(2\theta)(\lambda - 2)\,\mathrm{Im}(x_1 x_4).$$

If $\lambda \neq 2$, this implies that $\mathrm{Im}(x_1 x_4) = 0$, that is, $\{x_1, x_4\}$ must be linearly dependent. Recalling $y_3 = \lambda x_4$ and $y_1 = \cos\theta \sin\theta (x_1 + (\lambda - 1)x_4)$, we see that if $\lambda \neq 2$, then $\dim_{\mathbb{R}} \mathrm{span}_{\mathbb{R}}\{x_1, x_4, y_3, y_1\} = 1$.

We now show $\lambda = 2$ cannot occur. Assume for a contradiction that $\lambda = 2$. We first show this implies that the j and k components of x_2 and y_2 must vanish. We carry out the proof for the j component, as the proof for the k component is identical.

Given x_2 and x_4, Equation (6_j) determines the j component of x_1:

$$(x_1)_j = -\frac{\cos^2\theta\,(x_4)_j + 2\sqrt{3}(\cos\theta - 1)(x_2)_j}{\sin^2\theta}.$$

Substituting this into (9) and rearranging gives

$$(y_1)_j = -\frac{\cos\theta}{\sin\theta}(x_4)_j - \frac{2\sqrt{3}\cos\theta(\cos\theta - 1)}{\sin\theta(\cos^2\theta - \sin^2\theta)}(x_2)_j.$$

Then substituting this into (7_j), we determine

$$(y_2)_j = -\frac{-2\cos^2\theta(\cos\theta - 1)}{\sin\theta(\cos^2\theta - \sin^2\theta)}(x_2)_j.$$

On the other hand, from (10), $y_2 = -\tan\theta\,\bar{x}_2$, the j component of y_2 is determined in a different way by x_2. Thus, either $(x_2)_j = (y_2)_j$ or

$$-\frac{-2\cos^2\theta(\cos\theta - 1)}{\sin\theta(\cos^2\theta - \sin^2\theta)} = \frac{\sin\theta}{\cos\theta}. \qquad (12)$$

By clearing denominators and replacing $\sin^2\theta$ with $1 - \cos^2\theta$ everywhere, (12) is equivalent to $2\cos^3\theta - 3\cos^2\theta + 1 = 0$, which factors as

$$(\cos\theta - 1)^2(2\cos\theta + 1) = 0.$$

But this has no solutions $\theta \in (0, \tfrac{1}{6}\pi)$, since $0 < \cos\theta < 1$ on that interval. It follows that if $\lambda = 2$, then the j and k components of x_2 and y_2 vanish.

Because the j and k components of x_2 vanish, the proof of Proposition 3.3 shows that $|x_1| \geq |x_4|$ with equality only if $|x_1| = |x_4| = 0$.

Now, (9) gives $y_1 = \frac{1}{2}\tan(2\theta)(x_1 + x_4)$. Substituting this into (1), we get

$$\frac{1}{2}\tan(2\theta)(x_1^2 - x_4^2) = \tan\theta\,|x_2|^2.$$

Because x_1 is purely imaginary, $x_1^2 = -|x_1|^2$ and similarly for x_4, so this equation is equivalent to

$$\frac{1}{2}\tan(2\theta)(|x_4|^2 - |x_1|^2) = \tan\theta|x_2|^2. \tag{13}$$

For $\theta \in \left(0, \frac{1}{4}\pi\right)$, both tangents are positive, and so, by Proposition 3.3, the right side of (13) is positive.

On the other hand, since $|x_1| \geq |x_4|$, the left side is nonpositive. This contradiction implies $\lambda = 2$ cannot occur for any $\theta \in \left(0, \frac{1}{6}\pi\right)$. $\qquad\square$

Using Proposition 3.5 and the fact that $y_1 \neq 0$, we see that x_1, x_4, and y_3 are real multiples of y_1.

Proposition 3.6. *Suppose $\theta \in \left(0, \frac{1}{6}\pi\right)$. Then the i components of x_1, x_4, y_1, y_3 and x_3 are all zero.*

Proof. If $(y_1)_i = 0$, it follows from Proposition 3.5, together with the fact that $y_1 \neq 0$ (Proposition 3.3), that the i component of x_1, x_4, and y_3 are all 0 as well. Then (5_i) shows $(x_3)_i = 0$ as well. So, we need only show $(y_1)_i = 0$ when $\theta \in \left(0, \frac{1}{6}\pi\right)$.

So, assume for a contradiction that $(y_1)_i \neq 0$. Solving for y_3 in (7_i) and substituting into (9), we see

$$\cos\theta\sin\theta(x_1 - x_4) = \left(\cos^2\theta - \sin^2\theta - \cos\theta\sin\theta\frac{4\cos\theta\sin\theta}{2\sin^2\theta + 1}\right)y_1.$$

Since $\theta \in \left(0, \frac{1}{6}\pi\right)$, the coefficient on the right is positive. It follows that $x_1 - x_4$ is a positive multiple of y_1.

Now, note that (1), rearranged, takes the form $(x_1 - x_4)y_1 = \tan\theta\,|x_2|^2$. Since $\theta \in \left(0, \frac{1}{6}\pi\right)$, the right-hand side is positive. But since $x_1 - x_4$ is a positive multiple of y_1, the left-hand side is a positive multiple of y_1^2. The square of any purely imaginary number is nonpositive, so we have a contradiction. $\qquad\square$

We now show that x_3 must be nonzero. Suppose for a contradiction that $x_3 = 0$. By (5_j) and (5_k), x_2 has no j or k component. Since $y_2 = -\tan\theta\,\bar{x}_2$, the j and k components of y_2 vanish as well.

Now, (7_j) and (7_k) give $y_3 = -2\tan\theta\,y_1$. In particular, y_3 is a negative multiple of y_1. From (9), we now see $\cos\theta\sin\theta\,(x_1 - x_4)$ is a positive multiple of y_1. Then, just as in the proof of Proposition 3.6, this contradicts (1).

We also find that the j and k components of x_2 and y_2 are constrained.

Proposition 3.7. *Let x_2', y_2' denote the projection of x_2 and y_2 into the jk-plane.
Then $\dim_{\mathbb{R}} \mathrm{span}_{\mathbb{R}}\{x_1, x_4, y_1, y_3, x_2', y_2'\} = 1$.*

Proof. Recalling that x_1 and x_4 have no i component by Proposition 3.6, we see
that multiplying (6_j) by j and (6_k) by k and adding gives the equation

$$(x_2')(\cos\theta - 1)2\sqrt{3} + x_1 \sin^2\theta + x_4 \cos^2\theta = 0.$$

Thus, x_2' is dependent on x_1 and x_4. Since $y_2 = -\tan\theta\, \bar{x}_2$, we find that $y_2' = \tan\theta\, x_2'$
is also dependent on x_2'. The result follows. $\qquad\square$

Proposition 3.8. *Either $(x_2)_j = 0$ or $(x_2)_k = 0$, but not both.*

Proof. If both are zero, then (5_j) and (5_k) give $x_3 = 0$, which is not possible. We
now show at least one vanishes.

We begin by rearranging (2) into the form

$$\bar{x}_2(\tan\theta\, x_4 - y_1) = \tan\theta\, x_3 \bar{x}_2.$$

We write $x_2 = x_2'' + x_2'$ as a decomposition into the complex components, together
with the j and k components. That is, $x_2'' \in \mathbb{C}$ while $x_2' \in \mathrm{span}\{j, k\}$, as before. Then,
the left-hand side can be expanded as $x_2''(\tan\theta\, x_4 - y_1) + x_2'(\tan\theta\, x_4 - y_1)$. Recalling
that the i component of x_4 and y_1 vanishes by Proposition 3.6, $x_2''(\tan\theta\, x_4 - y_1) \in$
$\mathrm{span}\{j, k\}$.

Further, we see $x_2'(\tan\theta\, x_4 - y_1) \in \mathbb{R}$ because x_2' is dependent on both x_4 and y_1
by Proposition 3.7. It follows that $x_2(\tan\theta\, x_4 - y_1)$ has no i component.

Hence, the i component of the right-hand side, $\tan\theta\, x_3 \bar{x}_2$, must vanish as well.
Since $(x_3)_i = 0$ by Proposition 3.6, the i component of $x_3 \bar{x}_2$ is given by

$$0 = (x_3\bar{x}_2)_i i = (x_3)_j j(\bar{x}_2)_k k + (x_3)_k k(\bar{x}_2)_j j = (-(x_3)_j(x_2)_k + (x_3)_k(x_2)_j)i.$$

Now, using (5_j) and (5_k), we see $(x_3)_j = \sqrt{3}(x_2)_j$ and $(x_3)_k = -\sqrt{3}(x_2)_k$. Substi-
tuting yields $0 = -2\sqrt{3}(x_2)_j(x_2)_k$, so at least one of $(x_2)_j$ and $(x_2)_k$ vanishes. $\quad\square$

As we have already shown $\dim_{\mathbb{R}} \mathrm{span}\{x_1, x_4, y_1, y_3, x_2', y_2'\} = 1$ (Proposition 3.7),
it follows that either they all only have a k component, or they all only have
a j component. Equations (5_j) and (5_k) show that x_3 is also in the span of
$\{x_1, x_4, y_1, y_3, x_2', y_2'\}$.

Our next proposition will show that all the variables must commute.

Proposition 3.9. $\qquad\qquad (x_2)_i = (y_2)_i = 0.$

Proof. Since $y_2 = -\tan\theta\, \bar{x}_2$, it is enough to show that $(x_2)_i = 0$.

Equation (2) can be rearranged into the form

$$\tan\theta\, x_4 - y_1 = \frac{\tan\theta}{|x_2|^2} x_2 x_3 \bar{x}_2.$$

By Propositions 3.6, 3.7, and 3.8, we see that the left-hand side, x_3, and x_2 are all either a real multiple of j or a real multiple of k. For the remainder of the proof, we assume they are all multiples of j; the case where they are multiples of k is identical.

The right side is, up to multiple, given by conjugating x_3 by the unit quaternion $x_2/|x_2|$. Recall that a unit quaternion can be written as $q = (\cos\phi)q_0 + (\sin\phi)q_1$, where q_0 is real and q_1 is purely imaginary and $|q_0| = |q_1| = 1$. Then conjugation by q, viewed as a map from $\mathbb{R}^3 \cong \operatorname{Im}(\mathbb{H})$ to itself, is a rotation with axis given by q_1 and with rotation angle given by 2ϕ.

Since the j-axis is invariant under conjugation by x_2, we see one of two things happen. Either the j-axis is fixed point-wise, in which case $\operatorname{Im}(x_2)$ has only a j component, or the orientation of it is reversed. We now show the latter case cannot occur.

If the orientation is reversed, the rotation axis $\operatorname{Im}(x_2)$ must be perpendicular to j, so $\operatorname{Im}(x_2) \in \operatorname{span}\{i, k\}$. Because x_2' has no k part, so it follows that $x_2' = 0$. But then, using (5_j) and (6_j), we see that $x_3 = 0$, which is not possible. $\qquad\square$

It follows that $\operatorname{Im}(x_2) = x_2'$. Summarizing, we have now shown that at a point containing a 0-curvature plane with $\theta \in \left(0, \frac{1}{6}\pi\right)$ that $x_2' = \operatorname{Im}(x_2)$, $y_2' = \operatorname{Im}(y_2)$, $\dim_{\mathbb{R}} \operatorname{span}\{x_1, x_3, x_4, y_1, y_3, x_2', y_2'\} = 1$ and further, that each element in this set has vanishing i and j components or vanishing i and k components. In particular, the variables x_1, x_2, x_3, x_4, y_1, y_2, and y_3 all commute. Thus, we may replace (2) with the linear equation $\tan\theta\, x_4 - \tan\theta\, x_3 - y_1 = 0$ by substituting $y = -\tan\theta\, \bar{x}_2$ and canceling all occurrences of $\bar{x}_2$. We let $\ell \in \{j, k\}$ and set $\epsilon = 1$ if $\ell = j$ and $\epsilon = -1$ if $\ell = k$. Then, (2)–(7_k) are equivalent to the homogeneous system of linear equations

$$-\tan\theta\,(x_3)_\ell + \tan\theta\,(x_4)_\ell - (y_1)_\ell = 0,$$

$$\cos\theta\sin\theta\,(x_1)_\ell - \cos\theta\sin\theta\,(x_4)_\ell + (\sin^2\theta - \cos^2\theta)\,(y_1)_\ell + \cos\theta\sin\theta\,(y_3)_\ell = 0,$$

$$\tan\theta\,(x_2)_\ell - (y_2)_\ell = 0,$$

$$\sqrt{3}(x_2)_\ell + \epsilon(x_3)_\ell = 0,$$

$$\sin^2\theta\,(x_1)_\ell + 2\sqrt{3}(\cos\theta - 1)(x_2)_\ell + \cos^2\theta\,(x_4)_\ell = 0,$$

$$2\sin\theta\cos\theta\,(y_1)_\ell - 2\sqrt{3}\sin\theta\,(y_2)_\ell + \cos^2\theta\,(y_3)_\ell = 0.$$

Then one can easily compute that all solutions are given as real multiples of

$$
\begin{bmatrix} (x_1)_\ell \\ (x_2)_\ell \\ (x_3)_\ell \\ (x_4)_\ell \\ (y_1)_\ell \\ (y_2)_\ell \\ (y_3)_\ell \end{bmatrix}
=
\begin{bmatrix}
-3\cos\theta((2+\epsilon)\cos^2\theta - 4\cos\theta + 2) \\
-\sqrt{3}\cos\theta \\
3\epsilon\cos\theta \\
-3(\cos\theta - 1)((2+\epsilon)\cos^2\theta + (\epsilon - 2)\cos\theta - 2) \\
-3\tan\theta((2+\epsilon)\cos^3\theta - 4\cos^2\theta + 2) \\
-\sqrt{3}\sin\theta \\
6\tan^2\theta((2+\epsilon)\cos^3\theta - 4\cos^2\theta + 1)
\end{bmatrix}.
\tag{14}
$$

We now note that (1) is equivalent to $y_1(x_1 - x_4) = \tan\theta\,|x_2|^2$. In particular, (1) implies that $y_1(x_1 - x_4) > 0$. Thus, if we can show that for $\theta \in \left(0, \frac{1}{6}\pi\right)$, Equation (14) implies $y_1(x_1 - x_4) < 0$, we will have reached our final contradiction, showing N_9 is positively curved at points with $\theta \in \left(0, \frac{1}{6}\pi\right)$.

Proposition 3.10. *For $\theta \in \left(0, \frac{1}{6}\pi\right)$, $y_1(x_1 - x_4) < 0$.*

Proof. We first note that a simple calculation shows

$$(x_1)_\ell - (x_4)_\ell = 6 - (6 + 3\epsilon)\cos\theta.$$

We first prove $y_1(x_1 - x_4) < 0$ when $\ell = j$, that is, $\epsilon = 1$. In this case, $(x_1 - x_4)_j = 6 - 9\cos\theta$ and this is negative so long as $\cos\theta > \frac{2}{3}$. Of course, since $\cos\left(\frac{1}{6}\pi\right) > \frac{2}{3}$, we know that $(x_1 - x_4)_j < 0$ on $\left(0, \frac{1}{6}\pi\right)$.

Further, $(y_1)_j = -3\tan\theta\,(3\cos^3\theta - 4\cos^2\theta + 2)$. The polynomial $3x^3 - 4x^2 + 2$ is clearly positive on the interval $(\sqrt{3}/2, 1)$, so $(y_1)_j < 0$.

It follows that $y_1(x_1 - x_4) = (y_1)_j(x_1 - x_4)_j j^2 = -(y_1)_j(x_1 - x_4)_j < 0$.

Finally, we prove $y_1(x_1 - x_4) < 0$ when $\ell = k$, that is, $\epsilon = -1$. Then it is easy to see that $(y_1)_k$ is positive since the polynomial $x^3 - 4x^2 + 1$ is negative on the interval $(\sqrt{3}/2, 1)$. Further, if $\epsilon = -1$, then $(x_1)_k - (x_4)_k = 6 - 3\cos\theta > 0$.

Thus, $y_1(x_1 - x_4) = (y_1)_k(x_1 - x_4)_k k^2 = -(y_1)_k(x_1 - x_4)_k < 0$, as claimed. $\square$

Acknowledgement

References

[Aloff and Wallach 1975] S. Aloff and N. R. Wallach, "An infinite family of distinct 7-manifolds admitting positively curved Riemannian structures", *Bull. Amer. Math. Soc.* **81** (1975), 93–97. MR Zbl

[Barden 1965] D. Barden, "Simply connected five-manifolds", *Ann. of Math.* (2) **82** (1965), 365–385. MR Zbl

[Berger 1961] M. Berger, "Les variétés riemanniennes homogènes normales simplement connexes à courbure strictement positive", *Ann. Scuola Norm. Sup. Pisa* (3) **15** (1961), 179–246. MR Zbl

[Cheeger 1973] J. Cheeger, "Some examples of manifolds of nonnegative curvature", *J. Differential Geometry* **8** (1973), 623–628. MR Zbl

[Dearricott 2011] O. Dearricott, "A 7-manifold with positive curvature", *Duke Math. J.* **158**:2 (2011), 307–346. MR Zbl

[DeVito et al. 2014] J. DeVito, R. DeYeso, III, M. Ruddy, and P. Wesner, "The classification and curvature of biquotients of the form $Sp(3)//Sp(1)^2$", *Ann. Global Anal. Geom.* **46**:4 (2014), 389–407. MR Zbl

[Dickinson 2004] W. C. Dickinson, "Curvature properties of the positively curved Eschenburg spaces", *Differential Geom. Appl.* **20**:1 (2004), 101–124. MR Zbl

[Eschenburg 1982] J.-H. Eschenburg, "New examples of manifolds with strictly positive curvature", *Invent. Math.* **66**:3 (1982), 469–480. MR Zbl

[Eschenburg 1984] J.-H. Eschenburg, *Freie isometrische Aktionen auf kompakten Lie-Gruppen mit positiv gekrümmten Orbiträumen*, Schriftenreihe des Mathematischen Instituts der Universität Münster, 2. Serie **32**, Universität Münster, Mathematisches Institut, Münster, Germany, 1984. MR Zbl

[Eschenburg and Kerin 2008] J.-H. Eschenburg and M. Kerin, "Almost positive curvature on the Gromoll–Meyer sphere", *Proc. Amer. Math. Soc.* **136**:9 (2008), 3263–3270. MR Zbl

[Gromoll and Meyer 1974] D. Gromoll and W. Meyer, "An exotic sphere with nonnegative sectional curvature", *Ann. of Math.* (2) **100** (1974), 401–406. MR Zbl

[Grove et al. 2011] K. Grove, L. Verdiani, and W. Ziller, "An exotic $T_1 \mathbb{S}^4$ with positive curvature", *Geom. Funct. Anal.* **21**:3 (2011), 499–524. MR Zbl

[Kerin 2011] M. Kerin, "Some new examples with almost positive curvature", *Geom. Topol.* **15**:1 (2011), 217–260. MR Zbl

[Kerin 2012] M. Kerin, "On the curvature of biquotients", *Math. Ann.* **352**:1 (2012), 155–178. MR Zbl

[Kerr and Tapp 2014] M. M. Kerr and K. Tapp, "A note on quasi-positive curvature conditions", *Differential Geom. Appl.* **34** (2014), 63–79. MR Zbl

[O'Neill 1966] B. O'Neill, "The fundamental equations of a submersion", *Michigan Math. J.* **13** (1966), 459–469. MR

[Petersen and Wilhelm 1999] P. Petersen and F. Wilhelm, "Examples of Riemannian manifolds with positive curvature almost everywhere", *Geom. Topol.* **3** (1999), 331–367. MR Zbl

[Tapp 2003] K. Tapp, "Quasi-positive curvature on homogeneous bundles", *J. Differential Geom.* **65**:2 (2003), 273–287. MR Zbl

[Wallach 1972] N. R. Wallach, "Compact homogeneous Riemannian manifolds with strictly positive curvature", *Ann. of Math.* (2) **96** (1972), 277–295. MR Zbl

[Wilhelm 2001] F. Wilhelm, "An exotic sphere with positive curvature almost everywhere", *J. Geom. Anal.* **11**:3 (2001), 519–560. MR Zbl

[Wilking 2002] B. Wilking, "Manifolds with positive sectional curvature almost everywhere", *Invent. Math.* **148**:1 (2002), 117–141. MR Zbl

jdevito@ut.utm.edu   *Department of Mathematics and Statistics, University of Tennessee Martin, Martin, TN 38238, United States*

wesmart@utm.edu   *Department of Education, University of Tennessee Martin, Martin, TN 38238, United States*

Symmetric numerical ranges of four-by-four matrices

Shelby L. Burnett, Ashley Chandler and Linda J. Patton

(Communicated by Chi-Kwong Li)

Numerical ranges of matrices with rotational symmetry are studied. Some cases in which symmetry of the numerical range implies symmetry of the spectrum are described. A parametrized class of 4×4 matrices $K(a)$ such that the numerical range $W(K(a))$ has fourfold symmetry about the origin but the generalized numerical range $W_{K(a)^*}(K(a))$ does not have this symmetry is included. In 2011, Tsai and Wu showed that the numerical ranges of weighted shift matrices, which have rotational symmetry about the origin, are also symmetric about certain axes. We show that any 4×4 matrix whose numerical range has fourfold symmetry about the origin also has the corresponding axis symmetry. The support function used to prove these results is also used to show that the numerical range of a composition operator on Hardy space with automorphic symbol and minimal polynomial $z^4 - 1$ is not a disk.

1. Introduction

Let H be a Hilbert space with inner product $\langle \, \cdot \, , \cdot \, \rangle$ and T a bounded linear operator on H. The *numerical range* of T, denoted by $W(T)$, is the subset of the complex plane $\mathbb{C}$ defined by

$$W(T) = \{\langle Tv, v \rangle \mid v \in H, \|v\| = 1\}.$$

The Toeplitz–Hausdorff theorem states that the numerical range of any bounded linear operator on a Hilbert space is convex [Toeplitz 1918; Hausdorff 1919]. In addition, it follows immediately from the definition that $W(T)$ is unitarily invariant; that is, if R is a linear operator satisfying $R = UTU^*$ for a unitary operator U, then $W(R) = W(T)$. Other well-known results about the numerical range are listed below; these and many other properties of the numerical range appear in [Gustafson

and Rao 1997; Horn and Johnson 1991]. The set of $n \times n$ complex matrices is denoted by $M_n(\mathbb{C})$.

(I) The numerical range contains the spectrum $\sigma(T)$ of T.

(II) If the Hilbert space H is finite-dimensional, then $W(T)$ is compact.

(III) The numerical range $W(T)$ is bounded by $\|T\|$.

(IV) If A is a Hermitian matrix, then $W(A)$ is a real line segment with endpoints equal to the maximum and minimum eigenvalues of A.

(V) $W(A^*) = \{\bar{z} \mid z \in W(A)\}$.

(VI) If A is a normal matrix then $W(A)$ is the convex hull of the eigenvalues of A.

(VII) If $A \in M_2(\mathbb{C})$ then $W(A)$ is a (possibly degenerate) ellipse with foci equal to the eigenvalues of A.

In this paper, some 4×4 matrices with numerical ranges that have a strong type of symmetry are studied. A parametrized family of matrices $K(a)$ where $W(K(a))$ has fourfold symmetry about the origin but certain generalized numerical ranges of $K(a)$ are not symmetric are described; this class generalizes an example in [Deaett et al. 2013]. The relationship between symmetry of the numerical range and symmetry of the spectrum is discussed. In particular, we show that if an associated algebraic curve to an $n \times n$ matrix is irreducible, then symmetry of the numerical range implies symmetry of the spectrum; when $n = 4$, the irreducibility assumption can be dropped. Applications to symmetry about axes are included. The derivations of these results will use two closely related functions associated with the numerical range of a matrix, namely Kippenhahn's boundary-generating curve and the support function of the numerical range. Finally, we show that the numerical range of a composition operator on the Hardy space of the disk with automorphic symbol and minimal polynomial $q(z) = z^4 - 1$ is not a circular disk.

2. Boundary-generating curve and support function

Kippenhahn [1951; 2008] defined the boundary-generating curve for (the numerical range of) an $n \times n$ matrix A as follows. Let $H = (A + A^*)/2$ and $K = (A - A^*)/(2i)$, and let I_n denote the $n \times n$ identity matrix. The polynomial

$$f_A(x, y, z) = \det(x H + y K + z I_n) \tag{1}$$

is homogeneous of degree n with real coefficients. The domain of f_A is complex projective space $\mathbb{P}_2(\mathbb{C})$, which consists of all equivalence classes of points in $\mathbb{C}^3 \setminus \{(0, 0, 0)\}$ under the equivalence relation $\sim$; this relation is defined by

$$(x, y, z) \sim (x', y', z')$$

if and only if there is a nonzero $\alpha \in \mathbb{C}$ such that $(x, y, z) = \alpha(x', y', z')$. Any point $x + iy$ (with $x, y \in \mathbb{R}$) in the complex plane can be identified with the equivalence class of the point $(x, y, 1)$. The natural setting for the study of algebraic curves is $\mathbb{P}_2(\mathbb{C})$; see [Fischer 2001; Gibson 1998] for an introduction to this subject. However, properties of the numerical range primarily involve the restriction of the domain of f_A to points identified with the complex plane, because Kippenhahn showed that $W(A)$ is the convex hull of the curve C defined in line coordinates by $f_A(x, y, 1) = 0$ with $(x, y) \in \mathbb{R}^2$; that is, C is the real part of the zero set of f_A. Kippenhahn called C "the boundary-generating curve of the matrix A". Since C is defined in terms of line coordinates, the line consisting of all $(u, v) \in \mathbb{R}^2$ such that $ux + vy + 1 = 0$ is tangent to C if and only if $f_A(x, y, 1) = 0$. For convenience, if f is a homogeneous polynomial, we will set

$$V_{\mathbb{R}}(f) = \{(x, y) \in \mathbb{R}^2 \mid f(x, y, 1) = 0\}.$$

The polynomial f_A is *reducible* if there exist nonconstant polynomials g and h with real coefficients such that $f_A = gh$; if this occurs, g and h are necessarily homogeneous. A nonconstant polynomial is *irreducible* if it is not reducible. It suffices to consider irreducibility over the real numbers; if f_A was reducible over $\mathbb{C}$ and irreducible over $\mathbb{R}$, then $f_A = g\bar{g}$, where g is an irreducible polynomial with complex coefficients. The polynomials g, $\bar{g}$, and f_A have the same zero set in the complex plane so any arguments requiring irreducibility could be applied to g.

An $n \times n$ matrix A is *unitarily reducible* if there exist matrices B and C of sizes $r \times r$ and $s \times s$, respectively, where $r + s = n$ and $1 \leq r, s \leq n - 1$, and a unitary matrix $U \in M_n(\mathbb{C})$ such that

$$U^*AU = \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix}.$$

The matrix A is called *unitarily irreducible* if A is not unitarily reducible. Determinant properties show that if A is unitarily reducible, then f_A is reducible. However, the converse does not hold because, as shown in [Kippenhahn 1951; 2008], there exist unitarily irreducible matrices A such that f_A is reducible.

In addition to developing properties of the boundary-generating curve, Kippenhahn classified the numerical ranges of 3×3 matrices by showing that the shape of $W(A)$ depends on whether $f(x, y, z)$ is reducible or irreducible. He showed that $W(A)$ is either (1) the convex hull of the eigenvalues of A; (2) the convex hull of an ellipse and a point (reducing to an ellipse when the point is inside the ellipse); (3) a shape with one flat part on the boundary; (4) an ovular shape with no flat part. In [Keeler et al. 1997; Rodman and Spitkovsky 2005], Kippenhahn's classifications are used to derive straightforward tests in terms of the entries of a matrix A that determine the shape. Recently, Chien and Nakazato [2012] classified the numerical ranges of 4×4 matrices using the boundary-generating curve.

The boundary of the numerical range can also be described more directly in terms of its support lines. If S is a closed convex subset of $\mathbb{C}$, then for each point z on the boundary of S, there exists a line L such that $z \in L$ and S lies entirely in one half-plane determined by L. The line L is called a support line for S at z. See [Valentine 1964] for more background on convex sets.

When S is the numerical range of an $n \times n$ matrix A, the rightmost vertical support line $x = \lambda$ of $W(A)$ can be determined directly, because λ is the maximum real part of any complex number in $W(A)$. Straightforward calculations involving inner products produce the following equality:

$$\max\{\operatorname{Re}\langle Av, v\rangle \mid v \in \mathbb{C}^n, \ \|v\|=1\} = \max\{\langle \tfrac{1}{2}(A + A^*)v, v\rangle \mid v \in \mathbb{C}^n, \ \|v\|=1\}.$$

The set in braces on the right side of the equality is the numerical range of the Hermitian matrix $H = \tfrac{1}{2}(A + A^*)$, so by (IV) the maximum value in this set is the maximum eigenvalue of H. Hence the rightmost vertical support line of $W(A)$ is the line $x = \lambda$, where λ is the maximum eigenvalue of H.

Since $W(cA) = cW(A)$ for any complex scalar c, we can derive the support line in every direction by rotating A. The rightmost vertical support line of $W(e^{-i\theta} A)$ will be $x = p_A(\theta)$ where:

Definition 1. $p_A(\theta) = \max \sigma\left(\tfrac{1}{2}(e^{-i\theta} A + e^{i\theta} A^*)\right).$

Rotating this line back by an angle θ will yield the support line of the original numerical range $W(A)$ that is orthogonal to the line from 0 to $e^{i\theta}$.

Therefore the *support function* $p_A(\theta)$ completely determines the numerical range of the matrix A since it describes the support lines in every direction. When T is an operator on an infinite-dimensional Hilbert space, the analogously defined support function determines the closure of $W(T)$.

Note that for any real θ, we have $f_A(\cos(\theta), \sin(\theta), -p_A(\theta)) = 0$ because $(x, y, z) = (\cos(\theta), \sin(\theta), -p_A(\theta))$ satisfies $\det(xH + yK + zI) = 0$.

3. *n*-fold symmetry about the origin

As mentioned in the list of properties of $W(A)$ above, the numerical range of any matrix A contains the eigenvalues of A and when A is normal, $W(A)$ is the convex hull of $\sigma(A)$. In many cases, a plot of the eigenvalues of A along with $W(A)$ shows no obvious relationship except containment. However, a special class of generalized permutation matrices have numerical ranges consisting of a "fattened up" convex hull of the eigenvalues of A. These matrices, whose numerical ranges are studied in [Tsai and Wu 2011; Li and Tsing 1991] as discussed later, are weighted shifts. For consistency with some other references we will work with their adjoints, which by property (V) in the Introduction will produce equivalent results.
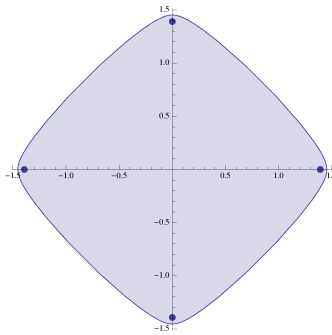
Figure 1. $W(A)$.

Definition 2. A matrix $A \in M_n(\mathbb{C})$ is an AWS (adjoint of weighed shift) matrix if $A = (a_{ij})$ with $a_{ij} = 0$ unless $i = j + 1$ or $i = 1$ and $j = n$.

In the 4×4 case, this yields

$$A = \begin{pmatrix} 0 & 0 & 0 & a_{14} \\ a_{21} & 0 & 0 & 0 \\ 0 & a_{32} & 0 & 0 \\ 0 & 0 & a_{43} & 0 \end{pmatrix}. \tag{2}$$

If A is $n \times n$ of class AWS and the entries of A which are not specified to be zero are in fact nonzero, then the eigenvalues of A are given by a common scalar multiple of the n-th roots of unity. It turns out that the numerical range $W(A)$ is symmetric about the origin in a similar manner.

For example, if A is the 4×4 matrix of class AWS given by

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 \\ 0 & \frac{5}{4} & 0 & 0 \\ 0 & 0 & \frac{3}{2} & 0 \end{pmatrix},$$

then the eigenvalues of A are $\{c, ci, -c, -ci\}$, where $c = 15^{1/4}/\sqrt{2}$.

The numerical range $W(A)$ and the eigenvalues are shown in Figure 1.

This motivates the following definition.

Definition 3. Let n be a positive integer. A subset S of the complex plane has n-fold symmetry about the origin (n-sato) if $z \in S$ implies $e^{2\pi i/n} z \in S$.

That is, S has n-sato if the set S' obtained by rotating S by $2\pi/n$ radians around the origin is equal to S. Clearly the numerical range and the spectrum in Figure 1 have 4-sato.

A result credited to Anderson, which appears in [Tam and Yang 1999; Wu 2011], provides an immediate result about numerical range symmetry.

Theorem 4 [Tam and Yang 1999; Wu 2011]. *Assume $N \geq 2$ and $A \in M_N(\mathbb{C})$. If $W(A)$ is contained in a circular disk and $\partial W(A)$ meets the boundary of the disk at more than N points, then $W(A)$ is equal to the circular disk.*

Corollary 5. *Assume $n > N \geq 2$. Assume $A \in M_N(\mathbb{C})$ is a nonzero matrix. If $W(A)$ has n-sato, then $W(A)$ is a circular disk centered at the origin.*

Proof. Assuming the hypotheses of the corollary, let z_0 be a point of $\partial W(A)$ where the numerical radius of A is attained. Note that $z_0 \neq 0$ and $W(A)$ is contained in the circular disk D with center at the origin and radius $|z_0|$. Since $W(A)$ has n-sato, the distinct points $e^{2\pi k i / n} z_0$ are on $\partial W(A)$ for $k = 0, \ldots, n-1$. Therefore $W(A)$ meets D in more than N points and hence $W(A) = D$. □

Symmetry results about numerical ranges of block AWS operators are proved in [Li and Tsing 1991]. In fact, they prove that much stronger symmetry results hold for AWS operators because the symmetry extends to certain generalized numerical ranges introduced in [Goldberg and Straus 1977]. This generalization is defined below.

Definition 6. Let A and C be in $M_n(\mathbb{C})$. The *C-numerical range of A* is the subset of $\mathbb{C}$ defined by

$$W_C(A) = \{\operatorname{tr}(CUAU^*) \mid U \in M_n(\mathbb{C}),\ U^*U = I\}.$$

Recall that the standard inner product on $M_n(\mathbb{C})$ is $\langle A, B \rangle = \operatorname{tr}(B^*A)$, so $\operatorname{tr}(B^*A)$ can be considered a scaled projection of A onto B. Hence $W_C(A)$ can be considered the projection of the collection of all matrices unitarily equivalent to A (this collection is called the *unitary orbit of A*) onto the matrix C^*. When $C = E_{11}$, the $n \times n$ matrix with 1 in the first row, first column entry and zeroes elsewhere, the generalized numerical range $W_{E_{11}}$ equals the classical numerical range. Unlike the classical numerical range, the C-numerical range is not convex in general [Westwick 1975] but it is always star-shaped [Cheung and Tsing 1996]. See [Li 1994] for more background and properties of the C-numerical range.

Li and Tsing [1991] showed that the Hilbert space operators for which all the (appropriately generalized) C-numerical ranges have n-sato are exactly those unitarily similar to a block form of the AWS. For convenience, we state below a special case of their results that is directly related to the results in this paper.

Theorem 7 (Li–Tsing, special case). *Let n be a positive integer and $A \in M_n(\mathbb{C})$. The following conditions are equivalent*:

(a) $W_C(A)$ *has n-sato for all $C \in M_n(\mathbb{C})$.*

(b) $W_{A^*}(A)$ *has n-sato.*

(c) *A is unitarily equivalent to an $n \times n$ AWS matrix.*

Thus the only $n \times n$ matrices for which all C-numerical ranges have n-sato are those unitarily equivalent to AWS matrices. Since the classical numerical range is one C-numerical range, it follows that the classical numerical range of any AWS matrix has n-sato. However, based on the Li–Tsing theorem, it is possible that there exists an $n \times n$ matrix A that is not unitarily equivalent to an AWS matrix but where $W(A)$ has n-sato. Of course for such a matrix A, there would exist C (in particular $C = A^*$) such that $W_C(A)$ does not have n-sato.

Results in [Tam and Yang 1999] provide conditions (some of which are in terms of associated graphs) that are necessary and sufficient for classes of matrices with the same zero or ray pattern as a given matrix A to have numerical ranges with n-sato or circular symmetry. In particular, conditions for a single matrix with nonnegative entries and a connected undirected graph to have a numerical range with n-sato are provided.

In the 2×2 case, however, it is straightforward to show that $W(A)$ has 2-sato if and only if the eigenvalues of A have 2-sato if and only if A is unitarily equivalent to a 2×2 AWS matrix. These facts follow from property (VII) in the Introduction, basic facts about ellipses, and unitary equivalence arguments. See [Horn and Johnson 1991].

In [Harris et al. 2011], it is shown that $W(A)$ has 3-sato (and is not a circular disk) if and only if A is unitarily similar to a matrix of class AWS.

In [Deaett et al. 2013], matrices in $M_n(\mathbb{C})$ for $n \geq 4$ such that $W(A)$ has n-sato are studied and the following result is proved.

Theorem 8 [Deaett et al. 2013]. *Assume A is a 4×4 matrix with complex entries whose eigenvalues have 4-fold symmetry about the origin. Assume $W(A)$ is not a circular disk. Then the numerical range $W(A)$ has 4-fold symmetry about the origin if and only if* $\mathrm{tr}(A^2 A^*) = 0$ *and* $\mathrm{tr}(A^3 A^*) = 0$.

A natural generalization of the trace condition in Theorem 8 that is sufficient to show $W(A)$ has n-sato for all integers $n \geq 4$ also appears in [Deaett et al. 2013].

The matrix

$$B = \begin{pmatrix} 1 & 1 & \frac{1}{3}(-18-5\sqrt{14}) & 1 \\ 0 & i & 2 & \frac{2}{3}(9+2\sqrt{14}) \\ 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & -i \end{pmatrix} \tag{3}$$

was also constructed in [Deaett et al. 2013]. The numerical range $W(B)$ has 4-sato; however, B is not unitarily equivalent to an AWS matrix of the form (2). Hence there exist 4×4 matrices C such that $W_C(A)$ does not have 4-sato. We will now use similar methods to produce a simpler collection of matrices whose numerical ranges have the same properties.
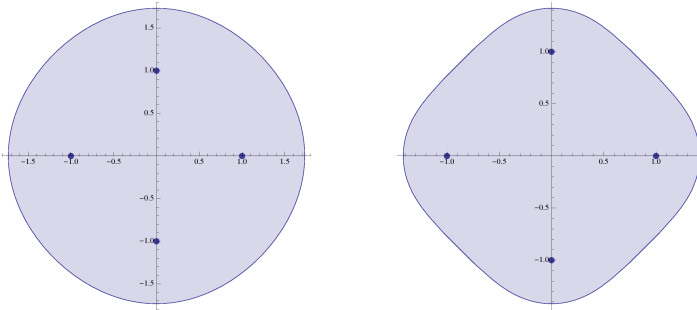
Figure 2. $W(K(a))$ for $a = 1$ (left) and $a = 0.1$ (right)

Let $a \in \mathbb{C}$ with $a \neq 0$ and define

$$K(a) = \begin{pmatrix} 1 & a & \sqrt{2|a|^2+4} & a \\ 0 & i & 0 & -2 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -i \end{pmatrix}. \tag{4}$$

Clearly the eigenvalues of $K(a)$ have 4-sato and since there are no repeated eigenvalues, $W(K(a))$ is not a disk; see [Wu 2011]. It is straightforward to check that $\text{tr}(K(a)^j K(a)^*) = 0$ for $j = 2, 3$. Hence $W(K(a))$ has 4-sato. However, $\text{tr}(K(a)^3(K(a)^*)^2) = 4|a|^2 \neq 0$. For any AWS matrix A, we have $\text{tr}(A^3(A^*)^2) = 0$. Since this trace is a unitary invariant, the matrix $K(a)$ is not unitarily equivalent to an AWS matrix. Therefore, there exists $C \in M_4(\mathbb{C})$ such that $W_C(K(a))$ does not have 4-sato. In particular, $W_{K(a)^*}(K(a))$ does not have 4-sato.

More generally, a similar analysis can be done for many matrices of the form

$$K = \begin{pmatrix} 1 & a & b & a \\ 0 & i & f & c \\ 0 & 0 & -1 & f \\ 0 & 0 & 0 & -i \end{pmatrix} \tag{5}$$

by fixing the "keystone" variable a and solving for b, c, and f to obtain the correct trace values.

Both (3) and (4) have the form (5).

A straightforward computation shows that the boundary-generating curve for $K(a)$ is

$$f_{K(a)}(u, v, w) = w^4 - w^2(u^2+v^2)(3+|a|^2) + (2+|a|^2)(u^4+v^4) + (5+2|a|^2)u^2v^2.$$

This polynomial is quadratic in $x = u^2$, $y = v^2$ and $z = w^2$. The Hessian of the resulting polynomial in x, y, and z is $H(f) = 2|a|^2(2+|a|^2) \neq 0$. Therefore this polynomial is irreducible so $f_{K(a)}$ does not factor into two quadratics. One can

also show that $f_{K(a)}$ cannot factor into a cubic and a linear factor since the linear factor would correspond to an eigenvalue of $K(a)$ and this leads to a contradiction. Consequently $f_{K(a)}$ is irreducible and thus the matrix $K(a)$ is unitarily irreducible.

We include plots of $W(K(a))$ for $a = 1$ and $a = 0.1$ in Figure 2. In general, the problem of plotting $W_C(A)$ is difficult.

4. Symmetry of the spectrum

Assume A is a 2×2 matrix and therefore $W(A)$ is an ellipse with foci equal to the eigenvalues of A. As mentioned earlier, it clearly follows that $W(A)$ has 2-sato if and only if the spectrum $\sigma(A)$ has 2-sato. In general, the spectrum of A can have n-sato even though $W(A)$ does not have n-sato. However, under an irreducibility condition on the boundary-generating curve, symmetry of $W(A)$ implies that of $\sigma(A)$. Proposition 10 below generalizes the $n = 3$ case that appeared in [Harris et al. 2011]. The following lemma is used in the proofs of Propositions 10 and 11.

Lemma 9. *Let n and N be positive integers and let g be an irreducible homogeneous polynomial of degree N. Then the polynomial $\hat{g}_n$ obtained by rotating each affine point $(x, y) = (x, y, 1)$ on $V_{\mathbb{R}}(g)$ through an angle $-\frac{2\pi}{n}$ about the origin is also irreducible of degree N. Hence if there are infinitely many points on $V_{\mathbb{R}}(g) \cap V_{\mathbb{R}}(\hat{g}_n)$ then g is a nonzero scalar multiple of $\hat{g}_n$.*

Proof. Since the transformation of rotation in the first two coordinates of (x, y, z) is an invertible transformation, it preserves irreducibility and degree of homogeneous polynomials. Therefore if the intersection $V_{\mathbb{R}}(g) \cap V_{\mathbb{R}}(\hat{g}_n)$ is infinite, Bézout's theorem shows that $g = c\hat{g}_n$ for some nonzero scalar c. □

Proposition 10. *Let N and n be integers with $n \geq 2$ and $N \geq 3$. Assume A is an $N \times N$ matrix such that f_A as defined in (1) is irreducible. If $W(A)$ has n-sato, then the spectrum $\sigma(A)$ has n-sato.*

Proof. Since f_A is irreducible, it follows that A is unitarily irreducible and hence the boundary of $W(A)$ is smooth [Kippenhahn 1951; Horn and Johnson 1991]. Since there are no corners of $\partial W(A)$, it is not possible that two flat parts on $\partial W(A)$ intersect. There are at most $(N-1)(N-2)/2$ flat parts on the boundary of the numerical range of an $N \times N$ matrix such that f_A is irreducible; see [Gau and Wu 2008]. Any of these finitely many flat parts are separated by a nonflat portion Γ of $\partial W(A)$ consisting of infinitely many points. The numerical range is the convex hull of the boundary-generating curve $C = \{x + iy \in \mathbb{C} \mid f_A(x, y, 1) = 0\}$ and therefore Γ is on a piece of C itself. If there are no flat portions of $\partial W(A)$, then Γ could be any infinite subset of C.

Let $\alpha = \frac{2\pi}{n}$ and $\omega = e^{i\alpha}$. The assumption that $W(A)$ has n-sato is equivalent to the statement that $W(A) = W(\omega A)$. Therefore $\partial W(\omega A)$ also contains Γ and as a nonflat portion of $\partial W(\omega A)$, it must by the argument above be a piece of $V_{\mathbb{R}}(f_{\omega A})$.

The polynomial $f_{\omega A}$ is equal to $(\hat{f}_A)_n$ in the notation of Lemma 9, so $f_A = c f_{\omega A}$ for some scalar c. The coefficient of z^N is 1 in both $f_A(x, y, z)$ and $f_{\omega A}(x, y, z)$; hence $f_A = f_{\omega A}$. Kippenhahn [1951] showed that the eigenvalues of a matrix A are the real foci of the curve f_A. Hence the eigenvalues of A and ωA are equal. Since the eigenvalues of ωA are obtained from those of A by rotating by α about the origin, this proves that $\sigma(A)$ has n-sato. □

The irreducibility condition on f_A, or at least a condition on the size of A, is necessary in Proposition 10. If $A = B \oplus C$, where $W(B)$ has n-sato and C is diagonal with any (i.e., nonsymmetrical) spectrum contained in $W(B)$, then $f_A = f_B f_C$ and the spectrum of A need not have n-sato. However, we can show that if $n = 4$ and A is a 4×4 matrix, noncircular symmetry of the numerical range implies symmetry of the spectrum.

Proposition 11. *Assume A is a 4×4 matrix and $W(A)$ has 4-sato but is not a circular disk. Then $\sigma(A)$ has 4-sato. Under these hypotheses, if $\sigma(A) = \{0\}$, then A is the zero matrix.*

Proof. Assume A is a 4×4 matrix and $W(A)$ has 4-sato but is not a circular disk. Let f_A be defined as in (1). If f_A is irreducible, then $\sigma(A)$ has 4-sato by Proposition 10. Therefore, assume f_A is reducible with the factorization $f_A(u, v, w) = g(u, v, w) h(u, v, w)$, where g is irreducible. In addition, assume the degree m of g is greater than or equal to the degree of every other factor of f_A. So m is either 1, 2, or 3. Note that since the coefficient of w^4 is 1 in the polynomial f_A, we may assume the coefficient of any monomial w^k in any degree-k factor of f_A is also 1.

<u>Case 1</u>: If $m = 1$, then f_A factors into four factors of degree 1; that is, $f_A = h_1 h_2 h_3 h_4$, where $h_j(u, v, w) = a_j u + b_j v + 1 w$ and $\lambda_j = a_j + i b_j$ is an eigenvalue of A for $j = 1, 2, 3, 4$. The numerical range $W(A)$ is a polygon (which could reduce to a line or point) which is the convex hull of these four points. In fact, $W(A)$ is the convex hull of its uniquely determined vertices, which could be a priori a proper subset of $\sigma(A) = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. If $0 \in \sigma(A)$ and 0 is the only vertex of $W(A)$, then $W(A) = \{0\}$ and consequently $\sigma(A) = \{0\}$, which has 4-sato. In this case, A is the zero matrix. Otherwise, let λ_ℓ be a nonzero element of $\sigma(A)$ which is a vertex of $W(A)$. Then $i\lambda_\ell, -\lambda_\ell$ and $-i\lambda_\ell$ are distinct and they are also vertices of $W(A)$ by the 4-sato assumption. This means $\sigma(A) = \{\lambda_\ell, i\lambda_\ell, -\lambda_\ell, -i\lambda_\ell\}$, which has 4-sato.

<u>Case 2</u>: If $m = 2$, then $f_A = gh$, where g is irreducible of degree 2 and the set $V_{\mathbb{R}}(g)$ is an ellipse E_1. If h has two factors of degree 1 then $h(u, v, w) = (a_1 u + b_1 v + w)(a_2 u + b_2 v + w)$, where $\lambda_1 = a_1 + i b_1$ and $\lambda_2 = a_2 + i b_2$ are eigenvalues of A. In this case, $W(A)$ is the convex hull of $E_1 \cup \{\lambda_1\} \cup \{\lambda_2\}$. If both λ_1 and λ_2 are inside the convex hull of E_1, then $W(A)$ is the convex hull of E_1, which does not have 4-sato unless it is a disk; this is precluded by hypothesis. If

one or both of λ_1 and λ_2 are outside E_1, then $\partial W(A)$ will have exactly one or two corners where lines intersect, which is impossible if $W(A)$ has 4-sato. Therefore h is also irreducible of degree 2 and hence $V_\mathbb{R}(h)$ is an ellipse E_2. If either E_1 or E_2 is contained inside the convex hull of the other, then $\partial W(A)$ is the outer ellipse and we are back at the impossible case where $W(A)$ is a circular disk. Therefore $\partial W(A)$ consists of portions of E_1, E_2, and flat portions connecting the two ellipses. In particular, there is a (nonuniquely determined) arc of E_j (denoted by γ_j) that is contained in $\partial W(A)$ for each $j = 1, 2$.

Notate $g(u, v, w) = a_1 u^2 + a_2 v^2 + w^2 + a_4 uv + a_5 uw + a_6 vw$ and $h(u, v, w) = b_1 u^2 + b_2 v^2 + w^2 + b_4 uv + b_5 uw + b_6 vw$. If $(u, v, 1) \in \gamma_1$, then $g(u, v, 1) = 0$. The assumption that $W(A)$ has 4-sato means that the point $(-v, u, 1)$ obtained by rotating $(u, v, 1)$ by $\frac{\pi}{2}$ radians is either on E_1 or E_2. If $(-v, u, 1)$ is in E_1 for infinitely many points on the arc γ_1, then $g(-v, u, 1) = 0$ for those points and the (irreducible) polynomials $g(u, v, w)$ and $g(-v, u, w)$ are the same. Matching coefficients of these polynomials shows that $a_1 = a_2$, $a_4 = -a_4$, $a_5 = a_6$, and $a_6 = -a_5$. Therefore $g(u, v, w) = a_1(u^2 + v^2) + w^2$, and $V_\mathbb{R}(g)$ is a circle centered at the origin. A similar analysis applied to points in γ_2 shows that either there are infinitely many points of $i\gamma_2$ on E_1 or else E_2 is also a circle centered at the origin. Since $W(A)$ is the convex hull of $E_1 \cup E_2$, both curves cannot be circles centered at the origin or else $W(A)$ will be the circular disk with the smaller radius.

In fact, neither E_1 nor E_2 can be a circle centered at the origin. To prove this, assume without loss of generality that E_1 is a circle centered at the origin. If infinitely many points of γ_2 rotate to land on E_1, then the rotated curve is a circle centered at the origin, so E_2 is also such a circle. But the argument above shows that if infinitely many points of γ_2 rotate to E_2, then E_2 is also a circle centered at the origin. Thus neither E_1 nor E_2 can be a circle centered at the origin.

So without loss of generality, there must be infinitely many points on the arc γ_1 such that the corresponding rotated points are on E_2. Thus $g(u, v, 1) = 0$ and $h(-v, u, 1) = 0$ for infinitely many (u, v). Therefore since g and h are irreducible, $g(u, v, w) = h(-v, u, w)$. Setting corresponding coefficients equal yields

$$h(u, v, w) = a_2 u^2 + a_1 v^2 + w^2 - a_4 uv - a_6 uw + a_5 vw.$$

We can rotate the points $(u, v, 1)$ on γ_1 again to obtain that either $h(-u, -v, 1) = 0$ or $g(-u, -v, 1) = 0$ for infinitely many points satisfying $g(u, v, 1) = 0$. The former means that $h(-u, -v, w) = g(u, v, w)$, which leads to $a_4 = a_5 = a_6 = 0$, which results in the circle contradiction. Consequently $g(-u, -v, w) = g(u, v, w)$, which results in $a_5 = a_6 = 0$. Therefore the original ellipse E_1 is centered at the origin and E_2 is described by $h(u, v, w) = g(-v, u, w) = 0$, which is the ellipse E_1 rotated by $\frac{\pi}{2}$. We conclude that the original boundary-generating curve f_A satisfies

$$f_A(u, v, w) = (a_1 u^2 + a_2 v^2 + w^2 + a_4 uv)(a_2 u^2 + a_1 v^2 + w^2 - a_4 uv).$$

Because we know the eigenvalues of A are precisely the values of $-w$ for which $f_A(1, i, w) = 0$, it follows that the eigenvalues of A are solutions to

$$0 = (a_1 - a_2 + ia_4 + w^2)(a_2 - a_1 - ia_4 + w^2) = (w^4 - (a_1 - a_2 + ia_4)^2).$$

Therefore the eigenvalues are the four fourth roots of a fixed complex number and thus have 4-sato. Note that if $\sigma(A) = \{0\}$, then $a_1 = a_2$ and $a_4 = 0$, which again leads to a circular numerical range and is thus impossible in this case by hypothesis.

<u>Case 3</u>: If $m = 3$, then $f_A = gh$, where g is irreducible of degree 3 and h has degree 1. As in Case 1, $h(u, v, w) = au + bv + w$, where $\lambda = a + ib$ is an eigenvalue of A. The numerical range is the convex hull of λ and the real part of the curve $V_\mathbb{R}(g)$ in line coordinates. If $\lambda \in \text{conv}(V_\mathbb{R}(g))$ then $W(A) = \text{conv}(V_\mathbb{R}(g))$. As in the proof of Proposition 10, there must be a nonflat portion of $\partial W(A)$ that consists of a portion γ of $V_\mathbb{R}(g)$ with infinitely many points. When $W(A)$ is rotated by $\frac{\pi}{2}$ radians, the rotation of γ is also on $\partial W(A)$. Hence by Lemma 9, $g(u, v, w) = g(-v, u, w)$ for all (u, v, w) in $\mathbb{P}_2(\mathbb{C})$. Since g has degree 3, it must be of the form

$$g(u, v, w) = c_1 u^3 + c_2 v^3 + w^3 + c_4 u^2 v + c_5 u^2 w$$
$$+ c_6 uvw + c_7 uv^2 + c_8 uw^2 + c_9 v^2 w + c_{10} w^2 v.$$

Setting equivalent coefficients of $g(u, v, w)$ and $g(-v, u, w)$ equal yields

$$c_1 = c_2, \quad c_2 = -c_1, \quad c_4 = -c_7, \quad c_5 = c_9, \quad c_6 = -c_6,$$
$$c_7 = c_4, \quad c_8 = c_{10}, \quad c_9 = c_5 \quad \text{and} \quad c_{10} = -c_8.$$

Therefore $g(u, v, w) = w^3 + c_5 u^2 w + c_5 v^2 w$. If $c_5 < 0$, then $W(A)$ is a circular disk, contradicting our hypothesis. If $c_5 > 0$, then $V_\mathbb{R}(g)$ is empty, contradicting the assumption that $\lambda \in \text{conv}(V_\mathbb{R}(g))$. If $c_5 = 0$, then $f_A(u, v, w) = w^4$, which contradicts our assumption that Case 3 holds.

If the eigenvalue λ is not in $\text{conv}(V_\mathbb{R}(g))$ then $\partial W(A)$ has a vertex at λ where two flat portions of the boundary must meet. By assumption, $\partial W(A)$ also has vertices at $i\lambda$, $-\lambda$, and $-i\lambda$. Note that these points are distinct; the assumption that $W(A)$ has 4-sato means that the origin is either the only point in $W(A)$ (precluded by this case) or in the interior of $W(A)$. Since any vertex on $\partial W(A)$ is an eigenvalue of A, this would immediately show that $\sigma(A)$ has 4-sato. However, this case will not even occur because the convex hull of the real part of the irreducible cubic g will not contain four vertices. □

5. Support function and symmetry about axes

If A is a 4×4 matrix such that $W(A)$ has 4-sato, we will use the support function for $W(A)$ to derive the numerical radius of A and we will provide an estimate that measures how far $W(A)$ is from a circular disk. We will also prove that $W(A)$ has a

particular type of axis symmetry. We will assume that $W(A)$ is noncircular; clearly if $W(A)$ is a circular disk centered at the origin, then $W(A)$ is also symmetric about every line through the origin and the numerical radius is the radius of the circle. Determining the support function involves a lot of calculation which was done in a special case in [Deaett et al. 2013], so we will use the support function from that special case to obtain the general case.

Accordingly, assume that A is a nonzero 4×4 matrix such that $W(A)$ has 4-sato but is not a circular disk. By Proposition 11, the eigenvalues of A have 4-sato. By Theorem 8, $\text{tr}(A^2 A^*) = \text{tr}(A^3 A^*) = 0$. Now rename this matrix B and assume we are in the special case where the eigenvalues of B are $1, i, -1, -i$. Then the proof of Theorem 3.1 in [Deaett et al. 2013] shows that the characteristic polynomial for $\text{Re}(e^{-i\theta} B)$ is

$$q_\theta(z) = z^4 - \tfrac{1}{4}\text{tr}(BB^*)z^2 - \tfrac{1}{4}\text{tr}\big(\tfrac{1}{16}(e^{-4i\theta}B^4 + 4(B^*)^2 B^2 + 2(B^*B)^2 + e^{4i\theta}(B^*)^4)\big)$$
$$+ \tfrac{1}{32}(\text{tr}(BB^*))^2.$$

Since $\text{Re}(e^{-i\theta} B)$ is Hermitian, all of the roots of q_θ are real. The support function is the maximum root of q_θ, so the formula for $p_B(\theta)$ follows directly from the equation above and each expression under a root is real and nonnegative for all θ.

$$p_B(\theta) = \frac{\sqrt{\text{tr}(BB^*) + \sqrt{8\cos(4\theta) + 4\,\text{tr}(B^{*2}B^2) + 2\,\text{tr}(B^*BB^*B) - (\text{tr}(BB^*))^2}}}{2\sqrt{2}}. \quad (6)$$

Now assume the general case where A is a nonzero 4×4 matrix such that $W(A)$ is noncircular and has 4-sato. By Proposition 11 the eigenvalues of A are $a, ai, -a, -ai$ for some nonzero $a \in \mathbb{C}$. Thus $A = aB$ for some B with eigenvalues $1, i, -1, -i$. Let $\alpha = \arg a$. It is straightforward to compute that $p_A(\theta) = |a| p_B(\theta - \alpha)$. Therefore, by (6), we obtain

$$p_A(\theta)$$
$$= \frac{|a|\sqrt{\text{tr}(BB^*) + \sqrt{8\cos(4\theta - 4\alpha) + 4\,\text{tr}(B^{*2}B^2) + 2\,\text{tr}(B^*BB^*B) - (\text{tr}(BB^*))^2}}}{2\sqrt{2}}$$
$$= \frac{\sqrt{\text{tr}(AA^*) + \sqrt{8|a|^4\cos(4\theta - 4\alpha) + 4\,\text{tr}(A^{*2}A^2) + 2\,\text{tr}(A^*AA^*A) - (\text{tr}(AA^*))^2}}}{2\sqrt{2}}. \quad (7)$$

The numerical radius of A is the maximum value of the support function. The previous discussion leads to the following result.

Proposition 12. *Assume A is a 4×4 matrix such that $W(A)$ has 4-sato and is not a circular disk. Assume $\sigma(A) = \{a, ai, -a, -ai\}$ for some nonzero complex*

number a. Then the numerical radius of A is

$$\omega(A) = p_A(\alpha) = \frac{\sqrt{\operatorname{tr}(AA^*) + \sqrt{8|a|^4 + 4\operatorname{tr}(A^{*2}A^2) + 2\operatorname{tr}(A^*AA^*A) - (\operatorname{tr}(AA^*))^2}}}{2\sqrt{2}},$$

the minimum value of the support function for A is

$$p_A\left(\alpha + \tfrac{\pi}{4}\right) = \frac{\sqrt{\operatorname{tr}(AA^*) + \sqrt{-8|a|^4 + 4\operatorname{tr}(A^{*2}A^2) + 2\operatorname{tr}(A^*AA^*A) - (\operatorname{tr}(AA^*))^2}}}{2\sqrt{2}},$$

and

$$-8|a|^2 + 4\operatorname{tr}(A^{*2}A^2) + 2\operatorname{tr}(A^*AA^*A) - (\operatorname{tr}(AA^*))^2 \geq 0. \tag{8}$$

We will now derive an expression that measures the "noncircularity" of the numerical range of a 4×4 matrix A (where $W(A)$ has 4-sato) in terms of $\operatorname{tr}(A^*A)$. Let g_A denote the difference between the maximum and minimum values of the support function of A. That is, $g_A = p_A(\alpha) - p_A\left(\alpha + \tfrac{\pi}{4}\right)$ as defined above. The quantity g_A measures the gap between the points on the boundary of $W(A)$ that are farthest from, and closest to, the origin.

To prove the following lower bound for g_A, we will produce some inequalities involving $\operatorname{tr}(A^*A)$ and traces of more complicated words in A and A^*. This proposition will be used in the next section to prove that the numerical range of a certain composition operator is not circular.

Proposition 13. *Assume A is a 4×4 matrix such that $W(A)$ has 4-sato and $\sigma(A) = \{1, -1, i, -i\}$. Let $\alpha = \operatorname{tr}(A^*A)$. Then*

$$g_A \geq \frac{8}{\sqrt{2}\left(\sqrt{\alpha + \sqrt{5\alpha^2 + 8}} + \sqrt{\alpha + \sqrt{5\alpha^2 - 8}}\right)\left(\sqrt{5\alpha^2 + 8} + \sqrt{5\alpha^2 - 8}\right)}.$$

Proof. Recall that

$$\langle A, B \rangle = \operatorname{tr}(B^*A)$$

defines an inner product on $M_n(\mathbb{C})$ and in particular on $M_4(\mathbb{C})$. Thus $\langle A, A \rangle = \|A\|_{\operatorname{tr}}^2 = \operatorname{tr}(A^*A)$.

The Cauchy–Schwarz inequality on this space shows that

$$|\operatorname{tr}(B^*A)| \leq \|A\|_{\operatorname{tr}} \|B\|_{\operatorname{tr}} = \sqrt{\operatorname{tr}(A^*A)\operatorname{tr}(B^*B)}. \tag{9}$$

The trace norm induced by this inner product is a matrix norm [Horn and Johnson 1991], so $\|AB\|_{\operatorname{tr}} \leq \|A\|_{\operatorname{tr}} \|B\|_{\operatorname{tr}}$, and therefore

$$\operatorname{tr}(A^*AA^*A) = \|A^*A\|_{\operatorname{tr}}^2 \leq \|A^*\|_{\operatorname{tr}}^2 \|A\|_{\operatorname{tr}}^2 = \|A\|_{\operatorname{tr}}^4 = (\langle A, A \rangle)^2 = \alpha^2. \tag{10}$$

Also,

$$\operatorname{tr}((A^*)^2A^2) = \|A^2\|_{\operatorname{tr}}^2 \leq \|A\|_{\operatorname{tr}}^4 = (\langle A, A \rangle)^2 = \alpha^2. \tag{11}$$

Define

$$d(A, A^*) = 4\operatorname{tr}((A^*)^2 A^2) + 2\operatorname{tr}(A^* A A^* A) - (\operatorname{tr}(A^* A))^2. \tag{12}$$

Combining (8), (10), and (11) yields

$$8 \le d(A, A^*) \le 4\alpha^2 + 2\alpha^2 - \alpha^2 = 5\alpha^2. \tag{13}$$

The assumptions on the spectrum of A imply that $\alpha^2 \ge 4$. The maximum value of the support function can be written in terms of $d(A, A^*)$ and α as

$$p_A(0) = \frac{\sqrt{\alpha + \sqrt{8 + d(A, A^*)}}}{2\sqrt{2}},$$

while the minimum value is

$$p_A\left(\tfrac{\pi}{4}\right) = \frac{\sqrt{\alpha + \sqrt{-8 + d(A, A^*)}}}{2\sqrt{2}}.$$

Therefore, the distance between the maximum value of the support function and the minimum value of the support function is

$$g_A = \frac{\left(\sqrt{\alpha + \sqrt{d(A, A^*) + 8}} - \sqrt{\alpha + \sqrt{d(A, A^*) - 8}}\right)}{2\sqrt{2}}.$$

We want to find a lower bound for g_A in terms of α.

By multiplying g_A by its algebraic conjugate in the numerator and denominator, we obtain

$$g_A = \frac{\sqrt{d(A, A^*) + 8} - \sqrt{d(A, A^*) - 8}}{2\sqrt{2}\left(\sqrt{\alpha + \sqrt{d(A, A^*) + 8}} + \sqrt{\alpha + \sqrt{d(A, A^*) - 8}}\right)}.$$

Now multiply numerator and denominator by the conjugate of the numerator to see that

$$g_A = \frac{8}{\sqrt{2}\left(\sqrt{\alpha + \sqrt{d(A, A^*) + 8}} + \sqrt{\alpha + \sqrt{d(A, A^*) - 8}}\right)\left(\sqrt{d(A, A^*) + 8} + \sqrt{d(A, A^*) - 8}\right)}.$$

Each term of each factor in the denominator of g_A is a positive increasing function of $d(A, A^*)$. Therefore, (13) implies that

$$g_A \ge \frac{8}{\sqrt{2}\left(\sqrt{\alpha + \sqrt{5\alpha^2 + 8}} + \sqrt{\alpha + \sqrt{5\alpha^2 - 8}}\right)\left(\sqrt{5\alpha^2 + 8} + \sqrt{5\alpha^2 - 8}\right)}. \qquad \square$$

Tsai and Wu [2011] proved a number of results about numerical ranges of weighted shift matrices. In particular, they show that the numerical range of any $n \times n$ weighted shift matrix A (and thus any AWS matrix) is symmetric about each of n lines through the origin that are determined by the entries of A. The angle between each pair of adjacent lines is $\frac{\pi}{n}$. We will show that if the numerical range

of a 4×4 matrix A has 4-sato, then $W(A)$ is similarly symmetric about four lines through the origin even if A is not unitarily equivalent to an AWS matrix.

Property (V) from the Introduction shows that for any $n \times n$ matrix A, the set $W(A^*)$ is the reflection of $W(A)$ about the real axis. For any angle θ, the matrix Re $e^{-i\theta} A$ is the same as Re $e^{i\theta} A^*$. Since the support function $p_A(\theta)$ is the maximum eigenvalue of Re $e^{-i\theta} A$, it is also true that $p_A(-\theta) = p_{A^*}(\theta)$ for all real θ.

Proposition 14. *Let A be an $n \times n$ matrix. The numerical range $W(A)$ is symmetric about the real axis if and only if the support function p_A is an even function.*

Proof. Assume $W(A)$ is symmetric about the real axis and consequently $W(A) = W(A^*)$. Hence $p_A(\theta) = p_{A^*}(\theta) = p_A(-\theta)$ for all real θ.

Now assume $p_A(\theta) = p_A(-\theta)$ for all θ. This implies that $p_A = p_{A^*}$ and consequently the numerical ranges $W(A)$ and $W(A^*)$ are equal, so $W(A)$ is symmetric about the real axis.   □

Corollary 15. *Let A be an $n \times n$ matrix such that the origin is in the numerical range $W(A)$. Let $\delta \in \mathbb{R}$. The numerical range is symmetric about the line ℓ through the origin and $e^{i\delta}$ if and only if the support function p_A for $W(A)$ satisfies $p_A(\theta + \delta) = p_A(-\theta + \delta)$.*

Proof. $W(A)$ is symmetric about ℓ if and only if the rotated set $e^{-i\delta} W(A)$ is symmetric about the real axis. The latter statement is equivalent to the numerical range of $e^{-i\delta} A$ having symmetry about the real axis. Since the definition of the support function shows that $p_{e^{-i\delta} A}(\theta) = p_A(\theta + \delta)$, the corollary follows from Proposition 14. □

The strong connection between n-fold symmetry about the origin and axis symmetry depends on the size of the matrix relative to n, as the following example due to Spitkovsky (personal communication, 2012) shows:

Example 16. Let A be the 8×8 diagonal matrix with diagonal entries 1, i, -1, $-i$, $0.9e^{\pi i/6}$, $0.9e^{2\pi i/3}$, $0.9e^{7\pi i/6}$, $0.9e^{5\pi i/3}$. Thus $W(A)$ is a polygon with vertices at the eight diagonal entries as shown in Figure 3. Clearly $W(A)$ has 4-sato, as does $\sigma(A)$, but $W(A)$ is not symmetric about any axis.

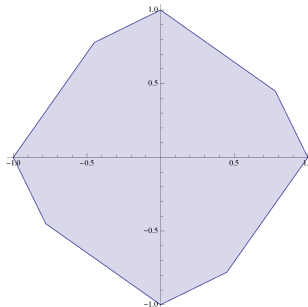**Figure 3.** 4-sato but no axis symmetry.

Theorem 17. *Assume A is a 4×4 nonzero matrix such that $W(A)$ has 4-sato and is noncircular. The eigenvalues of A are $a, ia, -a,$ and $-ia$ for $a \in \mathbb{C}$ with $a \neq 0$ and $\alpha = \arg a$. Let $\delta_n = \alpha + \frac{n\pi}{4}$ for $n = 0, 1, 2, 3$. The numerical range $W(A)$ is symmetric about the lines through 0 and $e^{i\delta_n}$ for $n = 0, 1, 2, 3$.*

Proof. Assume A is a 4×4 nonzero matrix satisfying the theorem hypotheses. Note that the form of $\sigma(A)$ follows from Proposition 11. The support function $p_A(\theta)$ for $W(A)$ only depends on θ through the term $\cos(4\theta - 4\alpha)$, as seen in (7). For each integer n with $0 \leq n < 4$ and each real θ,

$$\cos(4(\theta + \delta_n) - 4\alpha) = \cos(4\theta + n\pi)$$
$$= \cos(-4\theta - n\pi + 2n\pi)$$
$$= \cos(-4\theta + n\pi) = \cos(4(-\theta + \delta_n) - 4\alpha).$$

Therefore $p_A(\theta + \delta_n) = p_A(-\theta + \delta_n)$ for all θ and $n = 0, 1, 2, 3$, which means $W(A)$ has the stated symmetry by Corollary 15.   □

After submission of this paper, the authors learned of the preprint [Lentzos and Pasley 2017], one result in which provides an alternate proof of Theorem 17 by showing that any boundary-generating curve for a numerical range with n-sato can be associated with an AWS matrix, even if the original matrix is not unitarily equivalent to an AWS matrix.

6. Application to numerical range of composition operator

The Hardy–Hilbert space $H^2 = H^2(\mathbb{D})$ is the set of all analytic functions f on the unit disk $\mathbb{D}$ such that

$$\|f\|_{H^2}^2 = \sup_{0 < r < 1} \int_0^{2\pi} |f(re^{i\theta})|^2 \frac{d\theta}{2\pi} < \infty.$$

If φ is an analytic self-map of $\mathbb{D}$, the associated composition operator C_φ is defined for $f \in H^2$ by $C_\varphi f = f \circ \varphi$. On H^2, it can be shown that the operator C_φ is bounded for all analytic mappings φ from $\mathbb{D}$ to itself. See [Cowen and MacCluer 1995] for this and other properties of composition operators.

If φ is an automorphism of the disk $\mathbb{D}$, then there exist $\eta \in \partial \mathbb{D}$ and $p \in \mathbb{D}$ such that

$$\varphi(z) = \eta \frac{p - z}{1 - \bar{p}z}.$$

The automorphism φ can be classified as elliptical, hyperbolic, or parabolic depending on the locations of the fixed points of φ. If φ has one of its fixed points in the interior of $\mathbb{D}$ then it is elliptical. Bourdon and Shapiro [2000] determined the shape of the numerical range for many composition operators on $H^2(\mathbb{D})$ with automorphic symbols. In many cases the numerical range was a circular disk and

it was also determined whether the numerical range was open, closed, or neither. However, Bourdon and Shapiro noted that when the automorphic symbol satisfies $\varphi \circ \varphi(z) = z$, and hence C_φ^2 is the identity operator I, the numerical range is a noncircular ellipse. This fact holds more generally for all quadratic operators, as shown in [Tso and Wu 1999].

Bourdon and Shapiro conjectured that any composition operator on H^2 with automorphic symbol satisfying $\varphi^{(n)}(z) = z$ (where n is a positive integer and $\varphi^{(n)}$ denotes composition of φ with itself n times) has a noncircular numerical range. Unlike the case for quadratic operators, this fact does not generalize; for example, there exists an operator T on a Hilbert space such that $T^3 = I$ and $W(T)$ is a circular disk [Harris et al. 2011]. The third author showed that Bourdon and Shapiro's conjecture is true for $n = 3$. That is, a composition operator satisfying $C_\varphi^3 = I$ does not have a circular disk as its numerical range [Patton 2013]. The result follows because any composition operator C_φ with automorphic symbol satisfying $\varphi^{(n)}(z) = z$ is unitarily equivalent to a block Toeplitz matrix with Toeplitz symbol equal to an $n \times n$ matrix-valued polynomial of degree 1. That is, the symbol has the form $A(z) = A_0 + A_1 z$ and there is an orthonormal basis with respect to which C_φ has the matrix

$$\mathcal{M}(C_\varphi) = \begin{pmatrix} A_0 & 0 & 0 & \cdots \\ A_1 & A_0 & 0 & \cdots \\ 0 & A_1 & A_0 & \\ 0 & 0 & A_1 & \ddots \\ \vdots & \vdots & & \ddots \end{pmatrix}. \tag{14}$$

Bebiano and Spitkovsky [2012] showed that in general, the closure of the numerical range of a block Toeplitz matrix with matrix-valued symbol a is the convex hull of the set $\{W(A) : A \in R(a)\}$, where $R(a)$ is the essential range of the symbol on $\partial\mathbb{D}$. In the composition operator case above, this reduces to the following theorem.

Theorem 18 [Patton 2013]. *Let $\eta \in \partial\mathbb{D}$ and let $p \in \mathbb{D}$. Define the disk automorphism*

$$\varphi = \eta \frac{p - z}{1 - \bar{p}z}$$

and assume $\varphi^{(n)}(z) = z$. The numerical range of the composition operator C_φ satisfies

$$\operatorname{clos} W(C_\varphi) = \operatorname{conv}\{W(A_0 + A_1 z) \mid z \in \partial\mathbb{D}\},$$

where A_0 and A_1 are $n \times n$ matrices whose entries depend on η and p.

Formulas for the entries of A_0 and A_1 appear in [Patton 2013]. The matrix A_1 is a particularly simple rank-1 matrix. In the case where $n = 4$, the entries of A_0 and

A_1 are shown below:

$$A_0 = \begin{pmatrix} 1 & \frac{p\eta}{P_-} & 0 & 0 \\ 0 & -\eta & \frac{p(1+\eta)}{P_+} & 0 \\ 0 & -\frac{\eta\bar{p}}{P_+} & \frac{-1+\eta}{1+\bar{\eta}} & \frac{p}{P_+} \\ 0 & -\frac{(1+\eta)\bar{p}^2}{(P_+)^2} & \frac{(-1+\eta)\bar{p}}{P_+} & \frac{1-\bar{\eta}}{1+\bar{\eta}} \end{pmatrix}.$$

$$A_1 = -\frac{\bar{\eta}\,\bar{p}\,(1-\bar{\eta})}{1-\bar{\eta}|p|^2} \begin{pmatrix} 0 & \frac{\bar{p}^2}{P_-} & -\frac{\eta\,\bar{p}(1-\bar{\eta}|p|^2)}{P_+P_-} & \frac{P_-}{\eta-1} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where $P_- = \sqrt{1-|p|^2}$ and $P_+ = \sqrt{1+|p|^2}$. We will use the numerical radius estimates in the previous section to show that the conjecture of Bourdon and Shapiro holds for $n = 4$.

The assumption that $\varphi^{(4)}(z) = z$ implies that the parameters η and p satisfy the identity

$$2|p|^2 = \eta + \bar{\eta}, \tag{15}$$

and this immediately yields

$$|1 - \bar{\eta}|^2 = 2(1 - |p|^2), \quad |1 + \bar{\eta}|^2 = 2(1 + |p|^2), \quad \left|1 - \eta|p|^2\right|^2 = 1 - |p|^4, \tag{16}$$

and

$$4(1 + |p|^2)^2 = (\eta + 1)(\eta + 3) + (\bar{\eta} + 1)(\bar{\eta} + 3). \tag{17}$$

These identities can be used to rewrite the entries of A_0 and A_1 with only real quantities in the denominator, and we obtain

$$A(z) = \begin{pmatrix} 1 & \frac{p\eta}{P_-} - z\frac{(\bar{\eta}+1)\bar{p}^3}{(P_+)^2 P_-} & z\frac{\bar{p}^2(1-\bar{\eta})}{P_+P_-} & z\frac{\bar{p}(1-\bar{\eta}|p|^2)}{(P_+)^2 P_-} \\ 0 & -\eta & \frac{p(1+\eta)}{P_+} & 0 \\ 0 & -\frac{\eta\bar{p}}{P_+} & \frac{\eta|p|^2-1}{(P_+)^2} & \frac{p}{P_+} \\ 0 & -\frac{(1+\eta)\bar{p}^2}{(P_+)^2} & \frac{(-1+\eta)\bar{p}}{P_+} & \frac{\eta-\bar{\eta}}{2(P_+)^2} \end{pmatrix}.$$

Proposition 13 will be applied to $A(z)$ in order to show there is a fixed gap between the maximum and minimum value of the support function of $W(C_\varphi)$; this suffices to prove $W(C_\varphi)$ is not a circular disk. In order to apply the proposition, we must show that $W(A(z))$ has 4-sato for all z on the unit circle. By Theorem 8, it suffices to show that $\sigma(A(z)) = \{1, i, -1, -i\}$ and the traces of $A(z)^2 A(z)^*$ and $A(z)^3 A(z)^*$ are zero.

The condition $\varphi^{(4)}(z) = z$ shows that $C_\varphi^4 = I$, and thus $\mathcal{M}(C_\varphi)^4 = I$. The latter and (14) imply that $A_0^4 = I$, and the form of the matrix $A(z)$ guarantees that $\sigma(A(z)) = \sigma(A_0) = \{1, i, -1, -i\}$ for all $z \in \partial\mathbb{D}$.

Some tedious calculations that lead to the trace requirements are done next.

First, we obtain

$$A(z)A(z)^* =$$

$$\begin{pmatrix} \left(\frac{P_+}{P_-}\right)^2 + \frac{2\operatorname{Re}(-z\bar{\eta}(1+\bar{\eta})\bar{p}^4)}{(P_+P_-)^2} & -\frac{p}{P_-} + \frac{z(1+\bar{\eta})\bar{p}^3}{(P_+)^2 P_-} & \frac{-p^2+z\bar{p}^2\bar{\eta}}{P_+P_-} & \frac{(-p^3(\eta+1)+z\bar{p}(1+\bar{\eta})^2/2)}{(P_+)^2 P_-} \\ -\frac{\bar{p}}{P_-} + \frac{\bar{z}(1+\eta)p^3}{(P_+)^2 P_-} & 1+2|p|^2 & \frac{\bar{\eta}p}{P_+} & \frac{p^2(1+\bar{\eta})}{(P_+)^2} \\ \frac{-\bar{p}^2+\bar{z}p^2\eta}{P_+P_-} & \frac{\eta\bar{p}}{P_+} & 1 & \frac{p}{P_+} \\ \frac{(-\bar{p}^3(\bar{\eta}+1)+\bar{z}p(1+\eta)^2/2)}{(P_+)^2 P_-} & \frac{(1+\eta)\bar{p}^2}{(P_+)^2} & \frac{\bar{p}}{P_+} & 1 \end{pmatrix}.$$

Next we compute that

$$A(z)^2 A(z)^* =$$

$$\begin{pmatrix} 1+2|p|^2 & \frac{p\eta(\eta+1)}{P_-} - z\left(\frac{\bar{p}^3(1+\bar{\eta})^2}{(P_+)^2 P_-}\right) & \frac{z\bar{p}^2(1-\bar{\eta}^2)}{P_+P_-} & \frac{z\bar{p}(1+\bar{\eta})(1-\bar{\eta}^2)}{2(P_+)^2 P_-} \\ \frac{\bar{p}(\eta-|p|^2)}{(P_+)^2 P_-} & \frac{\eta(-1-2|p|^2-2|p|^4+\eta|p|^2)}{(P_+)^2} & \frac{\eta p}{P_+} & 0 \\ \frac{\bar{p}^2(1-\bar{\eta})}{P_+P_-} & -\frac{\bar{p}(1+\eta)^2}{(P_+)^3} & \frac{\eta|p|^2-1}{(P_+)^2} & 0 \\ \frac{\bar{p}^3(1+\bar{\eta})}{(P_+)^2 P_-} - \bar{z}\left(\frac{p}{P_-}\right) & \frac{\bar{p}^2(-3-2\eta-\bar{\eta})}{(P_+)^2} & \frac{\bar{p}(\eta^2-\bar{\eta}^2+\eta-3\bar{\eta}-2)}{2(P_+)^3} & -\frac{(\bar{\eta}+1)^2}{2(P_+)^2} \end{pmatrix}.$$

Straightforward calculations can be used to simplify $\operatorname{tr} A(z)^2 A(z)^*$ as follows:

$$\operatorname{tr} A(z)^2 A(z)^*$$
$$= 1+2|p|^2 + \frac{\eta(-1-2|p|^2-2|p|^4+\eta|p|^2)}{1+|p|^2} + \frac{\eta|p|^2-1}{1+|p|^2} - \frac{(\bar{\eta}+1)^2}{2(1+|p|^2)}$$
$$= \frac{\left(1+3|p|^2+2|p|^4-\eta-2\eta|p|^2-2\eta|p|^4+\eta^2|p|^2+\eta|p|^2-1-\frac{1}{2}\bar{\eta}^2-\bar{\eta}-\frac{1}{2}\right)}{1+|p|^2}.$$

Using the identities $2|p|^2 = \eta + \bar{\eta}$ and $4|p|^4 = \eta^2 + 2 + \bar{\eta}^2$ and grouping like powers of η and $\bar{\eta}$ proves that

$$\operatorname{tr} A(z)^2 A(z)^* = 0. \tag{18}$$

The value $\operatorname{tr} A(z)^3 A(z)^*$ has a constant term, to which all four diagonal terms contribute, and a z-term, which only occurs in the $(1, 1)$ entry of $A(z)^3 A(z)^*$. This

z-term has coefficient

$$\frac{-\bar{p}^4(\eta - |p|^2)(\bar{\eta} + 1)}{(1 - |p|^2)(1 + |p|^2)^2} + \frac{\bar{p}^4(1 - \bar{\eta})^2}{(1 - |p|^2)(1 + |p|^2)} + \frac{\bar{p}^4(1 + \bar{\eta})(1 - \bar{\eta}|p|^2)}{(1 - |p|^2)(1 + |p|^2)^2}.$$

Factoring out $\bar{p}^4/(1 - |p|^4)$ yields

$$\frac{\bar{p}^4}{(1 - |p|^4)}\left(-\frac{(\bar{\eta} + 1)(\eta - |p|^2)}{1 + |p|^2} + \frac{(1 - \bar{\eta}|p|^2)(1 + \bar{\eta})}{1 + |p|^2} + (1 - \bar{\eta})^2\right),$$

and after forming a common denominator for the terms inside the square brackets and rewriting everything in terms of η and $\bar{\eta}$ using $2|p|^2 = \eta + \bar{\eta}$, we obtain that the coefficient of z in the trace of $A(z)^3 A(z)^*$ is zero.

The constant term is more difficult to simplify; we work separately with each diagonal entry.

The $(2, 2)$ entry of $A(z)^3 A(z)^*$ simplifies to

$$\frac{\eta^3 + 2\eta^2 - 2 - \bar{\eta}}{4(1 + |p|^2)^2}.$$

The $(3, 3)$ entry of $A(z)^3 A(z)^*$ simplifies to

$$\frac{-(\eta + 1)^3 - (\bar{\eta} + 1)^3}{4(1 + |p|^2)^2}.$$

The $(4, 4)$ entry of $A(z)^3 A(z)^*$ is

$$\frac{-(\bar{\eta} + 1)^2(\eta - \bar{\eta})}{4(1 + |p|^2)^2}.$$

The constant term of the $(1, 1)$ entry of $A(z)^3 A(z)^*$ is

$$1 + 2|p|^2 + \frac{|p|^2(\eta - |p|^2)\eta}{1 - |p|^4} - \frac{|p|^2(1 - \bar{\eta}|p|^2)}{1 - |p|^4}.$$

The numerator of the latter expression over the common denominator $(1 - |p|^4)$ can be expressed as

$$1 + |p|^2\big(1 + \eta^2 + |p|^2(-1 - \eta - \bar{\eta}) - 2|p|^4\big),$$

and this simplifies to $1 - |p|^4$ using (15) and (16). Consequently the constant part of the $(1, 1)$ entry of $A(z)^3 A(z)^*$ is 1.

The simplified sum of the $(2, 2)$, $(3, 3)$, and $(4, 4)$ entries of the constant term is

$$\frac{-\bar{\eta}^2 - 4\bar{\eta} - 6 - 4\eta - \eta^2}{4(1 + |p|^2)^2} = -1,$$

where the equality follows from (17). Therefore, we obtain

$$\operatorname{tr} A(z)^3 A(z)^* = 0. \tag{19}$$

Equations (18) and (19) hold for all z on the unit circle; we also saw that the spectrum of $A(z)$ is $\{1, i, -1, -i\}$ for all such z. Therefore Theorem 8 shows that the numerical range of the matrix $A(z)$ has 4-sato for all $z \in \partial\mathbb{D}$.

The lemma below follows immediately from the calculated entries of $A(z)A(z)^*$.

Lemma 19. *If $A(z)$ is the 4×4 block matrix defined above at any value z on the unit circle, then*

$$\operatorname{tr}(A(z)A(z)^*) \leq \frac{4 + 4|p|^2 + 2|p|^4}{1 - |p|^4}.$$

Theorem 20. *If φ is an automorphism of the disk such that C_φ has minimal polynomial $z^4 - 1$, then $W(C_\varphi)$ is not a disk.*

Proof. The value $p = \varphi^{-1}(0)$ is in the open unit disk. Let α_p denote the upper bound for $\operatorname{tr}(A(z)A(z)^*)$ from Lemma 19. Combining this value with Proposition 13 shows that there is a uniform lower bound

$$g_p = \frac{8}{\sqrt{2}\left(\sqrt{\alpha_p + \sqrt{5\alpha_p^2 + 8}} + \sqrt{\alpha_p + \sqrt{5\alpha_p^2 - 8}}\right)\left(\sqrt{5\alpha_p^2 + 8} + \sqrt{5\alpha_p^2 - 8}\right)}$$

for the difference between the maximum and minimum values of the support functions of $A(z)$ because $g_{A(z)} \geq g_p$ for each z on the unit circle. Furthermore, Proposition 12 shows that for each z the maximum value of $p_{A(z)}(\theta)$ is attained at $\theta = 0$, while the minimum is attained at $\theta = \frac{\pi}{4}$. Since the numerical range of C_φ is the convex hull of all of these matrix numerical ranges as z ranges over the unit circle, it follows that the difference between the maximum and minimum values of the support function of C_φ is bounded below by g_p. Hence $W(C_\varphi)$ is not a circular disk. □

Recently, Heydari and Abdollahi [2015] showed there is a large class of finite-order elliptic composition operators such that $W(C_\varphi)$ is not a circular disk.

References

[Bebiano and Spitkovsky 2012] N. Bebiano and I. M. Spitkovsky, "Numerical ranges of Toeplitz operators with matrix symbols", *Linear Algebra Appl.* **436**:6 (2012), 1721–1726. MR Zbl

[Bourdon and Shapiro 2000] P. S. Bourdon and J. H. Shapiro, "The numerical ranges of automorphic composition operators", *J. Math. Anal. Appl.* **251**:2 (2000), 839–854. MR Zbl

[Cheung and Tsing 1996] W.-S. Cheung and N.-K. Tsing, "The C-numerical range of matrices is star-shaped", *Linear and Multilinear Algebra* **41**:3 (1996), 245–250. MR Zbl

[Chien and Nakazato 2012] M.-T. Chien and H. Nakazato, "Singular points of the ternary polynomials associated with 4-by-4 matrices", *Electron. J. Linear Algebra* **23** (2012), 755–769. MR Zbl

[Cowen and MacCluer 1995] C. C. Cowen and B. D. MacCluer, *Composition operators on spaces of analytic functions*, CRC Press, Boca Raton, FL, 1995. MR Zbl

[Deaett et al. 2013] L. Deaett, R. H. Lafuente-Rodriguez, J. Marin, Jr., E. Haller Martin, L. J. Patton, K. Rasmussen, and R. B. Johnson Yates, "Trace conditions for symmetry of the numerical range", *Electron. J. Linear Algebra* **26** (2013), 591–603. MR Zbl

[Fischer 2001] G. Fischer, *Plane algebraic curves*, Student Mathematical Library **15**, American Mathematical Society, Providence, RI, 2001. MR Zbl

[Gau and Wu 2008] H.-L. Gau and P. Y. Wu, "Line segments and elliptic arcs on the boundary of a numerical range", *Linear Multilinear Algebra* **56**:1-2 (2008), 131–142. MR Zbl

[Gibson 1998] C. G. Gibson, *Elementary geometry of algebraic curves: an undergraduate introduction*, Cambridge University Press, 1998. MR Zbl

[Goldberg and Straus 1977] M. Goldberg and E. G. Straus, "Elementary inclusion relations for generalized numerical ranges", *Linear Algebra and Appl.* **18**:1 (1977), 1–24. MR Zbl

[Gustafson and Rao 1997] K. E. Gustafson and D. K. M. Rao, *Numerical range: the field of values of linear operators and matrices*, Springer, 1997. MR Zbl

[Harris et al. 2011] T. R. Harris, M. Mazzella, L. J. Patton, D. Renfrew, and I. M. Spitkovsky, "Numerical ranges of cube roots of the identity", *Linear Algebra Appl.* **435**:11 (2011), 2639–2657. MR Zbl

[Hausdorff 1919] F. Hausdorff, "Der Wertvorrat einer Bilinearform", *Math. Z.* **3**:1 (1919), 314–316. MR Zbl

[Heydari and Abdollahi 2015] M. T. Heydari and A. Abdollahi, "The numerical range of finite order elliptic automorphism composition operators", *Linear Algebra Appl.* **483** (2015), 128–138. MR Zbl

[Horn and Johnson 1991] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge University Press, Cambridge, 1991. MR Zbl

[Keeler et al. 1997] D. S. Keeler, L. Rodman, and I. M. Spitkovsky, "The numerical range of 3×3 matrices", *Linear Algebra Appl.* **252** (1997), 115–139. MR Zbl

[Kippenhahn 1951] R. Kippenhahn, "Über den Wertevorrat einer Matrix", *Math. Nachr.* **6** (1951), 193–228. MR Zbl

[Kippenhahn 2008] R. Kippenhahn, "On the numerical range of a matrix", *Linear Multilinear Algebra* **56**:1-2 (2008), 185–225. MR Zbl

[Lentzos and Pasley 2017] K. Lentzos and L. Pasley, "Determinantal representations of invariant hyperbolic plane curves", preprint, 2017. arXiv

[Li 1994] C.-K. Li, "C-numerical ranges and C-numerical radii", *Linear and Multilinear Algebra* **37**:1-3 (1994), 51–82. MR Zbl

[Li and Tsing 1991] C.-K. Li and N.-K. Tsing, "Matrices with circular symmetry on their unitary orbits and C-numerical ranges", *Proc. Amer. Math. Soc.* **111**:1 (1991), 19–28. MR Zbl

[Patton 2013] L. J. Patton, "Some block Toeplitz composition operators", *J. Math. Anal. Appl.* **400**:2 (2013), 363–376. MR Zbl

[Rodman and Spitkovsky 2005] L. Rodman and I. M. Spitkovsky, "3×3 matrices with a flat portion on the boundary of the numerical range", *Linear Algebra Appl.* **397** (2005), 193–207. MR Zbl

[Tam and Yang 1999] B.-S. Tam and S. Yang, "On matrices whose numerical ranges have circular or weak circular symmetry", *Linear Algebra Appl.* **302/303** (1999), 193–221. MR Zbl

[Toeplitz 1918] O. Toeplitz, "Das algebraische Analogon zu einem Satze von Fejér", *Math. Z.* **2**:1-2 (1918), 187–197. MR Zbl

[Tsai and Wu 2011] M. C. Tsai and P. Y. Wu, "Numerical ranges of weighted shift matrices", *Linear Algebra Appl.* **435**:2 (2011), 243–254. MR Zbl

[Tso and Wu 1999] S.-H. Tso and P. Y. Wu, "Matricial ranges of quadratic operators", *Rocky Mountain J. Math.* **29**:3 (1999), 1139–1152. MR Zbl

[Valentine 1964] F. A. Valentine, *Convex sets*, McGraw-Hill, 1964. MR Zbl

[Westwick 1975] R. Westwick, "A theorem on numerical range", *Linear and Multilinear Algebra* **2** (1975), 311–315. MR Zbl

[Wu 2011] P. Y. Wu, "Numerical ranges as circular discs", *Appl. Math. Lett.* **24**:12 (2011), 2115–2117. MR Zbl

shelby.burnett@yahoo.com   *California Polytechnic State University, San Luis Obispo, CA, United States*

ashleyychandler@gmail.com   *California Polytechnic State University, San Luis Obispo, CA, United States*

lpatton@calpoly.edu   *California Polytechnic State University, San Luis Obispo, CA, United States*

Counting eta-quotients of prime level

Allison Arnold-Roksandich, Kevin James and Rodney Keaton

(Communicated by Kenneth S. Berenhaut)

It is known that a modular form on $SL_2(\mathbb{Z})$ can be expressed as a rational function in $\eta(z)$, $\eta(2z)$ and $\eta(4z)$. By using known theorems and calculating the order of vanishing, we can compute the eta-quotients for a given level. Using this count, knowing how many eta-quotients are linearly independent, and using the dimension formula, we can figure out a subspace spanned by the eta-quotients. In this paper, we primarily focus on the case where the level is $N = p$, a prime. In this case, we will show an explicit count for the number of eta-quotients of level p and show that they are linearly independent.

1. Introduction and statement of results

Modular forms and cusp forms encode important arithmetic information, and are therefore important to study. An easy way to accomplish this is to study the Dedekind eta-function:

$$\eta(z) := q^{1/24} \prod_{n \geq 1} (1 - q^n), \quad \text{where} \quad q = e^{2\pi i z}. \tag{1-1}$$

In particular, we focus on functions of the form

$$f(z) = \prod_{d \mid N} \eta^{r_d}(dz), \quad r_d \in \mathbb{Z}, \tag{1-2}$$

which we call eta-quotients, as they provide nice examples of modular forms.

The following theorem is the primary motivation behind this paper.

Theorem 1.1 [Ono 2004, Theorem 1.67]. *Every modular form on $SL_2(\mathbb{Z})$ may be expressed as a rational function in $\eta(z)$, $\eta(2z)$, and $\eta(4z)$.*

While the recent work of Rouse and Webb [2015] has shown that Theorem 1.1 does not generalize to all levels, the subspace of eta-quotients for fixed level at least 2 is still interesting. The goal of this paper is to look at the vector space of modular

forms with prime level, $M_k(\Gamma_1(p))$, and count the number of eta-quotients for fixed weight k and level p, and compare the span of these eta-quotients with $M_k(\Gamma_1(p))$. In other words, this paper focuses on explicitly counting the eta-quotients that are modular forms for the congruence subgroups $\Gamma_0(p)$ and $\Gamma_1(p)$, where p is a prime.

Theorem 1.2. *Let $p > 3$ be a prime and k be an integer. Then there exists $f(z) = \eta^{r_1}(z)\eta^{r_p}(pz)$ such that $f(z)$ is a weakly holomorphic modular form with weight k of level p if and only if k is divisible by $h = \frac{1}{2}\gcd(p-1, 24)$.*

This first theorem provides a condition on k that is necessary and sufficient for showing that the space of weakly holomorphic modular forms with weight k and level p contains eta-quotients. With some effort, we can create similar conditions to guarantee when $f(z)$ is in $M_k(\Gamma_1(p))$. The next theorem gives an explicit count of the number of eta-quotients that are cusp forms of weight k and level p.

Theorem 1.3. *Let $p > 3$ be a prime. Let $k = hk'$, where h is the needed divisor of k given by Theorem 1.2. Let $d = (p-1)/(2h)$, and let c be the smallest positive integer representative of $k'h/12$ modulo d:*

(1) *For $c = k(p+1)/12 - \lfloor k(p+1)/(12d)\rfloor d$, the number of eta-quotients in $S_k(\Gamma_1(p))$ is*
$$\frac{k(p+1)}{12d} - 1.$$

(2) *For $c < k(p+1)/12 - \lfloor k(p+1)/(12d)\rfloor d$, the number of eta-quotients in $S_k(\Gamma_1(p))$ is*
$$\left\lceil \frac{k(p+1)}{12d} \right\rceil.$$

(3) *For $c > k(p+1)/12 - \lfloor k(p+1)/(12d)\rfloor d$, the number of eta-quotients in $S_k(\Gamma_1(p))$ is*
$$\left\lfloor \frac{k(p+1)}{12d} \right\rfloor.$$

There are also eta-quotients in $M_k(\Gamma_1(p))$ that are not cusp forms that are given by the following theorem.

Theorem 1.4. *Let $p > 3$ be a prime. Then, $M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p))$ contains at least one eta-quotient if and only if $\frac{1}{2}(p-1) \mid k$. Furthermore, for $k > 0$ and $\frac{1}{2}(p-1) \mid k$, there are exactly two eta-quotients in $M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p))$, which are of the forms*
$$\frac{\eta^{2pk/(p-1)}(pz)}{\eta^{2k/(p-1)}(z)} \quad and \quad \frac{\eta^{2pk/(p-1)}(z)}{\eta^{2k/(p-1)}(pz)}.$$

Finally, the following theorem also tells us the size of the subspace spanned by eta-quotients.

Theorem 1.5. *Let $p > 3$ be a prime. Then, the eta-quotients in $M_k(\Gamma_1(p))$ given by the previous theorems are linearly independent.*

Section 2 of this paper provides the necessary background for the results. The background includes information on modular forms, the dimension formula, and eta-quotients. Section 3 provides the proofs of the results given in this section. Finally, Section 4 details still-open questions and some ideas of how to extend these results further.

2. Background

2A. *Modular forms.* In this section, we present some definitions and basic facts from the theory of modular forms. For further details, the interested reader is referred to [Koblitz 1993, Chapter 3].

Definition 2.1. The *modular group*, denoted by $SL_2(\mathbb{Z})$, is the group of all matrices of determinant 1 which have integral entries.

The modular group acts on the upper half-plane $\mathcal{H} = \{x + iy \mid x, y \in \mathbb{R}, y > 0\}$ by linear fractional transformations

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}.$$

Furthermore, if we define $\mathcal{H}^*$ to be the set $\mathcal{H} \cup \mathbb{Q} \cup \{i\infty\}$, then the action of $SL_2(\mathbb{Z})$ on $\mathcal{H}$ extends to an action on $\mathcal{H}^*$ [Koblitz 1993].

There are only certain specific subgroups of $SL_2(\mathbb{Z})$ which we will use for our purposes. They are

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \,\middle|\, \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \pmod{N} \right\},$$

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \,\middle|\, \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \pmod{N} \right\}.$$

Each of these subgroups is called a congruence subgroup of level N. Note that if $N = 1$, then $\Gamma_0(N) = \Gamma_1(N) = SL_2(\mathbb{Z})$. This brings us to our next definition.

Definition 2.2. Let $\Gamma \leq SL_2(\mathbb{Z})$ be a congruence subgroup and define an equivalence relation on $\mathbb{Q} \cup \{\infty\}$ by $z_1 \sim z_2$ if there is a $\gamma \in \Gamma$ such that $\gamma \cdot z_1 = z_2$. We call each equivalence class under this relation a *cusp of* Γ.

Now, for an integer k and a function $f : \mathcal{H}^* \to \mathbb{C}$ and a $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$ we define the weight-k slash operator by

$$f|_k\gamma(z) = (cz+d)^{-k} f(\gamma \cdot z).$$

Note that we will often suppress the weight from the notation when it is clear from context or irrelevant for our purposes.

We can now define the objects which will be of primary interest to us.

Definition 2.3. A function $f : \mathcal{H}^* \to \mathbb{C}$ is called a *weakly holomorphic modular form* of weight k and level Γ if

(1) f is holomorphic on $\mathcal{H}$,

(2) f is modular, i.e., for every $\gamma \in \Gamma$ and $z \in \mathcal{H}$ we have $f \mid \gamma(z) = f(z)$, and

(3) f is meromorphic at each cusp of Γ.

Furthermore, if we replace condition (3) by "f is holomorphic at each cusp of Γ", then we call f a modular form. If we further replace condition (3) with "f vanishes at each cusp of Γ", then we call f a cusp form.

Consider a form f of level N. We will clarify what we mean by a function being "holomorphic at a cusp". First, consider the cusp $\{i\infty\}$, which we call "the cusp at ∞". Note that the matrix

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

is an element of $\Gamma_1(N)$ and hence $\Gamma_0(N)$ for every N. As our function satisfies condition (2), we have $f(Tz) = f(z+1) = f(z)$; i.e., our function is periodic. It is a basic fact from complex analysis that such a function has a Fourier expansion of the form

$$f(z) = \sum_{n=-\infty}^{\infty} a_n q^n, \quad \text{where} \quad q := e^{2\pi i z}.$$

Using this, we say that f is meromorphic at $\{i\infty\}$ if there is some $c < 0$ such that $a_n = 0$ for all $n < c$. We say that f holomorphic at $\{i\infty\}$ if $a_n = 0$ for all $n < 0$, and we say that f vanishes at $\{i\infty\}$ if $a_n = 0$ for all $n \leq 0$. We call the smallest n such that $a_n \neq 0$ the order of vanishing on the cusp at ∞. To cover another cusp α, let $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ be such that $\gamma \cdot \infty = \alpha$. Then, we need

$$(cz+d)^{-k} f(\gamma \cdot z) = \sum_{n=-\infty}^{\infty} c_n q^n.$$

If this holds, then we say f is meromorphic at α if there is some $c < 0$ such that $c_n = 0$ for all $n < c$. We also similarly say that the smallest n such that $c_n \neq 0$ is the order of vanishing at α.

Now, we set some notation which we will use throughout. For $\Gamma \leq \mathrm{SL}_2(\mathbb{Z})$ we denote the spaces of weakly holomorphic modular forms, modular forms, and cusp forms of level Γ and weight k by $M_k^!(\Gamma)$, $M_k(\Gamma)$, and $S_k(\Gamma)$, respectively. Note that the spaces $S_k(\Gamma) \leq M_k(\Gamma)$ are finite-dimensional complex vector spaces.

Throughout, we will also need the notion of a modular form with an associated character. We define a Dirichlet character of modulus N as a map $\chi : \mathbb{Z} \to \mathbb{C}$ such that:

(1) $\chi(m) = \chi(m + N)$ for all $m \in \mathbb{Z}$.

(2) If $\gcd(m, N) > 1$ then $\chi(m) = 0$. If $\gcd(m, N) = 1$, then $\chi(m) \neq 0$.

(3) $\chi(mn) = \chi(m)\chi(n)$ for all integers m, n.

Furthermore, if we let c be the minimal integer such that χ factors through $(\mathbb{Z}/c\mathbb{Z})^\times$, then we say χ has conductor c.

Let $f \in M_k(\Gamma_1(N))$ and suppose further that f satisfies

$$f \mid \gamma(z) = \chi(d) f(z) \quad \text{for all } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N).$$

Then we say that f is a modular form of level N and character χ, and we denote the space of such functions by $M_k(N, \chi)$. Note that this is defined similarly for weakly holomorphic modular forms and cusp forms.

It is well known that we have the decomposition

$$M_k(\Gamma_1(N)) = \bigoplus_{\chi \bmod N} M_k(N, \chi),$$

where the direct sum is over all Dirichlet characters modulo N. We can further decompose $M_k(N, \chi)$ into

$$M_k(N, \chi) = S_k(N, \chi) \oplus E_k(N, \chi),$$

where $S_k(N, \chi)$ is the space of cusp forms and $E_k(N, \chi)$, called the Eisenstein subspace, is orthogonal complement of $M_k(N, \chi)$ with respect to the Petersson inner product.

2B. *Dimension formulas.*

In this section we present formulas for the dimensions of spaces of cusp and modular forms. For more details regarding dimension formulas, the interested reader is referred to [Stein 2007].

2B1. *The dimension formula for level $\Gamma_0(p)$ with trivial character.* We present a formula for the dimension of $E_k(\Gamma_0(p))$ and $S_k(\Gamma_0(p))$ for a rational prime $p \geq 5$.

First, we set

$$\mu_{0,2}(p) = \begin{cases} 0 & \text{if } p \equiv 3 \pmod 4, \\ 2 & \text{if } p \equiv 1 \pmod 4, \end{cases} \quad \mu_{0,3}(p) = \begin{cases} 0 & \text{if } p \equiv 2 \pmod 3, \\ 2 & \text{if } p \equiv 1 \pmod 3. \end{cases}$$

Then define

$$g_0(p) = \tfrac{1}{12}(p+1) - \tfrac{1}{4}\mu_{0,2}(p) - \tfrac{1}{3}\mu_{0,3}(p).$$

dim $S_k(\Gamma_0(p))$				
$k(12) \downarrow\ p(12) \rightarrow$	1	5	7	11
0	$\frac{1}{12}(u+2)$	$\frac{1}{12}(u-6)$	$\frac{1}{12}(u-4)$	$\frac{1}{12}(u-12)$
1	0	0	0	0
2	$\frac{1}{12}(u-26)$	$\frac{1}{12}(u-18)$	$\frac{1}{12}(u-20)$	$\frac{1}{12}(u-12)$
3	0	0	0	0
4	$\frac{1}{12}(u-6)$	$\frac{1}{12}(u-6)$	$\frac{1}{12}(u-12)$	$\frac{1}{12}(u-12)$
5	0	0	0	0
6	$\frac{1}{12}(u-10)$	$\frac{1}{12}(u-18)$	$\frac{1}{12}(u-4)$	$\frac{1}{12}(u-12)$
7	0	0	0	0
8	$\frac{1}{12}(u-14)$	$\frac{1}{12}(u-6)$	$\frac{1}{12}(u-20)$	$\frac{1}{12}(u-12)$
9	0	0	0	0
10	$\frac{1}{12}(u-18)$	$\frac{1}{12}(u-18)$	$\frac{1}{12}(u-12)$	$\frac{1}{12}(u-12)$
11	0	0	0	0

Table 1. The dimension of $S_k(\Gamma_0(p))$ with trivial character and $k > 2$. Note that $u = (p+1)(k-1)$.

Using this we have $\dim S_2(\Gamma_0(p)) = g_0(p)$ and $\dim E_2(\Gamma_0(p)) = 1$, and for even $k \geq 4$ we have $\dim E_k(\Gamma_0(p)) = 2$ and

$$\dim S_k(\Gamma_0(p)) = (k-1)(g_0(p) - 1) + (k-2) + \mu_{0,2}(p)\lfloor \tfrac{1}{4}k \rfloor + \mu_{0,3}(p)\lfloor \tfrac{1}{3}k \rfloor.$$

From this, we see that our formula depends on the congruence class which k and p lie in modulo 12, so compiling these different congruences together we have Table 1.

2B2. *The dimension formula for $\Gamma_0(p)$ with quadratic character.* We will consider the case that our level is $\Gamma_0(p)$ for some rational prime p and that our associated character is quadratic. Note that at the end of the section we compile all of our computations together in a table for convenience.

In Section 2B1 we considered the trivial character case, so we now set $\chi(\cdot) = \left(\frac{\cdot}{p}\right)$. We must compute the summations

$$\sum_{x \in A_4(p)} \chi(x) \quad \text{and} \quad \sum_{x \in A_3(p)} \chi(x),$$

where $A_4(N) = \{x \in \mathbb{Z}/N\mathbb{Z} : x^2 + 1 = 0\}$ and $A_3(p) = \{x \in \mathbb{Z}/p\mathbb{Z} : x^2 + x + 1 = 0\}$.

First, we will consider $\sum_{x \in A_4(p)} \chi(x)$. This is clearly zero if $A_4(p)$ is empty, which occurs precisely when $p \equiv 3 \pmod 4$. Also, it is immediate that our summation equals 1 when $p = 2$. Now suppose $p \equiv 1 \pmod 4$. Then $\#A_4(p) = 2$.

Note that if $r \in A_4(p)$ then $-r \in A_4(p)$ and $\chi(r) = \chi(-r)$ since $\chi(-1) = 1$. Furthermore, it is not hard to see that $\chi(r) = 1$ if and only if there is an element of order 8 in $(\mathbb{Z}/p\mathbb{Z})^{\times}$, i.e., $p \equiv 1 \pmod 8$. Thus, we have

$$\sum_{x \in A_4(p)} \chi(x) = \begin{cases} 1 & \text{if } p = 2, \\ 0 & \text{if } p \equiv 3 \pmod 4, \\ 2 & \text{if } p \equiv 1 \pmod 8, \\ -2 & \text{if } p \equiv 5 \pmod 8. \end{cases}$$

Now we consider the summation $\sum_{x \in A_3(p)} \chi(x)$. Similar to the above, we have $A_3(p)$ is empty if $p \equiv 2 \pmod 3$, in which case our summation is zero. Also, if $p = 3$ then our summation is 1. Now, suppose that $p \equiv 1 \pmod 3$. Note: it is immediate that if $r \in A_3(p)$ then so is r^2. Similar to the previous situation, we have $\chi(r) = 1$ if and only if there is an element of order 6 in $(\mathbb{Z}/p\mathbb{Z})^{\times}$, i.e., $p \equiv 1 \pmod 6$. Note that as p is prime, it follows that $p \equiv 1 \pmod 6$ is equivalent to $p \equiv 1 \pmod 3$. Thus, we have

$$\sum_{x \in A_3(p)} \chi(x) = \begin{cases} 1 & \text{if } p = 3, \\ 0 & \text{if } p \equiv 2 \pmod 3, \\ 2 & \text{if } p \equiv 1 \pmod 3. \end{cases}$$

We summarize our calculations in Table 2.

2C. *Eta-quotients.* We introduce the eta-function and present some results relating this to modular forms. For further details regarding the eta-function, the interested reader is referred to [Köhler 2011].

Recall Dedekind's eta-function given in (1-1). The eta-function satisfies the following transformation properties with respect to our matrices S, T defined in Section 2A:

$$\eta(Sz) = \eta(-z^{-1}) = \sqrt{-iz}\,\eta(z), \quad \eta(Tz) = \eta(z+1) = e^{2\pi i/24}\eta(z).$$

More generally, we have the following general transformation formula for the eta-function:

$$\eta(\gamma z) = \epsilon(\gamma)(cz+d)^{1/2}\eta(z) \quad \text{for all } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

where

$$\epsilon(\gamma) = \begin{cases} \left(\dfrac{d}{|c|}\right)e^{(2\pi i/24)((a+d)c-bd(c^2-1)-3c)} & \text{if } c \text{ is odd,} \\ (-1)^{(1/4)(\mathrm{sgn}(c)-1)(\mathrm{sgn}(d)-1)}\left(\dfrac{d}{|c|}\right)e^{(2\pi i/24)((a+d)c-bd(c^2-1)+3d-3-3cd)} & \text{if } c \text{ is even,} \end{cases}$$

and $\mathrm{sgn}(x) = x/|x|$. For a proof of this transformation formula, the reader is referred to [Knopp 1970, Theorem 10, Chapter 3].

$\dim S_k\big(p, \big(\frac{\cdot}{p}\big)\big)$				
$k(12)\downarrow\ p(24)\rightarrow$	1	5	7	11
0	$\frac{1}{12}(u+8)$	$\frac{1}{12}(u-12)$	0	0
1	0	0	$\frac{1}{12}u$	$\frac{1}{12}(u-6)$
2	$\frac{1}{12}(u-20)$	$\frac{1}{12}u$	0	0
3	0	0	$\frac{1}{12}(u+2)$	$\frac{1}{12}(u-6)$
4	$\frac{1}{12}u$	$\frac{1}{12}(u-12)$	0	0
5	0	0	$\frac{1}{12}(u-14)$	$\frac{1}{12}(u-6)$
6	$\frac{1}{12}(u-4)$	$\frac{1}{12}u$	0	0
7	0	0	$\frac{1}{12}u$	$\frac{1}{12}(u-6)$
8	$\frac{1}{12}(u-4)$	$\frac{1}{12}(u-12)$	0	0
9	0	0	$\frac{1}{12}(u+2)$	$\frac{1}{12}(u-6)$
10	$\frac{1}{12}(u-12)$	$\frac{1}{12}u$	0	0
11	0	0	$\frac{1}{12}(u-14)$	$\frac{1}{12}(u-6)$
$k(12)\downarrow\ p(24)\rightarrow$	13	17	19	23
0	$\frac{1}{12}(u-4)$	$\frac{1}{12}u$	0	0
1	0	0	$\frac{1}{12}u$	$\frac{1}{12}(u-6)$
2	$\frac{1}{12}(u-8)$	$\frac{1}{12}(u-12)$	0	0
3	0	0	$\frac{1}{12}(u+2)$	$\frac{1}{12}(u-6)$
4	$\frac{1}{12}(u-12)$	$\frac{1}{12}u$	0	0
5	0	0	$\frac{1}{12}(u-14)$	$\frac{1}{12}(u-6)$
6	$\frac{1}{12}(u+8)$	$\frac{1}{12}(u-12)$	0	0
7	0	0	$\frac{1}{12}u$	$\frac{1}{12}(u-6)$
8	$\frac{1}{12}(u-20)$	$\frac{1}{12}u$	0	0
9	0	0	$\frac{1}{12}(u+2)$	$\frac{1}{12}(u-6)$
10	$\frac{1}{12}u$	$\frac{1}{12}(u-12)$	0	0
11	0	0	$\frac{1}{12}(u-14)$	$\frac{1}{12}(u-6)$

Table 2. Dimension of $S_k\big(p, \big(\frac{\cdot}{p}\big)\big)$. Note that $u = (p+1)(k-1)$.

In addition to the eta-function, we will also need to consider the related function $\eta(\delta z)$ for a positive integer δ. If we set $f(z) = \eta(\delta z)$ then $f(z)$ satisfies

$$f(\gamma z) = \epsilon\left(\begin{pmatrix} a & \delta b \\ c/\delta & d \end{pmatrix}\right)(cz+d)^{1/2} f(z) \quad \text{for all } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(\delta).$$

Finally, we will need the transformation

$$f(Tz) = e^{2\pi i \delta/24} f(z).$$

Notice that this function is "almost" a modular form. With this in mind, we consider certain products of these functions with the goal of eliminating the "almost". This brings us to eta-quotients, which we defined in (1-2). We are interested in when these eta-quotients are modular forms. We have the following theorem which partially answers this question.

Theorem 2.4 [Ono 2004, Theorem 1.64]. *Define the eta-quotient*

$$f(z) = \prod_{\delta \mid N} \eta^{r_\delta}(\delta z),$$

and set

$$k = \frac{1}{2} \sum_{\delta \mid N} r_\delta \in \mathbb{Z}.$$

Suppose our exponents $r_1, \ldots, r_N$ *satisfy*

$$\sum_{\delta \mid N} \delta r_\delta \equiv 0 \pmod{24} \quad and \quad \sum_{\delta \mid N} \frac{N}{\delta} r_\delta \equiv 0 \pmod{24}.$$

Then,

$$f|_k \gamma(z) = \chi(d) f(z)$$

for all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$, *where*

$$\chi(n) = \left(\frac{(-1)^k s}{n} \right)$$

with $s = \prod_{\delta \mid N} \delta^{r_\delta}$.

This theorem provides conditions on when an eta-quotient is a weakly holomorphic modular form. However, to answer the question of when an eta-quotient is a modular form we need the following theorem, which provides information concerning the order of vanishing at the cusps of $\Gamma_0(N)$.

Theorem 2.5 [Ono 2004, Theorem 1.65]. *Let* $f(z)$ *be an eta-quotient satisfying the conditions of Theorem 2.4. Let* $c, d \in \mathbb{N}$ *with* $d \mid N$ *and* $(c, d) = 1$. *Then, the order of vanishing of* $f(z)$ *at the cusp* c/d *is*

$$v_d = \frac{N}{24} \sum_{\delta \mid N} \frac{(d, \delta)^2 r_\delta}{(d, N/d) d \delta}.$$

3. Proofs of results

We will provide the proofs for the results given in Section 1. We will assume that the eta-quotients being discussed always have $N = p > 3$, which is a prime, unless

otherwise stated. From Theorems 2.4 and 2.5, we have conditions that tell us when an eta-quotient is a holomorphic modular form. Thus, we will use the equations

$$\tfrac{1}{2}(r_1 + r_p) = k, \tag{3-1}$$

$$r_1 + p r_p \equiv 0 \pmod{24}, \tag{3-2}$$

$$p r_1 + r_p \equiv 0 \pmod{24}, \tag{3-3}$$

$$v_1 = \tfrac{1}{24}(p r_1 + r_p), \tag{3-4}$$

$$v_p = \tfrac{1}{24}(r_1 + p r_p), \tag{3-5}$$

where v_1 and v_p are the orders of vanishing at the two cusps of $\Gamma_1(p)$, $i\infty$ and $1/p$, respectively.

For a fixed prime p and a fixed weight k, we see that it is possible to express r_p in terms of r_1 by (3-1). It is convenient to rewrite (3-4) and (3-5) as

$$24 v_1 = 2k + (p-1) r_1, \tag{3-6}$$

$$24 v_p = 2kp + (1-p) r_1. \tag{3-7}$$

It is now clear that we can relate the orders of vanishing to the weight of an eta-quotient by

$$24(v_1 + v_p) = 2k(p+1). \tag{3-8}$$

We begin the discussion for counting eta-quotients of level $\Gamma_0(p)$ by looking at possible conditions on k. These conditions were stated in Theorem 1.2, which we restate here for convenience.

Theorem 1.2. *Let $p > 3$ be a prime and k be an integer. Then there exists $f(z) = \eta^{r_1}(z)\eta^{r_p}(pz)$ such that $f(z)$ is a weakly holomorphic modular form with weight k of level p if and only if k is divisible by $h = \tfrac{1}{2} \gcd(p-1, 24)$.*

Proof. ($\rightarrow$) Suppose that $f(z) \in M^!(\Gamma_1(p))$. We note that it suffices to show that we can satisfy (3-7) and (3-8) since (3-6) can be gained from these two.

From (3-8), we see that we want $\tfrac{1}{12}k(p+1)$ to be an integer, as the orders of vanishing, v_1 and v_p, are integers. From here we can find a divisor d of k that would make this possible. Then by (3-7), we know that we need $24 \mid (2kp - (p-1)r_1)$. This gives us

$$2pdn \equiv (p-1)r_1 \pmod{24}, \tag{3-9}$$

where $dn = k$. Let $\delta = \gcd(24, 2dp, p-1)$. Then, we get that $2dp/\delta$, $(p-1)/\delta \in (\mathbb{Z}/(24/\delta)\mathbb{Z})^\times$ and obtain our desired conclusion, where $d = h = \tfrac{1}{2}\gcd(p-1, 24)$; except for when p is congruent to 1 or 17 modulo 24.

Suppose that $p \equiv 1 \pmod{24}$. Then we can rewrite (3-9) as

$$12n \equiv 0 \pmod{24}.$$

This tells us that n must be even. Thus, we have $k \equiv 0 \pmod{12}$. We further note that $12 = \frac{1}{2}\gcd(24\ell, 24)$, therefore showing our result for this case.

Suppose that $p \equiv 17 \pmod{24}$. Rewriting (3-9), we get

$$68n \equiv 16r_1 \pmod{24}.$$

This tells us that

$$5n \equiv 4r_1 \pmod{6}.$$

Since $5 \in \mathbb{Z}/6\mathbb{Z}^\times$, we have

$$n \equiv 2r_1 \pmod{6}.$$

Therefore, n must be even, and we have $4 \mid k$. As $4 = \frac{1}{2}\gcd(24\ell + 16, 24)$, we reach our desired conclusion.

($\leftarrow$) Suppose that $h = \frac{1}{2}\gcd(p - 1, 24)$ divides k. We want to show that there exists $f(z) = \eta^{r_1}(z)\eta^{r_p}(pz)$ in $M_k^!(\Gamma_1(p))$. It is sufficient to show that there exists r_1 such that $r + p(2k - r_1) \equiv 0 \pmod{24}$. We can interpret this as

$$r_1(1 - p) + 2pk = 24N,$$

where $N \in \mathbb{Z}$. As $2h$ divides every term, we can get

$$-r_1 d + p\frac{2k}{2h} = \frac{24}{2h}N.$$

Therefore, we have

$$dr_1 \equiv p\frac{k}{h} \left(\bmod \frac{24}{2h}\right).$$

Since d and $24/(2h)$ are relatively prime, d has an inverse in $\mathbb{Z}/(24/(2h))\mathbb{Z}$. Thus, there exists a unique $r_1 \in \mathbb{Z}/(24/(2h))\mathbb{Z}$ such that

$$r_1 \equiv p\frac{k}{h}(d)^{-1} \left(\bmod \frac{24}{2h}\right). \qquad \square$$

As mentioned in Section 1, we can extend Theorem 1.2 to show when there exists $f(z) = \eta^{r_1}(z)\eta^{r_p}(pz) \in M_k(\Gamma_1(p))$. Before we do so, we need a lemma.

Lemma 3.1. *Let N be an integer such that $\gcd(N, 6) = 1$. Let $f(z)$ be given by*

$$f(z) = \prod_{d \mid N} \eta^{r_d}(dz).$$

If $f \in M_k(\Gamma_0(N), \chi)$, then it must be the case that

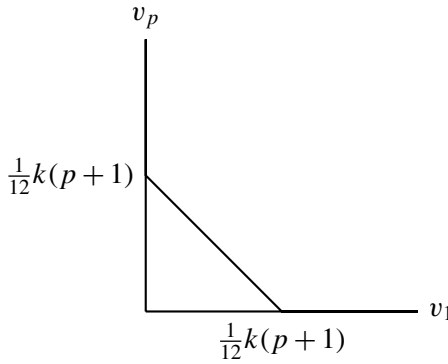$$\sum_{d \mid N} dr_d \equiv 0 \pmod{24} \quad and \quad \sum_{d \mid N} \frac{N}{d}r_d \equiv 0 \pmod{24}.$$

Figure 1. The line $v_1 + v_p = \frac{1}{12}k(p+1)$.

Proof. Since $f \in M_k(\Gamma_0(p), \chi)$, the q-series expansion of f about the cusp at infinity must look like

$$f(z) = \sum_{n \geq 0} c_n q^n.$$

Recall that $\eta(z) = q^{1/24} \prod_{n \geq 1}(1 - q^n)$. Thus, we would have

$$\prod_{d \mid N} \eta^{r_d}(dz) = q^{(\sum_{d \mid N} r_d d)/24} \prod_{n \geq 1} \left(\prod_{d \mid N} (1 - q^{dn})^{r_d} \right).$$

Therefore, we need 24 to divide

$$\sum_{d \mid N} dr_d.$$

We also note that for all primes $p \geq 5$, we have $p^2 \equiv 1 \pmod{24}$. Therefore, $Nd \equiv N/d \pmod{24}$. Thus, we have

$$0 \equiv N \sum_{d \mid N} dr_d \equiv \sum_{d \mid N} \frac{N}{d} r_d \pmod{24}. \qquad \square$$

As we wish to focus on holomorphic modular forms, we now want nonnegative orders of vanishing, i.e., $v_1, v_p \geq 0$. Using this condition and (3-8), we also have $v_1, v_p \leq k(p+1)/12$. We use Figure 1 to show the line that relates v_1 to v_p given a fixed k and p. We note that given (3-6), we can define v_1 in terms of r_1, and vice versa. Thus, to count the number of eta-quotients of our desired form, it suffices to count the number of possible orders of vanishing. As orders of vanishing are integer values, we only consider integer points on the line given in Figure 1.

Furthermore, from (3-6), we have

$$(p - 1)r_1 = 24v_1 - 2k.$$

This implies that $24v_1 - 2k \equiv 0 \pmod{p-1}$. In other words, we can write $24v_1 - 2k = (p-1)\ell$, where $\ell \in \mathbb{Z}$. Recall how we defined h in Theorem 1.2. Since $2h \mid 2k$ and $2h \mid 24$, we can write $24/(2h)v_1 - k' = d\ell$. We also know that $2h \mid (p-1)$. Therefore, we have

$$\frac{24}{2h}v_1 \equiv k' \pmod{d}.$$

Since we have $2h = \gcd(p-1, 24)$, we get $1 = \gcd(d, 24/(2h))$. This implies that we have a multiplicative inverse of $24/(2h)$ in $\mathbb{Z}/d\mathbb{Z}$. Thus, we have

$$v_1 \equiv \left(\frac{24}{2h}\right)^{-1} k' \pmod{d}. \tag{3-10}$$

From Theorem 2.4 and Lemma 3.1, we get that (3-10) becomes a necessary and sufficient condition for an eta-quotient with order of vanishing v_1 to be in $M_k^!(\Gamma_1(p))$. Now, we have the following corollary which follows from this explanation as well as Theorem 1.2 and Lemma 3.1.

Corollary 3.2. *Let $p \geq 5$ be a prime. There exists $f(z) = \eta^{r_1}(z)\eta^{r_p}(pz)$ in $M_k(\Gamma_1(p))$ if and only if $h = \frac{1}{2}\gcd(p-1, 24)$ divides k and $d \leq \frac{1}{12}k(p+1)$.*

We note that by definition, cusp forms occur on the interior of our line, and non-cuspidal modular forms occur at the end points. For this reason it is useful to perform the counts of cusp forms and noncuspidal modular forms separately. First, we prove the count of cusp forms given in Theorem 1.3, which we restate here for convenience.

Theorem 1.3. *Let $p > 3$ be a prime. Let $k = hk'$, where h is the needed divisor of k given by Theorem 1.2. Let $d = (p-1)/(2h)$, and let c be the smallest positive integer representative of $k'h/12$ modulo d.*

(1) *For $c = k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$, the number of eta-quotients in $S_k(\Gamma_1(p))$ is*

$$\frac{k(p+1)}{12d} - 1.$$

(2) *For $c < k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$, the number of eta-quotients in $S_k(\Gamma_1(p))$ is*

$$\left\lceil \frac{k(p+1)}{12d} \right\rceil.$$

(3) *For $c > k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$, the number of eta-quotients in $S_k(\Gamma_1(p))$ is*

$$\left\lfloor \frac{k(p+1)}{12d} \right\rfloor.$$

Proof. Since we are only considering cusp forms, we can assume that $v_1, v_p > 0$. The number of points on our line from Figure 1 which satisfy this inequality and the congruence from (3-10) is the number of eta-quotients. We now consider three cases.

Case 1: Suppose $c = 0 = k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$. Then, we have $v_1 \equiv 0 \pmod d$. Furthermore, we note that $d \mid k(p+1)/12$. Thus, we have the number of points which match our congruence is $k(p+1)/(12d)$. However, we note that one of these points gives us $v_p = 0$, which is not desired. Therefore, the number of eta-quotients that are in $S_k(\Gamma_1(p))$ is

$$\frac{k(p+1)}{12d} - 1.$$

Case 2: Suppose $c < k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$. Note that $\lfloor k(p+1)/(12d) \rfloor d$ is less than $k(p+1)/12$. However, since $c < k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$, we have another point to count that is between $\lfloor k(p+1)/(12d) \rfloor d$ and $k(p+1)/12$. Therefore, the number of eta-quotients that are in $S_k(\Gamma_1(p))$ is

$$\left\lceil \frac{k(p+1)}{12d} \right\rceil.$$

Case 3: Suppose $c > k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$. Note that $\lfloor k(p+1)/(12d) \rfloor d$ is less than $k(p+1)/12$. Since $c > k(p+1)/12 - \lfloor k(p+1)/(12d) \rfloor d$, we have no more points to count between $\lfloor k(p+1)/(12d) \rfloor d$ and $k(p+1)/12$. Therefore, the number of eta-quotients that are in $S_k(\Gamma_1(p))$ is

$$\left\lfloor \frac{k(p+1)}{12d} \right\rfloor. \qquad \square$$

Second, we prove the count of noncusp forms given in Theorem 1.4, which we restate here for convenience.

Theorem 1.4. *Let $p > 3$ be a prime. Then, $M_k(\Gamma_0(p)) \setminus S_k(\Gamma_1(p))$ contains at least one eta-quotient if and only if $\frac{1}{2}(p-1) \mid k$. Furthermore, for $k > 0$ and $\frac{1}{2}(p-1) \mid k$, there are exactly two eta-quotients in $M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p))$, which are of the form*

$$\frac{\eta^{2pk/(p-1)}(pz)}{\eta^{2k/(p-1)}(z)} \quad and \quad \frac{\eta^{2pk/(p-1)}(z)}{\eta^{2k/(p-1)}(pz)}.$$

Proof. ($\rightarrow$) Suppose $f(x) \in M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p))$ is an eta-quotient that satisfies Theorem 2.4. Then, we know that at least one of the orders of vanishing must be zero. Thus, we have two cases.

Case 1: Suppose $v_1 = 0$. Then, $pr_1 + r_p = 0$, which can be rewritten to get $(p-1)r_1 = -2k$. Therefore, we have $\frac{1}{2}(p-1) \mid k$. Furthermore, we can get that $r_1 = 2k/(p-1)$, and thus $r_p = 2pk/(p-1)$. When plugging these values into v_p we get

$$v_p = \frac{1}{24} \left(\frac{-2k}{p-1} + \frac{2pk}{p-1} \right) = \frac{2k}{24} > 0.$$

Case 2: Suppose $v_p = 0$. Then, $r_1 + pr_p = 0$, which can be rewritten to get $(1-p)r_1 = -2pk$. Therefore, $\frac{1}{2}(p-1) \mid k$ since $p \nmid (p-1)$ and therefore $p \mid r_1$.

Furthermore, we get that $r_1 = 2pk/(p-1)$, and thus $r_p = -2k/(p-1)$. When plugging these values into v_1 we get

$$v_1 = \frac{1}{24}\left(\frac{2pk}{p-1} + \frac{-2k}{p-1}\right) = \frac{2k}{24} > 0.$$

In both cases, the number needed to divide k is the same. Furthermore, both create a single eta-quotient for a fixed k. Therefore, we have $\frac{1}{2}(p-1) \mid k$. Furthermore, there are exactly two eta-quotients which result from looking at either of the orders of vanishing being zero, and they are

$$\frac{\eta^{2pk/(p-1)}(pz)}{\eta^{2k/(p-1)}(z)} \quad \text{and} \quad \frac{\eta^{2pk/(p-1)}(z)}{\eta^{2k/(p-1)}(pz)}.$$

($\leftarrow$) Suppose that $k = \frac{1}{2}(p-1)m > 0$ for some integer m. Also, suppose we have the two eta-quotients

$$\frac{\eta^{2pk/(p-1)}(pz)}{\eta^{2k/(p-1)}(z)} \quad \text{and} \quad \frac{\eta^{2pk/(p-1)}(z)}{\eta^{2k/(p-1)}(pz)}.$$

We consider each eta-quotient as its own case.

Case 1: Consider $\eta^{2pk/(p-1)}(pz)/\eta^{2k/(p-1)}(z)$. Note that $r_1 + r_p = -2k/(p-1) + 2pk/(p-1) = 2k$. Furthermore,

$$r_1 + pr_p = \frac{-2k}{p-1} + \frac{2pk}{p-1}p = (p^2-1)m \equiv 0 \pmod{24}$$

since p is relatively prime to 24, and

$$pr_1 + r_p = p\frac{-2k}{p-1} + \frac{2pk}{p-1} = 0 \equiv 0 \pmod{24}.$$

When looking at the orders of vanishing, we get

$$v_1 = \frac{1}{24}(pr_1 + r_p) = \frac{1}{24}\left(p\frac{-2k}{p-1} + \frac{2pk}{p-1}\right) = 0 \geq 0 \quad \text{and}$$

$$v_p = \frac{1}{24}(r_1 + pr_p) = \frac{1}{24}\left(\frac{-2k}{p-1} + \frac{2pk}{p-1}p\right) = \frac{1}{24}(p^2-1)m \geq 0.$$

Since our orders of vanishing are both ≥ 0 and one of them is equal to 0, we have $\eta^{2pk/(p-1)}(pz)/\eta^{2k/(p-1)}(z) \in M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p))$.

Case 2: Consider $\eta^{2pk/(p-1)}(z)/\eta^{2k/(p-1)}(pz)$. Note that $r_1 + r_p = 2pk/(p-1) + -2k/(p-1) = 2k$. Furthermore,

$$r_1 + pr_p = \frac{2pk}{p-1} + p\frac{-2k}{p-1} = 0 \equiv 0 \pmod{24},$$

and since p is relatively prime to 24,

$$pr_1 + r_p = p\frac{2pk}{p-1} + \frac{-2k}{p-1} = (p^2-1)m \equiv 0 \pmod{24}.$$

When looking at the orders of vanishing, we get

$$v_1 = \frac{1}{24}(pr_1+r_p) = \frac{1}{24}\left(p\frac{2pk}{p-1} + \frac{-2k}{p-1}\right) = \frac{1}{24}(p^2-1)m \geq 0 \quad \text{and}$$

$$v_p = \frac{1}{24}(r_1+pr_p) = \frac{1}{24}\left(\frac{2pk}{p-1} + \frac{-2k}{p-1}p\right) = 0 \geq 0.$$

Since our orders of vanishing are both ≥ 0 and one of them is equal to 0, we have

$$\frac{\eta^{2pk/(p-1)}(z)}{\eta^{2k/(p-1)}(pz)} \in M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p)).$$

Thus, $M_k(\Gamma_1(p)) \setminus S_k(\Gamma_1(p))$ contains exactly two eta-quotients. $\square$

From the eta-quotients given in the theorem, let $k = \frac{1}{2}(p-1)m$ where m is a positive integer. Then the eta-quotients have characters

$$\chi_1(n) = \left(\frac{(-1)^{(p-1)/2m}p^{pm}}{n}\right) \quad \text{and} \quad \chi_2(n) = \left(\frac{(-1)^{(p-1)/2m}p^m}{n}\right),$$

respectively. In the case where m is even, both of the characters are guaranteed to be the trivial character. When m is odd, we are guaranteed to have a quadratic character. In fact, both quadratic characters are the same.

Now that we know how many eta-quotients there are and can write down what they are if needed, we would like to know the dimension of the space spanned by these eta-quotients. This is provided by Theorem 1.5, which we restate here for convenience.

Theorem 1.5. *Let $p > 3$ be a prime. Then, the eta-quotients in $M_k(\Gamma_1(p))$ given by the previous theorems are linearly independent.*

Proof. Suppose that we are looking at eta-quotients in $M_k(\Gamma_1(p))$ for a prime $p > 3$. Without loss of generality, we look at the Fourier series about the cusp at ∞. By using the Sturm bound [Ono 2004], we get that we need to compare the first $\lfloor \frac{1}{12}pk \rfloor + 1$ terms of each Fourier series. We can pick a cusp and order the eta-quotients increasingly by looking at the order of vanishing. We can then create a matrix A where the i, j-th entry represents $a(j)$ in the i-th eta-quotient's Fourier series. Since all of the eta-quotients have different orders of vanishing and are in increasing order, we get that A is in echelon form. This tells us that all the rows are linearly independent. Thus all of the eta-quotients are linearly independent. $\square$

The following corollaries can all be obtained by comparing dimension formulas with our counts and applying the previous theorem.

Corollary 3.3. *Let $p \geq 5$ be a prime. Recall that $h = \frac{1}{2}\gcd(p-1, 24)$ from* Theorem 1.2. *Denote the space of level p, weight k eta-quotients by $\eta_k(p)$.*

(1) *If $p \equiv 3 \pmod 4$, then taking the limit over odd k in the appropriate congruence class from* Theorem 1.2 *gives*

$$\lim_{k \to \infty} \frac{\dim \eta_k(p)}{\dim S_k\left(p, \left(\frac{\cdot}{p}\right)\right)} = \frac{2h}{p-1}.$$

(2) *If $p \equiv 3 \pmod 4$, then taking the limit over even k in the appropriate congruence class from* Theorem 1.2 *gives*

$$\lim_{k \to \infty} \frac{\dim \eta_k(p)}{\dim S_k(\Gamma_0(p))} = \frac{2h}{p-1}.$$

(3) *If $p \equiv 1 \pmod 4$, then taking the limit over all k in the appropriate congruence class from* Theorem 1.2 *gives*

$$\lim_{k \to \infty} \frac{\dim \eta_k(p)}{\dim S_k(\Gamma_0(p)) + \dim S_k\left(p, \left(\frac{\cdot}{p}\right)\right)} = \frac{h}{p-1}.$$

Finally, we would like to consider the case that our v_1 and v_p are integral but do not correspond to integral r_1, r_p. To gain some intuition concerning the properties of the "eta-quotients" formed from these r_1, r_p, we consider the following example.

Example 3.4. Let $p = 11$ and $k = 6$. Note that in this situation we have that in order to have eta-quotients we must have $v_1 \equiv 3 \pmod 5$. So, we will investigate the properties of the function obtained by choosing $v_1 \not\equiv 3 \pmod 5$.

Consider $v_1 = 1$. This implies $v_p = 5$. Then,

$$\begin{pmatrix} r_1 \\ r_p \end{pmatrix} = \begin{pmatrix} 1 & 11 \\ 11 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 24 \\ 120 \end{pmatrix} = \begin{pmatrix} 54/5 \\ 6/5 \end{pmatrix}.$$

Now, we can use these to form the "eta-quotient"

$$f(z) = \eta^{54/5}(z)\eta^{6/5}(11z).$$

Using the transformation properties from Section 2C we have

$$f(Tz) = e^{27\pi i/30}\eta^{54/5}(z)e^{11\pi i/10}\eta^{6/5}(11z) = f(z).$$

Note that if we raise $f(z)$ to the fifth power to cancel denominators of r_1 and r_p then we can use Theorem 2.4 to verify that we obtain $f(z)^5 \in S_{30}(\Gamma_0(11))$, i.e., our lattice point corresponds to a "root" of an η quotient of higher weight.

Note that the remaining choices for v_1 give us similar results.

4. Conclusion and further questions

We detailed the number of eta-quotients in $M_k(\Gamma_0(p))$ of the form $\eta^{r_1}(z)\eta^{r_p}(pz)$. Further work in this project would involve generalizing these results for all levels as well as figuring out linear combinations of eta-quotients that would be possible on a given level.

Acknowledgments

References

[Knopp 1970] M. I. Knopp, *Modular functions in analytic number theory*, Markham, Chicago, 1970. MR Zbl

[Koblitz 1993] N. Koblitz, *Introduction to elliptic curves and modular forms*, 2nd ed., Graduate Texts in Mathematics **97**, Springer, 1993. MR Zbl

[Köhler 2011] G. Köhler, *Eta products and theta series identities*, Springer, 2011. MR

[Ono 2004] K. Ono, *The web of modularity: arithmetic of the coefficients of modular forms and q-series*, CBMS Regional Conference Series in Mathematics **102**, American Mathematical Society, Providence, RI, 2004. MR Zbl

[Rouse and Webb 2015] J. Rouse and J. J. Webb, "On spaces of modular forms spanned by eta-quotients", *Adv. Math.* **272** (2015), 200–224. MR Zbl

[Stein 2007] W. Stein, *Modular forms, a computational approach*, Graduate Studies in Mathematics **79**, American Mathematical Society, Providence, RI, 2007. MR Zbl

arnoldra@oregonstate.edu   *Department of Mathematics, Oregon State University, Corvallis, OR, United States*

kevja@clemson.edu   *Department of Mathematical Sciences, Clemson University, Clemson, SC, United States*

keatonr@etsu.edu   *Department of Mathematics and Statistics, East Tennessee State University, Johnson City, TN, United States*

The k-diameter component edge connectivity parameter

Nathan Shank and Adam Buzzard

(Communicated by Joshua Cooper)

We focus on a network reliability measure based on edge failures and considering a network operational if there exists a component with diameter k or larger. The *k-diameter component edge connectivity parameter* of a graph is the minimum number of edge failures needed so that no component has diameter k or larger. This implies each resulting vertex must not have a k-neighbor. We give results for specific graph classes including path graphs, complete graphs, complete bipartite graphs, and a surprising result for perfect r-ary trees.

1. Introduction

Network reliability and graph connectivity parameters have been studied for many years. The network reliability measure can vary greatly based on the type of application being considered. In particular networks, the vulnerabilities of particular pieces of the network often influence the parameter used to measure reliability. In particular cases, nodes or vertices may fail or become inoperable; in other cases, the edges or connections between vertices may fail or become inoperable and in some cases both the nodes and the edges may fail. See [Boesch et al. 2009] for a survey of recent results and techniques.

In general, network reliability measures are driven by two different yet connected concepts. First, we need to know what objects are prone to failure: edges, vertices, or both. Second, we need to know what the requirements are to make a network functional. Stated differently, we need to know what objects fail and what characterizes a failure state for a network.

Vertex connectivity and *edge connectivity* are two of the original network reliability measures which have been studied extensively. The vertex connectivity parameter is the minimum number of vertices that must be deleted so that the resulting graph is disconnected. Similarly the edge connectivity parameter measures

the minimum number of edges that must be deleted so that the resulting graph is disconnected. These parameters have been generalized to other reliability measures based on different characterizations of failure states for networks. For example, the *component order vertex connectivity parameter* is the minimum number of vertices that must be deleted so that the resulting graph has all components of order less than some value k (see [Boesch et al. 1998; 1999] for example). Similarly the *component order edge connectivity parameter* is the minimum number of edges that must be deleted so that the resulting graph has all components of order less than some value k (see [Boesch et al. 2006; 2007] for example).

Conditional connectivity was studied by Frank Harary [1983]. It requires each component of a disconnected graph to have a chosen property P. Thus if P is any property of a graph $G = (V, E)$ and $S \subset V(G)$, then the P-connectivity of G is the minimum $|S|$ such that $G - S$ is disconnected and every component of $G - S$ has property P. Similarly we can define the edge conditional connectivity parameter of G if we consider edge deletions rather than vertex deletions.

In this paper, we focus our attention on edge failures and consider a graph to be in a failure state if no vertex has a neighbor of a fixed distance. In other words, we study the minimum number of edges that can fail in order to produce a graph which has all components with a diameter less than some fixed value. In this particular case a network would be operational if there exists a component with a sufficiently large diameter.

One important application of such a parameter centers around the spread of disease or genetic traits. If a particular disease or genetic trait only becomes active after k successive transmissions, then we would want to stop the spread so that components in the network (tree) have diameter less than k. This will be explored more in Section 3B.

2. Background and definitions

Throughout this paper, let $G = (V, E)$ be a simple graph with vertex set V and edge set E. For any set A, let $|A|$ denote the cardinality of A. If $D \subset E$, let $G - D$ denote the subgraph of G containing the vertex set V and the edge set $E - D$. Thus $G - D = (V, E - D)$.

Throughout the paper, unless otherwise specified, we will assume that n, r, l, and k are all positive integers. We will also use the conventions of notation adapted from [West 1996]. A pair of vertices u, v are said to be k-neighbors if the distance between u and v is k, written as $d(u, v) = k$.

Definition 2.1. Let $G = (V, E)$ be a graph and k be a positive integer. A set $D \subseteq E$ is a *k-diameter component edge disconnecting set* if $G - D$ has all components of diameter less than k.

This means that an edge set D is a *k-diameter component edge disconnecting set* if no vertex in $G - D$ has a k-neighbor. If D is a *k-diameter component edge disconnecting set* then $G - D$ is said to be a *failure state.*

Definition 2.2. Given a graph $G = (V, E)$ and a positive integer k, the *k-diameter component edge connectivity parameter* of G, denoted by $\mathrm{CE}_k(G)$, is the size of the smallest k-diameter component edge disconnecting set.

Thus, the k-diameter component edge connectivity parameter is the size of the smallest edge set D such that $G - D$ is a failure state.

3. Results

When $k = 1$, a failure state will occur if no vertex has a 1-neighbor. In order for this to occur every edge must be removed. Thus $\mathrm{CE}_1(G) = |E|$ for every graph $G = (V, E)$. Therefore for the remainder of the paper we will assume that $k \geq 2$.

In Section 3A we will show some easy results for some simple graph classes, particularly path graphs, complete graphs, and complete bipartite graphs. In Section 3B1 we will consider perfect r-ary trees.

3A. *Simple graphs.*

3A1. *Path graphs.* The first type of graph we will consider is a path on n vertices, denoted by P_n. We can label the edges consecutively from 1 to $n - 1$ starting at a pendant edge. For a component to have a diameter less than k, it can have at most $k - 1$ edges. If we delete every edge whose label is a multiple of k, then the remaining components all have $k - 1$ edges, except for possibly one component which could have less than $k - 1$ edges. Therefore the diameter of each component will be less than k. Hence we see $\mathrm{CE}_k(P_n) \leq \lfloor (n-1)/k \rfloor$.

Since path graphs are trees, every edge deletion creates one new component. Since we cannot have components of length k in a failure state, we need at least one edge deletion in every k-edge disjoint connected subpath. Hence $\mathrm{CE}_k(P_n) \geq \lfloor (n-1)/k \rfloor$. These two observations imply the following:

Theorem 3.1. *For every positive integer n,*

$$\mathrm{CE}_k(P_n) = \left\lfloor \frac{n-1}{k} \right\rfloor.$$

3A2. *Complete graphs.* Since the diameter of K_n is 1, K_n is already a failure state. Thus we see the following obvious result:

Theorem 3.2. *For every positive integer n,*

$$\mathrm{CE}_k(K_n) = 0.$$

3A3. *Complete bipartite Graphs.* Consider the complete bipartite graph $K_{a,b} = (V, E)$ with parts A and B, where $V = A \cup B$, $A \cap B = \varnothing$, $|A| = a > 0$ and $|B| = b > 0$. Recall that the diameter of a complete bipartite graph is 2 unless $a = b = 1$, in which case the diameter is 1. If $k > 2$, then $K_{a,b}$ is already a failure state. If $k = 2$, then the size of the largest subgraph in a failure state is the size of the maximum matching in $K_{a,b}$, which is $\min\{a, b\}$. So the number of edges that must be deleted to produce a failure state is $\min\{a, b\}$ less than the total number of edges. Therefore we have the following theorem:

Theorem 3.3. *For every pair of positive integers $a \leq b$,*

$$CE_k(K_{a,b}) = \begin{cases} 0 & \text{if } k > 2, \\ a(b-1) & \text{if } k = 2. \end{cases}$$

3B. *Trees.*

3B1. *Perfect r-ary trees.* We will now consider perfect r-ary trees.

Definition 3.4. Let $T_{r,l} = (V, E)$ denote a perfect r-ary tree with height l, where

$$V = \{v_{i,j} : 1 \leq i \leq l + 1, \ 1 \leq j \leq r^{(l+1)-i}\}, \quad \text{and}$$

$$E = \{(v_{i,j}, v_{i-1,m}) : 2 \leq i \leq l + 1, \ 1 \leq j \leq r^{(l+1)-i}, \ (j-1)r + 1 \leq m \leq jr\}.$$

We will say that vertex $v_{i,h} \in V(T_{r,l})$ is on *level i*. Notice we are using the unconventional notation that the root vertex of the full complete tree is on level $l + 1$ and the leaves are on level 1.

In order to separate the tree into failure states we need to know the distance between vertices. The following lemma shows a lower bound for the distance between two vertices in the same level.

Lemma 3.5. *Assume $T_{r,l} = (V, E)$ and $v_{i,j}, v_{i,j+pr^{n-1}} \in V$ for some positive integers i, j, n, and p. Then*

$$d(v_{i,j}, v_{i,j+pr^{n-1}}) \geq 2n.$$

Proof. We will proceed by induction on n. Consider the case when $n = 1$.

Since $v_{i,j}$ and $v_{i,j+p}$ are both on level i, they are not adjacent. Since any two vertices of a tree are connected by a path, we conclude $d(v_{i,j}, v_{i,j+p}) \geq 2$.

Assume there exists a positive integer n such that for any pair $v_{a,b}, v_{a,b+pr^{n-1}} \in V$,

$$d(v_{a,b}, v_{a,b+pr^{n-1}}) \geq 2n.$$

Consider a pair of vertices, $v_{i,j}, v_{i,j+qr^n} \in V$ for some positive integer q. The unique path from $v_{i,j}$ to $v_{i,j+qr^n}$ must contain vertices

$$v_{i+1,\lceil j/r \rceil} \quad \text{and} \quad v_{i+1,\lceil j/r \rceil + qr^{n-1}}.$$

By induction we know

$$d(v_{i+1,\lceil j/r \rceil}, v_{i+1,\lceil j/r \rceil + qr^{n-1}}) \geq 2n.$$

Therefore by the uniqueness of paths in trees we see

$$d(v_{i,j}, v_{i,j+qr^n}) = d(v_{i+1,\lceil j/r \rceil}, v_{i+1,\lceil j/r \rceil + qr^{n-1}}) + 2 \geq 2(n+1). \qquad \square$$

To find $CE_k(T_{r,l})$ we will find a set of vertices $V' \subset V$ such that the distance between any two vertices in V' is at least k, therefore finding a lower bound $|V'|$ for the number of components in a failure state for $T_{r,l}$. We will then show that you can make $T_{r,l}$ a failure state by removing $|V'|$ edges.

The following lemma produces a set V' of vertices such that the distance between any two vertices in V' is at least k.

Lemma 3.6. *Let* $k \in \mathbb{Z}^+$. *Suppose* $T_{r,l} = (V, E)$ *and* $V' \subseteq V$ *such that*

$$V' = \left\{ v_{yk+1, 1+zr^{\lfloor (k-1)/2 \rfloor}} : 0 \leq y \leq \left\lfloor \frac{l}{k} \right\rfloor, \ 0 \leq z \leq \left\lceil \frac{r^{l-yk}}{r^{\lfloor (k-1)/2 \rfloor}} \right\rceil - 1 \right\}.$$

Then for all distinct $u, v \in V'$,

$$d(u, v) \geq k.$$

Proof. Assume $u, v \in V'$. Consider the following two cases:

Case 1: Assume u and v are distinct vertices in the same level of $T_{r,l}$. Thus there exist some integers i, a, and b such that $u = u_{i,1+ar^{\lfloor (k-1)/2 \rfloor}}$ and $v = v_{i,1+(a+b)r^{\lfloor (k-1)/2 \rfloor}}$. Then, by Lemma 3.5,

$$
\begin{aligned}
d(u, v) &= d(u_{i,1+ar^{\lfloor (k-1)/2 \rfloor}}, v_{i,1+(a+b)r^{\lfloor (k-1)/2 \rfloor}}) \\
&= d(u_{i,1+ar^{\lfloor (k-1)/2 \rfloor}}, v_{i,1+ar^{\lfloor (k-1)/2 \rfloor} + br^{(\lfloor (k-1)/2 \rfloor + 1) - 1}}) \\
&\geq 2\left(\left\lfloor \tfrac{1}{2}(k-1) \right\rfloor + 1 \right) \geq k.
\end{aligned}
$$

Case 2: Assume $u = u_{i,j}$ and $v = v_{i',j'}$ for some $i \neq i'$. Since $u, v \in V'$, we know $|i - i'| \geq k$. Therefore $d(u, v) \geq k$. $\qquad \square$

Now that we know the distance between any two vertices in V' is at least k, we need to find $|V'|$.

Lemma 3.7. *Suppose* $T_{r,l} = (V, E)$ *and* $V' \subseteq V$ *such that*

$$V' = \left\{ v_{yk+1, 1+zr^{\lfloor (k-1)/2 \rfloor}} : 0 \leq y \leq \left\lfloor \frac{l}{k} \right\rfloor, \ 0 \leq z \leq \left\lceil \frac{r^{l-yk}}{r^{\lfloor (k-1)/2 \rfloor}} \right\rceil - 1 \right\}.$$

Then,

$$|V'| = \begin{cases} 1, & l \leq \lfloor \frac{1}{2}(k-1) \rfloor, \\ \dfrac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \dfrac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})} + 1, & nk \leq l \leq nk + \lfloor \frac{1}{2}(k-1) \rfloor, \\ \dfrac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \dfrac{1-(r^{-k})^{\lfloor l/k \rfloor+1}}{1-(r^{-k})}, & else, \end{cases}$$

where n is a positive integer.

Proof. Summing over all possible choices for y and z we see

$$|V'| = \sum_{y=0}^{\lfloor l/k \rfloor} \sum_{z=0}^{\lceil R_y \rceil - 1} 1,$$

where $R_y = r^{l-yk}/r^{\lfloor (k-1)/2 \rfloor}$. Consider the following three cases:

Case 1: If $l \leq \lfloor \frac{1}{2}(k-1) \rfloor$, then $\lfloor l/k \rfloor = 0$ which implies y can only be zero. Thus

$$\lceil R_y \rceil - 1 = \lceil R_0 \rceil - 1 = 0.$$

Therefore

$$|V'| = \sum_{y=0}^{0} \sum_{z=0}^{0} 1 = 1.$$

Case 2: Assume there exists a positive integer n such that $nk \leq l \leq nk + \lfloor \frac{1}{2}(k-1) \rfloor$.

If $y = n$, then $\lceil R_y \rceil = \lceil R_n \rceil = 1$ since $0 \leq l - nk \leq \lfloor \frac{1}{2}(k-1) \rfloor$.

If $y < n$, then $y + 1 \leq n$, which implies $k(y+1) \leq kn \leq l$. Therefore $k \leq l - yk$, which implies

$$\lceil R_y \rceil = R_y.$$

Since $\lfloor l/k \rfloor = n$,

$$|V'| = \sum_{y=0}^{n} \sum_{z=0}^{\lceil R_y \rceil - 1} 1 = \sum_{y=0}^{n-1} \sum_{z=0}^{R_y - 1} 1 + \sum_{z=0}^{\lceil R_n \rceil - 1} 1$$

$$= \left(\sum_{y=0}^{n-1} R_y \right) + 1 = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})} + 1.$$

Case 3: Assume $nk + \lfloor \frac{1}{2}(k-1) \rfloor + 1 \leq l \leq (n+1)k - 1$ for some nonnegative integer n.

Note that $\lfloor \frac{1}{2}(k-1) \rfloor \leq l - nk$. Then for all $y \leq n$,

$$\lceil R_y \rceil = R_y.$$

Since $\lfloor l/k \rfloor = n$,

$$|V'| = \sum_{y=0}^{n} \sum_{z=0}^{\lceil R_y \rceil - 1} 1 = \sum_{y=0}^{n} \sum_{z=0}^{R_y - 1} 1$$

$$= \sum_{y=0}^{n} R_y = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1 - (r^{-k})^{\lfloor l/k \rfloor + 1}}{1 - (r^{-k})}. \qquad \square$$

We have now constructed a set of vertices which must be in separate components in order for $T_{r,l}$ to be a failure state (Lemma 3.6) and calculated the size of this vertex set (Lemma 3.7). We will now construct a set of edges that, when deleted, ensure these vertices are in different components. The idea is not to create perfect r-ary subtrees as we might expect. Instead we allow a perfect r-ary subtree but allow its root vertex to have a path up $T_{r,l}$ until the maximum diameter allowed is achieved. This propagates up the tree so that we do not have to remove entire rows of edges very often. This "saves" edges from being deleted by creating failure components which are larger than a perfect r-ary tree of diameter $k - 1$.

Lemma 3.8. *Fix r, l, and k and suppose $T_{r,l} = (V, E)$. For each integer $0 \le m \le \lceil l/k \rceil - 1$ define the sets*

$$A_m = \left\{ (v_{i,j}, v_{i+1,\lceil j/r \rceil}) \in E : mk + \left\lfloor \tfrac{1}{2}(k-1) \right\rfloor + 1 \le i \le (m+1)k - 1, \right.$$
$$\left. j \not\equiv 1 \bmod r \right\},$$

and

$$B_m = \left\{ (v_{(m+1)k,j}, v_{(m+1)k+1,\lceil j/r \rceil}) \in E : 1 \le j \le r^{l+1-(m+1)k} \right\}.$$

Then

$$|A_m| = r^{l+1}(r^{-mk - \lfloor (k-1)/2 \rfloor - 1} - r^{-(m+1)k}) \quad and \quad |B_m| = r^{l+1-(m+1)k}.$$

Proof. Fix $0 \le m \le \lceil l/k \rceil - 1$. First notice the number of edges of the form $(v_{i,a}, v_{i+1,\lceil a/r \rceil})$ is the number of vertices in level i, which is r^{l+1-i}.

Now consider A_m. The total number of edges of the form $(v_{i,a}, v_{i+1,\lceil a/r \rceil})$ is r^{l+1-i}, and of these, $r^{l+1-(i+1)}$ are of the form $(v_{i,j}, v_{i+1,\lceil j/r \rceil})$, where $j \equiv 1 \bmod r$. Thus

$$|A_m| = \sum_{i=mk+\lfloor (k-1)/2 \rfloor + 1}^{(m+1)k-1} r^{l+1-i} - r^{l+1-(i+1)}$$

$$= r^{l+1}(r^{-mk - \lfloor (k-1)/2 \rfloor - 1} - r^{-(m+1)k}).$$

Next consider B_m. The set B_m contains all edges of the form $(v_{(m+1)k,j}, v_{(m+1)k+1,\lceil j/r \rceil})$. Thus

$$|B_m| = r^{l+1-(m+1)k}. \qquad \square$$

Now we are ready to use A_m and B_m to find $\mathrm{CE}_k(T_{r,l})$.

Theorem 3.9. *If r, l, and k are positive integers, then*

$$\mathrm{CE}_k(T_{r,l}) = \begin{cases} 0, & l \leq \lfloor \frac{1}{2}(k-1) \rfloor, \\[2mm] \dfrac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \dfrac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})}, & nk \leq l \leq nk + \lfloor \frac{1}{2}(k-1) \rfloor, \\[2mm] \dfrac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \dfrac{1-(r^{-k})^{\lfloor l/k \rfloor + 1}}{1-(r^{-k})} - 1, & \text{else,} \end{cases}$$

where n is a positive integer.

Proof. Fix r, l, and k. Let $T_{r,l} = (V, E)$. There are three cases to consider:

Case 1: Assume $l \leq \lfloor \frac{1}{2}(k-1) \rfloor$.

Notice that the diameter of $T_{r,l}$ is $2l$. If $l \leq \lfloor \frac{1}{2}(k-1) \rfloor$, then $2l \leq 2\lfloor \frac{1}{2}(k-1) \rfloor < k$, and therefore $T_{r,l}$ is already a failure state. Hence, $\mathrm{CE}_k(T_{r,l}) = 0$.

For the following two cases, consider $V' \subseteq V$ as defined in Lemma 3.6. As shown in Lemma 3.6, $d(u, v) \geq k$ for all $u, v \in V'$. Therefore, to produce a failure state, no two vertices in V' can be in the same component. Since every edge cut in a tree produces one new component, there must be at least $|V'| - 1$ edge cuts to ensure no two vertices in V' are connected. Hence $\mathrm{CE}_k(T_{r,l}) \geq |V'| - 1$.

Case 2: Assume $nk \leq l \leq nk + \lfloor \frac{1}{2}(k-1) \rfloor$ for some positive integer n.

By Lemma 3.7,

$$|V'| - 1 = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})}.$$

Hence,

$$\mathrm{CE}_k(T_{r,l}) \geq \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})}.$$

For each integer $0 \leq m \leq \lfloor l/k \rfloor - 1$, define A_m and B_m as in Lemma 3.8.

Let $E' = \bigcup_{m=0}^{\lfloor l/k \rfloor - 1}(A_m \cup B_m)$. We will show that $G - E'$ is a failure state. Assume by way of contradiction that $G - E'$ is not a failure state. Thus there exists a path of length k in $G - E'$.

Case 2a: Assume there exists a path in $G - E'$ from a vertex in level i to a vertex in level $i + k$. Let $P = v_{i,j_0}, v_{i+1,j_1}, v_{i+2,j_2}, \ldots, 1, v_{i+k,j_k}$ be such a path of length k in $G - E'$, where $j_z = \lceil j_{z-1}/r \rceil$ and $(m-1)k < i \leq mk$ for some $2 \leq m \leq n-1$. Then, $mk < i + k \leq (m+1)k$ and $i \leq mk < i + k$.

Since $i \leq mk < i + k$, there exist a vertex of the form $v_{mk,j_{mk-i}} \in P$ and a vertex of the form $v_{mk+1,j_{mk-i+1}} \in P$ which are adjacent. However, $(v_{mk,j_{mk-i}},$

$v_{mk+1,j_{mk-i+1}}) \in B_{m-1}$. Consequently, $(v_{mk,j_{mk-i}}, v_{mk+1,j_{mk-i+1}}) \notin G - E'$, so P is not a path in $G - E'$.

Case 2b: Let $P = v_{i_0,j_0}, v_{i_1,j_1}, \ldots, v_{i_k,j_k}$ be the path of length k in $G - E'$. By the definition of B_m, we know there do not exist any edges in $G - E'$ joining $v_{(m-1)k,j}$ and $v_{(m-1)k+1,\lceil j/r \rceil}$ for any integer m where $2 \leq m \leq n$. Therefore we can assume there exists an integer $2 \leq m \leq n$ such that for all $0 \leq p \leq k$, we have $(m-1)k + 1 \leq i_p \leq mk$. In other words, all the vertices of path P fall between level $(m-1)k + 1$ and level mk inclusively.

Since there are only k distinct levels between level $(m-1)k + 1$ and level mk and P has $k + 1$ vertices, this implies there exists a subpath of P of the form $v_{a,b}, v_{a+1,c}, v_{a,b'}$, where $(c-1)r + 1 \leq b$, $b' \leq cr$, and $b \neq b'$. Since P is of length k, we can assume without loss of generality that $d(v_{i_0,j_0}, v_{a+1,c}) \geq \lfloor \frac{1}{2}(k-1) \rfloor + 1$. This implies that $i_0 + \lfloor \frac{1}{2}(k-1) \rfloor + 1 \leq a + 1$.

Since $(m-1)k + 1 \leq i_0 \leq mk$, we see

$$(m-1)k + \left\lfloor \tfrac{1}{2}(k-1) \right\rfloor + 1 + 1 \leq i_0 + \left\lfloor \tfrac{1}{2}(k-1) \right\rfloor + 1 \leq a + 1,$$

which implies

$$(m-1)k + \left\lfloor \tfrac{1}{2}(k-1) \right\rfloor + 1 \leq a.$$

Also, since $a \leq mk - 1$, we can see

$$(m-1)k + \left\lfloor \tfrac{1}{2}(k-1) \right\rfloor + 1 \leq a \leq mk - 1.$$

Since $(c-1)r + 1 \leq b$, $b' \leq cr$ and $b \neq b'$, we know $b \not\equiv 1 \bmod r$ or $b' \not\equiv 1 \bmod r$. Consequently, $(v_{a,b}, v_{a+1,c}) \in A_{m-1}$ or $(v_{a,b'}, v_{a+1,c}) \in A_{m-1}$, or both are in A_{m-1}. In either case, path P is not a path in $G - E'$ since it contains an edge in A_{m-1}. Hence, $G - E'$ is a failure state.

By Lemma 3.8,

$$|E'| = \sum_{m=0}^{n-1} (|A_m| + |B_m|)$$

$$= \sum_{m=0}^{n-1} (r^{l+1}(r^{-mk-\lfloor (k-1)/2 \rfloor - 1} - r^{-(m+1)k}) + r^{l+1-(m+1)k})$$

$$= \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1 - (r^{-k})^{\lfloor l/k \rfloor}}{1 - (r^{-k})}.$$

Therefore, since $G - E'$ is a failure state, we see

$$\mathrm{CE}_k(T_{r,l}) \leq \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1 - (r^{-k})^{\lfloor l/k \rfloor}}{1 - (r^{-k})}.$$

Since

$$\mathrm{CE}_k(T_{r,l}) \geq |V'| - 1 = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})},$$

we see

$$\mathrm{CE}_k(T_{r,l}) = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})}.$$

Case 3: Assume $nk + \lfloor \frac{1}{2}(k-1) \rfloor + 1 \leq l \leq (n+1)k - 1$ for some positive integer n. By Lemma 3.7,

$$|V'| - 1 = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor + 1}}{1-(r^{-k})} - 1.$$

Let $A_{n^*} = \{(v_{i,j}, v_{i+1,\lceil j/r \rceil}) : nk + \lfloor \frac{1}{2}(k-1) \rfloor + 1 \leq i \leq l, \ j \not\equiv 1 \bmod r\}$. Then,

$$|A_{n^*}| = \sum_{p=nk+\lfloor (k-1)/2 \rfloor + 1}^{l} r^{l+1-p} - r^{l+1-(p+1)} = r^{l+1}(r^{-nk-\lfloor (k-1)/2 \rfloor - 1} - r^{-l-1}).$$

Let $E' = \bigcup_{m=0}^{\lfloor l/k \rfloor - 1} (A_m \cup B_m) \cup A_{n^*}$. We will show that $T_{r,l} - E'$ is a failure state.

First, notice $G \subset T_{r,(n+1)k}$. Let $T_{r,(n+1)k} = (V^*, E^*)$. Let $E'' \subseteq E^*$ such that

$$E'' = \bigcup_{m=0}^{\lfloor (n+1)k/k \rfloor - 1} (A_m \cup B_m) = \bigcup_{0}^{\lfloor l/k \rfloor - 1} (A_m \cup B_m) \cup A_n \cup B_n.$$

As shown above in Case 2, $T_{r,(n+1)k} - E''$ is a failure state.

Note that

$$A_n = \{(v_{i,j}, v_{i+1,\lceil j/r \rceil}) : nk + \lfloor \tfrac{1}{2}(k-1) \rfloor + 1 \leq i \leq (n+1)k - 1, \ j \not\equiv 1 \bmod r\}.$$

Then, since $l \leq (n+1)k - 1$, we know $A_{n^*} \subseteq A_n$. Hence $E' \subseteq E''$ and $T_{r,l} - E' \subseteq T_{r,(n+1)k} - E''$. If there exists a path of length k in $T_{r,l} - E'$, then there must also exist a path of length k in $T_{r,(n+1)k} - E''$. However, $T_{r,(n+1)k} - E''$ is a failure state and therefore has no paths of length k. Therefore $T_{r,l} - E'$ has no paths of length k and is a failure state.

Thus,

$$|E'| = \sum_{m=0}^{n-1} (|A_m| + |B_m|) + |A_n^*|$$

$$= \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor}}{1-(r^{-k})} + r^{l+1}(r^{-nk-\lfloor (k-1)/2 \rfloor - 1} - r^{-l-1})$$

$$= \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1-(r^{-k})^{\lfloor l/k \rfloor + 1}}{1-(r^{-k})} - 1.$$

Therefore,

$$\mathrm{CE}_k(T_{r,l}) \leq \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1 - (r^{-k})^{\lfloor l/k \rfloor + 1}}{1 - (r^{-k})} - 1,$$

which implies

$$\mathrm{CE}_k(T_{r,l}) = \frac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \frac{1 - (r^{-k})^{\lfloor l/k \rfloor + 1}}{1 - (r^{-k})} - 1.$$

Combining all three of these cases, we see that

$$\mathrm{CE}_k(T_{r,l}) = \begin{cases} 0, & l \leq \lfloor \frac{1}{2}(k-1) \rfloor, \\[2mm] \dfrac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \dfrac{1 - (r^{-k})^{\lfloor l/k \rfloor}}{1 - (r^{-k})}, & nk \leq l \leq nk + \lfloor \frac{1}{2}(k-1) \rfloor, \\[3mm] \dfrac{r^l}{r^{\lfloor (k-1)/2 \rfloor}} \cdot \dfrac{1 - (r^{-k})^{\lfloor l/k \rfloor + 1}}{1 - (r^{-k})} - 1, & \text{else,} \end{cases}$$

where n is a positive integer.   □

3B2. *General trees.* Although finding a solution for general trees is too difficult, the general principles for perfect r-ary trees will still hold for general trees. Since each edge removal creates a new component, we need to remove edges that create components of diameter less than k which have as large an order as possible. Some bounds could easily be created based on minimum and maximum degree. Other special trees including caterpillar graphs, lobster graphs, and binary trees could be computed using the techniques outlined for the perfect r-ary tree.

References

[Boesch et al. 1998] F. Boesch, D. Gross, and C. Suffel, "Component order connectivity", *Congr. Numer.* **131** (1998), 145–155. MR Zbl

[Boesch et al. 1999] F. Boesch, D. Gross, and C. Suffel, "Component order connectivity: a graph invariant related to operating component reliability", pp. 109–116 in *Combinatorics, graph theory, and algorithms, I* (Kalamazoo, MI, 1996), edited by Y. Alavi et al., New Issues Press, Kalamazoo, MI, 1999. MR

[Boesch et al. 2006] F. Boesch, D. Gross, L. W. Kazmierczak, C. Suffel, and A. Suhartomo, "Component order edge connectivity: an introduction", *Congr. Numer.* **178** (2006), 7–14. MR Zbl

[Boesch et al. 2007] F. Boesch, D. Gross, L. W. Kazmierczak, C. Suffel, and A. Suhartomo, "Bounds for the component order edge connectivity", *Congr. Numer.* **185** (2007), 159–171. MR Zbl

[Boesch et al. 2009] F. Boesch, A. Satyanarayana, and C. Suffel, "A survey of some network reliability analysis and synthesis results", *Networks* **54**:2 (2009), 99–107. MR Zbl

[Harary 1983] F. Harary, "Conditional connectivity", *Networks* **13**:3 (1983), 347–357. MR Zbl

[West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR Zbl

shank@math.moravian.edu　　　*Mathematics and Computer Science, Moravian College, Bethlehem, PA, United States*

stawb01@moravian.edu　　　*Mathematics and Computer Science, Moravian College, Bethlehem, PA, United States*

msp

Time stopping for Tsirelson's norm

Kevin Beanland, Noah Duncan and Michael Holt

(Communicated by Stephan Garcia)

Tsirelson's norm $\|\cdot\|_T$ on c_{00} is defined as the limit of an increasing sequence of norms $(\|\cdot\|_n)_{n=1}^{\infty}$. For each $n \in \mathbb{N}$ let $j(n)$ be the smallest integer satisfying $\|x\|_{j(n)} = \|x\|_T$ for all x with max supp $x = n$. We show that $j(n)$ is $O(n^{1/2})$. This is an improvement of the upper bound of $O(n)$ given by P. Casazza and T. Shura in their 1989 monograph on Tsirelson's space.

In 1974 B. Tsirelson [1974] constructed a remarkable reflexive Banach space not containing an isomorphic copy of ℓ_p for any $1 < p < \infty$. T. Figiel and W. B. Johnson [1974] gave an analytic description of the dual Tsirelson's space that was subsequently used to discover many new types of Banach spaces and was very influential in solving many old problems in the isomorphic theory of Banach spaces. A monograph of P. Casazza and T. Shura [1989] contains a detailed analysis of many structural properties of Tsirelson's space and played a critical role in the developments in the mid-1990s. In the last chapter in that book, the authors present FORTRAN code that computes the Tsirelson norm of finite length vectors. In the discussion of this code they state several problems and lines of research that to our knowledge are still open or unexplored. The authors of the current paper became interested in these questions since they relate to the well-known open problem of whether Tsirelson's space is arbitrarily distortable and the "polymath" problem [Gowers 2009], which asks whether every "explicitly defined" Banach space must contain ℓ_p or c_0. Our main result is the first nontrivial step toward finding the computational time for computing the Tsireslon's norm. We should note that although Casazza and Shura's book was written almost 30 years ago, there are still many problems and constructions related to Tsireslon's space that are currently attracting attention. For example, the reader should consult the papers [Argyros et al. 2013; Argyros and Motakis 2014; 2016; Khanaki 2016; Ojeda-Aristizabal 2013; Tan 2012] and the aforementioned blog post of W. T. Gowers.

The dual of Tsirelson's space T is the completion of c_{00}, the space of all eventually zero scalar sequences, with respect to a norm $\|\cdot\|_T$. This norm is defined as the supremum of an increasing sequence of recursively defined norms $(\|\cdot\|_n)_{n=1}^\infty$. We recall all precise definitions in the next section. Casazza and Shura introduced the following *time-stopping* function.

Definition 1. For n a positive integer, let $j(n)$ be the smallest nonnegative integer such that for all $x \in c_{00}$ with max supp $x \le n$ we have $\|x\|_{j(n)} = \|x\|_T$.

In [Casazza and Shura 1989, Problem 2(a)], the authors ask for a "reasonably tight" upper bound for the quantity $j(n)$ and offer the upper bound $\lfloor \frac{1}{2}n \rfloor$ as a starting point. Our main theorem is the following improvement on this upper bound.

Theorem A. *For each $n \in \mathbb{N}$ we have $j(n) \le \lfloor 2\sqrt{n}+4 \rfloor$. That is, $j(n)$ is $O(n^{1/2})$.*

In a forthcoming paper we provide a lower bound on the order of $\log_2(n)$. The upper bound on $j(n)$ determines the computation time of the vector of length n. Indeed it is shown by Casazza and Shura that the computational time it takes to go from the n norm to the $n+1$ norm is the same for every n. Therefore if t is that computational time, our theorem shows that the computation time required to calculate the norm of a vector of length n is bounded above by $t\sqrt{n}$.

1. Main result

Let (e_i) and (e_i^*) both denote the standard unit vectors in c_{00}. For $E \subset \mathbb{N}$ and $x = \sum_{i=1}^\infty a_i e_i \in c_{00}$ let $Ex = \sum_{i \in E} a_i e_i$. If E, F are subsets of $\mathbb{N}$ we write $E < F$ if max $E < \min F$. A set $E \subset \mathbb{N}$ is in $\mathcal{S}_1$ if $\min E \ge |E|$ (the cardinality of E). If $\sum_{i=1}^\infty a_i e_i \in c_{00}$ then supp $x = \{i : a_i \neq 0\}$. For $n \in \mathbb{N}$ we say that a sequence $(E_i)_{i=1}^n$ of subsets of $\mathbb{N}$ is called admissible if $E_1 < E_2 < \cdots < E_n$ and $(\min E_i)_{i=1}^n \in \mathcal{S}_1$. We define the norm of Tsirelson's space by defining a certain subset of c_{00} to be the norming functionals for the space. The set W_T is the union of the following subsets of c_{00}. A sequence $(f_i)_{i=1}^d \subset c_{00}$ is called admissible if $(\text{supp } f_i)_{i=1}^d$ is admissible. Let $W_0 = \{\pm e_i^* : i \in \mathbb{N}\}$ and for $k \ge 0$ let

$$W_{k+1} = W_k \cup \left\{ \frac{1}{2} \sum_{i=1}^d Ef_i : d \in \mathbb{N}, (f_i)_{i=1}^d \subset W_k \text{ is admissible}, E \subset \mathbb{N} \right\}.$$

Then $W_T = \bigcup_{k=1}^\infty W_k$.

The intermediate norms are defined by $\|x\|_n = \sup\{f(x) : f \in W_n\}$. Here $f(x)$ is the usual inner product of f with x. Tsirelson's norm is defined by

$$\|x\| = \max_n \|x\|_n = \sup\{f(x) : x \in W_T\}.$$

Tsirelson's space is the completion of c_{00} with respect to the above norm, which satisfies the following implicit equation for $x \in c_{00}$:

$$\|x\| = \|x\|_\infty \vee \sup\left\{\frac{1}{2}\sum_{i=1}^{n}\|E_i x\| : n \in \mathbb{N}, (E_i)_{i=1}^{n} \text{ is admissible}\right\}. \qquad (1)$$

The following remarks follow from the definition of W_T.

Remark 1.1. Let $f \in W_T$. Then $f \in W_T \setminus W_n$ if and only if there is a $k \in \mathbb{N}$ with $0 < |f(e_k)| \le 1/2^{n+1}$.

Remark 1.2. If $f \in W_T$ then either $f = \pm e_i^*$ for some $i \in \mathbb{N}$ or $f \in W_n \setminus W_0$ and there is an admissible sequence $(f_i)_{i=1}^{d} \subset W_{n-1}$ with $f = \frac{1}{2}\sum_{i=1}^{d} f_i$. In particular, if $f \in W_T \setminus W_0$ then $|f(e_k)| \le \frac{1}{2}$ for all $k \in \mathbb{N}$.

Based on the above remark it is easy to see that each functional has a decomposition into a "tree" of functionals. The functionals in the tree are naturally enumerated by tuples in $\mathbb{N}$. Let $\mathbb{N}^{<\mathbb{N}} = \bigcup_{n=1}^{\infty} \mathbb{N}^n \cup \{\varnothing\}$. For $\sigma \in \mathbb{N}^{<\mathbb{N}}$, if $\sigma = (\sigma(1), \dots, \sigma(k))$, we set $|\sigma| = k$.

Definition 2 (tree index set and decomposition). For each $f \in W_T$ there is a set $\mathcal{T}_f \subseteq \mathbb{N}^{<\mathbb{N}} \cup \{\varnothing\}$ called the *tree index set* and a collection of functionals $(f_\alpha)_{\alpha \in \mathcal{T}_f} \subset W_T$ called a *tree decomposition* of f satisfying:

(1) $\varnothing \in \mathcal{T}_f$ and $f_\varnothing = f$.

(2) $\sigma \in \mathcal{T}_f$ is called a terminal node if $\sigma \frown 1 \notin \mathcal{T}_f$. A node $\sigma \in \mathcal{T}$ is a terminal if and only if $f_\sigma = \pm e_i^*$ for some $i \in \mathbb{N}$.

(3) If $\sigma \in \mathcal{T}_f$ is not a terminal node, then

$$f_\sigma = \frac{1}{2}\sum_{\{k:\sigma \frown k \in \mathcal{T}_f\}} f_{\sigma \frown k},$$

where $\{k : \sigma \frown k \in \mathcal{T}_f\} = \{1, \dots, d_\sigma\}$ for some $d_\sigma \in \mathbb{N}$. Moreover $(f_{\sigma \frown k})_{k=1}^{d_\sigma}$ is admissible.

If $\sigma = (n_1, n_2, \dots, n_k) \in \mathcal{T}_f$ then $\beta = (n_1, n_2, \dots, n_{k-1})$ is the immediate predecessor of σ and σ is an immediate successor of β. To set notation let $E_\sigma = \text{supp } f_\sigma$ for each $\sigma \in \mathcal{T}_f$.

The fact that each $f \in W_T$ has a (not necessarily unique) tree index set $\mathcal{T}_f$ and decomposition follows from the definition of an arbitrary $f \in W_T$.

Lemma 1.3. *Let $f \in W_T$. Then $f \in W_n$ if and only if there is a tree decomposition $(f_\alpha)_{\alpha \in \mathcal{T}_f} \subset W_T$ of f such that $|\sigma| \le n$ for all $\sigma \in \mathcal{T}_f$.*

Proof. Let $f \in W_n$. If $f \in W_0$ then any index set contains only the empty set. Thus we assume $f = \frac{1}{2} \sum_{i=1}^{d} f_i$, where $(f_i)_{i=1}^{d}$ is an admissible block sequence in W_{n-1}. Let $\mathcal{T}_{f_i}$ be the tree index set of f_i for each $i \in \{1, \ldots, d\}$ such that $|\sigma| \leq n - 1$ for each $\sigma \in \bigcup_{i=1}^{d} \mathcal{T}_{f_i}$. The tree index set of f is defined by

$$\mathcal{T}_f = \left\{ i \frown \sigma : i \in \{1, \ldots, d\}, \ \sigma \in \bigcup_{i=1}^{d} \mathcal{T}_{f_i} \right\} \cup \{\varnothing\}.$$

Alternatively, we proceed by induction. The base case is trivial and so assume the claim for some $n - 1 \geq 0$. We will establish the claim for n. Suppose there is a tree decomposition $(f_\alpha)_{\alpha \in \mathcal{T}_f} \subset W_T$ of some $f \in W_T$ so that $\sigma \leq n$ for all $\sigma \in \mathcal{T}_f$. Let $d \in \mathbb{N}$ so that $\{k : (k) \in \mathcal{T}_f\} = \{1, \ldots, d\}$. For each $1 \leq i \leq d$, set $\mathcal{T}_{f_{(i)}} = \{\sigma : i \frown \sigma \in \mathcal{T}_f\}$ and let $\{g_\sigma = f_{i \frown \sigma} : \sigma \in \mathcal{T}_{f_{(i)}}\}$ be a tree decomposition for $f_{(i)}$. Then for each $\sigma \in \mathcal{T}_{f_{(i)}}$ with $i \in \{1, \ldots, d\}$ we have $|\sigma| \leq n - 1$. Thus, $f_{(i)} \in W_{n-1}$ and $f \in W_n$ as desired. $\qquad \square$

The following is a simple but critical definition for our purposes. For a given $x \in c_{00}$ there may be many functionals in W_T that norm x. The support of some of these functionals may not even be a subset of the support of x, while other norming functionals may have supports disjoint from one another. Our goal is to prove an upper bound on $j(n)$ by minimizing the maximum node length of a tree decomposition for a functional that norms an arbitrary x with max supp $x \leq n$. In order to minimize this quantity, we discard the parts of a functional that are not required to norm a given vector. To this end, we define for each $x \in c_{00}$ a minimal set for x and a functional that *minimally norms* x. We can then restrict our attention to counting the maximum node length of a tree decomposition for a minimally norming functional for a given x.

Definition 3. Let $x \in c_{00}$. Then a set $E \subset \mathbb{N}$ is minimal for x if $\|Ex\| = \|x\|$ and for each $E' \subsetneq E$, we have $\|E'x\| < \|x\|$.

Let us note that minimal sets need not be unique.

Lemma 1.4. *Suppose $f \in W_T$ norms $x \in c_{00}$ and supp $f \subset$ supp x. Then for all $\alpha \in \mathcal{T}_f$, we have f_σ norms $E_\sigma x$.*

Proof. Assume, via contradiction, we can find a minimal length node $\sigma \in \mathcal{T}_f$ so that $f_\sigma(E_\sigma x) < \|E_\sigma x\|$. By assumption $\sigma \neq \varnothing$ (recall that $E_\varnothing = $ supp f). Find the unique predecessor $\beta \in \mathcal{T}_f$ of σ. Let $i_0 \in \{1, \ldots, d_\beta\}$ so that $\sigma = \beta \frown i_0$. Then $f_{\beta \frown i}(E_{\beta \frown i}x) \leq \|E_{\beta \frown i}x\|$ for $i \neq i_0$ and $f_{\beta \frown i_0}(E_{\beta \frown i_0}x) < \|E_{\beta \frown i_0}x\|$; however, $f_\beta(E_\beta x) = \|E_\beta x\|$ by the minimality of σ. This leads to the following contradiction:

$$\|E_\beta x\| = f_\beta(E_\beta x) = \frac{1}{2} \sum_{i=1}^{d_\beta} f_{\beta \frown i}(E_{\beta \frown i}x) < \frac{1}{2} \sum_{i=1}^{d_\beta} \|E_{\beta \frown i}x\| \leq \|E_\beta x\|.$$

The last inequality follows from the implicit equation (1) for the norm, noting that $(E_{\beta \frown i})_{i=1}^{d_\beta}$ is admissible. □

Definition 4. Let $x \in c_{00}$. We say that $f \in W_T$ *minimally norms* x if supp $f = E$ is minimal for x and $f(x) = \|x\|$.

Note that if f minimally norms x then supp $f \subset$ supp x. Also, if E is a minimal set for x there is a $f \in W_T$ that minimally norms x with supp $f = E$.

Lemma 1.5. *Let $x \in c_{00}$ and suppose that f minimally norms x. Then for each $\sigma \in \mathcal{T}_f$, we have E_σ is a minimal set for $E_\sigma x$ and f_σ minimally norms $E_\sigma x$.*

Proof. Using Lemma 1.4 we know that $f_\sigma(E_\sigma x) = \|E_\sigma x\|$ for each $\sigma \in \mathcal{T}_f$. Find a minimal length node $\sigma \in \mathcal{T}_f$ so that E_σ is not a minimal set for $E_\sigma f$. Again it follows from the hypothesis that $\sigma \neq \varnothing$. Let β be the immediate predecessor of σ and $i_0 \in \{1, \ldots, d_\beta\}$ with $\sigma = \beta \frown i_0$. Using our assumption, we can find a $E'_{\beta \frown i_0} \subsetneq E_{\beta \frown i_0}$ with $\|E'_{\beta \frown i_0} x\| = \|E_{\beta \frown i_0} x\|$. Let

$$E'_\beta = \left(\bigcup_{i=1,\, i \neq i_0}^{d_\beta} E_{\beta \frown i} \right) \cup E'_{\beta \frown i_0} \subsetneq E_\beta.$$

We can now show that $\|E_\beta x\| \leq \|E'_\beta x\|$ as follows:

$$\|E_\beta x\| = f_\beta(E_\beta x) = \frac{1}{2} \sum_{i=1}^{d_\beta} f_{\beta \frown i}(E_{\beta \frown i} x) = \frac{1}{2} \sum_{i=1}^{d_\beta} \|E_{\beta \frown i} x\|$$

$$= \frac{1}{2}\left(\sum_{i=1}^{i_0-1} \|E_{\beta \frown i} x\| + \|E'_{\beta \frown i_0} x\| + \sum_{i=i_0+1}^{d_\beta} \|E_{\beta \frown i} x\| \right)$$

$$\leq \|E'_\beta x\|. \tag{2}$$

The last inequality uses that

$$(E_{\beta \frown 1}, E_{\beta \frown 2}, \ldots, E_{\beta \frown (i_0-1)}, E'_{\beta \frown i_0}, E_{\beta \frown (i_0+1)}, \ldots, E_{\beta \frown d_\beta})$$

is admissible. This contradicts the minimality of σ.

Therefore for each $\sigma \in \mathcal{T}_f$, we have E_σ is a minimal set for $E_\sigma x$. The fact that f_σ minimally norms $E_\sigma x$ follows from Lemma 1.4. □

Lemma 1.6. *Let $x \in c_{00}$ and suppose $f \in W_T$ minimally norms x and supp $f \in \mathcal{S}_1$. Then $f \in W_1$.*

Proof. If $f \in W_T \setminus W_1$ then there is a $k \in$ supp f with $0 < |f(e_k)| \leq \frac{1}{4}$. However since supp $f \in \mathcal{S}_1$, we know that $g = \frac{1}{2} \sum_{i \in \text{supp } f} \text{sign}(e_i^*(x)) e_i^* \in W_1$. But $g(x) > f(x) = \|x\|$. This is a contradiction. □

Lemma 1.7. *Let $x \in c_{00}$ and suppose $f \in W_T$ minimally norms x. Suppose further that $f \in W_T \setminus W_1$. If $f = \frac{1}{2} \sum_{i=1}^{d} f_i$, where $(f_i)_{i=1}^{d}$ is admissible, then* $\min \operatorname{supp} f = d$.

Proof. By definition, $d \leq \min \operatorname{supp} f =: m$, and so it suffices to show that equality holds. Suppose towards a contradiction that $d < m$. Our goal is to build a functional $f' \in W_T$ so that $f'(x) > f(x)$. This will contradict the assumption that $f(x) = \|x\|$. Since $f \in W_T \setminus W_1$ there is some $i_0 \in \{1, \ldots, d\}$ with $f_{i_0} \notin W_0$. By appealing to Remark 1.1 and Lemma 1.6 we may assume that the supp f_{i_0} has more than one element. Let $k_0 = \min \operatorname{supp} f_{i_0}$. Then $0 < |f(e_{k_0})| \leq \frac{1}{4}$. In particular

$$f(e_{k_0}) = \operatorname{sign} e_{k_0}^*(x) \frac{1}{2^n} \quad \text{for some } n > 1.$$

Set $f_{i_0}^1 = \operatorname{sign}(e_{k_0}^*(x))e_{k_0}^*$ and $f_{i_0}^2 = f_{i_0}|_{[k_0+1,\infty)}$. Since $d < m$ and $f_{i_0}^1, f_{i_0}^2 \in W_T$ are successive with $\operatorname{supp} f_{i_0}^1 \cup \operatorname{supp} f_{i_0}^2 = \operatorname{supp} f_{i_0}$, we have that

$$f' = \frac{1}{2}\left(\sum_{i=1}^{i_0-1} f_i + f_{i_0}^1 + f_{i_0}^2 + \sum_{i=i_0+1}^{d} f_i \right) \in W_T.$$

The above holds since $(f_1, \ldots, f_{i_0-1}, f_{i_0}^1, f_{i_0}^2, f_{i_0+1}, \ldots, f_d)$ is admissible. However, f' has the same coordinates as f except at the k_0 position where

$$f'(e_{k_0}) = \operatorname{sign} e_{k_0}^*(x) \frac{1}{2} \quad \text{and} \quad f(e_{k_0}) = \operatorname{sign} e_{k_0}^*(x) \frac{1}{2^n} \text{ for } n > 1.$$

Since $k_0 \in \operatorname{supp} x$ we have that $f'(x) - f(x) = \left(\frac{1}{2} - \frac{1}{2^n}\right)|e_{k_0}^*(x)| > 0$. Therefore $f'(x) > f(x)$. Since $f' \in W_T$ and $f(x) = \|x\|$, this is the desired contradiction. $\square$

The next lemma is the critical observation that allows us to prove the main theorem. It is essentially an averaging argument that allows us to restrict our attention to a smaller collection of norming functions therefore enabling an upper bound on $j(n)$.

Lemma 1.8. *Let $x \in c_{00}$ and suppose $f \in W_T$ minimally norms x, with $\operatorname{supp} f = E$. Suppose that $(f_\sigma)_{\sigma \in \mathcal{T}_f}$ is a tree decomposition for f. If $\sigma \in \mathcal{T}_f$ with $|\sigma| \geq 2$ so that there is a k with $\sigma(k-2) = \sigma(k-1) = 1$ then $|\sigma| \leq k$.*

The above lemma roughly states that for any vector x there is a norming functional f for x so that its tree decomposition has very few consecutive 1s. In particular, if a node σ has two consecutive 1s then they must be contained in the last three coordinates of the node.

Proof. Fix $x \in c_{00}$ and fix $f \in W_T$ that minimally norms x, with $\operatorname{supp} f = E$. Let $(f_\sigma)_{\sigma \in \mathcal{T}_f}$ be a tree decomposition for f and fix $\sigma \in \mathcal{T}_f$ with $|\sigma| \geq 2$ so that there is a k with $\sigma(k-2) = \sigma(k-1) = 1$. For convenience let $g = f_{\sigma|_{k-3}}$ ($\sigma|_{k-3}$ is σ restricted to its first $k - 3$ coordinates). In the case that $k = 3$, we have $\sigma|_{k-3} = \varnothing$. Let

$g_i = f_{\sigma|_{k-3}\frown i}$ for $i \in \{1, \ldots, d_{\sigma|_{k-3}}\}$ and $g_{(i,j)} = f_{\sigma|_{k-3}\frown(i,j)}$ for $i \in \{1, \ldots, d_{\sigma|_{k-3}}\}$ and $j \in \{1, \ldots, d_{\sigma|_{k-3}\frown i}\}$.

Set $m = \min \operatorname{supp} g$. Suppose first that $\max \operatorname{supp} g_{(1,1)} < 2m - 1$. This implies that $\operatorname{supp} g_{(1,1)} \in \mathcal{S}_1$. Since f minimally norms x and $g_{(1,1)}$ is a functional in the tree decomposition of f, Lemma 1.4 yields that $g_{(1,1)}$ norms $E_{g_{(1,1)}}x$ (where $\operatorname{supp} g_{(1,1)} = E_{g_{(1,1)}}$). Applying Lemma 1.6 for $g_{(1,1)}$ and $E_{g_{(1,1)}}x$, we conclude that $g_{(1,1)} \in W_1$. Therefore $|\sigma| \le k$. Therefore we may consider the case $\max \operatorname{supp} g_{(1,1)} \ge 2m - 1$. Set $d = d_{\sigma|_{k-3}}$ and $r = d_{\sigma|_{k-3}\frown 1}$. Note that $d, r \le m$. Define

$$h_1 := \tfrac{1}{2}(g_{(1,2)} + \cdots + g_{(1,d)} + g_{(2)} + \cdots + g_{(r)}),$$
$$h_2 := \tfrac{1}{2}(g_{(1,1)} + g_{(2)} + \cdots + g_{(r)}).$$

Observe that $h_2 \in W_T$ and since $d + r - 1 \le 2m \le \min \operatorname{supp} g_{1,2}$ we have $h_1 \in W_T$. Again, by Lemma 1.4, we have that $g(E_g x) = \|E_g x\|$, where $E_g = \operatorname{supp} g$. Since f minimally norms x, Lemma 1.5 yields that g minimally norms $E_g x$. Therefore since $\operatorname{supp} h_1 \subsetneq \operatorname{supp} g$ and $\operatorname{supp} h_2 \subsetneq \operatorname{supp} g$, we know that $h_1(x) < g(x)$ and $h_2(x) < g(x)$. This implies that

$$(g_{(1,2)} + \cdots + g_{(1,d)})(x) < g_{(1)}(x) \quad \text{and} \quad g_{(1,1)}(x) < g_{(1)}(x).$$

However, by definition, $g_{(1)}(x) = \tfrac{1}{2}(g_{(1,1)}(x)) + \tfrac{1}{2}(g_{(1,2)} + \cdots + g_{(1,d)})(x)$. This is a contradiction. Therefore the case $\max \operatorname{supp} g_{(1,1)} \ge 2m - 1$ is not possible. □

Corollary 1.9. *For each $x \in c_{00}$ there is an $f \in W_T$ that minimally norms x with a tree decomposition $(f_\sigma)_{\sigma \in \mathcal{T}_f}$ such that for each $\sigma \in \mathcal{T}_f$ either σ has no consecutive 1s, the third-to-last and second-to-last coordinates are 1 or the final two coordinates are 1.*

Let $\mathcal{T}_a$ be the set of all $\sigma \in \mathbb{N}^{<\mathbb{N}}$ that satisfy the conclusion of Corollary 1.9. For example $\sigma = (1, 1, 2, 3) \notin \mathcal{T}_a$ but $(2, 3, 1, 1, 2) \in \mathcal{T}_a$.

Lemma 1.10. *For each $\sigma \in \mathcal{T}_a$ with $|\sigma| = k \ge 3$,*

$$\left(\sum_{i=1}^{k-1} [k-i][\sigma(i) - 1] \right) \ge \left(\frac{k-3}{2} \right)^2. \tag{3}$$

Proof. Suppose $|\sigma| = k = 2d + 1$ for $d \in \mathbb{N}$. By replacing $\sigma(1)$ by 1 and $\sigma(2)$ by 2, the quantity on the left-hand side of (3) does not increase. This new element is still in $\mathcal{T}_a$. Continuing in this manner, we see that the above is minimized by $\sigma = (1, 2, 1, 2, \ldots, 2, 1, 1, 1)$ — that is, $d - 2$ many 2s. If $k = 2d$ we may do the same procedure described previously to see that the quantity is minimized by $\sigma = (1, 2, 1, 2, \ldots, 2, 1, 1)$. Plugging these in the above yields $\sum_{i=2}^{d} 2i = d^2 - 1$ in the odd case and $\sum_{i=1}^{d-1} 2i + 1 = d^2 - d$. Both of these quantities are larger than $\tfrac{1}{4}(k-3)^2$, as desired. □

The next corollary follows from combining Corollary 1.9 and Lemma 1.10.

Corollary 1.11. *For each $x \in c_{00}$ there is an $f \in W_T$ that minimally norms x having a tree decomposition $(f_\sigma)_{\sigma \in \mathcal{T}_f}$ such that*

$$\min\left\{ \sum_{i=1}^{|\sigma|-1} [|\sigma|-i][\sigma(i)-1] : \sigma \in \mathcal{T}_f \right\} \geq \left(\frac{|\sigma|-3}{2} \right)^2. \tag{4}$$

We need one more technical lemma before proceeding to the proof of the main theorem.

Lemma 1.12. *Suppose that $f \in W_T$ and $\max \operatorname{supp} f \leq n$. Suppose further that f minimally norms x for some $x \in c_{00}$. Then for $\sigma \in \mathcal{T}_f$ with $f_\sigma \in W_T \setminus W_1$ we have*

$$|\operatorname{supp} f_\sigma| \leq n - \left(\sum_{i=1}^{|\sigma|-1} [|\sigma|-i][\sigma(i)-1] \right). \tag{5}$$

We postpone the proof of Lemma 1.12 to the end of paper. We now recall Theorem A and give its proof.

Theorem 1.13. *For $n \in \mathbb{N}$ and $x \in c_{00}$ with $\max \operatorname{supp} x = n$ we have $\|x\|_{\lfloor 2\sqrt{n}+4 \rfloor} = \|x\|$. That is, $j(n)$ is $O(n^{1/2})$.*

Proof. Let $x \in c_{00}$ with $\max \operatorname{supp} x = n$. Suppose further that f minimally norms x. Suppose that $\sigma \in \mathcal{T}_f$ with $|\sigma| \geq \lfloor 2\sqrt{n}+3 \rfloor$. If $f_\sigma \in W_T \setminus W_1$ then by combining Lemma 1.12 and Corollary 1.11, we know that

$$|\operatorname{supp} f_\sigma| \leq n - \left(\sum_{i=1}^{|\sigma|-1} [|\sigma|-i][\sigma(i)-1] \right) \leq n - \left(\frac{|\sigma|-3}{2} \right)^2$$

$$\leq n - \left(\frac{2\sqrt{n}+3-3}{2} \right)^2 = 0. \tag{6}$$

Therefore no such σ exists. Thus if $|\sigma| \geq \lfloor 2\sqrt{n}+3 \rfloor$ we have $f_\sigma \in W_1$. Therefore $\max\{|\sigma| : \sigma \in \mathcal{T}_f\} \leq \lfloor 2\sqrt{n}+4 \rfloor$, which implies that $f \in W_{\lfloor 2\sqrt{n}+4 \rfloor}$. Since $f(x) = \|x\|$, this is the desired result. $\qquad \square$

We conclude by proving Lemma 1.12.

Proof. Let $x \in c_{00}$ and suppose $f \in W_T$ minimally norms x with $\max \operatorname{supp} f \leq n$. Let $\sigma \in \mathcal{T}_f$ with $f_\sigma \in W_T \setminus W_1$. Set $\ell = \min \operatorname{supp} f$. We will prove the following inequality, which is stronger than the desired estimate:

$$|\operatorname{supp} f_\sigma| \leq n - (|\sigma|+1)(\ell-1) - \left(\sum_{i=1}^{|\sigma|-1} [|\sigma|-i][\sigma(i)-1] \right). \tag{7}$$

First we need the inequality

$$\min \text{supp } f_\sigma \geq \ell + s(\sigma) - |\sigma|. \tag{8}$$

Here $s(\sigma) = \sum_{i=1}^{k} \sigma(i)$ for $|\sigma| = k$. To prove (8), we let $|\sigma| = k$ and use induction on k. Let $\sigma|_{k-1} = (n_1, \ldots, n_{k-1})$ if $\sigma = (n_1, \ldots, n_{k-1}, n_k)$. In the base case of $|\sigma| = 1$, we know $\min \text{supp } f_\sigma \geq \ell + s(\sigma) - 1$, since there are at least $(s(\sigma) - 1)$-many values from ℓ to f_σ's beginning index (worst case being that all prior functionals are in W_0). Now we assume $\min \text{supp } f_\sigma \geq \ell + s(\sigma) - |\sigma|$ for some $|\sigma| = k \in \mathbb{N}$ and show the same inequality holds for $|\sigma| = k + 1$:

$$\begin{aligned}
\min \text{supp } f_\sigma &\geq \min \text{supp } f_{\sigma|_k} + \sigma(k+1) - 1 \\
&\geq \ell + s(\sigma|_k) - |\sigma|_k| + \sigma(k+1) - 1 \\
&= \ell + s(\sigma) - (k+1).
\end{aligned}$$

The first inequality relies on the fact that f_σ can have the same minimum support value as $f_{\sigma|_k}$ if $\sigma(k+1) = 1$. The second inequality above follows from the inductive hypothesis. The lone equality above follows from the facts that $s(\sigma|_k) + \sigma(k+1) = s(\sigma)$ and $|\sigma|_k| = k$. Thus, (8) holds.

The proof of the inequality (8) begins with the observation that for all σ with $|\sigma| = k$ we have

$$|\text{supp } f_\sigma| \leq |\text{supp } f_{\sigma|_{k-1}}| - \#\{\text{immediate successor of } \sigma|_{k-1}\} + 1.$$

The fact that f minimally norms x combined with Lemma 1.7 implies that for each $\sigma \in \mathcal{T}$ with $f_\sigma \in W_T \setminus W_1$, the number of immediate successor nodes of $\sigma|_{k-1}$ equals $\min \text{supp } f_{\sigma|_{k-1}}$. Therefore in this case

$$|\text{supp } f_\sigma| \leq |\text{supp } f_{\sigma|_{k-1}}| - (\min \text{supp } f_{\sigma|_{k-1}}) + 1.$$

Now let $|\sigma| = k$ and use induction on k. It follows from the induction hypothesis and rearranging terms that

$$\begin{aligned}
|\text{supp } f_\sigma| &\leq |\text{supp } f_{\sigma|_{k-1}}| - (\min \text{supp } f_{\sigma|_{k-1}}) + 1 \\
&\leq n - k(\ell - 1) - \left(\sum_{i=1}^{k-2} [(k-1) - i][\sigma(i) - 1] \right) - \ell - s(\sigma|_{k-1}) + (k-1) + 1 \\
&\leq n - (k+1)(\ell - 1) - \left(\sum_{i=1}^{k-2} [(k-1) - i][\sigma(i) - 1] \right) - \sum_{i=1}^{k-1} [\sigma(i) - 1] \\
&= n - (k+1)(\ell - 1) - \sum_{i=1}^{k-1} [k - i][\sigma(i) - 1].
\end{aligned}$$

This is the desired estimate. $\square$

Acknowledgment

The authors acknowledge Professor Ben Grannan at Furman University for his help with computations that led the authors to prove Lemma 1.8.

References

[Argyros and Motakis 2014]   S. A. Argyros and P. Motakis, "A reflexive hereditarily indecomposable space with the hereditary invariant subspace property", *Proc. Lond. Math. Soc.* (3) **108**:6 (2014), 1381–1416. MR Zbl

[Argyros and Motakis 2016]   S. A. Argyros and P. Motakis, "A dual method of constructing hereditarily indecomposable Banach spaces", *Positivity* **20**:3 (2016), 625–662. MR Zbl

[Argyros et al. 2013]   S. A. Argyros, K. Beanland, and P. Motakis, "Strictly singular operators in Tsirelson like spaces", *Illinois J. Math.* **57**:4 (2013), 1173–1217. MR Zbl

[Casazza and Shura 1989]   P. G. Casazza and T. J. Shura, *Tsirelson's space*, Lecture Notes in Mathematics **1363**, Springer, Berlin, 1989. MR Zbl

[Figiel and Johnson 1974]   T. Figiel and W. B. Johnson, "A uniformly convex Banach space which contains no ℓ_p", *Compos. Math.* **29** (1974), 179–190. MR Zbl

[Gowers 2009]   W. T. Gowers, "Must an 'explicitly defined' Banach space contain c_0 or ℓ_P?", blog entry, 2009, available at http://tinyurl.com/gowersmust.

[Khanaki 2016]   K. Khanaki, "$\aleph_0$-categorical spaces contain ℓ_p or c_0", preprint, 2016. arXiv

[Ojeda-Aristizabal 2013]   D. Ojeda-Aristizabal, "A norm for Tsirelson's Banach space", *Extracta Math.* **28**:2 (2013), 235–245. MR Zbl

[Tan 2012]   D.-N. Tan, "Isometries of the unit spheres of the Tsirelson space T and the modified Tsirelson space T_M", *Houston J. Math.* **38**:2 (2012), 571–581. MR Zbl

[Tsirelson 1974]   B. S. Tsirelson, "Not every Banach space contains an imbedding of ℓ_p or c_0", *Funkcional. Anal. i Priložen.* **8**:2 (1974), 57–60. In Russian; translated in *Funct. Anal. Appl.* **8**:2 (1974), 138–141. MR

beanlandk@wlu.edu   *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

naduncan16@gmail.com   *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

holtm11493@gmail.com   *Department of Mathematics, Washington and Lee University, Lexington, VA, United States*

Enumeration of stacks of spheres

Lauren Endicott, Russell May and Sienna Shacklette

(Communicated by Glenn Hurlbert)

As a three-dimensional generalization of fountains of coins, we analyze stacks of spheres and enumerate two particular classes, so-called "pyramidal" stacks and "Dominican" stacks. Using the machinery of generating functions, we obtain exact formulas for these types of stacks in terms of the sizes of their bases.

1. Introduction

Odlyzko and Wilf [1988] analyzed fountains of coins. An (n, k) fountain is an arrangement of n coins into rows so that the bottom row consists of k contiguous coins and each coin in higher rows sits on two coins in the row beneath it. Figure 1 shows a $(25, 12)$ fountain. Two fountains are different if in any row and any position in the row, one fountain has a coin, but the other does not. Their goal was to enumerate the numbers $f_{n,k}$ of (n, k) fountains, and their main result was that the bivariate generating function $F(x, y) = \sum_{n,k} f_{n,k} x^n y^k$ was the continued fraction

$$F(x, y) = \cfrac{1}{1 - \cfrac{xy}{1 - \cfrac{x^2 y}{1 - \cfrac{x^3 y}{\ddots}}}}. \tag{1}$$

If the fountains are enumerated only by the number of coins in the bottom row, $g_k = \sum_n f_{n,k}$, then the generating function $G(y) = \sum_k g_k y^k = F(1, y)$ is much simpler. It is straightforward from (1) that the generating function satisfies $G(y) - yG^2(y) = 1$, and so the g_k are the Catalan numbers. Wilf [2006, Example 2.12] also considered a restricted class of "block" fountains having the property that each row must be a contiguous block of coins. If b_k is the number of block fountains with k coins in the bottom row, then $B(y) = \sum_k b_k y^k$ turns out to be $(1 - 2x)/(1 - 3x + x^2)$, which is the generating function for the Fibonacci numbers with odd indices.
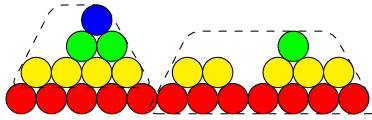
Figure 1. A (25, 12) fountain of coins with subfountains around the first missing coin in the second row.
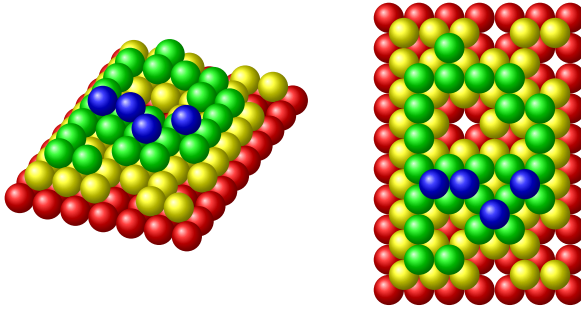

Figure 2. A (148, 7, 10) stack of spheres from oblique and top views.

Figure 3. Models for stacks: Pyramid of the Sun at Teotihuaca, Mexico (credit: [Lneuw 2006]), and the flag of the Dominican Republic.

As a three-dimensional variant of fountains of coins, we consider stacks of spheres. An (ℓ, m, n) stack of spheres is an arrangement of ℓ spheres into levels so that the bottom level consists of spheres in an $m \times n$ rectangular grid and each sphere in higher levels sits on four spheres in the level beneath it. Figure 2 shows a (148, 7, 10) stack of spheres. In grocery stores, fruits like oranges and cantaloupes are often arranged into such stacks.

Our goal is to analyze two classes of stacks, *pyramidal* and *Dominican*, and obtain generating functions and exact formulas for the number of stacks in terms of the sizes of their bases. Pyramidal stacks have the property that every level consists of a single rectangular grid of spheres, much like the Pyramid of the Sun at Teotihuaca, shown in Figure 3. Dominican stacks are closer in spirit to general stacks. Their inductive definition closely resembles the color scheme of solid regions and stripes in the flag of the Dominican Republic, also shown in Figure 3.

2. Basics of generating functions

Generating functions are a bridge between the discrete world of combinatorics and the continuous world of calculus and complex analysis. Wilf [2006] embraces a five-step method for describing sequences with generating functions:

(1) Find a recurrence relation for the sequence.

(2) Define the generating function.

(3) Convert the recurrence relation to a relation about the generating function.

(4) Solve for the generating function.

(5) Extract or approximate the coefficients of the generating function.

We use a handful of well-known generating functions, based on the geometric series and its derivatives:

$$\sum_{n \geq 0} 1 x^n = \frac{1}{1-x}, \tag{2a}$$

$$\sum_{n \geq 0} n x^n = \frac{x}{(1-x)^2}, \tag{2b}$$

$$\sum_{n \geq 0} \binom{n+k}{n} x^n = \frac{1}{(1-x)^{1+k}}. \tag{2c}$$

We also use the product rule: if $f(x) = \sum_{n \geq 0} a_n x^n$ and $g(x) = \sum_{n \geq 0} b_n x^n$, then $f(x) \cdot g(x) = \sum_{n \geq 0} c_n x^n$, where $c_n = \sum_{k=0}^{n} a_k \cdot b_{n-k}$. An important special case of the product rule is the partial sum rule: $\sum_{n \geq 0} (a_0 + a_1 + \cdots + a_n) x^n = 1/(1-x) f(x)$. We also need the bivariate version of the product rule; namely if $f(x, y) = \sum_{m,n \geq 0} a_{m,n} x^m y^n$ and $g(x, y) = \sum_{m,n \geq 0} b_{m,n} x^m y^n$, then $f(x, y) \cdot g(x, y) = \sum_{m,n \geq 0} c_{m,n} x^m y^n$, where $c_{m,n} = \sum_{m',n'} a_{m',n'} \cdot b_{m-m',n-n'}$. Following standard notation, we define the bivariate coefficient extraction operator as

$$[x^m y^n] \sum_{m,n \geq 0} a_{m,n} x^m y^n = a_{m,n}.$$

3. Pyramidal stacks of spheres

One of the simplest types of fountains of coins is a block fountain, defined by the property that each row consists of a single contiguous block of coins. We define a corresponding three-dimensional variant, a *pyramidal stack of spheres*, to be a stack where each level consists of a single rectangular grid of spheres. An example of a pyramidal stack is depicted in Figure 4. Unlike arbitrary stacks, pyramidal stacks are constrained to have only a single spire. We would like to enumerate the pyramidal stacks by the size of their bases. For $m, n \geq 1$, let $p_{m,n}$
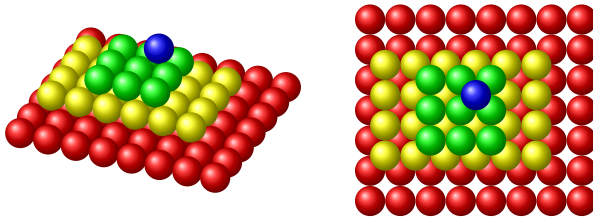
Figure 4. An 8×7 pyramidal stack of spheres from oblique and top views.

$p_{m,n}$	$n=1$	2	3	4	5	6	7
$m=1$	1	1	1	1	1	1	1
2	1	2	4	7	11	16	22
3	1	4	11	24	46	81	134
4	1	7	24	63	143	294	561
5	1	11	46	143	376	881	1894
6	1	16	81	294	881	2317	5534
7	1	22	134	561	1894	5534	14545

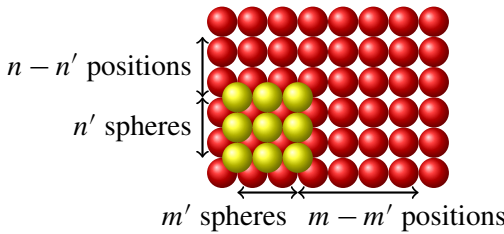**Table 1.** Values of $p_{m,n}$ for $m, n \leq 7$.

Figure 5. Possible positions of an $m' \times n'$ pyramid on top of an $m \times n$ base.

be the number of pyramidal stacks of spheres whose base consists of an $m \times n$ grid of spheres. For convenience, let $p_{m,0} = p_{0,n} = 0$ for all $m, n \geq 0$, and note that by symmetry $p_{m,n} = p_{n,m}$. Then define the bivariate generating function $P(x, y) = \sum_{m,n \geq 0} p_{m,n} x^m y^n$. By hand calculation and assistance from Maple, we computed $p_{m,n}$ for $m, n \leq 7$, shown in Table 1. A pyramidal stack with an $m \times n$ base can either contain nothing on the second level or support another pyramidal stack with an $m' \times n'$ base, where $1 \leq m' < m$ and $1 \leq n' < n$. If the second level is nonempty, it can be shifted horizontally to $m - m'$ positions and vertically to $n - n'$ positions to form different stacks, as shown in Figure 5. Therefore, we have the

recurrence relation for pyramidal stacks for $m, n \geq 1$, given as

$$p_{m,n} = 1 + \sum_{\substack{1 \leq m' \leq m-1 \\ 1 \leq n' \leq n-1}} (m - m')(n - n') p_{m',n'} = 1 + \sum_{\substack{0 \leq m' \leq m \\ 0 \leq n' \leq n}} (m - m')(n - n') p_{m',n'}.$$

The bounds on the sum can be extended from 0 to m and n since $p_{0,n} = p_{m,0} = 0$ and $(m - m') = (n - n') = 0$ when $m' = m$ and $n' = n$. Then, by use of the generating functions in (2a) and (2b) and the product rule we get

$$P(x, y) = \frac{x}{1 - x} \frac{y}{1 - y} + P(x, y) \frac{xy}{(1 - x)^2 (1 - y)^2}.$$

Solving this equation results in the rational generating function

$$P(x, y) = \frac{xy(1 - x)(1 - y)}{(1 - x)^2 (1 - y)^2 - xy}. \tag{3}$$

To obtain an exact expression for $p_{m,n}$, we first view $P(x, y)$ as a geometric series:

$$P(x, y) = \frac{\frac{xy}{(1-x)(1-y)}}{1 - \frac{xy}{(1-x)^2(1-y)^2}} = \sum_{\ell \geq 0} \frac{x^{\ell+1} y^{\ell+1}}{(1 - x)^{1+2\ell} (1 - y)^{1+2\ell}}$$

Then, using (2c),

$$\begin{aligned}
p_{m,n} &= [x^m y^n] P(x, y) \\
&= \sum_{\ell \geq 0} [x^{m-\ell-1} y^{n-\ell-1}] \frac{1}{(1 - x)^{1+2\ell} (1 - y)^{1+2\ell}} \\
&= \sum_{\ell \geq 0} \binom{m+\ell-1}{m-\ell-1} \binom{n+\ell-1}{n-\ell-1}.
\end{aligned}$$

This exact expression for $p_{m,n}$ is a sum with $\min(m - 1, n - 1)$ terms, a significant improvement over the recursion that requires $O(mn)$ computations. Also, note that g_m, the number of block fountains with m coins in the bottom row, is equivalent to the $(2m+1)$-th Fibonacci number, which is also expressible as the sum $\sum_{\ell} \binom{m+\ell-1}{m-\ell-1}$. Therefore, pyramidal stacks of spheres can be viewed as a direct generalization of block fountains of coins.

4. Dominican stacks

Pyramidal stacks, having only a single spire, form an extremely restricted class, just as block fountains are to general fountains of coins. We would like to analyze a more robust class that is closer in spirit to general stacks. Dominican stacks are a three-dimensional generalization of arbitrary two-dimensional fountains of coins.

In order to motivate the definition of Dominican stacks, let's review general fountains of coins. A fountain with m coins in the bottom row can be uniquely
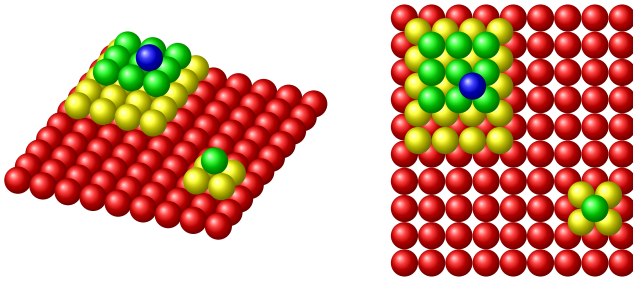
Figure 6. A 9×10 Dominican stack of spheres from oblique and top views.

decomposed into two subfountains by locating the first position in the second row, say at m', where a coin is missing. For instance, the second row of the fountain in Figure 1 has its first missing coin in the fifth position. Thus, a general fountain consists of the subfountain on the left with a base of m' coins, whose second row is full and so consists of an even smaller subfountain with a diminished base of $m' - 1$ coins, and a subfountain on the right with a base of $m - m'$ coins. So, the recurrence relation $g_m = \sum_{m'} g_{m'-1} \cdot g_{m-m'}$ holds for fountains of coins.

We make an analogous definition by induction for stacks of spheres. A *Dominican* stack of spheres is defined by the following cases:

Base case: A single level of spheres in a rectangular grid.

Inductive case: A multilevel stack of spheres with an $m \times n$ base built from smaller stacks, as follows. It is required that, when viewed from the top, there exist a (necessarily unique) column at position m' and row at position n' devoid of spheres so that the following conditions hold:

> *Bottom left*: Every position in the second level above positions $[1, \ldots, m'] \times [1, \ldots, n']$ has a sphere, and the stack with diminished base of size $(m' - 1) \times (n' - 1)$ from the second level and above is Dominican.
>
> *Top right*: The stack above positions $[m' + 1, \ldots, m] \times [n' + 1, \ldots, n]$ from the first level and above is Dominican.
>
> *Bottom right and top left*: The stacks above positions $[m'+1, \ldots, m] \times [1, \ldots, n']$ and $[1, \ldots, m'] \times [n' + 1, \ldots, n]$ consist solely of rectangular grids of spheres on the first level with nothing above.

An example of a Dominican stack is shown in Figure 6. Informally, a stack of spheres is called Dominican because of its resemblance to the flag of the Dominican Republic. As a visualization depicted in Figure 7, imagine placing a version of the Dominican Republic's flag under a stack of spheres and looking at the stack from above. On the second level, the white stripes of the flag appear in the row and

$d_{m,n}$	$n=1$	2	3	4	5	6	7
$m=1$	1	1	1	1	1	1	1
2	1	2	3	4	5	6	7
3	1	3	7	12	18	25	33
4	1	4	12	28	52	85	128
5	1	5	18	52	122	239	416
6	1	6	25	85	239	564	1147
7	1	7	33	128	416	1147	2723

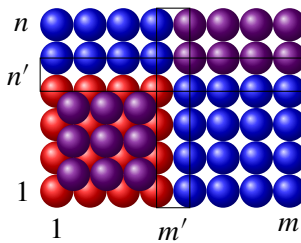**Table 2.** Values of $d_{m,n}$ for $m, n \leq 7$.

Figure 7. Indexing of a Dominican stack with regions of the bottom layer in red and blue and smaller Dominican stacks in purple.

column which are empty but the red region to the bottom-left is full; the substack above this red region is a smaller Dominican stack. The portion of the stack on the top-right above the other red region of the flag also forms a smaller Dominican stack. The portions of the stack above the blue regions are single layers of spheres. Note that if the first empty column ($m' = 1$ from the definition) in the second level forms the boundary, then the corresponding row must also be the first ($n' = 1$), and likewise for the converse. Therefore, the emblem of the flag where the white stripes cross may either be in position $[1, 1]$ of the second level or in a position $[m', n']$ with $2 \leq m' \leq m$ and $2 \leq n' \leq n$.

Let $d_{m,n}$ be the number of Dominican stacks with an $m \times n$ base. For convenience, let $d_{0,n} = d_{m,0} = 1$ for all $m, n \geq 0$, and note that by symmetry $d_{m,n} = d_{n,m}$. We define the bivariate generating function $D(x, y) = \sum_{m,n \geq 0} d_{m,n} x^n y^n$. By hand calculation and assistance from Maple, we computed $d_{m,n}$ for $m, n \leq 7$, shown in Table 2. In order to derive an expression for this generating function, we begin with the recurrence relation for the $d_{m,n}$ that follows immediately from the inductive definition of Dominican stacks. Equation (4b) is just a reindexing of (4a), and (4c) accounts for the inclusion of the terms $m' = 0$ or $n' = 0$ in the sum; see the

following:

$$d_{m+1,n+1} = d_{m,n} + \sum_{\substack{2 \le m' \le m+1 \\ 2 \le n' \le n+1}} d_{m'-1,n'-1}\, d_{m+1-m',n+1-n'} \quad (m, n \ge 0) \tag{4a}$$

$$= d_{m,n} + \sum_{\substack{1 \le m' \le m \\ 1 \le n' \le n}} d_{m',n'} d_{m-m',n-n'} \tag{4b}$$

$$= 2d_{m,n} + \sum_{\substack{0 \le m' \le m \\ 0 \le n' \le n}} d_{m',n'} d_{m-m',n-n'} - \sum_{0 \le m' \le m} d_{m',n} - \sum_{0, \le n' \le n} d_{m,n'}. \tag{4c}$$

Using the generating functions in (2a) and (2b), along with the product rule and the partial sum rule, we obtain the corresponding relation for $D(x, y)$, namely

$$D(x, y) - \frac{x}{1-x} - \frac{y}{1-y} - 1$$
$$= 2xy D(x, y) + xy D^2(x, y) - \frac{xy}{1-x} D(x, y) - \frac{xy}{1-y} D(x, y).$$

This relation is quadratic in $D(x, y)$ with coefficients that are polynomials in x, y. After some algebra, we get a ratio of polynomials with a radical for the generating function:

$$D(x, y)$$
$$= \frac{(1-xy)(1-x-y+2xy) - \sqrt{(1-xy)^2(1-x-y+2xy)^2 - 4xy(1-x)(1-y)(1-xy)}}{2xy(1-x)(1-y)}. \tag{5}$$

This generating function agrees with the values of $d_{m,n}$ from the recurrence relation. $D(x, y)$ can be viewed as a generalization of the generating function for the Catalan numbers. Unfortunately, asymptotic analysis even of rational bivariate generating functions is difficult, so analysis of this generating function with a radical will require further investigation.

5. Further problems

Stacking of spheres lends itself to many one-parameter combinatorial classes. For example, the class of pyramidal stacks can be restricted by the condition that each level forms a *square* grid of spheres. If s_n is the number of such pyramids with an $n \times n$ base, it follows quickly that the generating function $\sum s_n x^n$ is $x(1-x)^3/(1-4x+2x^2-x^3)$. Many geometrical variants with triangles, hexagons, etc., can be formed in this manner and result in single-variable generating functions.

Another restriction leading to single-variable generating functions is to fix the width of the base of a *general* stack. For each base width $m \ge 1$ and length $n \ge 1$, let $a_{m,n}$ be the number of general stacks with an $m \times n$ base and define

$A_m(x) = \sum_n a_{m,n} x^n$. It turns out that

$$A_1(x) = \frac{1}{1-x}, \quad A_2(x) = \frac{x}{1-2x}, \quad A_3(x) = \frac{x(1-x)}{1-5x+3x^2}.$$

The A_m for $m \geq 4$ are more difficult to compute, but would surely shed light on the general case.

However, the main problem of enumerating general stacks of spheres remains unsolved. While it is hoped that recursive methods similar to the pyramidal and Dominican cases will ultimately work out, it is entirely possible that altogether different machinery may be required to enumerate general stacks. The most direct reason that recursive methods may fail is that partitions of rectangles into collections of subrectangles are unwieldy. Additionally, since the generating function of a general fountain of coins (based on the total number of coins) in two dimensions is already significant in complexity as a continued fraction, one must expect the generating function of a general stack of spheres in three dimensions to be an order of difficulty harder.

A more tractable problem may be an asymptotic approximation of the coefficients of the generating functions in (3) and (5). Recent results in [Pemantle and Wilson 2013] may provide insight.

References

[Lneuw 2006] Lneuw, "Pyramid of the Sun Teotihuacán, Mexico, taken from the Pyramid of the Moon", photo, 2006, available at https://tinyurl.com/TeoSun. Public domain.

[Odlyzko and Wilf 1988] A. M. Odlyzko and H. S. Wilf, "The editor's corner: n coins in a fountain", *Amer. Math. Monthly* **95**:9 (1988), 840–843. MR Zbl

[Pemantle and Wilson 2013] R. Pemantle and M. C. Wilson, *Analytic combinatorics in several variables*, Cambridge Studies in Advanced Mathematics **140**, Cambridge University Press, 2013. MR Zbl

[Wilf 2006] H. S. Wilf, *generatingfunctionology*, 3rd ed., A K Peters, Wellesley, MA, 2006. MR Zbl

lkendicott@moreheadstate.edu   *The Craft Academy for Excellence in Science and Mathematics, Morehead State University, Morehead, KY 40351, United States*

r.may@moreheadstate.edu   *Department of Mathematics and Physics, Morehead State University, Morehead, KY 40351, United States*

slshacklette@moreheadstate.edu   *The Craft Academy for Excellence in Science and Mathematics, Morehead State University, Morehead, KY 40351, United States*

Rings isomorphic to their nontrivial subrings

Jacob Lojewski and Greg Oman

(Communicated by Scott T. Chapman)

Let G be a nontrivial group, and assume that $G \cong H$ for every nontrivial subgroup H of G. It is a simple matter to prove that $G \cong \mathbb{Z}$ or $G \cong \mathbb{Z}/\langle p \rangle$ for some prime p. In this note, we address the analogous (though harder) question for rings; that is, we find all nontrivial rings R for which $R \cong S$ for every nontrivial subring S of R.

1. Introduction

The notion of "same structure" is ubiquitous in mathematics. Indeed, the concept appears as early as high school geometry, where congruence of angles and similarity of triangles are studied. One then learns the analogous concept for groups in a first course on modern algebra, where two groups G and H have the same structure if there is a bijection $f \colon G \to H$ with the property that $f(xy) = f(x)f(y)$ for all $x, y \in G$. Such an f is called an *isomorphism* from G to H; if such an f exists, then we say that G and H are *isomorphic*, and write $G \cong H$. There exist many groups which are isomorphic to a proper subgroup. For example, the group $(\mathbb{Z}, +)$ is isomorphic to $(E, +)$, where E is the subgroup of $\mathbb{Z}$ consisting of the even integers. More generally, since every nontrivial subgroup of an infinite cyclic group is also infinite cyclic, and since every infinite cyclic group is isomorphic to $(\mathbb{Z}, +)$, it follows that the group $(\mathbb{Z}, +)$ is inordinately homogeneous in the sense that all nontrivial subgroups are isomorphic.

More generally, a mathematical structure $\mathfrak{M}$ is called κ-*homogeneous* (κ an infinite cardinal of size at most $|\mathfrak{M}|$) provided any two substructures of cardinality κ are isomorphic [Droste 1989; Oman 2009; 2011]. A related mathematical object called a *Jónsson group* is an infinite group G such that every proper subgroup of G has smaller cardinality than G; in this case, note that G is $|G|$-homogeneous. It is well known, see [Scott 1952], that the only abelian Jónsson groups are the quasicyclic groups $\mathbb{Z}(p^\infty)$, p a prime, which is isomorphic to the subgroup of the factor group $\mathbb{Q}/\mathbb{Z}$ consisting of those elements whose order is a power of p.

If one does not assume G to be abelian, then the situation becomes much more complicated. Saharon Shelah was the first to construct an example of a *Kurosh monster*, which is a group of size $\aleph_1$ in which all proper subgroups are countable. It is still an open problem to determine whether a Jónsson group of size $\aleph_\omega$ can be shown to exist in Zermelo–Fraenkel set theory with choice (ZFC); we refer the reader to the excellent survey [Coleman 1996] for more details.

Laffey [1974] characterized the countably infinite rings R for which every proper subring of R is finite. An infinite ring R with the property that every proper subring of R has smaller cardinality than R is called a *Jónsson ring*. It is known that any uncountable Jónsson ring is necessarily a noncommutative division ring. The existence of such a ring has yet to be established [Coleman 1996]. It is apparently a very difficult problem to classify all rings R for which $R \cong S$ for every subring S of size $|R|$, since doing so would automatically classify the Jónsson rings. In view of these results, we take a more modest approach in this paper and consider the problem of classifying those nontrivial rings R for which $R \cong S$ for every nontrivial subring S of R.

2. Results

We begin by fixing terminology. First, all rings will be assumed to be associative, but not necessarily commutative or unital. Indeed, commutativity of the rings studied in this paper can be deduced rather quickly (so it need not be assumed), and many important and well-studied classes of rings do not contain an identity. For example, Leavitt path algebras on graphs with infinitely many vertices *never* contain an identity; see [Abrams et al. 2017, Lemma 1.2.12(iv)]. If R is a ring, then a *subring* of R is a nonempty subset S of R which is closed under addition, multiplication, and negatives. It is important to note that in this article, we do *not* require a subring of a unital ring to contain an identity. For the purposes of this note, say that a ring R (respectively, group G) is *homogeneous* if R is nontrivial and $R \cong S$ for all nontrivial subrings S of R (respectively, if G is nontrivial and $G \cong H$ for every nontrivial subgroup H of G).

We begin our investigation by first classifying the homogeneous groups.

Lemma 1. *Let G be a group. Then G is homogeneous if and only if $G \cong \mathbb{Z}/\langle p \rangle$ for some prime p or $G \cong \mathbb{Z}$.*

Proof. Because (by Lagrange's theorem) $\mathbb{Z}/\langle p \rangle$ has no proper, nontrivial subgroups (that is, $\mathbb{Z}/\langle p \rangle$ is *simple*), we see that $\mathbb{Z}/\langle p \rangle$ is trivially homogeneous. As for the additive group $\mathbb{Z}$ of integers, if H is a nontrivial subgroup of $\mathbb{Z}$, then H is an infinite cyclic group; hence $H \cong \mathbb{Z}$. We deduce that $\mathbb{Z}$ is a homogeneous group.

Conversely, suppose that G is a homogeneous group. Let g be a nonidentity element of G. Then $G \cong \langle g \rangle$, and thus G is cyclic. If G is infinite, then $G \cong \mathbb{Z}$.

Thus suppose that G is finite. If H is a proper subgroup of G, then $|H| < |G|$; thus $H \ncong G$. As G is homogeneous, it follows that G is simple. It is well known that the only nontrivial simple abelian groups are the groups $\mathbb{Z}/\langle p \rangle$ where p is a prime. To keep the paper self-contained, we give the argument. We have already noted above that for a prime p, the group $\mathbb{Z}/\langle p \rangle$ is simple. Conversely, suppose that G is simple, and let $g \in G \backslash \{e\}$ be arbitrary. The simplicity of G implies that $G = \langle g \rangle$, and so G is cyclic. Because $\mathbb{Z}$ has proper, nontrivial subgroups, we deduce that G is a finite cyclic group, say of order $n > 1$. It remains to show that n is prime. If $n = rs$ for some integers r and s with $1 < r, s < n$, then $\langle g^r \rangle$ is a proper, nontrivial subgroup of G, contradicting that G is simple. This concludes the proof. $\square$

We arrive at the main result of this note, which classifies the homogeneous rings. As the reader will see, the argument we give to prove the ring version of Lemma 1 is more complicated than the argument just given above.

Theorem 1. *Let R be a ring. Then R is homogeneous if and only if one of the following holds*:

(i) $R \cong \mathbb{F}_p$, *where $\mathbb{F}_p$ is the field of p elements and p is a prime number*,

(ii) $R \cong \mathbb{Z}/\langle p \rangle$ *with trivial multiplication* (*that is, $xy = 0$ for all x and y*), *or*

(iii) $R \cong \mathbb{Z}$ *with trivial multiplication*.

Proof. Consider first the field $\mathbb{F}_p$, where p is prime. If S is a nontrivial subring of $\mathbb{F}_p$, then under addition, S is a nontrivial subgroup of $(\mathbb{F}_p, +)$. By Lagrange's theorem, $S = \mathbb{F}_p$, and thus $S \cong \mathbb{F}_p$ as rings. The same argument shows that $\mathbb{Z}/\langle p \rangle$ with trivial multiplication is homogeneous. As for (iii), suppose that S is a nontrivial subring of $\mathbb{Z}$ (with trivial multiplication). Then additively, S is a nontrivial subgroup of $(\mathbb{Z}, +)$. By Lemma 1, $(S, +) \cong (\mathbb{Z}, +)$; let $f \colon S \to \mathbb{Z}$ be an additive isomorphism. Because the multiplication on $\mathbb{Z}$ is trivial, it follows that f is also a ring isomorphism. We have verified that the rings in (i)–(iii) are homogeneous.

We now work toward establishing the converse. For $m \in \mathbb{Z}$, let $m\mathbb{Z}$ be the subring of $\mathbb{Z}$ consisting of all integer multiples of m. We claim that

$$\text{the ring } m\mathbb{Z} \text{ is not homogeneous for any } m \in \mathbb{Z}. \qquad (2\text{-}1)$$

If $m = 0$, then $m\mathbb{Z} = \{0\}$; thus is not homogeneous by definition. If $|m| = 1$, then observe that $m\mathbb{Z} = \mathbb{Z} \ncong 2\mathbb{Z}$ since the ring $\mathbb{Z}$ has an identity but the ring $2\mathbb{Z}$ does not. Now suppose that $|m| > 1$. Then $m\mathbb{Z}$ has a nonzero element α (namely m) such that $\alpha^2 = m\alpha$, yet the subring $m^2\mathbb{Z}$ does not possess such an element. To see this, suppose that $\beta \in m^2\mathbb{Z} \backslash \{0\}$ is such that $\beta^2 = m\beta$. We have $\beta = m^2 n$ for some $n \in \mathbb{Z} \backslash \{0\}$. Thus $m^4 n^2 = \beta^2 = m\beta = m(m^2 n)$. But then $mn = 1$, and m is a unit of $\mathbb{Z}$, which is impossible because $|m| > 1$. We conclude that $m\mathbb{Z} \ncong m^2\mathbb{Z}$. This completes the verification of (2-1).

Next, for a nonzero element r of a ring R, let

$$r\mathbb{Z}[r] := \{m_1 r + m_2 r^2 + \cdots + m_k r^k : k \in \mathbb{Z}^+, m_i \in \mathbb{Z}\}$$

be the subring of R generated by r. If $f : r\mathbb{Z}[r] \to R$ is a ring isomorphism, then one can see that $R = f(r)\mathbb{Z}[f(r)]$. Hence:

If R is a homogeneous ring, then $R = r\mathbb{Z}[r]$ for some $r \in R \backslash \{0\}$. Thus R is commutative. (2-2)

Now let D be a commutative domain with identity $1 \neq 0$, and let $D[X^2, X^3]$ be the ring generated by D, X^2, and X^3, where X is an indeterminate which commutes with the members of D. Consider the ideal $\langle X^2, X^3 \rangle$ of $D[X^2, X^3]$ generated by X^2 and X^3. We claim that

$$\langle X^2, X^3 \rangle \text{ is not a principal ideal of } D[X^2, X^3]. \quad (2\text{-}3)$$

Note first that

$$X \notin D[X^2, X^3], \quad (2\text{-}4)$$

lest X be a unit of $D[X]$. Suppose by way of contradiction that $\langle X^2, X^3 \rangle = \langle f(X) \rangle$ for some $f(X) \in D[X^2, X^3]$. Then $X^2 \mid f(X)$ and $f(X) \mid X^2$ in the ring $D[X]$. We deduce that $f(X) = uX^2$ for some unit $u \in D$. Because $f(X) \mid X^3$ in the ring $D[X^2, X^3]$, we have $uX^2 g(X) = X^3$ for some $g(X) \in D[X^2, X^3]$. But then $X = u \cdot g(X) \in D[X^2, X^3]$, contradicting (2-4). We have now established (2-3). Next, let $XD[X]$ be the subring of $D[X]$ consisting of all $f(X) \in D[X]$ for which $f(0) = 0$. We prove that

$$XD[X] \text{ is not homogeneous.} \quad (2\text{-}5)$$

Suppose otherwise, and let R be the subring of $XD[X]$ generated by X^2 and X^3. Then R is also homogeneous, and by (2-2), there is $f(X) \in R$ such that $R = f(X)\mathbb{Z}[f(X)]$. Next, let I be the ideal of $D[X^2, X^3]$ generated by R. Then it follows that $I = \langle X^2, X^3 \rangle = \langle f(X) \rangle$, and we have a contradiction to (2-3) above.

Finally, we are ready to classify the homogeneous rings. Toward this end, let R be an arbitrary homogeneous ring. We shall prove that one of (i)–(iii) holds. Suppose first that R possesses a nonzero nilpotent element α. Let $n > 1$ be least such that $\alpha^n = 0$. Setting $\beta := \alpha^{n-1}$, we have $\beta \neq 0$, yet $\beta^2 = 0$. Let $S := \{m\beta : m \in \mathbb{Z}\}$. One checks at once that S is a nonzero subring of R with trivial multiplication. Because R is homogeneous, $R \cong S$; hence R is a nontrivial ring with trivial multiplication. But then every subgroup of R is a subring of R. The homogeneity of R gives $H \cong K$ for any nontrivial subgroups H and K of $(R, +)$. Applying Lemma 1, we see that either (ii) or (iii) holds.

Thus we assume that

$$R \text{ is reduced; that is, } R \text{ has no nonzero nilpotent elements.} \quad (2\text{-}6)$$

Our next assertion is that

$$R \text{ has no nonzero zero divisors.} \tag{2-7}$$

Suppose by way of contradiction that $r_0 \in R \setminus \{0\}$ is a zero divisor. Let $T_1 := r_0 \mathbb{Z}[r_0]$ and $S_1 := \{r \in R : rT_1 = \{0\}\}$. We have seen that T_1 is a nonzero subring of R. As R is commutative by (2-2) and r_0 is a zero divisor, S_1 is a *nonzero* subring of R. Because R is reduced, it follows immediately that

$$S_1 \cap T_1 = \{0\}, \quad \text{and} \quad xy = 0 \quad \text{for all } x \in S_1 \text{ and } y \in T_1. \tag{2-8}$$

As R is homogeneous, $R \cong S_1$. We conclude that there exist nonzero subrings S_2 and T_2 of S_1 such that $S_2 \cap T_2 = \{0\}$ and $xy = 0$ for all $x \in S_2$ and $y \in T_2$. Continuing recursively and setting $S_0 := T_0 := R$, we obtain sequences $\{S_n : n \geq 0\}$ and $\{T_n : n \geq 0\}$ of nonzero subrings of R such that for every $n \geq 0$, S_{n+1} and T_{n+1} are nonzero subrings of S_n such that $S_{n+1} \cap T_{n+1} = \{0\}$ and $xy = 0$ for all $x \in S_{n+1}$ and $y \in T_{n+1}$. Next, we establish that for all positive integers k,

$$\begin{aligned} &\text{if } n_1, \ldots, n_k > 0 \text{ are distinct, and } t_1 + \cdots + t_k = 0 \text{ with } t_i \in T_{n_i}, \\ &\quad \text{then } t_i = 0 \text{ for } i = 1, \ldots, k. \end{aligned} \tag{2-9}$$

To prove this, we induct on k. Note that the base case of the induction is the assertion that if $t_1 = 0$ and $t_1 \in T_{n_1}$, then $t_1 = 0$, which is true. Suppose that the claim holds for some $k > 0$, and let $0 < n_1 < n_2 < \cdots < n_{k+1}$ and $t_1, \ldots, t_{k+1}$ be such that $t_1 + \cdots + t_{k+1} = 0$ with $t_i \in T_{n_i}$ for all $1 \leq i \leq k$. One checks that $t_2, \ldots, t_{k+1} \in S_{n_1}$; set $\alpha := t_2 + \cdots + t_{k+1}$. Then $t_1 + \alpha = 0$, $t_1 \in T_{n_1}$, and $\alpha \in S_{n_1}$. Since $S_{n_1} \cap T_{n_1} = \{0\}$, it follows that $t_1 = \alpha = 0$. Applying the inductive hypothesis, we see that $t_2 = \cdots = t_{k+1} = 0$, and (2-9) is verified. We further claim that

$$\text{if } 0 < n < m \text{ and } x \in T_n, y \in T_m, \text{ then } xy = 0. \tag{2-10}$$

This is straightforward: as above, $y \in S_n$, and the result follows. We deduce from (2-9), (2-10), and the homogeneity of R that R is isomorphic to the internal direct sum of the rings T_n, $n > 0$. More compactly,

$$R \cong \bigoplus_{n>0} T_n. \tag{2-11}$$

Thus $\bigoplus_{n>0} T_n$ is homogeneous. By (2-2), there is $(r_n) := r \in \bigoplus_{n>0} T_n$ such that $\bigoplus_{n>0} T_n = r\mathbb{Z}[r]$. Now, $r_i = 0$ for almost all i. Thus there is a k such that if $r_i \neq 0$, then $i \in \{1, \ldots, k\}$. But then for *every* $(\alpha_n) := \alpha \in r\mathbb{Z}[r]$, if $\alpha_i \neq 0$, then $i \in \{1, \ldots, k\}$. Since $\bigoplus_{n>0} T_n = r\mathbb{Z}[r]$, we deduce that the same is true of every member of $\bigoplus_{n>0} T_n$. But of course, this is absurd: recall that each T_i is a *nonzero* ring, so for every $k \in \mathbb{Z}^+$ there exists a sequence $(t_n : n \in \mathbb{N}) \in \bigoplus_{n>0} T_n$ such that $t_k \neq 0$. Finally, we have proven (2-7).

We pause to take inventory of what we have established thus far. By (2-2) and (2-7), R is a commutative domain, though we have not yet proven that R has a multiplicative identity. Let

$$K := \{a/b : a \in R, \ b \in R \setminus \{0\}\}$$

be the quotient field of R. It is well known that R embeds into K via the map $r \mapsto (rd)/d$, where $d \in R$ is some fixed nonzero element of R. We identity R with its image in K. Now let D be the subring of K generated by 1. Fix some nonzero $r \in R$. One checks at once that $rD[r]$ is a nonzero subring of R, whence

$$R \cong rD[r]. \tag{2-12}$$

The map $\varphi : XD[X] \to rD[r]$ defined by $\varphi(Xg(X)) := rg(r)$ is a surjective ring map. We apply (2-12) to conclude that $rD[r]$ is homogeneous. Therefore, (2-5) implies that the kernel of φ is nonzero. Choose a nonzero polynomial $Xf(X) := d_1 X + d_2 X^2 + \cdots + d_n X^n \in XD[X]$ of minimal degree n for which $rf(r) = 0$. We claim that

$$d_1 \neq 0. \tag{2-13}$$

If $n = 1$, this follows since $Xf(X) \neq 0$. Suppose now that $n > 1$. If $d_1 = 0$, then we have $d_2 r^2 + \cdots + d_n r^n = 0$. Recalling that R is a domain and $r \neq 0$, this equation reduces to $d_2 r + \cdots + d_n r^{n-1} = 0$, and this contradicts the minimality of n. So we have

$$d_1 r + d_2 r^2 + \cdots + d_n r^n = 0 \quad \text{and} \quad d_1 \neq 0. \tag{2-14}$$

Viewing the above equation in the quotient field K of R, we may divide through by r to get $d_1 + d_2 r + \cdots + d_n r^{n-1} = 0$. Solving the equation for d_1, we see that

$$d_1 \in R. \tag{2-15}$$

Recall that $d_1 \in D$, the ring generated by 1_K (the multiplicative identity of K). Thus $d_1 = m \cdot 1_K$ for some $m \in \mathbb{Z}$. Because K is a field, either $D \cong \mathbb{Z}$ or $D \cong \mathbb{Z}/\langle p \rangle$ for some prime p. In the former case, it follows from (2-13), (2-15), and the homogeneity of R that $R \cong m\mathbb{Z}$ for some $m \in \mathbb{Z}$. However, this is precluded by (2-1). We deduce that $D \cong \mathbb{F}_p$ for some prime p. But then by (2-15), we see that (up to isomorphism) $d_1 \in (\mathbb{F}_p \setminus \{0\}) \cap R$. Applying homogeneity a final time, we see that R is isomorphic to the ring generated by d_1. Thus, as $d_1 \neq 0$, we have $R \cong \mathbb{F}_p$, and the proof is complete. $\qquad\square$

We conclude the paper with the following corollary, which characterizes the fields of order p.

Corollary 1. *Let R be a ring with nontrivial multiplication. Then R is a field with p elements (p a prime) if and only if any two nontrivial subrings of R are isomorphic.*

Acknowledgment

We thank the referee for an extremely thorough reading of our paper and for offering numerous helpful suggestions.

References

[Abrams et al. 2017] G. Abrams, P. Ara, and M. Siles Molina, *Leavitt path algebras*, Lecture Notes in Mathematics **2191**, Springer, 2017. Zbl

[Coleman 1996] E. Coleman, "Jonsson groups, rings and algebras", *Irish Math. Soc. Bull.* 36 (1996), 34–45. MR Zbl

[Droste 1989] M. Droste, "k-homogeneous relations and tournaments", *Quart. J. Math. Oxford Ser.* (2) **40**:157 (1989), 1–11. MR Zbl

[Laffey 1974] T. J. Laffey, "Infinite rings with all proper subrings finite", *Amer. Math. Monthly* **81** (1974), 270–272. MR Zbl

[Oman 2009] G. Oman, "More results on congruent modules", *J. Pure Appl. Algebra* **213**:11 (2009), 2147–2155. MR Zbl

[Oman 2011] G. Oman, "On elementarily κ-homogeneous unary structures", *Forum Math.* **23**:4 (2011), 791–802. MR Zbl

[Scott 1952] W. R. Scott, "Groups and cardinal numbers", *Amer. J. Math.* **74** (1952), 187–197. MR Zbl

jlojewsk@uccs.edu   *Department of Mathematics, University of Colorado, Colorado Springs, CO, United States*

goman@uccs.edu   *Department of Mathematics, University of Colorado, Colorado Springs, CO, United States*

On generalized MacDonald codes

Padmapani Seneviratne and Lauren Melcher

(Communicated by Joshua Cooper)

We show that the generalized q-ary MacDonald codes $C_{n,u,t}(q)$ with parameters $\left[t\begin{bmatrix} n \\ 1 \end{bmatrix} - \begin{bmatrix} u \\ 1 \end{bmatrix}, n, tq^{n-1} - q^{u-1}\right]$ are two-weight codes with nonzero weights $w_1 = tq^{n-1}$, $w_2 = tq^{n-1} - q^{u-1}$ and determine the complete weight enumerator of these codes. This leads to a family of strongly regular graphs with parameters $\langle q^n, q^n - q^{n-u}, q^n - 2q^{n-u}, q^n - q^{n-u} \rangle$. Further, we show that the codes $C_{n,u,t}(q)$ satisfy the Griesmer bound and are self-orthogonal for $q = 2$.

1. Introduction

Two-weight codes are an interesting family of error-correcting codes. They are closely related to many other areas, including strongly regular graphs, partial geometries and finite projective spaces. The relationship between two-weight codes and projective sets was first studied by Delsarte [1972]. Calderbank and Kantor [1986], and later van Lint and Schrijver [1981], did extensive surveys on the subject. Most of these constructions used projective spaces and hence the constructed codes were projective codes. More recently, some cyclic two-weight codes were constructed in [Vega 2008; Vega and Wolfmann 2007].

The MacDonald codes, introduced in [MacDonald 1960] for binary codes, with the definition extended for q-ary codes [Bhandari and Durairajan 2003; Patel 1975], are punctured simplex codes of length $(q^n - q^u)/(q - 1)$ for any n and $1 \le u \le n-1$. They have parameters $[(q^n - q^u)/(q - 1), n, q^{n-1} - q^{u-1}]_q$ and are two-weight codes with nonzero words of weights $q^{n-1} - q^{u-1}$ and q^{n-1}. Following [Bhandari and Durairajan 2003], we denote these codes by $C_{n,u}(q)$.

The generalized MacDonald codes $C_{n,u,t}(q)$ were introduced in [Dodunekov and Simonis 1998] as an example of a projective multiset. The codes $C_{n,u,t}(q)$ are a direct sum of $t - 1$ q-ary simplex codes $C_n(q)$ with a MacDonald code $C_{n,u}(q)$ and have parameters $\left[t\begin{bmatrix} n \\ 1 \end{bmatrix} - \begin{bmatrix} u \\ 1 \end{bmatrix}, n, tq^{n-1} - q^{u-1}\right]$, where $\begin{bmatrix} n \\ 1 \end{bmatrix} = (q^n - 1)/(q - 1)$

is the q-ary Gaussian coefficient. Properties of the codes $C_{n,u,t}(q)$ were hardly studied, except for the uniqueness of these codes [Tamari 1984; Dodunekov and Simonis 1998].

In this article, we find the complete weight enumerator of the codes $C_{n,u,t}(q)$ and show that they are two-weight codes. We prove that the codes $C_{n,u,t}(q)$ are maximum minimum-distance and hence satisfy the Griesmer bound. Further, we show that both classes of codes $C_{n,u}(q)$ and $C_{n,u,t}(q)$ are self-orthogonal for $q = 2$. Later, we extend these codes to $C_{n,u,s,t}(q)$ codes by taking the direct sum of t simplex codes with s MacDonald codes.

We describe our notation and provide some background definitions in Section 2 and the prove the results on properties of $C_{n,u,t}(q)$ codes in Section 3. In Section 4, we find the parameters of the $C_{n,u,s,t}(q)$ codes and prove their properties.

2. Background and terminology

Codes. A linear $[n, k, d]_q$ code C is a k-dimensional subspace of an n-dimensional vector space over a finite field $\mathbb{F}_q$, where $q = p^m$ and p is a prime. Vectors in C are called codewords. The weight $\text{wt}(x)$ of a vector x in $\mathbb{F}_q^n$ is the number of nonzero entries of x. The distance $d(x, y)$ between two vectors x and y in $\mathbb{F}_q^n$ is the number of entries where x and y differ. Therefore, for a linear code $d(x, y) = \text{wt}(x - y)$. A code C is said to be an $[n, k, d]_q$ code if d is the minimum nonzero weight in C. A code C is said to be t-error correcting if $t = \lfloor (d - 1)/2 \rfloor$.

For vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ in $\mathbb{F}_q^n$, the Euclidean inner product (dot product) is defined to be $x \cdot y = \sum_{i=1}^{n} x_i y_i$. The dual code $C^\perp$ of C is defined as $C^\perp = \{x \in \mathbb{F}_q^n \mid x \cdot c = 0 \text{ for all } c \in C\}$. Then $C^\perp$ is an $[n, n-k]$ linear code over $\mathbb{F}_q$. A code C is called self-orthogonal if $C \subseteq C^\perp$.

The weight enumerator $W_C(x, y)$ of C is the polynomial

$$W_C(x, y) = \sum_{i=0}^{n} A_i x^{n-i} y^i,$$

where A_i is the number of codewords of weight i. C is called a two-weight code if all nonzero codewords have weights w_1 or $w_2 (w_1 < w_2)$ for some w_1 and w_2. A linear code C over $\mathbb{F}_q$ is called a projective code if any two of its coordinates are linearly independent, i.e., if the dual code $C^\perp$ has minimum distance ≥ 3.

For a q-ary $[n, k, d]$ code, the Griesmer bound is given by

$$n_q(k, d) \geq \sum_{i=0}^{k-1} \left\lceil \frac{d}{q^i} \right\rceil, \tag{1}$$

where $n_q(k, d)$ denotes the minimum length n for which an $[n, k, d]$ linear code, over $\mathbb{F}_q$, exists.

Strongly regular graphs.

Definition 1. A simple, undirected graph Γ is called strongly regular, with parameters v, k, λ, μ, if Γ has v vertices and

(1) Γ is regular with valency k,

(2) if the vertices x and y are adjacent then there are exactly λ vertices adjacent to both x and y,

(3) if the distinct vertices x and y are not adjacent then there are exactly μ vertices adjacent to both x and y.

It is easy to verify that the complement of a strongly regular graph is strongly regular. A graph Γ is described by its $(0, 1)$ adjacency matrix $A = (a_{i,j})$ of size v given by $a_{i,j} = 1$ if vertices i and j are adjacent and $a_{i,j} = 0$ if not. We quote the following theorems about strongly regular graphs.

Theorem 1. *If Γ is a graph with v vertices and adjacency matrix A then Γ is strongly regular if and only if there are numbers k, r, s and μ such that $AJ = kJ$ and $(A - rI)(A - sI) = \mu J$, where J is the $v \times v$ matrix of ones and I is the identity matrix of size v.*

Accordingly, it is easy to see that A has eigenvalues k (with a multiplicity of 1), r and s. We will denote the multiplicities of r and s by f and g, respectively. The following is an immediate consequence of Theorem 1.

Theorem 2. *If Γ is a regular graph with adjacency matrix A and A has only three eigenvalues then Γ is a strongly regular graph.*

The parameters of strongly regular graphs are not independent and are related.

Theorem 3. *Let Γ be a strongly regular graph with parameters $\langle v, k, \lambda, \mu \rangle$; then*

$$k(k - \lambda - 1) = (v - k - 1)\mu. \tag{2}$$

3. $C_{n,u,t}(q)$ codes

The MacDonald codes $C_{n,u}(q)$ can be considered as punctured simplex codes. The generator matrix of the $C_{n,u}(q)$ code can be expressed in the form

$$G_{n,u} = \left[G_n \setminus \left(\frac{\mathbf{0}}{G_u} \right) \right],$$

where $[A \setminus B]$ denotes the matrix obtained from the matrix A by deleting the columns of the matrix B and G_i is the generator matrix for the i-dimensional simplex code.

The generalized MacDonald codes $C_{n,u,t}(q)$ are defined by adding (via direct sum) $t - 1$ simplex codes to a MacDonald code. Hence, we can represent the

generator matrix, $G_{n,u,t}$ of a generalized MacDonald code by the generator matrix of a simplex code G_n and generator matrix of a MacDonald code $G_{n,u}$. We have

$$G_{n,u,t} = \big[\underbrace{G_n \mid G_n \mid \cdots \mid G_n}_{t-1} \mid G_{n,u}\big].$$

Theorem 4. *The binary MacDonald codes $C_{n,u}(2) = [2^n - 2^u, n, 2^{n-1} - 2^{u-1}]$ are self-orthogonal for $3 \leq u \leq n-1$.*

Proof. Let $\mathcal{G}_n$ be the matrix consisting of all column vectors of the vector space $\mathbb{F}_2^n$. Then $\mathcal{G}_n$ is an $n \times 2^n$ matrix. Let $\mathcal{G}_u$ be the matrix consisting of all column vectors of the vector space $\mathbb{F}_2^u$. Then $\mathcal{G}_u$ is a $u \times 2^u$ matrix. Let

$$\mathcal{G}_{u,0} = \begin{pmatrix} \mathbf{0} \\ \mathcal{G}_u \end{pmatrix},$$

where $\mathbf{0}$ is the $(n-u) \times 2^u$ zero matrix with elements in $\mathbb{F}_2$. Then the generator matrix $\mathcal{G}_{n,u}$ of the binary MacDonald code is given by $[\mathcal{G}_n \setminus \mathcal{G}_{n,0}]$. Therefore, we can write $\mathcal{G}_n = [\mathcal{G}_{n,u} \mid \mathcal{G}_{u,0}]$. We know that $\mathcal{G}_n \mathcal{G}_n^T = \mathbf{0}$ for $n \geq 3$. Now,

$$\mathcal{G}_n \mathcal{G}_n^T = [\mathcal{G}_{n,u} \mid \mathcal{G}_{u,0}][\mathcal{G}_{n,u} \mid \mathcal{G}_{u,0}]^T = [\mathcal{G}_{n,u} \mid \mathcal{G}_{u,0}]\begin{bmatrix} \mathcal{G}_{n,u}^T \\ \mathcal{G}_{u,0}^T \end{bmatrix} = \mathcal{G}_{n,u}\mathcal{G}_{n,u}^T + \mathcal{G}_{u,0}\mathcal{G}_{u,0}^T.$$

Therefore, we have $\mathcal{G}_{n,u}\mathcal{G}_{n,u}^T + \mathcal{G}_{u,0}\mathcal{G}_{u,0}^T = \mathbf{0}$. Further, $\mathcal{G}_{u,0}\mathcal{G}_{u,0}^T = \mathbf{0}$ for $u \geq 3$. Hence $\mathcal{G}_{n,u}\mathcal{G}_{n,u}^T = \mathbf{0}$ for $u \geq 3$. This implies that the binary $C_{n,u}(2)$ codes are self-orthogonal for $u \geq 3$. $\qquad \square$

Similar to MacDonald codes, the generalized MacDonald codes are also self-orthogonal for $3 \leq u \leq n$.

Theorem 5. *The generalized MacDonald codes $C_{n,u,t}(q)$ are self-orthogonal for $3 \leq u \leq n-1$ and $q = 2$.*

Proof. Consider the generator matrix $G_{n,u,t}$ of the generalized MacDonald code $C_{n,u,t}(q)$. Then we have

$$G_{n,u,t} = \big[\underbrace{G_n \mid G_n \mid \cdots \mid G_n}_{t-1} \mid G_{n,u}\big],$$

where G_n is the generator matrix for a simplex code and $G_{n,u}$ is the generator matrix for a MacDonald code. Next, consider matrix product

$$G_{n,u,t}G_{n,u,t}^T = \big[\underbrace{G_n \mid G_n \mid \cdots \mid G_n}_{t-1} \mid G_{n,u}\big]\big[\underbrace{G_n \mid G_n \mid \cdots \mid G_n}_{t-1} \mid G_{n,u}\big]^T$$

$$= \underbrace{G_n G_n^T + \cdots + G_n G_n^T}_{t-1} + G_{n,u}G_{n,u}^T. \qquad (3)$$

Let $\mathcal{G}_n$ be the matrix obtained from the simplex matrix G_n by adding the zero column vector. That is, $\mathcal{G}_n = [\mathbf{0} \mid G_n]$. This is the same matrix obtained from

column vectors of $\mathbb{F}_q^n$. Now, consider

$$\mathcal{G}_n\mathcal{G}_n^T = [\mathbf{0} \mid G_n][\mathbf{0} \mid G_n]^T = \mathbf{00}^T + G_nG_n^T = \mathbf{0}_n + G_nG_n^T = G_nG_n^T.$$

Hence, we can rewrite (3) as

$$G_{n,u,t}G_{n,u,t}^T = \underbrace{\mathcal{G}_n\mathcal{G}_n^T + \cdots + \mathcal{G}_n\mathcal{G}_n^T}_{t-1} + G_{n,u}G_{n,u}^T.$$

From the proof of Theorem 4, we have $\mathcal{G}_n\mathcal{G}_n^T = \mathbf{0}$ and $G_{n,u}G_{n,u}^T = \mathbf{0}$ for $3 \le u \le n-1$. Therefore, we have $G_{n,u,t}G_{n,u,t}^T = \mathbf{0}$ for $3 \le u \le n-1$. Hence, the generalized MacDonald codes $C_{n,u,t}(q)$ are self-orthogonal for $3 \le u \le n-1$. $\square$

The complete weight enumerator of q-ary MacDonald codes is known [Calderbank and Kantor 1986]. Here we will state the result, as it is essential for Theorem 7.

Theorem 6. *The q-ary MacDonald code $C_{n,u}(q)$ is a $[(q^n - q^u)/(q-1), n, q^{n-1} - q^{u-1}]$ is a two-weight code with nonzero weights $w_1 = q^{n-1} - q^{u-1}$ and $w_2 = q^{n-1}$ with weight enumerator coefficients $A_{w_1} = q^n - q^{n-u}$ and $A_{w_2} = q^{n-u} - 1$.*

In the following theorem, we show that the generalized MacDonald codes are also two-weight codes with the same weight enumerator as the MacDonald codes, but with different weights.

Theorem 7. *The generalized MacDonald code $C_{n,u,t}(q)$ is a $\left[t{n \brack 1} - {u \brack 1}, tq^{n-1} - q^{u-1}\right]_q$ code with nonzero weights $w_1 = tq^{n-1} - q^{u-1}$ and $w_2 = tq^{n-1}$ and weight enumerator coefficients $A_{w_1} = q^n - q^{n-u}$ and $A_{w_2} = q^{n-u} - 1$.*

Proof. Consider the generator matrix $G_{n,u,t}$ of the code $C_{n,u,t}(q)$. We can represent $G_{n,u,t}$ by

$$G_{n,u,t} = \left[\underbrace{G_n|G_n|\cdots|G_n}_{t-1}|G_{n,u}\right]. \tag{4}$$

The first $t-1$ simplex matrices contribute $(t-1)q^{n-1}$ weights to a nonzero codeword and the MacDonald matrix $G_{n,u}$ contributes q^{n-1} and $q^{n-1} - q^{u-1}$ weights. Therefore, the weights of $C_{n,u,t}(q)$ are $w_1 = tq^{n-1} - q^{u-1}$ and $w_2 = tq^{n-1}$. From (4), it is easy to see that the number of words of weights w_1 and w_2 depend only on the last MacDonald matrix $G_{n,u}$. Hence, the weight enumerator of $C_{n,u,t}(q)$ is the same as the weight enumerator of $C_{n,u}(q)$. Therefore, $A_{w_1} = q^n - q^{n-u}$ and $A_{w_2} = q^{n-u} - 1$. $\square$

An important property of MacDonald codes is that they are maximum minimum-distance codes; i.e., they satisfy the Griesmer bound. In the next theorem, we show that the generalized MacDonald codes are also maximum minimum-distance codes.

Theorem 8. *The codes $C_{n,u,t}(q)$ satisfy the Griesmer bound.*

Proof. Let $C_{n,u,t}(q)$ be a $\left[t\begin{bmatrix} n \\ 1 \end{bmatrix} - \begin{bmatrix} u \\ 1 \end{bmatrix}, tq^{n-1} - q^{u-1} \right]$ code. Consider the right-hand side of the Griesmer bound (1). Then

$$
\sum_{i=0}^{k-1} \left\lceil \frac{d}{q^i} \right\rceil = \sum_{i=0}^{k-1} \left\lceil \frac{tq^{n-1} - q^{u-1}}{q^i} \right\rceil = \sum_{i=0}^{u-1} \left\lceil \frac{tq^{n-1} - q^{u-1}}{q^i} \right\rceil + \sum_{i=u}^{k} \left\lceil \frac{tq^{n-1} - q^{u-1}}{q^i} \right\rceil
$$

$$
= \left\{ \lceil tq^{n-1} - q^{u-1} \rceil + \lceil tq^{n-2} - q^{u-2} \rceil + \cdots + \lceil tq^{n-u} - 1 \rceil \right\}
$$
$$
+ \left\{ \left\lceil tq^{n-u-1} - \frac{q^{u-1}}{q^u} \right\rceil + \left\lceil tq^{n-u-2} - \frac{q^{u-1}}{q^{u+1}} \right\rceil + \cdots + \left\lceil t - \frac{q^{u-1}}{q^{n-1}} \right\rceil \right\}
$$

$$
= \left\{ (tq^{n-1} - q^{u-1}) + (tq^{n-2} - q^{u-2}) + \cdots + (tq^{n-u} - 1) \right\}
$$
$$
+ \left\{ (tq^{n-u-1}) + (tq^{n-u-2}) + \cdots + (tq^0) \right\}
$$

$$
= \left\{ t(q^{n-1} + q^{n-2} + \cdots + q^{n-u} + q^{n-u-1} + \cdots + q^0) \right\}
$$
$$
- \left\{ q^{u-1} + q^{u-2} + \cdots + q^0 \right\}
$$

$$
= t \left(\frac{1 - q^n}{1 - q} \right) - \left(\frac{1 - q^u}{1 - q} \right) = t \begin{bmatrix} n \\ 1 \end{bmatrix} - \begin{bmatrix} u \\ 1 \end{bmatrix} = n_q(k, d). \qquad \square
$$

We can obtain a strongly regular graph from a two-weight code C with weights w_1 and w_2 as follows [Calderbank and Kantor 1986]. Take codewords as vertices of Γ and join two codewords $\boldsymbol{x}$ and $\boldsymbol{y}$ by an edge if and only if $d(\boldsymbol{x}, \boldsymbol{y}) = w_1$. The strongly regular graph Γ is said be associated with C.

Theorem 9. *Let $\Gamma_{n,u,t}$ be the strongly regular graph associated with the generalized MacDonald code $C_{n,u,t}(q)$. Then $\Gamma_{n,u,t}$ has parameters $\langle q^n, q^n - q^{n-u}, q^n - 2q^{n-u}, q^n - q^{n-u} \rangle$.*

Proof. The number of vertices of $\Gamma_{n,u,t}$ is equal to the number of codewords of $C_{n,u,t}(q)$; hence $v = q^n$. Let W_1 be the set of codewords of weight $w_1 = tq^{n-1} - q^{u-1}$ and W_2 be the set of codewords of weight $w_2 = tq^{n-1}$ of $C_{n,u,t}(q)$. By the construction, we know $\Gamma_{n,u,t}$ is a regular graph. The zero-vector $\boldsymbol{0}$, as a vertex, has degree $|W_1|$, as $d(\boldsymbol{0}, \boldsymbol{x}) = w_1$ for all $\boldsymbol{x} \in W_1$. Therefore, from Theorem 7, we get $k = |W_1| = q^n - q^{n-u}$.

To obtain the value of μ, consider the zero-vector $\boldsymbol{0}$. Pick any codeword $\boldsymbol{u}$ from W_2; then $d(\boldsymbol{0}, \boldsymbol{u}) = w_2$, which implies $\boldsymbol{0}$ is nonadjacent to all the codewords in W_2. Let $\boldsymbol{v}$ be an arbitrary codeword in W_1. Then $d(\boldsymbol{u}, \boldsymbol{v}) = w_1$; otherwise $\boldsymbol{v} \in W_2$, which contradicts our assumption that $\boldsymbol{v} \in W_1$. Since $\boldsymbol{v} \in W_1$ is arbitrary, $\boldsymbol{u} \in W_2$ is adjacent to all the codewords in W_1. Therefore, the codeword $\boldsymbol{u}$ is adjacent to $|W_1| = q^n - q^{n-u}$ vertices and hence $\mu = q^n - q^{n-u}$.

We will use (2) to determine the value of λ from the other three parameters. Consider $k(k - \lambda - 1) = (v - k - 1)\mu$, but $\mu = k$ implies $(k - \lambda - 1) = (v - k - 1)$, and then $\lambda = 2k - v = 2(q^n - q^{n-u}) - q^n = q^n - 2q^{n-u}$. This leads to $\langle v, k, \lambda, \mu \rangle = \langle q^n, q^n - q^{n-u}, q^n - 2q^{n-u}, q^n - q^{n-u} \rangle$. $\qquad \square$

4. $C_{n,u,s,t}(q)$ codes

In this section, we extend the definition of generalized MacDonald codes $C_{n,u,t}(q)$ to that of $C_{n,u,s,t}(q)$ codes. Define $C_{n,u,s,t}(q)$ codes by adding t simplex codes to s MacDonald codes.

The generator matrices of $C_{n,u,s,t}(q)$ codes can be defined similarly to generator matrices of generalized MacDonald codes $C_{n,u,t}(q)$. Let $G_{n,u,s,t}$ be the generator matrix of the code $C_{n,u,s,t}(q)$. Then

$$G_{n,u,s,t} = \Big[\underbrace{G_n \mid G_n \mid \cdots \mid G_n}_{t} \mid \underbrace{G_{n,u} \mid G_{n,u} \mid \cdots \mid G_{n,u}}_{s}\Big],$$

where G_n and $G_{n,u}$ are the generator matrices of simplex codes and MacDonald codes, respectively.

The parameters of the $C_{n,u,s,t}(q)$ codes can be easily deduced from that of the codes $C_{n,u}(q)$ and $C_{n,u,t}(q)$. By the form of the generator matrix $G_{n,u,s,t}$, the weight enumerator of $C_{n,u,s,t}(q)$ is the same as that of the MacDonald codes.

Theorem 10. $C_{n,u,s,t}(q)$ is a $\left[(t+s)\begin{bmatrix} n \\ 1 \end{bmatrix} - s\begin{bmatrix} u \\ 1 \end{bmatrix}, n, (t+s)q^{n-1} - sq^{u-1}\right]_q$ code with nonzero weights $w_1 = (t+s)q^{n-1} - sq^{u-1}$ and $w_2 = (t+s)q^{n-1}$ with weight enumerator coefficients $A_{w_1} = q^n - q^{n-1}$ and $A_{w_2} = q^{n-u} - 1$.

Similar to MacDonald and generalized MacDonald codes, the binary codes $C_{n,u,s,t}(q)$ are self-orthogonal for $u \geq 3$.

Theorem 11. $C_{n,u,s,t}$ codes are self-orthogonal for $q = 2$ and $3 \leq u \leq n - 1$.

Proof. We will show that $G_{n,u,s,t}G_{n,u,s,t}^T = \mathbf{0}$ for $3 \leq u \leq n - 1$:

$$G_{n,u,s,t}G_{n,u,st}^T = \Big[\underbrace{G_n \mid \cdots \mid G_n}_{t} \mid \underbrace{G_{n,u} \mid \cdots \mid G_{n,u}}_{s}\Big]\Big[\underbrace{G_n \mid \cdots \mid G_n}_{t} \mid \underbrace{G_{n,u} \mid \cdots \mid G_{n,u}}_{s}\Big]^T$$

$$= \underbrace{G_nG_n^T + \cdots + G_nG_n^T}_{t} + \underbrace{G_{n,u}G_{n,u}^T + \cdots + G_{n,u}G_{n,u}^T}_{s}$$

$$= tG_nG_n^T + sG_{n,u}G_{n,u}^T$$

$$= t\mathcal{G}_n\mathcal{G}_n^T + sG_{n,u}G_{n,u}^T.$$

For $3 \leq u \leq n - 1$, from the proofs of Theorems 4 and 5, we have $\mathcal{G}_n\mathcal{G}_n^T = \mathbf{0}$ and $G_{n,u}G_{n,u}^T = \mathbf{0}$. $\qquad \square$

Since these codes have the same weight enumerator as that of MacDonald codes, parameters of the strongly regular graphs generated by them are the same as the strongly regular graphs generated by the MacDonald codes.

5. Conclusion

In this work, we have described the weight enumerators of generalized MacDonald codes $C_{n,u,t}(q)$ and the codes $C_{n,u,s,t}(q)$ and showed that these are two-weight codes. Further, we have shown that the codes $C_{n,u}(q)$, $C_{n,u,t}(q)$ and $C_{n,u,s,t}(q)$ are self-orthogonal for $3 \leq u \leq n - 1$. All three classes have the same weight enumerator and hence generate the same strongly regular graph.

All the codes in this work were constructed as a direct sum of a one-weight code (simplex code) with a two-weight code (MacDonald code). It might be interesting to study other such constructions arising from one- and two-weight codes.

Acknowledgments

References

[Bhandari and Durairajan 2003] M. C. Bhandari and C. Durairajan, "A note on covering radius of MacDonald codes", pp. 221–225 in *International conference on information technology: coding and computing* (Las Vegas, 2003), IEEE, Piscataway, NJ, 2003.

[Calderbank and Kantor 1986] R. Calderbank and W. M. Kantor, "The geometry of two-weight codes", *Bull. London Math. Soc.* **18**:2 (1986), 97–122. MR Zbl

[Delsarte 1972] P. Delsarte, "Weights of linear codes and strongly regular normed spaces", *Discrete Math.* **3** (1972), 47–64. MR Zbl

[Dodunekov and Simonis 1998] S. Dodunekov and J. Simonis, "Codes and projective multisets", *Electron. J. Combin.* **5** (1998), art. id. 37. MR Zbl

[van Lint and Schrijver 1981] J. H. van Lint and A. Schrijver, "Construction of strongly regular graphs, two-weight codes and partial geometries by finite fields", *Combinatorica* **1**:1 (1981), 63–73. MR Zbl

[MacDonald 1960] J. E. MacDonald, "Design methods for maximum minimum-distance error-correcting codes", *IBM. J. Res. Develop.* **4** (1960), 43–57. MR

[Patel 1975] A. M. Patel, "Maximal q-nary linear codes with large minimum distance", *IEEE Trans. Information Theory* **21**:1 (1975), 106–110. MR Zbl

[Tamari 1984] F. Tamari, "On linear codes which attain the Solomon–Stiffler bound", *Discrete Math.* **49**:2 (1984), 179–191. MR Zbl

[Vega 2008] G. Vega, "Two-weight cyclic codes constructed as the direct sum of two one-weight cyclic codes", *Finite Fields Appl.* **14**:3 (2008), 785–797. MR Zbl

[Vega and Wolfmann 2007] G. Vega and J. Wolfmann, "New classes of 2-weight cyclic codes", *Des. Codes Cryptogr.* **42**:3 (2007), 327–334. MR Zbl

padmapani.seneviratne@tamuc.edu   *Department of Mathematics, Texas A&M University–Commerce, Commerce, TX, United States*

lmelcher@leomail.tamuc.edu   *Department of Mathematics, Texas A&M University–Commerce, Commerce, TX, United States*

A simple proof characterizing interval orders with interval lengths between 1 and k

Simona Boyadzhiyska, Garth Isaak and Ann N. Trenk

(Communicated by Glenn Hurlbert)

A poset $P = (X, \prec)$ has an interval representation if each $x \in X$ can be assigned a real interval I_x so that $x \prec y$ in P if and only if I_x lies completely to the left of I_y. Such orders are called *interval orders*. Fishburn (1983, 1985) proved that for any positive integer k, an interval order has a representation in which all interval lengths are between 1 and k if and only if the order does not contain $(\boldsymbol{k+2})\boldsymbol{+1}$ as an induced poset. In this paper, we give a simple proof of this result using a digraph model.

1. Introduction

1.1. *Posets and interval orders.* A poset P consists of a set X of *points* and a relation $\prec$ that is irreflexive and transitive, and therefore antisymmetric. It is sometimes convenient to write $y \succ x$ instead of $x \prec y$. If $x \prec y$ or $y \prec x$, we say that x and y are *comparable*, and otherwise we say they are *incomparable*, and denote the incomparability by $x \parallel y$. An *interval representation* of a poset $P = (X, \prec)$ is an assignment of a closed real interval I_v to each $v \in X$ so that $x \prec y$ if and only if I_x is completely to the left of I_y. A poset with such a representation is called an *interval order*. It is well known that the classes studied in this paper are the same if open intervals are used instead of closed intervals; e.g., see Lemma 1.5 in [Golumbic and Trenk 2004].

The poset $\mathbf{2+2}$ shown in Figure 1 consists of four elements $\{a, b, x, y\}$ and the only comparabilities are $a \prec x$ and $b \prec y$. The following elegant theorem characterizing interval orders was anticipated by Wiener in 1914, see [Fishburn and Monjardet 1992], and shown by Fishburn [1970]: poset P is an interval order if and only if it contains no induced $\mathbf{2+2}$. Posets that have an interval representation in which all intervals are the same length are known as *unit interval orders* or
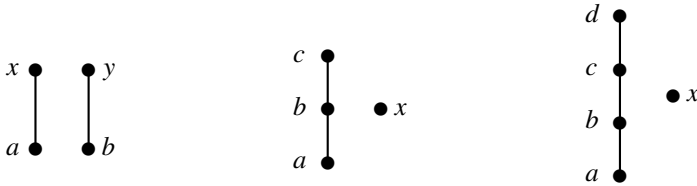
Figure 1. The posets **2+2** (left), **3+1** (middle), and **4+1** (right).

semiorders. Scott and Suppes [1958] characterized unit interval orders as those posets with no induced **2+2** and no induced **3+1**. Figure 1 shows the posets **2+2**, **3+1**, and **4+1**. More generally, the poset $n+1$ consists of a chain of n distinct elements $a_1 \prec a_2 \prec \cdots \prec a_n$ and an additional element that is incomparable to each a_i.

In this paper, we consider an intermediate class between the extremes of interval orders (no restrictions on interval lengths) and unit interval orders (all intervals the same length). In particular, we allow interval lengths to range from 1 to k, where k is a positive integer. Fishburn [1983; 1985] characterized this class as those posets with no induced **2+2** and no induced $(k+2)+1$, generalizing the result of Scott and Suppes. In fact, Fishburn characterized those posets that have an interval representation by intervals whose lengths are between m and n for any relatively prime integers m, n in terms of what he calls *picycles*. Fishburn's proof uses a set of inequalities similar to those in our proof of Theorem 4. His proof is technical, and it does not immediately yield a forbidden poset characterization in the general case. Doignon [1987; 1988] introduced the idea of using potentials in a digraph model to solve a related interval representation problem. (Pages 91–93 of [Pirlot and Vincke 1997] contain an English version of the main result in [Doignon 1988].)

We use a different digraph model, one that appears in [Isaak 2009], to give a shorter and more accessible proof of Fishburn's result for the case $m = 1$, $n = k$. This digraph model uses two vertices for each element, one for each of the endpoints of an interval representing the element. Our digraph model and the equivalence of statements (1) and (3) in Theorem 4 can easily be extended to general m, n. It is also natural to consider allowing the interval lengths to vary between 1 and any real value. Fishburn and Graham [1985] studied the classes $C(\alpha)$ of interval graphs that have a representation by intervals with lengths between 1 and α for any real $\alpha \geq 1$, showing that the points where $C(\alpha)$ expands are the rational values of α. The problem of characterizing posets that have an interval representation in which the possible interval lengths come from a discrete set (rather than from an interval) is more challenging, and we consider two variants of this question in [Boyadzhiyska et al. 2017].

1.2. *Digraphs and potentials.* A *directed graph*, or *digraph*, is a pair $G = (V, E)$, where V is a finite set of *vertices*, and E is a set of ordered pairs (x, y), with $x, y \in V$, called *arcs*. A *weighted digraph* is a digraph in which each arc (x, y) is

assigned a real number weight w_{xy}. We sometimes denote the arc (x, y) by $x \to y$, and in a weighted digraph by $x \xrightarrow{w_{xy}} y$. A *potential function* $p : V \to \mathbb{R}$, defined on the vertices of a weighted digraph, is a function satisfying $p(y) - p(x) \leq w_{xy}$ for each arc (x, y). Theorem 1 is a well known result that specifies precisely which digraphs have potential functions.

A *cycle* in digraph G is a subgraph with vertex set $\{x_1, x_2, x_3, \dots, x_t\}$ and arc set $\{(x_i, x_{i+1}) : 1 \leq i \leq t - 1\} \cup \{(x_t, x_1)\}$. In a weighted digraph, the *weight* of cycle C, denoted by $\mathrm{wgt}(C)$, is the sum of the weights of the arcs of C. A cycle with negative weight is called a *negative cycle*. The following theorem is well known, see Chapter 8 of [Schrijver 2003] for example, and we provide a proof in [Boyadzhiyska et al. 2017].

Theorem 1. *A weighted digraph has a potential function if and only if it contains no negative cycle.*

2. Orders with a $[1, k]$-interval representation

We say that poset P has an $[a, b]$-interval representation if it has a representation by intervals whose lengths are between a and b (inclusive). When $a = b > 0$, the posets with such a representation are the unit interval orders. Because representations can be scaled, for any $b > 0$, all interval orders have a $[0, b]$-interval representation. This motivates us to consider the lower bound $a = 1$, and in particular, posets that have a $[1, k]$-interval representation where k is a positive integer. Fishburn [1983] characterized this class by showing the equivalence of (1) and (2) in Theorem 4; however, the proof is quite technical. Using the framework in [Isaak 2009], we construct a weighted digraph $G_{P,k}$ associated with poset P and show that P has a $[1, k]$-interval representation if and only if $G_{P,k}$ has no negative cycle. This allows for a more accessible proof of Theorem 4. We choose the value of ϵ appearing as a weight in $G_{P,k}$ so that $0 < \epsilon < 1/(2|X|)$.

Definition 2. Let $P = (X, \prec)$ be a partial order. Define $G_{P,k}$ to be the weighted digraph with vertices $\{\ell_x, r_x\}_{x \in X}$ and the arcs

- (ℓ_y, r_x) with weight $-\epsilon$ for all $x, y \in X$ with $x \prec y$,
- (r_x, ℓ_y) with weight 0 for all $x, y \in X$ with $x \| y$,
- (r_x, ℓ_x) with weight -1 for all $x \in X$,
- (ℓ_x, r_x) with weight k for all $x \in X$.

It is helpful to think of the arcs of $G_{P,k}$ as coming in two categories: $\ell \to r$ and $r \to \ell$. We list the arcs by category in Table 1 for easy reference.

Any negative cycle in $G_{P,k}$ with a minimum number of arcs will have at most $2|X|$ arcs since $G_{P,k}$ has $2|X|$ vertices. Since ϵ satisfies $0 < \epsilon < 1/(2|X|)$, the

type	arc	weight	(x, y) relation
$\ell \to r$	(ℓ_y, r_x)	$-\epsilon$	$y \succ x$
	(ℓ_x, r_x)	k	
$r \to \ell$	(r_x, ℓ_y)	0	$x \parallel y$
	(r_x, ℓ_x)	-1	

Table 1. The arcs of the weighted digraph $G_{P,k}$.

arcs of weight $-\epsilon$ will have combined weight w, where $-1 < w \leq 0$. We record a consequence of this observation in the following remark.

Remark 3. If C is a negative weight cycle in $G_{P,k}$ containing the minimum number of arcs, then C contains at least k arcs of weight -1 for every arc of weight k.

Theorem 4. *Let $P = (X, \prec)$ be a partial order and let $k \in \mathbb{Z}_{\geq 1}$. The following are equivalent*:

(1) *P has a $[1, k]$-interval representation.*

(2) *P contains no induced $\mathbf{2+2}$ or $\mathbf{(k+2)+1}$.*

(3) *The weighted digraph $G_{P,k}$ contains no negative cycle.*

Proof. (1) $\Rightarrow$ (3): Suppose that P has an interval representation $\mathcal{I} = \{I_x\}_{x \in X}$, where $I_x = [L(x), R(x)]$, and for each $x \in X$ we have $1 \leq |I_x| \leq k$. Choose $\epsilon = \min\{1/(2|X|+1), \delta\}$, where δ is the smallest distance between unequal endpoints in the representation $\mathcal{I}$. By the definition of an interval representation and the conditions on the interval lengths, we have

(i) $R(x) - L(y) \leq -\epsilon$ for all $x, y \in X$ with $x \prec y$,

(ii) $L(y) - R(x) \leq 0$ for all $x, y \in X$ with $x \| y$,

(iii) $L(x) - R(x) \leq -1$ for all $x \in X$,

(iv) $R(x) - L(x) \leq k$ for all $x \in X$.

Now define the function p on the vertex set of $G_{P,k}$ as follows. For each $x \in X$ let $p(r_x) = R(x)$ and $p(\ell_x) = L(x)$. So p satisfies

(a) $p(r_x) - p(l_y) \leq -\epsilon$ for all $x, y \in X$ with $x \prec y$,

(b) $p(l_y) - p(r_x) \leq 0$ for all $x, y \in X$ with $x \| y$,

(c) $p(l_x) - p(r_x) \leq -1$ for all $x \in X$,

(d) $p(r_x) - p(l_x) \leq k$ for all $x \in X$.

Thus, for all $(u, v) \in E(G_{P,k})$, we have $p(v) - p(u) \leq w_{uv}$. Hence p is a potential function on $G_{P,k}$ and by Theorem 1, $G_{P,k}$ has no negative cycle.

(3) $\Rightarrow$ (1): Given $G_{P,k}$ has no negative cycle, by Theorem 1, there exists a potential function p on $G_{P,k}$, and by definition, p satisfies (a), (b), (c), (d). For each $x \in X$, let $L(x) = p(\ell_x)$ and $R(x) = p(r_x)$. By (c) we know $L(x) + 1 \le R(x)$, so $I_x = [L(x), R(x)]$ is indeed an interval with $|I_x| \ge 1$. By (d), the length of interval I_x satisfies $|I_x| \le k$, and by (a) and (b), $x \prec y$ in P if and only if $R(x) < L(y)$. Thus the set of intervals $\{I_x\}_{x \in X}$ forms a representation of P in which each interval has length between 1 and k.

(3) $\Rightarrow$ (2): If P contains an induced $\mathbf{2+2}$, denoted by $(x \succ a) \parallel (y \succ b)$, then

$$\ell_x \xrightarrow{-\epsilon} r_a \xrightarrow{0} \ell_y \xrightarrow{-\epsilon} r_b \xrightarrow{0} \ell_x$$

is a cycle in $G_{P,k}$ with weight -2ϵ. Similarly, if P contains an induced $(k+2)+1$, denoted by $x \parallel (a_{k+2} \succ a_{k+1} \succ \cdots \succ a_2 \succ a_1)$, then $G_{P,k}$ contains the cycle

$$r_x \xrightarrow{0} \ell_{a_{k+2}} \xrightarrow{-\epsilon} r_{a_{k+1}} \xrightarrow{-1} \ell_{a_{k+1}} \xrightarrow{-\epsilon} r_{a_k} \xrightarrow{-1} \ell_{a_k} \xrightarrow{-\epsilon}$$
$$\cdots \xrightarrow{-\epsilon} r_{a_2} \xrightarrow{-1} \ell_{a_2} \xrightarrow{-\epsilon} r_{a_1} \xrightarrow{0} l_x \xrightarrow{k} r_x,$$

whose weight is $(-1)k + k + (-\epsilon)(k+1) < 0$. In either case, we obtain a negative cycle in P, a contradiction.

(2) $\Rightarrow$ (3): Now assume P contains no induced $\mathbf{2+2}$ or $(k+2)+1$. For a contradiction, assume that $G_{P,k}$ contains a negative cycle, and let C be a negative cycle in $G_{P,k}$ containing the minimum number of arcs. By the definition of $G_{P,k}$, the arcs in C must alternate between arcs of type $\ell \to r$ and arcs of type $r \to \ell$, thus C has the form $\ell_{x_1} \to r_{x_2} \to \ell_{x_3} \to \cdots \to r_{x_n} \to \ell_{x_1}$ for some $x_1, x_2, \ldots, x_n \in X$, not necessarily distinct. The cycles in $G_{P,k}$ that contain exactly two arcs have nonnegative weight; hence $n \ge 4$. Furthermore, since vertices of a cycle are distinct, we know that $x_i \ne x_{i+2}$ for $1 \le i \le n$, where the indices are taken modulo n.

Next we show $\text{wgt}(C) \le -2\epsilon$. Since $x_i \ne x_{i+2}$ for $1 \le i \le n$ (indices taken modulo n), the arcs of C immediately before and after a weight-k arc must have weight 0. If C has at most one arc of weight $-\epsilon$, then the remaining $\ell \to r$ arcs have weight k, resulting in a positive weight for C, a contradiction. Thus C contains at least two arcs of weight $-\epsilon$, and Remark 3 implies that $\text{wgt}(C) \le -2\epsilon$.

We next claim that C does not contain a segment of three consecutive arcs of weights $-\epsilon, 0, -\epsilon$. For a contradiction, suppose C contains the segment

$$S_1 : \ell_a \xrightarrow{-\epsilon} r_b \xrightarrow{0} \ell_c \xrightarrow{-\epsilon} r_d.$$

Then by the definition of $G_{P,k}$, we have $a \succ b$, $b \parallel c$, and $c \succ d$. If $d \succ a$, we get $c \succ d \succ a \succ b$, contradicting $b \parallel c$. If $a \parallel d$, then the elements a, b, c, d induce in P the poset $\mathbf{2+2}$, a contradiction. Otherwise, $a \succ d$ and we can replace the segment S_1 by $\ell_a \xrightarrow{-\epsilon} r_d$ to yield a shorter cycle C' with $\text{wgt}(C') = \text{wgt}(C) + \epsilon \le -2\epsilon + \epsilon = -\epsilon < 0$. This contradicts the minimality of C.

We now consider two cases depending on whether or not C contains an arc of weight k.

Case 1: C has no arc of weight k. In this case, C alternates between arcs with weight $-\epsilon$ and arcs with weight in the set $\{0, -1\}$. Since C has at least four arcs and no segment of the form $(-\epsilon, 0, -\epsilon)$, there must be an arc of weight -1. Without loss of generality, choose a starting point for C so that it begins with the segment

$$S_2 : \ell_{x_1} \xrightarrow{-\epsilon} r_{x_2} \xrightarrow{-1} \ell_{x_3} \xrightarrow{-\epsilon} r_{x_4}.$$

By the definition of $G_{P,k}$ we have $x_1 \succ x_2 = x_3 \succ x_4$, so $x_1 \succ x_4$. Replace segment S_2 by $\ell_{x_1} \xrightarrow{-\epsilon} r_{x_4}$ to obtain a cycle C' whose weight is also negative since it contains no arcs of weight k. Since C' has fewer arcs than C, this contradicts the minimality of C.

Case 2: C contains an arc of weight k. By Remark 3, there is a segment of C that starts with an arc of weight k and has at least k arcs of weight -1 before the next arc of weight k. Thus this segment of C contains at least $2k$ arcs. Without loss of generality, we can choose the starting point of C so that it begins with the segment

$$\ell_{x_1} \xrightarrow{k} r_{x_2} \longrightarrow \ell_{x_3} \xrightarrow{-\epsilon} r_{x_4} \longrightarrow \cdots \xrightarrow{-\epsilon} r_{x_{2k}} \longrightarrow \ell_{x_{2k+1}}.$$

If the arc (r_{x_2}, ℓ_{x_3}) has weight -1, then $x_1 = x_2 = x_3$, a contradiction since $x_1 \neq x_3$. Thus, the arc (r_{x_2}, ℓ_{x_3}) has weight 0 and C begins with the segment

$$\ell_{x_1} \xrightarrow{k} r_{x_2} \xrightarrow{0} \ell_{x_3} \xrightarrow{-\epsilon} r_{x_4}.$$

If any of the next k arcs of the type $r \to \ell$ on C had weight 0, then C would contain a segment of the form $(-\epsilon, 0, -\epsilon)$, contradicting our earlier claim. Thus each of these arcs has weight -1 and C starts with the segment

$$\ell_{x_1} \xrightarrow{k} r_{x_2} \xrightarrow{0} \ell_{x_3} \xrightarrow{-\epsilon} r_{x_4} \xrightarrow{-1} \ell_{x_5} \xrightarrow{-\epsilon} r_{x_6} \xrightarrow{-1} \cdots \xrightarrow{-\epsilon} r_{x_{2k+2}} \xrightarrow{-1} \ell_{x_{2k+3}}.$$

Note that the arcs $\ell_{x_{2k+1}} \to r_{x_{2k+2}} \to \ell_{x_{2k+3}}$ are included since there must be k arcs of weight -1 before the next arc of weight k.

By the definition of $G_{P,k}$, we have the following relations in P:

$$x_1 = x_2 \parallel x_3 \succ x_4 = x_5 \succ x_6 = x_7 \succ \cdots = x_{2k+1} \succ x_{2k+2} = x_{2k+3}.$$

If $x_1 = x_{2k+3}$, then by transitivity, $x_1 \prec x_3$, contradicting the relation $x_1 = x_2 \parallel x_3$. Thus C contains at least two more arcs $(\ell_{x_{2k+3}}, r_{x_{2k+4}})$ and $(r_{x_{2k+4}}, \ell_{x_{2k+5}})$. If arc $(\ell_{x_{2k+3}}, r_{x_{2k+4}})$ had weight k, then $x_{2k+2} = x_{2k+3} = x_{2k+4}$, a contradiction since $x_{2k+2} \neq x_{2k+4}$. Thus arc $(\ell_{x_{2k+3}}, r_{x_{2k+4}})$ has weight $-\epsilon$, and $x_{2k+3} \succ x_{2k+4}$ in P, and C starts with the segment

$$S : \ell_{x_1} \xrightarrow{k} r_{x_2} \xrightarrow{0} \ell_{x_3} \xrightarrow{-\epsilon} r_{x_4} \xrightarrow{-1} \ell_{x_5} \xrightarrow{-\epsilon} r_{x_6} \xrightarrow{-1} \cdots \xrightarrow{-\epsilon} r_{x_{2k+2}} \xrightarrow{-1} \ell_{x_{2k+3}} \xrightarrow{-\epsilon} r_{x_{2k+4}}.$$

Finally, we consider the relation between x_1 and x_{2k+4} in P. If $x_1 \prec x_{2k+4}$, then by transitivity, $x_1 \prec x_3$, a contradiction. If $x_1 \succ x_{2k+4}$, we can replace segment S by $\ell_{x_1} \xrightarrow{-\epsilon} r_{x_{2k+4}}$ to obtain a shorter cycle C' in $G_{P,k}$. As noted earlier, the combined weight of the arcs of C that have weight $-\epsilon$ is strictly greater than -1, so C' also has negative weight, contradicting the minimality of C. Hence $x_1 \parallel x_{2k+4}$ and the $k+3$ elements in the set $\{x_1, x_3, x_5, \ldots, x_{2k+3}, x_{2k+4}\}$ induce a $(\boldsymbol{k}+\boldsymbol{2})+\boldsymbol{1}$ in P, a contradiction. $\square$

We end by describing an algorithm that constructs a $[1, k]$-interval representation of a poset P if one exists and otherwise produces a forbidden poset, either $\boldsymbol{2}+\boldsymbol{2}$ or $(\boldsymbol{k}+\boldsymbol{2})+\boldsymbol{1}$. Use a standard shortest-paths algorithm such as the Bellman–Ford or the matrix multiplication method on $G_{P,k}$ to compute the weight of a minimum-weight path between each pair of vertices or detect a negative cycle. If there is a negative cycle, these algorithms detect one with a minimum number of arcs. If such a negative cycle exists in $G_{P,k}$, then as in the proof of (2) $\Rightarrow$ (3) of Theorem 4, either the cycle contains the segment $-\epsilon, 0, -\epsilon$, and a $\boldsymbol{2}+\boldsymbol{2}$ is detected in P, or else as in Case 2 of that proof, a $(\boldsymbol{k}+\boldsymbol{2})+\boldsymbol{1}$ is detected in P. If there is no negative cycle, Theorem 1 ensures that a potential function p exists for $G_{P,k}$. Indeed, setting $p(v)$ to be the minimum weight of a walk ending at v produces a potential function. As we showed in the proof of (3) $\Rightarrow$ (1), the intervals $[p(\ell_x), p(r_x)]$ provide a $[1, k]$-interval representation of P. Thus there is a polynomial-time certifying algorithm.

References

[Boyadzhiyska et al. 2017] S. Boyadzhiyska, G. Isaak, and A. N. Trenk, "Interval orders with two interval lengths", preprint, 2017. arXiv

[Doignon 1987] J.-P. Doignon, "Threshold representations of multiple semiorders", *SIAM J. Algebraic Discrete Methods* **8**:1 (1987), 77–84. MR Zbl

[Doignon 1988] J.-P. Doignon, "Sur les représentations minimales des semiordres et des ordres d'intervalles", *Math. Sci. Humaines* **101** (1988), 49–59. MR Zbl

[Fishburn 1970] P. C. Fishburn, "Intransitive indifference with unequal indifference intervals", *J. Mathematical Psychology* **7**:1 (1970), 144–149. MR Zbl

[Fishburn 1983] P. C. Fishburn, "Threshold-bounded interval orders and a theory of picycles", *SIAM J. Algebraic Discrete Methods* **4**:3 (1983), 290–305. MR Zbl

[Fishburn 1985] P. C. Fishburn, *Interval orders and interval graphs: a study of partially ordered sets*, John Wiley and Sons, Chichester, 1985. MR Zbl

[Fishburn and Graham 1985] P. C. Fishburn and R. L. Graham, "Classes of interval graphs under expanding length restrictions", *J. Graph Theory* **9**:4 (1985), 459–472. MR Zbl

[Fishburn and Monjardet 1992] P. Fishburn and B. Monjardet, "Norbert Wiener on the theory of measurement (1914, 1915, 1921)", *J. Math. Psych.* **36**:2 (1992), 165–184. MR Zbl

[Golumbic and Trenk 2004] M. C. Golumbic and A. N. Trenk, *Tolerance graphs*, Cambridge Studies in Advanced Mathematics **89**, Cambridge University Press, 2004. MR Zbl

[Isaak 2009] G. Isaak, "Interval order representations via shortest paths", pp. 303–311 in *The mathematics of preference, choice and order*, edited by S. J. Brams et al., Springer, 2009. MR Zbl

[Pirlot and Vincke 1997] M. Pirlot and P. Vincke, *Semiorders: properties, representations, applications*, Theory and Decision Library **36**, Kluwer, Dordrecht, 1997. MR Zbl

[Schrijver 2003] A. Schrijver, *Combinatorial optimization: polyhedra and efficiency*, Algorithms and Combinatorics **24**, Springer, 2003. MR Zbl

[Scott and Suppes 1958] D. Scott and P. Suppes, "Foundational aspects of theories of measurement", *J. Symb. Logic* **23** (1958), 113–128. MR Zbl

s.boyadzhiyska@fu-berlin.de   *Berlin Mathematical School, Freie Universität Berlin, Berlin, Germany*

gi02@lehigh.edu   *Department of Mathematics, Lehigh University, Bethlehem, PA, United States*

atrenk@wellesley.edu   *Department of Mathematics, Wellesley College, Wellesley, MA, United States*

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LATEX but submissions in other varieties of TEX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTEX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. 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.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve