THE HEISENBERG GROUP HAS RATIONAL GROWTH IN ALL GENERATING SETS

MOON DUCHIN AND MICHAEL SHAPIRO

Abstract. A group presentation is said to have rational growth if the generating series associated to its growth function represents a rational function. A long-standing open question asks whether the Heisenberg group has *rational growth for all finite generating sets*, and we settle this question affirmatively. We also establish *almost-convexity for all finite generating sets*. Previously, both of these properties were known to hold for hyperbolic groups and virtually abelian groups, and there were no further examples in either case. Our main method is a close description of the relationship between word metrics and associated Carnot-Carathéodory Finsler metrics on the ambient Lie group. We provide (non-regular) languages of geodesics in any word metric that suffice to represent all group elements.

1. Introduction

Growth functions of finitely-generated groups count the number of elements that can be spelled as words in a generating alphabet, as a function of spelling length. Though the functions themselves depend on a choice of generating set, they become group invariants under the standard equivalence relation that allows affine rescaling of domain—in particular, this preserves the property of having polynomial growth of a particular degree.

It has been known since the early 1970s that all nilpotent groups have growth functions in the polynomial range, in fact bounded above and below by polynomials of the same degree, and the degree was computed by Bass and Guivarc'h independently [1, 16, 17]. A breakthrough theorem of Gromov states that in fact any group with growth bounded above by a polynomial is virtually nilpotent [14].

One can still wonder, however, whether the growth function is precisely polynomial. This turns out to be a bit too much to ask for nilpotent groups. Virtually abelian groups, for instance, have a more general property called *rational growth*: no matter what finite generating set is chosen, the power series associated to the growth function represents a rational function.

Hyperbolic groups have rational growth for all generators—this is an important theorem from the early 1980s for which credit can be shared among Cannon, Thurston, and Gromov [7, 8, 12, 15]. (This has an interesting history: Cannon's argument for fundamental groups of closed hyperbolic manifolds directly generalized to hyperbolic groups once that definition was in place. And Thurston's definition of automatic groups was partly motivated by these ideas.) At almost the same time, Benson established the same result for virtually abelian groups [2]. Given the work

Date: September 20, 2017.
2010 *Mathematics Subject Classification.* Primary 20F65; Secondary 20F18, 20F69.
Key words and phrases. Growth of goups, Heisenberg group.

at the time understanding the growth of nilpotent groups, it was a natural question to ask whether nilpotent groups also have rational growth, which was open even for the simplest non-abelian nilpotent group, the *integer Heisenberg group*. This question was posed or referred to by many authors, including [18, 13, 3, 24, 20, 25]. By the late 1980s, Benson and Shapiro had independently established one piece of this: the Heisenberg group has rational growth in its standard generators. We settle the full question here.

Theorem 1. The Heisenberg group has rational growth for all generating sets.

In the process of establishing this fact, we will get quite precise information about the combinatorial geometry of Heisenberg geodesics (Theorem 25) that will be useful in the further study of the geometric group theory of $H(\mathbb{Z})$, and should therefore have applications to complex hyperbolic lattices with Heisenberg cusps. We give remarks, applications (including almost-convexity), and open questions in the last section.

1.1. **Literature.** Let us review what is known about rationality of growth in groups and classes of groups.

For all S	For at least one S	For no S
hyperbolic groups	some automatic groups	unsolvable word problem
virtually abelian groups	Coxeter groups, standard S	intermediate growth
Heisenberg group H	~~H, standard S~~	
	H_5, cubical S	
	$BS(1,n)$, standard S	

Note that the inclusion of group with unsolvable word problem in the table above is made under the assumption that the group is recursively presented. It is apparently an open question whether a group that is not recursively presented can have rational growth. A result either way should prove fascinating.

Automatic groups have rational growth when the automatic structure consists of geodesics. In this case, there is a regular language of geodesics that bijects to the group; this is used in [21] to study groups that act geometrically finitely on hyperbolic space. There are more examples belonging in the middle category—known to have rational growth in a special generating set—found in work of Barré (quotients of triangular buildings), Alonso (amalgams), Brazil and Freden et al (other Baumslag-Solitar groups), Johnson (wreath products and torus knot groups), and others. For references and an excellent survey, see [13].

The nilpotent cases go as follows. As mentioned above, [3, 24] show that H has rational growth in standard generators. In [25], Stoll proves the following remarkable result: the higher Heisenberg group H_5 has transcendental growth in its standard generators, but rational growth in a certain dual generating set, which we will call *cubical generators*. (See Sec 3.3 for a definition of H_5.) On the other hand, Stoll establishes the following theorem to use as a criterion for non-rational growth.

Theorem 2 (Stoll [25])**.** If $\frac{\beta(n)}{\alpha \cdot n^d} \to 1$ and α is an irrational (resp. transcendental) number, then $\mathbb{B}(x) = \sum \beta(n) x^n$ is a non-rational (resp. transcendental) function.

A volume computation gives $\alpha = \frac{6027+2\ln 2}{65610}$, establishing that (H_5, std) has transcendental growth. Over fifteen years later, this (with small variations explained

by Stoll) still provides the only known example of a group with both rational and irrational growth series.

Acknowledgments. The authors have been thinking about this problem for a long time and have many people to thank for interesting ideas and stimulating conversations. Particular thanks go to Christopher Mooney, to Cyril Banderier, and to Laurent Bartholdi for initially communicating the interest of this problem. We thank Dylan Thurston for stimulating discussions of computational aspects, including calculations of periods and coefficients of quasipolynomiality.

The first author is supported by NSF grants DMS-1207106 and DMS-1255442.

MS wishes to thank MD for inviting him on such a grand adventure.

2. Outline

2.1. **Geometric overview.** In this paper, we will give a way to compare geodesics in the Cayley graph of $H(\mathbb{Z})$ with geodesics in a geometrically much simpler continuous metric. We regard $H(\mathbb{Z}) \leq H(\mathbb{R})$ as a subset and take $\mathbb{R}^3$ coordinates on $H(\mathbb{R})$. We can then treat spellings in the word metric as piecewise linear paths in $\mathbb{R}^3$. We will show that every group element in $H(\mathbb{Z})$ is reached by a geodesic spelling that is boundedly close to a path from one of two families that are completely understood. Furthermore, since these comparison paths in $\mathbb{R}^3$ are determined by their projection into $\mathbb{R}^2$, this ultimately allows us to use planar pictures to understand geodesics in $H(\mathbb{Z})$. The comparison geodesics are polygonal paths in normed planes and are characterized by the relationship between their length and the area that they enclose.

By an important theorem of Pansu [22], any word metric on the Heisenberg group $H(\mathbb{Z})$ is asymptotic to a left-invariant metric on its ambient Lie group $H(\mathbb{R})$, known as a Carnot-Carathéodory (cc) Finsler metric, which admits $\mathbb{R}^3$ coordinates. (Pansu's theorem is much more general, and was further generalized by Breuillard in [4].) Several authors have studied the geodesics in these cc metrics, including Krat, Stoll, Breuillard, and Duchin–Mooney, and we will stay close to the notation of [11]. It is to Pansu's cc geodesics, which come in "regular" and "unstable" types, that we will compare our word geodesics. (See Section 3 for full background.) In this language, the main geometric result of the paper (Theorem 25) is that for every generating set on $H(\mathbb{Z})$, there is a bound so that *every group element has a geodesic spelling that is boundedly close to a* cc *geodesic.*

To show this, an important ingredient is to identify curves in normed planes that are short relative to their enclosed area. A classical theorem of Busemann identifies which the optimal polygonal curves, and we will need to analyze (in Section 4) how the area falls off from optimality when the proportions of the polygonal curves are perturbed.

2.2. **Language-theoretic overview.** The conclusion of the strategy can be described in language-theoretical terms. Given a generating set S for $H(\mathbb{Z})$, we describe two languages $\mathcal{L}_{\mathcal{S}}, \mathcal{L}_{\mathcal{P}} \subset S^*$ (corresponding to the special families of geodesics from the geometric description above) which have the following properties with respect to $H(\mathbb{Z})$ and its abelianization $\mathbb{Z}^2$.

The language $\mathcal{L}_{\mathcal{S}}$ is made up of images of finitely many "shapes," which are functions ω from domains in $\mathbb{Z}^M$ to S^*. The language $\mathcal{L}_{\mathcal{P}}$ is made up of finitely many "patterns" $\mathtt{w}$ that likewise map from domains in $\mathbb{Z}^M$ to spellings in S^*. In

both cases, the length ℓ of the spelling and its image in abelianization $(a, b) \in \mathbb{Z}^2$ are each affine functions of the $\mathbb{Z}^M$ input. Properties of shapes and patterns are established in Sections 5 and 6, respectively. This allows us to show in Section 7 that every group element has a geodesic spelling produced by these languages. That is, if $\mathcal{G}$ is the language of geodesics in $H(\mathbb{Z})$ with respect to S, then $(\mathcal{L}_\mathcal{S} \cup \mathcal{L}_\mathcal{P}) \cap \mathcal{G}$ surjects onto $H(\mathbb{Z})$ by evaluation.

Finally, we will prove a series of *rational competition lemmas* in Section 8. the length and abelianization tell us when two shapes or patterns are candidates to produce the same group element geodesically; when two candidates compete, the winner can be identified by a linear inequality.

Putting these facts together, we find that the domain on which a shape or pattern winningly represents a group element is determined by linear equalities, inequalities, and congruences in $\mathbb{Z}^M$, and so counting the spellings enumerated by the shapes amounts to solving congruences in rational polyhedra. By a marvelous theorem of Benson [3], given here as Theorem 4, we are then done establishing the rationality of the growth, because *enumeration over rational polyhedra yields a rational function.*

The final section records some other applications of this geometric/combinatorial machinery, and offers some open questions.

2.3. **Example: Shapes in $\mathbb{Z}^2$.** To illustrate the idea of shapes of geodesics, consider the example of $\mathbb{Z}^2$, first with standard generators a, b. Here, we will introduce four shapes: $a^m b^{-n}$, $a^{-m} b^n$, $a^m b^n$ and $a^{-m} b^{-n}$.

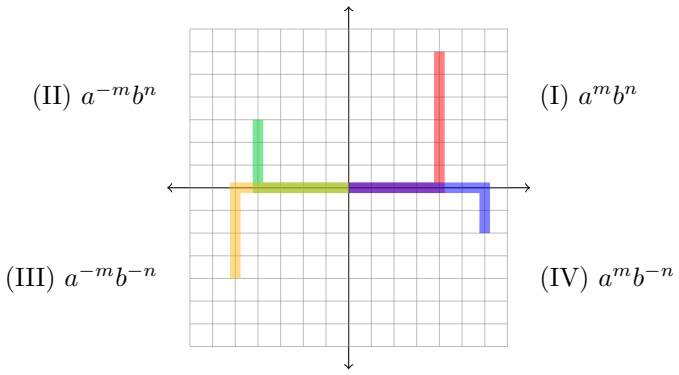


FIGURE 1. Four shapes of geodesics for $(\mathbb{Z}^2, \mathsf{std})$. (Take $m, n \geq 0$ in each case.)

One quickly observes a few basic properties:

- There are finitely many shapes.
- Each shape is a language, and a map from a subset of some $\mathbb{Z}^M$ to S^*. (Here, $M = 2$ for each shape, and the domain is the first quadrant of $\mathbb{Z}^2$.)
- Every group element admits a geodesic spelling by at least one shape (even though not every geodesic is realized this way).

This case is too simple to capture some features of the situation, so consider the slightly more complicated case of $\mathbb{Z}^2$ with *chess-knight* generators $\{(\pm 2, \pm 1), (\pm 1, \pm 2)\}$. Consider the case of geodesically spelling the group element $(100, 100)$. If we label the generators counterclockwise with $a_1 = (2, 1)$ and $a_2 = (1, 2)$, then $a_1^{33} a_2^{33}$

reaches the adjacent position $(99, 99)$ in $\mathbb{Z}^2$, but an exact spelling—in fact a geodesic spelling—requires two more letters: $(a_3a_8)a_1^{33}a_2^{33}$. (This is because of the well-known property of chess-knights that it takes several moves to arrive at an adjacent square on the chessboard.) These correction terms never have more than three letters, so we can arrive at a finite list of shapes: every shape has the form $x \cdot a_i^{n_1} a_{i+1}^{n_2}$, where x is in the ball of radius three, a_i and a_{i+1} are cyclically successive generators, and $n_1, n_2 \geq 0$. For a given shape, the length of the spelling is $\ell(x) + n_1 + n_2$, and the position of its endpoint is the position of x plus the n_1, n_2 combination of the positions of a_i and a_{i+1}. Also note that for large enough words, two shapes can only compete to represent the same group element if their indices i and i' differ by at most one. Between two competitors, the winner can be determined by checking whether the positions are equal and which was attained at shorter length. (Assume that ties are broken for the lower-indexed shape in some arbitrary order.)

Now we can add to the list of properties:

- A shape may not evaluate to a geodesic for every value of its arguments.
- The spelling lengths and positions of the evaluations of any shape or pattern into $\mathbb{Z}^2$ are affine functions of n_1, n_2.
- Each shape has only finitely many possible competitors, and the positions reached winningly by a shape at a given length are described by finitely many linear equalities and inequalities.

This serves as a good roadmap for the sequence of features we will establish in the Heisenberg group.

3. Background

3.1. **Growth of groups.** Suppose a group G is generated by the finite symmetric generating set $S = S^{-1}$. We take S^n to be the set of all (unreduced) strings of length n in the elements of S (sometimes called *spellings*) and $S^* = \cup_{n=0}^{\infty} S^n$ to be the set of all spellings of any finite length. This S^* comes equipped with two important maps, *spelling length* and *evaluation* into G. Length, denoted $\ell(\gamma)$, is defined on $\gamma \in S^n \subset S^*$ via $\ell(\gamma) = n$. Evaluation into G is given by the monoid homomorphism which carries concatenation in S^* to group multiplication in G. An element of S^* can be thought of as a path in the Cayley graph $\mathsf{Cay}(G, S)$ from e to the evaluation of γ.

We define the *word length* of a group element $g \in G$ by

$$|g| = |g|_S = \min\{\ell(\gamma) \mid \gamma \in S^* \text{ and } \overline{\gamma} = g\},$$

i.e., the shortest spelling length of any spelling.

The *sphere* and *ball* of radius n are denoted S_n, B_n respectively, and the associated growth functions are

$$\sigma(n) := \#S_n = \#\{g \in G : |g| = n\} \ ;$$
$$\beta(n) := \#B_n = \#\{g \in G : |g| \leq n\} \ ,$$

related of course by $\sigma(n) = \beta(n) - \beta(n-1)$. Then we can form associated generating functions, called the *spherical growth series* and the *growth series* of (G, S), as follows:

$$\mathbb{S}(x) := \sum_{n=0}^{\infty} \sigma(n) x^n \ ; \qquad \mathbb{B}(x) := \sum_{n=0}^{\infty} \beta(n) x^n.$$

Since $\sigma(n) \leq \beta(n) \leq \sum_{i=0}^{n} |S^i| = \sum_{i=0} |S|^i$, the coefficients are bounded above by an exponential, ensuring a positive radius of convergence for both series. We recall that the growth rate of a group is well defined up to an equivalence relation $\asymp$ given by linear rescaling of domain and range. This sets the stage for one of the breakthrough theorems in the current explosion of geometric group theory, which characterizes nilpotent groups by their growth: Gromov established in the early 1980s [14] that $\beta(n) \asymp n^d$ if and only if the group is virtually nilpotent.

Instead of the rate of growth of $\beta(n)$, it can also be productive to study its finer arithmetic properties. We say that (G, S) has *rational growth* if the growth series are rational functions: $\mathbb{S}(x), \mathbb{B}(x) \in \mathbb{Q}(x)$. (That is, if each is a ratio of polynomials in x.) Note that the relationship between σ and β implies that $(1-x)\mathbb{B}(x) = \mathbb{S}(x)$, so either is rational iff the other is.

It is a standard fact that rationality of a generating function $F(x) = \sum f(n)x^n$ is equivalent to the property that the values $f(n)$ satisfy a finite-depth linear recursion for $n \gg 1$, i.e., there exist N_0 and P such that for $n > N_0$,

$$f(n+P) = a_0 \cdot f(n) + a_1 \cdot f(n+1) + \cdots + a_{P-1} \cdot f(n+P-1).$$

(Here, the coefficients a_i come from the same base field as the polynomials in the rational function.)

The growth of a regular language is well known to be rational with integer coefficients, and therefore the values $\sigma(n)$ satisfy an integer recursion. In fact, this recursion can be described in terms of the finite-state automaton which accepts the language, and therefore can be written with non-negative integer coefficients in the recursion. In the case of groups, if there is a generating set for which there is a regular language of geodesics which bijects to the group, then the corresponding growth function is rational. This can be used to prove rational growth for free abelian groups and for word hyperbolic groups.

In this paper we focus on the integer Heisenberg group $H = H(\mathbb{Z})$ and consider its growth functions with various finite generating sets. Shapiro 1989 [24] shows that for the standard Heisenberg generators $S = \mathsf{std}$, there is no regular language of geodesics for (H, std). Nevertheless, the growth function is rational [24, 3]. In this paper, we will show the same holds for arbitrary generating sets.

3.2. **Rational families.** We now review material from Max Benson's papers [2, 3], articulating the principle that *counting in polyhedra is rational*.

Suppose we have a parameter n which we will take to lie in the non-negative integers and we consider sets of points $E(n) \subset \mathbb{Z}^d$ defined by finitely many equalities, inequalities, and congruences

$$\begin{cases} \mathsf{a}_i \cdot \mathsf{x} = b_i(n) \ ; \\ \mathsf{a}_j \cdot \mathsf{x} \leq b_j(n) \ ; \\ \mathsf{a}_k \cdot \mathsf{x} \equiv b_k(n) \pmod{c_k} \ , \end{cases}$$

where each a_i, a_j and a_k are in $\mathbb{Z}^d$, and each b_i, b_j and b_k is an affine function of n with integer coefficients. Such a sequence of sets $\{E(n)\}$ is called an *elementary family*. Benson defines a *polyhedral family* $\{P(n)\}$ to be a finite union of finite intersections of elementary families. If each $P(n)$ is bounded, then $\{P(n)\}$ is called a *bounded polyhedral family*.

For example, for the chess-knight language $\{a_3a_8a_1^{n_1}a_2^{n_2} : n_1, n_2 \geq 0\}$ described in Section 2.3, the set of positions reached geodesically at length n is the (x, y) such that $x+y = 3n+1$ and $0 \leq x/2 \leq y \leq 2x$, which is a (bounded) elementary family.

Lemma 3. The class of polyhedral families is closed under complementation, union, intersection, and set difference.

Proof. This is clear for union, from the definition, and for intersection, by taking the combined system of defining equalities, inequalities and congruences.

We now consider complementation. The complement of the solution set of an equation is the disjoint union of the solution sets of two inequalities. For an inequality, the complement of its solution set is given by a single inequality. For a congruence mod r, the complement of its solution set is the disjoint union of solutions to $r-1$ congruences.

Finally, set difference can be built with intersection and complementation. □

Note also that the class of polyhedral families is closed under affine push-forward; if $\{P(n)\}$ is a polyhedral family in $\mathbb{Z}^d$ and $g : \mathbb{Z}^d \to \mathbb{Z}^m$ is an affine map, then $\{g(P(n))\}$ is a polyhedral family in $\mathbb{Z}^m$. The push-forward of a bounded family is bounded. The pull-back of a bounded family is bounded if the linear part has trivial kernel.

Theorem 4 (Counting over polyhedral families [2, 3])**.** Suppose that $\{P(n)\}$ is a bounded rational family in $\mathbb{Z}^d$ and $f : \mathbb{Z}^d \to \mathbb{Z}$ is a polynomial with integer coefficients. Then

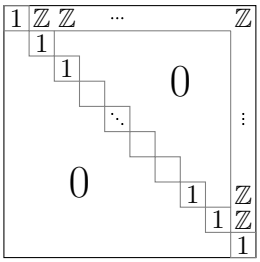
$$F(x) = \sum_{n=0}^{\infty} \sum_{\mathsf{v} \in P(n)} f(\mathsf{v})\, x^n$$

is a rational function of x.

The hypotheses let you apply an arbitrary polynomial to the points of a family of sets defined linearly; the special case that f is constant is the lattice point count in the polyhedral family, so this theorem can be thought of as generalizing Ehrhart's theorem about lattice point counts in polyhedral dilates. We will need to apply the theorem with a quadratic f, so we need its full strength. (We note that the linearity in $P(n)$ and the polynomiality in f can not be exchanged: enumeration over regions defined by quadratics would *not* in general give a rational series.)

Benson's counting theorem is a powerful tool and one of the few known methods for establishing rationality of growth series. He used this theorem in [2] to show that virtually abelian groups have rational growth with respect to arbitrary generating sets, and he made partial progress extending his analysis to the Heisenberg group in [3].

3.3. **The Heisenberg groups.** Most of this paper will focus on the Heisenberg group $H(\mathbb{Z})$, which is also the first in the family H_k, $k = 3, 5, 7, \ldots$ of two-step nilpotent groups realized as



inside the $N \times N$ matrices, where $N = \frac{k+3}{2}$. (This parametrization has k as the number of integer parameters in each matrix.) For $i = 1, 2, \ldots, N-2$, let a_i be the $(1, i+1)$ elementary matrix, let b_i be the $(N, i+1)$ elementary matrix, and write c for top-right elementary matrix. Then we have the commutator relations $[a_i, b_i] = c$ and all other commutators are trivial. Thus for any k, the commutator subgroup is $\langle c \rangle$, so that the lower central series is

$$1 \trianglelefteq \mathbb{Z} \trianglelefteq H_k.$$

The well-known Bass-Guivarc'h formula for the degree of polynomial growth in nilpotent groups tells us that the growth function of H_k satisfies $\beta(n) \asymp n^d$ for $d = (k-1)\cdot 1 + 1 \cdot 2 = k+1$.

For the Heisenberg group $H(\mathbb{Z})$, we will drop the subscripts and write the elementary matrices as $\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3$, so that $[\mathsf{e}_1, \mathsf{e}_2] = \mathsf{e}_3$ and $[\mathsf{e}_1^m, \mathsf{e}_2^n] = \mathsf{e}_3^{mn}$. The standard generating set for $H(\mathbb{Z})$ is $\{\mathsf{e}_1, \mathsf{e}_2\}^{\pm 1}$, and from the above formula we know that the growth function in these generators is bounded above and below by fourth-degree polynomials.

3.4. **Geometric model, spelling paths, and boost.** We will use the *exponential coordinates* on $H(\mathbb{Z}) \leq H(\mathbb{R})$ given by the following representation:

$$(a, b, c) \leftrightarrow \begin{pmatrix} 1 & a & c + \frac{1}{2}ab \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix}.$$

These coordinates have the property that $(a, b, c)^n = (na, nb, nc)$, and in this notation $\mathsf{e}_1 = (1, 0, 0)$, $\mathsf{e}_2 = (0, 1, 0)$, and $\mathsf{e}_3 = (0, 0, 1)$.

For integers a and b, define $\epsilon(a, b)$ to be $1/2$ if a and b are both odd, and 0 otherwise. In these coordinates, $H(\mathbb{Z})$ looks just like the standard lattice $\mathbb{Z}^3 \subset \mathbb{R}^3$ shifted by ϵ in the z direction, and the Haar measure on $H(\mathbb{R})$ is Lebesgue measure in $\mathbb{R}^3$.

The real Heisenberg group is equipped with a natural projection $H(\mathbb{R}) \to \mathbb{R}^2$ via the quotient by its center. In coordinates, this takes $(x, y, z) \mapsto (x, y)$. We will use $\mathfrak{m}$ to denote $\mathbb{R}^2$ with a norm, sitting in the tangent bundle $TH(\mathbb{R})$.

Definition 5. A *spelling path* is a string of letters from any given generating set S, i.e., an element of S^*, regarded as a path in the Cayley graph that represents a group element from $H(\mathbb{Z})$. (Assume these paths are based at the identity unless specified otherwise.) Define ℓ, (a, b), and z to be the *length*, *horizontal position*, and *height* of γ, respectively: if the group element represented by γ is (a, b, c), then $\ell(\gamma)$ is the spelling length of the string, $(a, b) \in \mathfrak{m}$ is the projection of the endpoint, and the *height* of the group element and hence the path is $z(\gamma) = c$. Also let the *shadow*,

denoted $\pi(\gamma)$, be the projection to $\mathfrak{m}$ (the path in $\mathfrak{m}$ obtained by concatenating the projections of the generators to $\mathfrak{m}$ in the order of appearance in γ).

Define the area of a spelling path γ, denoted $z_A(\gamma)$, to be the balayage area of its projection, that is, the signed area of the concatenation of $\pi(\gamma)$ with the chord between its endpoint and 0. The *boost* of a generating letter a_i is its height $z(a_i)$. Then the boost of a spelling, denoted $z_b(\gamma)$, is the sum of the boosts of the letters in the spelling.

Note that the height of a spelling path is equal to its balayage area plus its boost: $z(\gamma) = z_b(\gamma) + z_A(\gamma)$.

Here is a brief example to track through the definitions in this section. Suppose a generating set includes the letters $a_1 = (1, 2, 1)$ and $a_2 = (0, 1, 2)$. Then the spelling path $\gamma = a_1 a_2$ evaluates to $(1, 3, 3.5)$, as one can check by matrix multiplication:

$$\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 5 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}.$$

The path has length 2, horizontal position $(1, 3)$, and height 3.5. Of that height, 3 units come from boost $z_b(\gamma) = z_b(a_1) + z_b(a_2) = 1 + 2$ and the remaining height is the area $A(\gamma) = .5$ of the triangle with vertices $(0, 0), (1, 2), (1, 3)$.

3.5. CC **metrics and Pansu's theorem.** As mentioned above, Pansu's theorem states that the large-scale structure of the Cayley graph (H, S) gives a metric on $H(\mathbb{R})$. This is not a Riemannian metric, but rather a *sub-Finsler metric* called a CC metric. See [4, 11] for some explicit descriptions of the geometry of the limit metric, and [10] for general background on sub-Riemannian geometry and the Heisenberg group. We collect a few salient features here.

The CC metrics are defined as follows. Let $\mathfrak{m}$ denote the horizontal subspace of the Lie algebra $\mathfrak{h}$ of $H(\mathbb{R})$; that is, the span of the tangent vectors $X = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$ and $Y = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{smallmatrix}\right)$ at the identity, and identify $\mathfrak{m}$ with the xy–plane in $\mathbb{R}^3$ in exponential coordinates. We can regard $\mathfrak{m}$ as a copy of $\mathbb{R}^2$ and make use of the linear projection $\pi : H(\mathbb{R}) \to \mathfrak{m}$ given by $(a, b, c) \mapsto (a, b)$.

Fix a centrally symmetric convex polygon $L \subset \mathfrak{m}$; this uniquely defines a norm $\|\cdot\|_L$ on $\mathfrak{m}$ for which L is the unit sphere. The push-forwards of $\mathfrak{m}$ by left multiplication give *admissible planes* $\langle dL_g(X), dL_g(Y)\rangle$ at every point $g \in H(\mathbb{R})$, which are similarly normed; the plane field is a sub-bundle of the tangent bundle to $H(\mathbb{R})$. We say that a curve in $H(\mathbb{R})$ is *admissible* if it is piecewise differentiable and all of its tangent vectors lie in these normed planes. The length of an admissible curve is simply the integral of the lengths of its tangent vectors, and it is easily verified that this is the same as the length in the L–norm of the projection $\pi(\gamma)$, and that any two points are connected by an admissible path. Then the CC distance $d_{\text{CC}}(x, y)$ is (well-)defined as the infimal length of an admissible path between x and y. One easily checks that this is a geodesic metric.

In exponential coordinates, all CC metrics are equipped with a dilation $\delta_t(a, b, c) = (ta, tb, t^2 c)$ that is a metric similarity, scaling lengths and distances by t, areas in $\mathfrak{m}$ by t^2, and volumes by t^4.

Pansu also tells us which polygon L is induced by a generating set S: namely, L is the boundary of the convex hull of the projection $\pi(S)$ of the generators to $\mathfrak{m}$. For example, the two most basic generating sets for $H(\mathbb{Z})$ are $\{\mathsf{e}_1, \mathsf{e}_2\}^{\pm}$ and $\{\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3\}^{\pm}$. In either case, the CC metric is induced by the L^1 norm on $\mathfrak{m}$. By

contrast, if one took the nonstandard generators $\{\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_1\mathsf{e}_2\}^{\pm}$, the polygon would be a hexagon $L = \hexagon$.

In this language, we can state this special case of Pansu's theorem as follows: for any finite symmetric generating set S of $H(\mathbb{Z})$,

$$\lim_{x\to\infty} \frac{d_{\text{CC}}(x,0)}{|x|_S} \to 1.$$

While Pansu's result extends to a statement for all nilpotent groups, there is a substantial strengthening due to Krat [19] which was shown only in the case of $H(\mathbb{Z})$: there is a global bound (depending on S) in the *additive* difference between word and CC lengths: $\sup_x \left| d_{\text{CC}}(x,0) - |x|_S \right| < \infty$. In Section 5.2, we will give a new proof of Krat's (and therefore Pansu's) result for $H(\mathbb{Z})$. We note that Breuillard ([4]) has shown that bounded difference does *not* hold for all 2-step nilpotent groups, though on the other hand he has explained to us that arguments from [5] can be adapted to show bounded difference for all of the higher Heisenberg groups.

3.6. **Significant directions, isoperimetrices, structure of CC geodesics.** It is a standard fact in Heisenberg geometry that for any admissible path γ based at the origin $\mathsf{0} \in H(\mathbb{R})$, the height or z coordinate of $\gamma(t)$ is equal to its *balayage area*: the signed (Lebesgue) area enclosed by the concatenation of the curve's shadow $\pi(\gamma)$ with a straight line segment connecting its endpoints. This is a simple consequence, via Stokes' Theorem, of the fact that the description of the horizontal planes ensures that $\gamma_3' = \frac{1}{2}(\gamma_1\gamma_2' - \gamma_2\gamma_1')$.

As a consequence of the connection between height and balayage area, we have a criterion for geodesity in the CC metric: a curve γ in $\mathfrak{m}$ based at $(0,0)$ lifts to a geodesic in $H(\mathbb{R})$ iff its L-length is minimal among all curves with the same endpoints and enclosing the same area. As a result, to classify geodesics one uses the solution to the isoperimetric problem in the normed plane $(\mathfrak{m}, \|\cdot\|_L)$. By a classical theorem of Busemann from 1947 [6], the solution is described in terms of a polygon which he called the *isoperimetrix*.

Definition 6. For a finite symmetric generating set S, let $Q = \text{CHull}(\pi(S))$ be the convex hull of the projection of S to $\mathfrak{m}$ and let L be its boundary polygon, as above. The *polar dual* of Q is defined as $Q^* = \{v \in \mathfrak{m} : v \cdot x \leq 1 \quad \forall x \in Q\}$ with respect to the standard dot product. Busemann's *isoperimetrix* is the polygon $\partial(e^{i\pi/2}Q^*)$, obtained by rotating the polar dual of Q through a right angle.

Definition 7. The vertices of the polygon L will be labelled cyclically as $\mathsf{a}_1, \ldots, \mathsf{a}_{2k}$ and these vectors will be called *significant directions*. For the significant directions, we will extend the subscripts periodically by defining a_m to equal a_n if $m \equiv n \pmod{2k}$.

Each significant direction is the shadow of at least one *significant generator* in S and we will label the generators projecting to a_i as a_i, a_i', a_i'', etc. Elements of S which project to the edges of L are called *edge letters* and those that project properly inside L are called *interior letters*.

Remark. We will maintain this font distinction as much as possible to mark the difference between group elements $a \in H$ and their corresponding projections $\mathsf{a} \in \mathfrak{m}$, the latter thought of as vectors in the plane.

With this terminology, Busemann's theorem can be stated as follows. Use Lebesgue measure on $\mathbb{R}^2$ for area; use length in the Minkowski norm for perimeter. Then up to dilation and translation *the isoperimetrix is the unique closed curve realizing the maximal value of area divided by perimeter-squared.* The following properties follow from Busemann's construction.

- If the vertices of Q have rational coordinates, then the same is true of the vertices of the isoperimetrix.
- The edges of the isoperimetrix are parallel to the significant directions.

It follows that by clearing common denominators we can find positive integers $\sigma_1, \ldots, \sigma_{2k}$ with gcd $= 1$ and an integer λ such that the edge vectors of $\lambda\partial(e^{i\pi/2}Q^*)$ are $\sigma_i \mathsf{a}_i$.

Definition 8. Define the *standard isoperimetrix* to be the closed polygon $\mathbf{I} = \mathbf{I}(S)$ having vertices

$$\mathsf{0}, \quad \sigma_1\mathsf{a}_1, \quad \sigma_1\mathsf{a}_1 + \sigma_2\mathsf{a}_2, \quad \ldots$$

(This is a translated and scaled copy of Busemann's curve.)

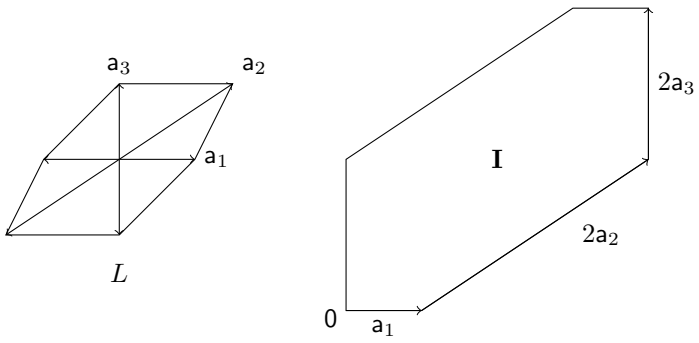


FIGURE 2. This example shows an isoperimetrix which is twice the rotated polar dual of the original polygon. Here $\mathbf{I}$ has perimeter 10, measured in the L–norm, and it is the unique shape maximizing area at that perimeter (up to translation).

CC geodesics based at $\mathsf{0}$ are classified in [11] into two kinds: *regular geodesics*, which project to $\mathfrak{m}$ as an arc of an isoperimetrix, and *unstable geodesics*, which project to $\mathfrak{m}$ as geodesic in the L–norm.

Fix a polygon L in $\mathfrak{m}$, which determines a CC metric on $H(\mathbb{R})$. Then for any $(a, b) \in \mathfrak{m}$, each length $\ell \geq \|(a, b)\|_L$ uniquely determines a height $c = c(a, b, \ell) \geq 0$ so that there exists a regular geodesic connecting $\mathsf{0}$ to (a, b, c) at length ℓ. That is, for each ℓ there exists a scale s and a translation vector q so that $s\mathbf{I} + \mathsf{q}$ passes through $\mathsf{0}$ and (a, b); the subarc between those two points has length ℓ with respect to L and encloses area c, and it lifts to a CC geodesic. On the other hand if $\ell = \|(a, b)\|_L$, there are L–norm geodesics connecting $\mathsf{0}$ to (a, b) with length ℓ, and these can enclose any area in an interval of possibilities.

The term "stable geodesics" is borrowed from the notion of "stability of geodesics" meaning that for each pair of endpoints, geodesics between the endpoints fellow travel with some fixed constant (as in the Morse Lemma in hyperbolic geometry). With limitted exceptions, our stable geodesics will fellow travel in projection. Here,

CC geodesics of the second type are called "unstable" because they are highly non-unique and geodesics for nearby points, or even the same point, can fail to fellow travel arbitrarily badly.

4. Paths in planar norms

Below, we will use multiplicative vector notation for polygonal paths in the plane, so that for instance $\mathsf{v}_1^t\mathsf{v}_2\mathsf{v}_1$ denotes the concatenated path obtained by starting with the vector $t\mathsf{v}_1$ followed by the vector v_2 followed by the vector v_1, with total displacement vector $(t+1)\mathsf{v}_1+\mathsf{v}_2$ from beginning to end. In this path notation, the exponents need not be integers. We will be considering areas of polygonal paths, where we will define the area of a not-necessarily-closed path to be its balayage area: the signed area enclosed by concatenating the path with the chord from its endpoint to its start point.

The path $P = \mathsf{v}_1\mathsf{v}_2\dots\mathsf{v}_r$ represents a closed polygon if $\sum \mathsf{v}_i = \mathsf{0}$; it is convex if the vectors $\mathsf{v}_1,\dots,\mathsf{v}_r$ are cyclically ordered (that is, if their arguments proceed in a monotone fashion once around the circle) and strictly convex if and only if no two cyclically successive vectors are parallel. For a fixed polygonal norm, the associated isoperimetrix can be written in this notation with $r = 2k$ and $\mathsf{v}_i = \mathsf{a}_i^{\sigma_i}$, where the a_i are extreme points in the polygon that is the norm's unit ball and the σ_i are integers described in the previous section. If $P = \mathsf{v}_1\mathsf{v}_2\dots\mathsf{v}_r$ arises as an isoperimetrix, then it is strictly convex, centrally symmetric, and its sides have lengths induced by the norm, which we will denote $\ell_i = \|\mathsf{v}_i\|$. In this section we will make several elementary geometric arguments about polygons and polygonal paths, particularly those arising as isoperimetric polygons.

Given a polygonal norm and its isoperimetrix $P = \mathsf{v}_1\dots\mathsf{v}_r$, define a *Busemann arc* to be a simple path based at $\mathsf{0}$ which is a subarc of some scaled and translated copy of P. With the convention of reading indices mod r, these are precisely given by polygonal paths of the form $\tau = \mathsf{v}_{\mathrm{I}}^{s^-}\mathsf{v}_{\mathrm{I}+1}^{s}\dots\mathsf{v}_{\mathrm{J}-1}^{s}\mathsf{v}_{\mathrm{J}}^{s^+}$ with at most $r+1$ terms, satisfying $0 \le s^-, s^+ \le s$ (and $s^- + s^+ \le s$ if there are $r+1$ terms).

In this case we will say that τ has *scale s and (combinatorial) type* (I,J). There are two possible ambiguities: first, if τ is one- or two-sided, the scale is underdetermined, so we take s to be the maximum exponent. Second, if s^- or s^+ equals 0 or s, then the arc is of more than one combinatorial type; for instance, $\mathsf{v}_3^{100}\mathsf{v}_4^{100}\mathsf{v}_5^{32}$ is of types $(2,5)$ and $(3,5)$. Note that the length of τ in the norm is $s^-\ell_{\mathrm{I}} + s\ell_{\mathrm{I}+1} + \dots + s^+\ell_{\mathrm{J}}$.

In the special case of a Busemann arc $\tau = \mathsf{v}_{\mathrm{I}}^{s^-}\mathsf{v}_{\mathrm{I}+1}^{s}\dots\mathsf{v}_{\mathrm{I}-1}^{s}\mathsf{v}_{\mathrm{I}}^{s^+}$ of type (I,I), a *weight-shifted arc* is any $\widehat{\tau} = \mathsf{v}_{\mathrm{I}}^{t^-}\mathsf{v}_{\mathrm{I}+1}^{s}\dots\mathsf{v}_{\mathrm{I}-1}^{s}\mathsf{v}_{\mathrm{I}}^{t^+}$ where $t^- + t^+ = s^- + s^+$. For a closed P–arc $\tau = \mathsf{v}_{\mathrm{I}}^{s}\mathsf{v}_{\mathrm{I}+1}^{s}\dots\mathsf{v}_{\mathrm{I}-2}^{s}\mathsf{v}_{\mathrm{I}-1}^{s}$ of type $(\mathrm{I},\mathrm{I}-1)$, a *cyclic permutation* of τ is $\bar{\tau} = \mathsf{v}_{\mathrm{J}}^{s}\mathsf{v}_{\mathrm{J}+1}^{s}\dots\mathsf{v}_{\mathrm{I}-2}^{s}\mathsf{v}_{\mathrm{I}-1}^{s}\mathsf{v}_{\mathrm{I}}^{s}\mathsf{v}_{\mathrm{I}+1}^{s}\dots\mathsf{v}_{\mathrm{J}-1}^{s}$ for any J, which can be achieved by iterated weight shifting. Note that these moves preserve area and endpoint (see Figure 3).

These definitions enable us to observe, as a consequence of Busemann's theorem, that isoperimetric arcs in a norm (arcs that enclose maximum area among all arcs of a fixed length) are Busemann arcs, and if any two exist with the same length and endpoint, then they must differ by weight shifting or cyclic permutation. (This follows from the convexity of P; see Duchin–Mooney for details.)

For a polygonal path $P = \mathsf{v}_1\mathsf{v}_2\dots\mathsf{v}_r$, we can consider the *parallel family*

$$\{P_{\mathsf{s}} = \mathsf{v}_1^{s_1}\mathsf{v}_2^{s_2}\dots\mathsf{v}_r^{s_r} \quad : \quad \mathsf{s} = (s_1,\dots,s_r) \in \mathbb{R}^r\}\,.$$

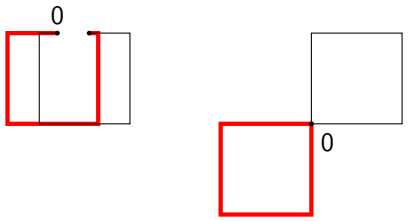


FIGURE 3. Weight shifting (left) and the special case of cyclic permutation (right).

We say that $P = \mathsf{v}_1 \dots \mathsf{v}_r$ is in *positive position* with respect to λ and w if there is a Busemann arc in the parallel family $\{P_\mathsf{s} : s_i \geq 0\}$. (As long as $\lambda \geq \|\mathsf{w}\|$, this can always be ensured by cyclically permuting the order of the vectors if necessary.) Within a parallel family, the Busemann arc for λ, w is unique if it exists, because there are never two paths in the same family that differ by nontrivial weight-shifting.

We can define a displacement vector $\mathsf{w}(P_\mathsf{s}) = \sum s_i \mathsf{v}_i$, and if we are given weights $\ell_i > 0$ (such as the norm-lengths above) we can also define a weighted perimeter $p(P_\mathsf{s}) = \sum s_i \ell_i$. Restricting the parallel family by perimeter $\lambda > 0$ and displacement $\mathsf{w} \in \mathbb{R}^2$, we write

$$\overline{M}(\lambda, \mathsf{w}) := \{\mathsf{s} \in \mathbb{R}^r : p(P_\mathsf{s}) = \lambda,\ \mathsf{w}(P_\mathsf{s}) = \mathsf{w}\}\ ; \qquad M(\lambda, \mathsf{w}) := \overline{M}(\lambda, \mathsf{w}) \cap [0, \infty)^r.$$

Note that $\overline{M}$ is an affine subspace of codimension three in $\mathbb{R}^r$ (the perimeter and the two coordinates of the displacement describe affine hyperplanes). M is therefore a polytope (in fact a simplex) where on each boundary face, some of the $s_i = 0$. If P is a closed polygon, then $M(\lambda, \mathsf{0})$ consists of closed polygons with the same perimeter obtained by moving the sides of P parallel to themselves, so that whenever a pair of successive sides have positive lenths, they have the same interior angles. Thus, for a square, $\{P_\mathsf{s} \mid \mathsf{s} \in M(\lambda, 0)\}$ consists of a set of rectangles, though some of these are degenerate.

Lemma 9 (Isoperimetric optimality)**.** If $P = \mathsf{v}_1\mathsf{v}_2 \dots \mathsf{v}_r$ is an isoperimetrix with $\ell_i = \|\mathsf{v}_i\|$ in the associated norm, then for any $\lambda > 0$ the function $\mathsf{Area}(P_\mathsf{s})$ is quadratic on $\overline{M}(\lambda, \mathsf{0})$ and concave down about a unique local maximum on the interior of $M(\lambda, \mathsf{0})$. If P is in positive position with respect to some λ, w, then the same is true on $M(\lambda, \mathsf{w})$; that is, there are affine coordinates $\mathsf{x} = \mathsf{x}(\mathsf{s})$ such that the area can be written as a negative-definite pure quadratic polynomial plus a constant:

$$\mathsf{Area}(P_\mathsf{s})\big|_{\overline{M}(\lambda,\mathsf{w})} = A - \sum_{i=1}^{r-3} \beta_i x_i^2.$$

Here $A = A(\lambda, \mathsf{w})$ but the $\beta_i > 0$ are independent of λ and w.

Proof. To see that the function is quadratic, note that any P_s can be triangulated by drawing chords from each vertex to the origin, and that the triangle with side v_i has area given by half the determinant of the matrix with columns $\sum_{j=1}^{i-1} s_j \mathsf{v}_j$ and $\sum_{j=1}^{i} s_j \mathsf{v}_j$. By bilinearity, this expands to a quadratic in the s_i.

First consider the case $\mathsf{w} = \mathsf{0}$. Busemann's theorem ensures that the unique global max in any $M(\lambda, \mathsf{0})$ occurs at $\mathsf{s}^{\max}_{\lambda,\mathsf{0}} = (\frac{\lambda}{r\ell_1}, \dots, \frac{\lambda}{r\ell_r})$, which is on the interior because no coordinate is zero. The Hessian of area is symmetric, so diagonalizable,

and this induces a change of basis so that the area is a function of the squares of the new variables. The existence of a unique maximum on $M(\lambda, \mathsf{0})$ ensures that it can be written $\mathsf{Area}(P_\mathsf{s})|_{M(\lambda,\mathsf{0})} = A - \sum \beta_i x_i^2$ for $\beta_i > 0$, where $\mathsf{x} = \mathsf{x}(\mathsf{s})$ is an $(r-3)$-dimensional coordinate system centered around this maximum. Since $\overline{M}(\lambda, \mathsf{0})$ is the smallest affine subspace containing $M(\lambda, \mathsf{0})$, the restriction $\mathsf{Area}(P_\mathsf{s})|_{\overline{M}(\lambda,\mathsf{0})}$ is given by the same quadratic function.

Note also that since the $\overline{M}(\lambda, \mathsf{0})$ are mutually parallel affine subspaces we can use the same coordinate directions x on each $\overline{M}(\lambda, \mathsf{w})$. To see this, consider $\overline{M}(\rho\lambda, \mathsf{0}) = \rho\overline{M}(\lambda, \mathsf{0})$. Let $\mathsf{s}_0 = \mathsf{s}_{\lambda,0}$. Then $\rho\mathsf{s}_{\lambda,\mathsf{0}}^{\max} = \mathsf{s}_{\rho\lambda,\mathsf{0}}^{\max}$ realizes the maximum area on $\overline{M}(\rho\lambda, \mathsf{0})$. We have

$$\mathsf{Area}(P_{\rho\mathsf{s}})|_{\overline{M}(\rho\lambda,\mathsf{0})} = \rho^2 \left(\mathsf{Area}(P_\mathsf{s})|_{\overline{M}(\lambda,\mathsf{0})}\right) = \rho^2 A - \sum \beta_i (\rho x_i)^2.$$

Thus the constant term $A_{\lambda,\mathsf{0}}$ scales quadratically but the eigenvalues and eigendirections surrounding the maximum are preserved.

Now let $\mathsf{w} \in \mathbb{R}^2$ be arbitrary and let β be the Busemann arc of length λ ending at w. Choose γ such that the concatenation $\beta.\gamma$ equals $tP + \mathsf{k}$, a complete copy of the isoperimetrix. Then for any path P_s reaching w with perimeter λ, the area $\mathsf{Area}(P_\mathsf{s}) + \mathsf{Area}(\gamma)$ is uniquely optimized at $P_\mathsf{s} = \beta$ and it is quadratic in s and thus concave down in affine coordinates $\mathsf{x} = \mathsf{x}(\mathsf{s})$ centered at this maximum. We conclude that $\mathsf{Area}(P_\mathsf{s})\big|_{\overline{M}(\lambda,\mathsf{w})} = A - \sum \beta_i x_i^2$. On the other hand, $\mathsf{Area}(P_\mathsf{s}.\gamma) = \mathsf{Area}(P_\mathsf{s}) + \mathsf{Area}(\gamma)$. Thus $\mathsf{Area}(P_\mathsf{s})$ on $\overline{M}(\lambda, \mathsf{w})$ differs by a constant from the area in the closed family containing $P_\mathsf{s}.\gamma$, so we can appeal to the closed case $\lambda' = \lambda + \ell(\gamma)$, $\mathsf{w}' = \mathsf{0}$ to conclude the proof. □

For example, let P be the unit square with all $\ell_i = 1$. Then with no constraint on perimeter or closedness, we would have

$$\mathsf{Area}(P_\mathsf{s}) = \frac{1}{2}\left(s_1 s_2 + s_2 s_3 - s_1 s_4 + s_3 s_4\right).$$

Closedness forces $s_1 = s_3, s_2 = s_4$, and the perimeter constraint $\lambda = 4$ forces $s_2 = 2 - s_1$, giving $\mathsf{Area}(P_\mathsf{s}) = 2s_1 - s_1^2$. This has unique max at $s_1 = 1$, and in coordinates centered at the max (namely $x = s_1 - 1$) the area is $1 - x^2$.

It will also be useful below to allow the parallel family to deviate from the isopermetrix by a finite amount at the corners. Given $P = \mathsf{v}_1\mathsf{v}_2 \dots \mathsf{v}_r$, we can fix a list $\mathcal{C} = (\mathsf{c}_0, \dots, \mathsf{c}_r)$ of finite-length polygonal paths c_i called *corner paths*. We can write w_i for the displacement vector of path c_i from beginning to end, and write p_i for a constant associated to each (if P is the isoperimetrix for a norm, then take p_i to be the length of c_i in the norm). Then define the *general parallel family* of P relative to $\mathcal{C}$ to be

$$\{\gamma_\mathsf{s} = \mathsf{c}_0\mathsf{v}_1^{s_1}\mathsf{c}_1\mathsf{v}_2^{s_2}\dots\mathsf{c}_{r-1}\mathsf{v}_r^{s_r}\mathsf{c}_r \quad : \quad \mathsf{s} \in \mathbb{R}^r\}$$

and define displacement vectors $\mathsf{w}(\gamma_\mathsf{s}) = (\sum s_j \mathsf{v}_j) + (\sum \mathsf{w}_j)$ and perimeters $p(\gamma_\mathsf{s}) = \sum s_i\ell_i + \sum p_i$. With $\mathcal{C}$ fixed we will use the same notation $\overline{M}(\lambda, \mathsf{w})$ and $M(\lambda, \mathsf{w})$ from above to define the feasible regions with $s_i \in \mathbb{R}$ and $s_i \geq 0$ respectively.

To apply our convex geometry results back to the discrete group, we will need to restrict to integer parameters s_i. Our next goal will be to establish that even allowing for corner paths and with an arbitrary linear boost added to area, optimal area in the integer lattice occurs with proportions that are close to those from P.

If a path γ_{s} has the form

$$\gamma_{\mathsf{s}} = \mathsf{c}_{\mathrm{I}-1}\mathsf{v}_{\mathrm{I}}^{s_{\mathrm{I}}}\mathsf{c}_{\mathrm{I}}\mathsf{v}_{\mathrm{I}+1}^{s_{\mathrm{I}+1}}\dots\mathsf{c}_{\mathrm{J}-1}\mathsf{v}_{\mathrm{J}}^{s_{\mathrm{J}}}\mathsf{c}_{\mathrm{J}}$$

with $\mathrm{J}-\mathrm{I}\geq 2$ and the other $s_i=0$, we will call it *balanced* if there exists $t>0$ such that $s_{\mathrm{I}}, s_{\mathrm{J}} \leq s_{\mathrm{I}+1} = s_{\mathrm{I}+2} = \cdots = s_{\mathrm{J}-1} = t$; this is the case in which, if the corner words were removed, the path has the right proportions to belong to tP as a sub-arc. More generally γ_{s} is *K–almost balanced* if all of these equalities and inequalities hold within K, i.e., $|s_i - s_j| \leq K$ for $\mathrm{I} < i, j < \mathrm{J}$ and $s_{\mathrm{I}}, s_{\mathrm{J}} \leq s_i + K$ for all $\mathrm{I} < i < \mathrm{J}$.

Lemma 10 (Balancing paths)**.** Fix an isoperimetrix $P = \mathsf{v}_1\mathsf{v}_2\dots\mathsf{v}_r$ with integer vertices, any vector $(k_1,\dots,k_r)\in\mathbb{R}^r$, and corner paths $\mathsf{c}_0,\dots,\mathsf{c}_r$, and consider all of the possible P_{s}. Then there is a constant K such that for any λ and w with $\mathbb{Z}^r\cap M(\lambda,\mathsf{w})\neq\emptyset$, the maximum value of $f(\mathsf{s}) = \mathsf{Area}(\gamma_{\mathsf{s}}) + \sum_{i=1}^r k_i s_i$ over $\mathsf{s}\in\mathbb{Z}^r\cap\overline{M}(\lambda,\mathsf{w})$ occurs at paths that are K–almost balanced.

Proof. We claim that the case of no boost ($k_i=0$) and no corners follows quickly Lemma 9: Since $\mathsf{Area}(P_{\mathsf{s}}) = A - \sum\beta_i x_i^2$, the level sets in each feasible region $\overline{M}(\lambda,\mathsf{w})$ are ellipsoids of the same eccentricity; that is, there is a ρ such that ellipsoids of diameter ρd have an inscribed ball of radius d. Also, the ℓ_i and the coordinates of the v_i are all rational, so since $\mathbb{Z}^r$ was assumed to intersect the affine subspace, it must do so in a lattice, and we can choose fundamental domains having a diameter Δ that is independent of λ,w. By choosing $d\geq\Delta$, we can be sure that the ellipsoid contains a lattice point. This may not be the smallest ellipsoid containing a lattice point, but it ensures that all lattice points on the smallest such ellipsoid are at distance no greater than $\rho\Delta/2$ from the center of coordinates.

Next we allow a nonzero linear term $f(\mathsf{s}) = \mathsf{Area}(\gamma_{\mathsf{s}}) + \sum k_i s_i$. Given an expression $\beta s^2 + ks$, "completing the square" via the shift $x = s - \frac{k}{2\beta}$ transforms it to $\beta x^2 - \frac{k^2}{4\beta}$. Thus, the maximum for $-\beta s^2 + ks$ occurs at $x=0$. The β_i and k_i are fixed. Hence these x coordinates differ from the s coordinates by a bounded amount, so this maximum occurs at a bounded distance from the maximum of $\mathsf{Area}(P_{\mathsf{s}})$. Note that the nearest integer point may have some negative coordinates, but comparing the ellipse eccentricities to the fundamental domains as above shows that the best point with non-negative coordinates is again boundedly far away. Since area is maximized at a balanced arc, we have shown that the non-negative integer max for $f(\mathsf{s})$ is nearly balanced.

Finally, adding the corner paths changes very little. Each interior corner c_i can be "straightened" (replacing it with whatever $\alpha_i\mathsf{v}_i + \alpha_i'\mathsf{v}_{i+1}$ has the same displacement vector) and it contributes a fixed amount to area relative to its straightening, independent of s. As for the initial and final corners: if w is the displacement vector of $\mathsf{c}_{\mathrm{I}-1}\gamma\mathsf{c}_{\mathrm{J}}$, then $\mathsf{w} - \mathsf{c}_{\mathrm{I}-1} - \mathsf{c}_{\mathrm{J}}$ is the displacement vector of γ, and the difference in area is independent of γ. All corners also make a fixed contribution to perimeter. Therefore the solution for the path with corners is gotten by shifting the solution for a corner-free path with adjusted perimeter and displacement. □

Remark. This is the first of several places where something is shown to be bounded with reference to a constant K. To avoid proliferating notation, we will maintain the symbol K in each successive place that a constant bound is derived, enlarging it each time as necessary. No earlier statement will be hurt by subsequent enlargement, so that in the end one value of K depending only on S will suffice for all applications.

Finally we establish a lemma on the combinatorial types of arcs.

Definition 11. Fix $P = \mathsf{v}_1\mathsf{v}_2 \dots \mathsf{v}_{r-1}\mathsf{v}_r$. A *$P$–arc τ of scale s* is a path $\mathsf{v}_{\mathrm{I}}^{s^-}\mathsf{v}_{\mathrm{I}+1}^{s} \dots \mathsf{v}_{\mathrm{J}-1}^{s}\mathsf{v}_{\mathrm{J}}^{s^+}$ (with the indices considered cyclically) where $0 \le s^-, s^+ \le s$, and its *combinatorial length* is the sum of the exponents, $L(\tau) = s^- + (\mathrm{J} - \mathrm{I} - 1)s + s^+$. Note that τ begins at the origin and lies on a scaled and translated copy of P, i.e., $\tau \subset sP + \mathsf{r}$. The *combinatorial type* of τ is the pair (I, J) of starting and ending sides.

Given $K > 0$, we say that the arc *K–almost* has combinatorial type (I, J) if it can be modified to an arc of combinatorial type (I, J) by adjusting s^+ and s^- by at most K (possibly making them equal either 0 or s to change type). If there is (I, J) so that τ and τ' both K–almost have combinatorial type (I, J), we say that they *K–almost have the same combinatorial type*.

Lemma 12 (Combinatorial types of nearby arcs)**.** Fix $K_1, K_2 > 0$. Then there are K_3, K_4 with the following property. If τ and τ' are P–arcs (based at the origin) whose combinatorial lengths are within K_1 and whose endpoints are within distance K_2, then their scales differ by at most K_3. Further, after possible weight-shifting and cyclic permutation, the arcs K_4–fellow travel.

Proof. Suppose τ is a P–arc with endpoint $(a, b) \in \mathbb{R}^2$, and the length of τ is L. Using convexity of P, we will see that there are only very limited ways to find (a, b) as a chord with given arclength in a scaled P, and this restricts the shape of τ.

If $(a, b) = (0, 0)$ and τ is nonempty, it can only be a translate of P containing the origin. Thus for endpoints near $(0, 0)$, the arc is either very short or is nearly of type (I, I) for some I, with a scale roughly determined by the length. If one is of nearly of type (I, I) and the other is of type (J, J), then weight-shifting followed by cyclic permutation suffices to make them fellow-travelers.

If (a, b) is a nonzero multiple of some v_{I} and L is sufficiently long relative to (a, b), then the P–arc must be of type (I, I). In this case there is clearly a family of polygons with the same (a, b, L) and type (I, I) obtained by shifting weight between s^- and s^+, and these are the only solutions to the chord problem. So if one of the arcs, say τ, has an endpoint precisely on the v_{I} direction, then it admits a weight-shifted family of P–arcs and by choosing the right one we can match τ'.

The only remaining cases are that (a, b) is a nonzero vector which is not parallel to any v_i or that L is short relative to (a, b). In either of these cases the triple (a, b, L) uniquely determines not only s but determines τ completely (by convexity of P), and the starting side and ending side are different ($\mathrm{I} \neq \mathrm{J}$). Within a combinatorial type, the scale s is a linear function of (a, b, L), and indeed it is piecewise linear (and continuous) across combinatorial types as (a, b) varies over the sector between any successive $\mathsf{v}_i, \mathsf{v}_{i+1}$. (See Duchin–Mooney for details and examples.) Being far from the origin forces arcs with nearby endpoints to be nearly of the same combinatorial type, and so they fellow travel. □

5. Shapes

5.1. **Simple shapes and highest height.** Suppose $u, v \in H(\mathbb{Z})$ project to integer vectors $\mathsf{u}, \mathsf{v} \in \mathfrak{m}$. We write $\mathsf{u} \wedge \mathsf{v}$ to denote the determinant of the matrix with those column vectors, i.e., the area of the parallelogram they define. Then when letters u and v are exchanged, the effect on area is given by the wedge: $z(uv) = z(vu) + \mathsf{u} \wedge \mathsf{v}$. For instance, $z(\mathsf{e}_1\mathsf{e}_2) = \frac{1}{2}$; $z(\mathsf{e}_2\mathsf{e}_1) = -\frac{1}{2}$; and $\mathsf{e}_1 \wedge \mathsf{e}_2 = 1$. Note that two group elements commute if and only if they project to the same direction in the plane.

To keep track of all the possible effects of rearranging letters, we once and for all define

$$N = N(S) := \mathrm{lcm}\{\mathsf{u} \wedge \mathsf{v} : u, v \in S\}.$$

Definition 13. For $(a, b) \in \mathfrak{m}$, define $n_0(a, b) = |(a, b)|_{\pi(S)}$, so that the fiber $(a, b, *)$ can be reached by a spelling path of length n if and only if $n \geq n_0$. Then for $n \geq n_0$, we define the *highest height at length n over (a, b)* to be the largest z coordinate reachable with at most n letters,

$$w_n = w_n(a, b) := \max\{t : \ell(a, b, t) \leq n\}.$$

A spelling (or a group element) will be called *highest-height* if it realizes (a, b, w_n) at length n. Let $W(a, b) := w_{n_0}(a, b)$ be first non-negative w_n.

Note that $w_n < w_{n+2}$. To see this simply replace some generator a_{I} by $a_{\mathrm{J}} a_{\mathrm{I}} a_{\mathrm{J}}^{-1}$. However there may be no spellings at all of a certain parity reaching (a, b), in which case $w_n = w_{n+1}$.

Definition 14. Given a constant K, let $C(K) = \bigcup_{i=0}^{K} S^i$ be the strings in S whose length is at most K (so that the evaluation map sends $C(K)$ onto the ball of radius K in the word metric). Then a *break word* is an element $c \in C(K)$ and a *break vector* is a tuple of break words $\mathsf{c} = (c_0, \ldots, c_{2k})$.

A *simple shape* is a tuple $\omega = (\mathrm{I}, \mathrm{J}, \mathsf{b}, \mathsf{c})$, where c is a break vector, I, J are indices $(1 \leq \mathrm{I}, \mathrm{J} \leq 2k)$, and $\mathsf{b} = (b_1, \ldots, b_{2k})$ is a vector of integers called *exponent corrections*. The *simple shape domain* is $\mathrm{Cone} := \{(s^-, s, s^+) \in \mathbb{Z}^3 : 0 \leq s^-, s^+ \leq s\}$ and the *restricted domain* is $\mathrm{Cone}_0 := \{(s^-, s, 0)\} \subset \mathrm{Cone}$. (Compare to Lemma 12.)

Each such shape induces a map from the shape domain to spellings in the group. That is, define the evaluation of a simple shape to be

$$\omega(s^-, s, s^+) = c_{\mathrm{I}-1} \cdot \widehat{a}_{\mathrm{I}}^{\,s^- + b_{\mathrm{I}}} \cdot c_{\mathrm{I}} \cdot \widehat{a}_{\mathrm{I}+1}^{\,s + b_{\mathrm{I}+1}} \cdot c_{\mathrm{I}+1} \cdots \widehat{a}_{\mathrm{J}-1}^{\,s + b_{\mathrm{J}-1}} \cdot c_{\mathrm{J}-1} \cdot \widehat{a}_{\mathrm{J}}^{\,s^+ + b_{\mathrm{J}}} \cdot c_{\mathrm{J}},$$

recalling that $\widehat{a}_i = a_i^{\sigma_i}$ is defined so that $\mathbf{I} = \widehat{a}_1 \cdots \widehat{a}_{2k}$. Further, we take the convention that if $\mathrm{I} = \mathrm{J}$, i.e., if the shape starts and ends with the same generator, then the domain is restricted to Cone_0.

Example 15. Consider the nonstandard generators for $H(\mathbb{Z})$ given by $S = \{a, b, A, B\}^{\pm}$, where a, b are the standard generators and A, B are big generators $A = a^3, B = b^3$, and a bar denotes the inverse of an element. Then the word $A^5 a B^9 b \bar{a} \bar{A}^{10} \bar{b} \bar{B}^3$ is given by evaluating the shape with $\mathsf{c} = (e, a, b\bar{a}, \bar{b}, e)$, $\mathsf{b} = (0, 0, 1, 0)$, $\mathrm{I} = 1$, $\mathrm{J} = 4$ at $\mathsf{s} = (5, 9, 3)$.

Remark. There are other shapes with other data that evaluate to the same path.

5.2. **Bounded difference between word and CC metrics.**

Proposition 16 (Form for highest-height geodesics)**.** Given a finite generating set S, there is a number $K = K(S)$ such that any highest-height spelling path is the evaluation of some simple shape with break words from $C(K)$ separating runs of significant letters given by integer values $s^-, s^+ \leq s$ with exponent corrections $0 \leq b_i \leq K$.

That is, in a very strong sense, highest-height spellings track along an arc of a canonical polygon (Busemann's isoperimetrix), which has a spelling of the form $\widehat{a}_{\mathrm{I}}^{\,s^-} \widehat{a}_{\mathrm{I}+1}^{\,s} \widehat{a}_{\mathrm{I}+2}^{\,s} \cdots \widehat{a}_{\mathrm{J}-1}^{\,s} \widehat{a}_{\mathrm{J}}^{\,s^+}$ with $s^-, s^+ \leq s$, not necessarily integers. The highest-height spellings only differ by bounded break words appearing in the corners, and by bounded deviation in run lengths.

Proof. We suppose that γ is a highest-height geodesic over (a, b) and that its length is n. We claim that the letters of γ are in cyclic order. If not, we produce γ' by putting its letters into cyclic order. This changes neither the horizontal endpoint (a, b) nor the boost $z_b(\gamma)$. If two of the letters which we move past each other in this process do not lie in the same direction in projection, then $z(\gamma') > z(\gamma)$, contradicting our assumption. Thus the letters appearing in γ are arranged in cyclic order in projection. Also if there are multiple letters a_i, a_i', a_i'' projecting to the same significant a_i, then clearly γ must use the one with greatest boost to achieve highest height.

We now claim that there is a bound K on the total exponent of any non-significant generator. To see this, suppose that u is a non-significant generator appearing with large exponent, as a subword u^m. Supposing a_i and a_{i+1} are the significant generators whose directions bound the sector that u lies in, there must be integers p, q, r so that $q\mathsf{u} = p\mathsf{a}_i + r\mathsf{a}_{i+1}$, with $q \geq p + r$. We can then replace u^{kq} by $a_i^{kp} a_{i+1}^{kr}$. The area gained by this operation is quadratic in k while any boost lost is linear in k. Consequently, if m is sufficiently large, this operation increases height. So if the total exponent of u in γ is m, then the reshuffling which brings all powers of u together and then performs the subword replacements above will produce a path over (a, b) with no greater length and with higher height, contradicting the assumption.

It follows now that γ consists of corner words of bounded length between ordered runs of highest-boost significant letters. That is, we have

$$\gamma = c_{\mathrm{I}-1} \cdot \widehat{a}_{\mathrm{I}}^{n_{\mathrm{I}}} \cdot c_{\mathrm{I}} \cdot \widehat{a}_{\mathrm{I}+1}^{n_{\mathrm{I}+1}} \dots c_{\mathrm{J}-1} \cdot \widehat{a}_{\mathrm{J}-1}^{n_{\mathrm{J}}} \cdot c_{\mathrm{J}}.$$

The statement now follows from an application of the Balancing Lemma (Lemma 10). □

Corollary 17 (Bounded difference)**.** For each generating set S, there exists a constant $K = K(S)$ with the following property. If $w_n(a, b) < c \leq w_{n+1}(a, b)$ then $n < |(a, b, c)| \leq n + K$, and if $0 \leq c \leq W = w_{n_0}(a, b)$, then $n_0 \leq |(a, b, c)| \leq n_0 + K$.

Consequently, there exists a constant $K = K(S)$ such that

$$d_{\mathrm{cc}}(\mathsf{x}, \mathsf{0}) - K \leq |\mathsf{x}|_S \leq d_{\mathrm{cc}}(\mathsf{x}, \mathsf{0}) + K.$$

Put differently, the embedding of $H(\mathbb{Z})$ with generating set S into $H(\mathbb{R})$ with the corresponding CC metric is a $(1, K)$ quasi-isometry.

Proof. By definition (a, b, w_n) is the highest-height element of the fiber over (a, b) which can be reached by a spelling of length less than or equal to n, so $n < |(a, b, c)|$.

Let $\omega(s^-, s, s^+)$ and $\omega'(t^-, t, t^+)$ be shapes evaluating to geodesic spellings for $g = (a, b, w_n)$ and $g' = (a, b, w_{n+1})$. Let τ and τ' be **I**-arcs which fellow-travel these in projection (whose existence is guaranteed by the previous result). If τ and τ' are of almost the same combinatorial type, then the polygonal paths $\beta = \pi(\omega(s^-, s, s^+))$ and $\beta' = \pi(\omega'(t^-, t, t^+))$ fellow-travel. Consider the sequence of paths $\beta' = \beta_0, \beta_1, \dots, \beta_{n+1} = \beta$ formed as follows. For $i = 1, \dots n$, let β_i be the path starting along β until $\beta(i)$, taking a geodesic from $\beta(i)$ to $\beta'(i)$, and continuing along β'. Since β and β' K_0–fellow-travel for some K_0, the connecting geodesics have bounded length, so each β_i has length at most $n+1+K_0$. Take γ_i to be the lift of β_i. These γ_i end at group elements (a, b, c_i) with $|c_{i+1} - c_i| \leq 2K_0 + 2$. Thus any value (a, b, c) in the range in question can be reached by tacking a bounded-length

path on to the end of an appropriate γ_i. It follows that there is K such that for each c with $w_n < c \leq w_{n+1}$, $|(a,b,c)| \leq n + K$ as required.

If τ and τ' are not of almost the same combinatorial type, then by Lemma 12 (a,b) is close to the origin and τ and τ' almost complete the entire boundary of an isoperimetrix. It follows that we can replace $\omega'(t^-,t,t^+)$ by a spelling path which fellow-travels $\omega(s^-,s,s^+)$ in projection and is only boundedly longer than $\omega'(t^-,t,t^+)$. (To be concrete, the blocks of significant letters and the corners can be preserved but reordered to correspond to the combinatorics of τ'.)

Note that for an **I**–arc of spelling length n, its length the L–norm is n, so its lift has CC length n as well and it is geodesic. Therefore $d_{\text{CC}}((a,b,w_n),\mathsf{0})$ is boundedly close to n and we are done with the case $w_n < c \leq w_{n+1}$.

For heights below W, we begin with a highest-height spelling realizing (a,b,w_{n_0}). By permuting the letters, we can lower the height in bounded increments down to some minimum. Suppose it can be lowered to a non-positive height. Then since the intermediate heights can be reached by appending a bounded-length correction word, we have $n_0 \leq |(a,b,c)| \leq n_0 + K$. On the other hand, the CC distance from $\mathsf{0}$ is constant in the (a,b) fiber up to the first height reached by a regular geodesic, which is boundedly close to (a,b,W). We enlarge the constant K from the first statement in the Lemma to be sufficient for the second statement.

On the other hand, it may be that every permutation of the letters in the spelling has positive height, for instance if the spelling is simply a single repeated letter with positive boost. In this case, suppose that $a_{\text{I}+1}$ is the first significant letter in the spelling and choose some significant letter u such that $\mathsf{u} \wedge \mathsf{a}_{\text{I}+1} < 0$. It follows that there is some power of u such that the conjugate $u^k a_{\text{I}+1} u^{-k}$ has height below zero. From this modified word, complete the proof as before with successive permutations.

Finally, observe that the map $g \mapsto g^{-1}$ is a length-preserving bijection which carries (a,b,c) to $(-a,-b,-c)$, so the $c<0$ case is similar. □

This gives a new proof of Krat's result. And in particular, since Krat's theorem (bounded difference) has a stronger conclusion than Pansu's theorem (ratio goes to 1), our argument also gives a direct geometric proof of Pansu's theorem for the special case of arbitrary word metrics on $H(\mathbb{Z})$.

5.3. **Simplification.** We will see below that every regular element has a geodesic which is close to a simple shape. To this end, we show that we can modify paths to become simple shapes while staying in the same fiber and increasing height in a controlled manner. Recall that we have taken N to be the lcm of the areas of generator swaps.

Lemma 18 (Simplifying paths)**.** There is a constant $K = K(S)$ so that for each spelling path γ there exists a refined path γ_1 with the following properties.

- If (a,b,c) is the evaluation of γ, then γ_1 evaluates to $(a,b,c+kN)$ for some $k \geq 0$;
- the length of γ_1 is less than or equal to the length of γ;
- $\gamma_1 = \omega(s^-,s,s^+)$ for some simple shape ω with corners from $C(K)$.

Proof. We start by replacing significant generators which do not have highest boost. We do this replacing a multiple of N of each type, i.e., each projection and boost, so that when we are done we have increased height by a multiple of N and are left with boundedly many letters which do not have highest boost.

Now suppose that u, v are any two letters appearing in γ such that $\mathsf{u} \wedge \mathsf{v} > 0$ (so that u comes before v in the cyclic ordering of their projections, and replacing vu with uv increases area). Let $\Lambda_{u,v} = \Lambda_{u,v}(\gamma)$ be the sum of all of the exponents k appearing in distinct subwords $vu^k w$ of γ with $w \in S$, $w \neq u$. Then we can make generator swaps of u and v letters to change the height by any multiple of $\mathsf{u} \wedge \mathsf{v}$ less than or equal to $\Lambda_{u,v} \cdot \mathsf{u} \wedge \mathsf{v}$. By rounding $\Lambda_{u,v} \cdot \mathsf{u} \wedge \mathsf{v}$ down to the nearest multiple of N, we can perform generator swaps to obtain γ_1, so that $\Lambda_{u,v}(\gamma_1) \leq N$ for all pairs u, v.

Notice that we may have to perform this procedure many times. A single application of this procedure reduces $\Lambda_{u,v}$ to be less than N, but may increase $\Lambda_{u,v'}$. We can perform this procedure whenever there is some pair u, v so that $\Lambda_{u,v} \cdot \mathsf{u} \wedge \mathsf{v} > N$. We claim that repeated applications of procedure must eventually terminate with a spelling where there is no such pair. To see this, consider the total number of pairs of letters in the spelling which are out of order. This total number decreases at every application of the procedure, and hence we must terminate with $\Lambda_{u,v} \cdot \mathsf{u} \wedge \mathsf{v} < N$ for every pair u and v.

Next, we will cash in any big blocks of non-significant letters for significant letters. Recall that significant letters project to corner points of the polygon L, while edge letters project to other boundary points and interior letters project to the interior. That is, for an edge letter u and an interior letter v with projections in the sector between a_i and a_{i+1}, we have $q\mathsf{u} = p\mathsf{a}_i + r\mathsf{a}_{i+1}$ and $q'\mathsf{v} = p'\mathsf{a}_i + r'\mathsf{a}_{i+1}$ such that $q = p + r$ while $q' > p' + r'$.

Consider the subword replacements $u^{kNq} \to a_i^{kNp} a_{i+1}^{kNr}$, or $v^{kNq'} \to a_i^{kNp'} a_{i+1}^{kNr'}$. As above, the new paths reach the same endpoint in $\mathfrak{m}$ while either preserving or reducing the total spelling length of the path, gaining area by an amount proportional to k^2, and reducing boost by an amount proportional to k. We perform these replacements in every instance where k is large enough to produce is a net height increase; and note that the height change is a multiple of N.

Repeat the reordering and the replacement steps one after the other until neither can be performed any further. Then $z(\gamma_1) \geq z(\gamma)$, and they differ by a linear combination of the wedges; namely, $\Delta z = z(\gamma_1) - z(\gamma) = \sum_{u,v} k_{uv}(\mathsf{u} \wedge \mathsf{v})$, for integers $k_{uv} \geq 0$ with $k_{uv} \equiv 0 \pmod{N}$. At this stage, the path γ_1 has well-defined sides with mostly a_i letters and only boundedly many exceptions.

Next, we set things up to push the remaining "out of place" letters to the corners so that the a_i side is mostly a single long block of the a_i letter. So far we have a spelling γ_1 that contains boundedly many non-significant letters and boundedly many significant letters on the wrong side. Consider a side which consists of significant generator a_i with a bounded number of letters which are not a_i. For each letter u on the a_i side, the sign of $\mathsf{u} \wedge \mathsf{a}_i$ tells us whether replacing $a_i u$ with $u a_i$ is height-increasing or height-decreasing (note that the case $\mathsf{u} \wedge \mathsf{a}_i = 0$ is the case that u and a_i commute). We swap each u past a_i^N in the height-increasing direction (or an arbitrary direction if they commute) until we create a block $w = u' a_i^m u$ with $m < N$. This w itself can be commuted with a_i^N to the left or right, not decreasing height. Since there are boundedly many out of place letters on each side, this process ends with all these letters within a bounded distance of a corner, so we merge them with the corner words. At each move we have increased height by a multiple of N.

Finally, we balance the side lengths of γ_1. To do this we apply the balancing lemma (Lemma 10) to the lattice of integer tuples which differ from the original $(t_{\mathrm{I}}, t_{\mathrm{I}+1}, \ldots, t_{\mathrm{J}})$ by multiples of N in each coordinate. This ensures that area and boost, and therefore height, changes by a multiple of N.

This final step has produced a modified path, again called γ_1, which still has the same (a, b) endpoint as γ and may have higher height by a multiple of N. Now there are bounded-size exceptional corner words between the sides, and the exponents of significant blocks differ only by a bounded amount, so this is the evaluation of a simple shape. □

5.4. **General shapes.** Beyond simple shapes, we will need a construction of shapes with break words not only at the corners: runs of significant generators can be separated by finitely many other break words.

Definition 19. Given a generating set S for which the isoperimetrix has $2k$ sides, a *general shape* with parameter $K \geq 1$ is a tuple $\omega = (\mathrm{I}, \mathrm{J}, \mathsf{b}, \chi)$, where

- $1 \leq \mathrm{I}, \mathrm{J} \leq 2k$ are a starting and ending side;
- $\mathsf{b} = (b_1, \ldots, b_{2k})$ is a vector of integers $0 \leq b_i \leq K$;
- χ is a $(K-1) \times 2k$ matrix whose entries are break words from $C(K)$.

Let Shape_K be the set of all such shapes, clearly a finite set for each value K. We will evaluate each shape at a matrix $X \in M_{K \times 2k}$. Let $\Lambda : \mathrm{Shape}_K \times M_{K\times 2k} \to \mathbb{Z}^{2k}$ be given by $\Lambda(\omega, X) = (\lambda_1, \ldots, \lambda_{2k})$, where $\lambda_i := \left(\sum_{j=1}^{K} x_{ji}\right) - b_j$. Then the *shape domain* $\mathrm{Dom}_K(\omega)$, for $\omega \in \mathrm{Shape}_K$, is the set of $K \times 2k$ matrices X of non-negative integers satisfying a condition on the image of Λ, namely:

- $\lambda_i = \lambda_j$ for all $\mathrm{I} < i, j < \mathrm{J}$;
- $\lambda_{\mathrm{I}}, \lambda_{\mathrm{J}} \leq \lambda_i$;
- $\lambda_t = 0$ for the t that are not between I and J;
- if $\mathrm{I} = \mathrm{J}$, then $\lambda_J J = 0$.

With slight abuse of notation, we will then write $\Lambda : \mathrm{Shape}_K \times \mathrm{Dom}_K \to \mathrm{Cone}$ given by $\Lambda(\omega, X) = (s^-, s, s^+)$ where $s^- = \lambda_{\mathrm{I}}$, $s = \lambda_{\mathrm{I}+1} = \cdots = \lambda_{\mathrm{J}-1}$, and $s^+ = \lambda_{\mathrm{J}}$. (Note that the last condition in the definition ensures that the map lands in Cone_0 in the $\mathrm{I} = \mathrm{J}$ case.)

It is immediate from this definition that $\mathrm{Dom}_K(\omega)$ is given by pulling back a rational family under an affine map.

The matrix X is to be thought of as a matrix of *run lengths*. The evaluation of a shape, $\omega(X)$, is the concatenation of the break words with the runs of significant generator blocks of length prescribed by X. The b vector records the failure of the column sums to be equal, i.e., the failure of the shadow to be balanced in terms of its side lengths. (Since its entries are bounded, the column sums are nearly equal, which means that the spelling will track close to an isoperimetrix.) Simple shapes are a subset of general shapes for which the break words only appear at the corners.

Remark. Note that the triple (a, b, ℓ) associated to a spelling $\omega(X)$ factors through Λ. That is, as X ranges over $\mathrm{Dom}_K(\omega)$, the three integers $\Lambda(\omega, X) = (s^-, s, s^+)$ determine the horizontal position and the word length of the evaluation word. Thus we can regard this as a map $\omega : \mathrm{Cone} \to \mathbb{Z}^3$ that is affine and injective.

Remark. If significant generators include several options with same projection and different boost, then we also need Y, a matrix specifying for each side how many of

each different boost level get used, and in this case the evaluation will be $\omega(X, Y)$. This makes no meaningful difference anywhere in the argument below.

5.5. **Unsimplification.** We describe a *2-sided surgery* and a *3-sided surgery* for paths and then explain how to use them algorithmically to begin with a path described by a simple shape and produce a path ending lower in the same fiber and still described by a general shape. In both of these moves, we will suppose that a_1, a_2, a_3 are successive significant generators and that p, q, r are the values with gcd $= 1$ so that where $q\mathsf{a}_2 = p\mathsf{a}_1 + r\mathsf{a}_3$. (In the special case that $\mathsf{a}_1 = -\mathsf{a}_3$ (the parallel case), we have such a surgery with $p = r = 1$, $q = 0$.) Here we describe the surgeries on side a_2.

2–sided surgery. Here, a subword of the form $a_1^{s_1} c_1 a_2^{s_2}$ is replaced by $a_1^{s_1 - 3Np} c_1 w a_2$, where w is a permutation of the letters in $a_1^{3Np} a_2^{s_2 - 1}$.

3–sided surgery. Here, a subword of the form $a_1^{s_1} c_1 a_2^{s_2} c_2 a_3^{s_3}$ is replaced by $a_1^{s_1 - 2Np} c_1 a_2^{s_2 + 2Nq} c_2 a_3^{s_3 - 2Nr}$

Note that since $q < p + r$, 3–sided surgery is length reducing.

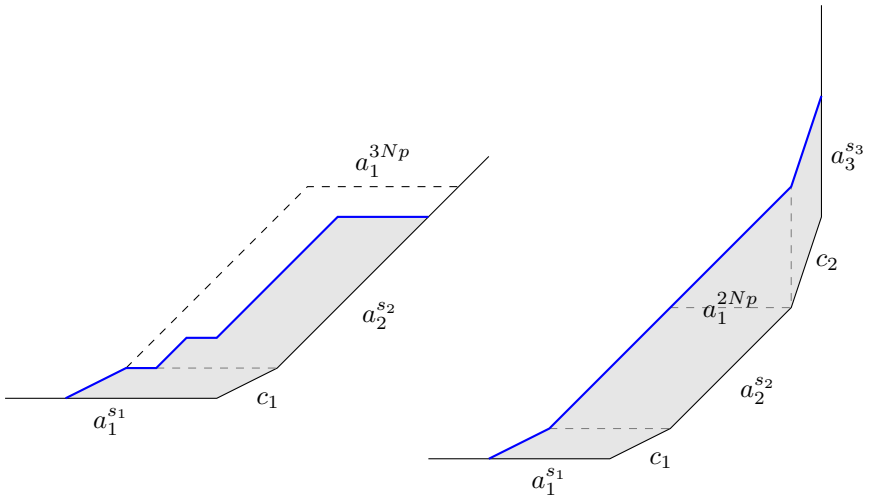


FIGURE 4. Examples of 2-sided and 3-sided surgery with corners. If the length s_2 of the second side is long enough, then 2-sided surgery can make a larger change to area because it is thicker: the width of the surgery is proportional to $3N$ rather than $2N$.

Lemma 20 (Unsimplification for shapes)**.** Given a starting word γ, let γ_1 be the simplification described above and suppose the height difference $\Delta z = z_1 - z_0$ is sufficiently large. Then for any full side of γ_1, a sequence of (possibly zero) 3-sided surgeries on that side followed by at most one 2-sided surgery on that side produces a word γ_2 which evaluates to the same group element as γ.

Note that if there are fewer than three sides (so that there is no well-defined "full side"), then we can appeal to the unstable (pattern) case presented in the next section.

Proof. The change in area for each application of three-sided surgery equals

$$(3SS) := 2Np(\mathsf{a}_1 \wedge \mathsf{c}_1) + 2Nps_2(\mathsf{a}_1 \wedge \mathsf{a}_2) + 2N^2pr(\mathsf{a}_1 \wedge \mathsf{a}_3) + 2Nr(\mathsf{c}_2 \wedge \mathsf{a}_3).$$

We note that since the wedges are all integers, this is divisible by N and therefore also by $\mathsf{a}_1 \wedge \mathsf{a}_2$.

On the other hand, the area difference from performing two-sided surgery depends on the permutation parameter; the area change equals

$$(2SS)_k := 3Np(\mathsf{a}_1 \wedge \mathsf{c}_1) + k(\mathsf{a}_1 \wedge \mathsf{a}_2),$$

where k is an arbitrary integer, $0 \leq k \leq 3Np(s_2 - 1)$.

The lemma's assumption that the height difference is large enough can be taken to precisely mean that $\Delta z = z_3 - z_0 > (2SS)_0$.

Perform (3SS) repeatedly, updating z_3 each time, until

$$\Delta z < (2SS)_0 + (3SS).$$

Then we must show that there exists k such that $\Delta z = (2SS)_k$. We know that Δz is a multiple of N and therefore of $\mathsf{a}_1 \wedge \mathsf{a}_2$. On the other hand, Δz is greater than $(2SS)_0$, and $(2SS)_k$ achieves all multiples of $\mathsf{a}_1 \wedge \mathsf{a}_2$ past that threshold and up to its maximum. Thus it is enough to show that $\Delta z < (2SS)_{\max}$. Since we saw above that $\Delta z < (2SS)_0 + (3SS)$, this amounts to showing that $(2SS)_{\max} - (2SS)_0 > (3SS)$. Since all the wedges of vectors and the values p, q, r are fixed by the choice of side, it suffices to take s_2 sufficiently large: since the left-hand side has a term $3Nps_2(\mathsf{a}_1 \wedge \mathsf{a}_2)$ and the right-hand side has a term $2Nps_2(\mathsf{a}_1 \wedge \mathsf{a}_2)$, eventually the difference between these overwhelms all the other fixed terms. □

Remark. We note for future reference that we are now in one of three situations. Either

- Δz is small, in which case not many letters were moved in producing γ_1 from γ and thus these two fellow travel in projection, or
- Δz is large and we used a 3–sided surgery, in which case we reduced length, contradicting the assumption that γ is geodesic, or
- We used only a single 2–sided surgery, in which case the resulting word fellow travels γ_1 in projection.

6. Patterns

Recall that CC geodesics are classified into two kinds (see Sec. 3.6), regular and unstable. In a particular fiber $(a, b, *)$, only unstable geodesics reach positive heights below a certain threshold height and only regular geodesics reach above that level. We will consider the corresponding situation for word geodesics.

We defined $W = W(a, b)$ to be the highest height reached by a spelling path of length $n_0 = |(a, b)|_{\pi(S)}$. In each fiber $\{(a, b, *)\}$, the general shapes defined in the previous section will reach the elements $\{(a, b, c) : c > W\}$, which may be called *regular* elements. In this section we turn to the growth of the *unstable* elements. Here we will consider the unstable elements $\{(a, b, c) : 0 \leq c \leq W\}$ at non-negative heights. (Later, we will appeal to the map $g \mapsto g^{-1}$ which carries (a, b, c) to $(-a, -b, -c)$ to deal with the negative heights.)

Definition 21. A *pattern* is a tuple $\mathsf{w} = (\mathrm{I}, c_1, c_2, c_3)$, where each $c_i \in C(K)$ is a break word (a string of length at most K), and $1 \leq \mathrm{I} \leq 2k$ picks out a sector between successive significant directions a_I, $\mathsf{a}_{\mathrm{I}+1}$. The (finite) set of all

such patterns will be denoted Patt_K. Each pattern w gives a map $\mathbb{N}^2 \to S^*$ via $\mathsf{w}(n_1, n_2) = c_1 a_{\mathrm{I}}^{n_1} c_2 a_{\mathrm{I}+1}^{n_2} c_3$.

Lemma 22 (Simplifying to a pattern)**.** Let (a, b) lie in the sector between a_{I} and $\mathsf{a}_{\mathrm{I}+1}$, and let N be the lcm of all possible area swaps, as usual. Then there is $K = K(S)$ with the following property. If γ is a geodesic for an unstable element (a, b, c), then there is a pattern $\mathsf{w} \in \mathrm{Patt}_K$ and $n_1, n_2 \in \mathbb{N}$ such that the spelling path $\tau = \mathsf{w}(n_1, n_2) = c_1 a_{\mathrm{I}}^{n_1} c_2 a_{\mathrm{I}+1}^{n_2} c_3$ has the following properties:

- the paths τ and γ have the same length;
- the path τ evaluates to an element $(a, b, c + kN)$ with $k \geq 0$ and $c + kN \leq W + K(n_1 + n_2)$; and
- the letters a_{I} and $a_{\mathrm{I}+1}$ are the highest-boost generators projecting to a_{I} and $\mathsf{a}_{\mathrm{I}+1}$, respectively.

Proof. First note that if that γ is a geodesic for an unstable element (a, b, c) where (a, b) lies in the sector between a_i and a_{i+1}, then all but boundedly many letters in γ project to convex combinations of a_i and a_{i+1} . This is because, by Bounded Difference (Corollary 17), there is K such that if (a, b, c) is unstable, then $n_0 \leq |(a, b, c)| \leq n_0 + K$, so that the projection $\pi(\gamma)$ must reach (a, b) in at most $n_0 + K$ letters. This means $\pi(\gamma)$ can only use boundedly many letters that are not on the edge between those points (i.e., convex combinations of a_i and a_{i+1})—to see this, just consider orthogonal projection to the normal of that edge, so every time any other letter is used, the projection falls behind by a definite amount.

We now carry out the simplification procedure used above (Lemma 18), making a few extra observations as we go. We note that the length of the path in this case will be maintained and not shortened, because there are only boundedly many interior letters and so we need not cash them in for significant letters.

If a_i is the highest-boost lift and a_i' is another letter projecting to a_i, then we can replace any $(a_i')^N$ by a_i^N. The remaining (boundedly many) a_i', which commute with a_i, can be pushed to the corner position.

Finally, $\pi(\tau)$ fellow-travels the L–norm geodesic $\mathsf{a}_{\mathrm{I}}^{n_1} \mathsf{a}_{\mathrm{I}+1}^{n_2}$, and therefore fellow-travels any geodesic achieving (a, b, W), so $|z(\tau) - W|$ is bounded by a constant multiple of $n_1 + n_2$. We enlarge K if necessary to complete the lemma. □

On the other hand, by controlled rearrangement of letters, patterns can produce a range of group elements in the same fiber.

Definition 23. For a pattern $\mathsf{w} = c_1 a_1^{n_1} c_2 a_2^{n_2} c_3$ evaluating to (a, b, c), define a process of rearrangements as follows. Consider letters $b_1, \ldots, b_k$ appearing in the word c_2. For $j = 1, \ldots, k$, let $d_j = \frac{N}{\mathsf{a}_{\mathrm{I}+1} \wedge \mathsf{b}_j}$, so that commuting $a_{\mathrm{I}+1}^{d_j}$ through b_j decreases height by N. We greedily perform commutations to move $a_{\mathrm{I}+1}$ letters past c_2, then continue if possible by commuting groups of $a_{\mathrm{I}+1}$ letters through a_{I} letters. Consider the set of (a, b, c') achievable by this process for which $0 \leq c' \leq W$, and let the *height interval* of the pattern, denoted $\mathcal{I}_{\mathsf{w}}(a, b)$, be the z coordinates in this set. Note that by construction $\mathcal{I}_{\mathsf{w}}(a, b)$ is the intersection of an interval with a residue class.

For example, for the generators $\{a, b, A, B\}^{\pm}$ described above in Example 15, if $\mathsf{w} = aA^*aB^*b$, then $\mathcal{I}_{\mathsf{w}}(52, 131) = \{6, 106, 206, \ldots, 3406\}$. Here $W(52, 131) = 3406$ and $N = 100$.

Lemma 24 (Unsimplification for patterns)**.** Let $\mathsf{w}(n_1, n_2)$ evaluate to (a, b, C_{w}), and define $C'_{\mathsf{w}} = \max \mathcal{I}_{\mathsf{w}}$ and $C''_{\mathsf{w}} = \min \mathcal{I}_{\mathsf{w}}$, so that $C_{\mathsf{w}}, C'_{\mathsf{w}}, C''_{\mathsf{w}}$ are functions of n_1, n_2 (or equivalently of a, b) representing the possible heights of rearrangements of patterns. Then there is a partition of $\mathbb{N}^2$ given by finitely many linear equations, inequalities, and congruences such that $C_{\mathsf{w}}, C'_{\mathsf{w}}, C''_{\mathsf{w}}$ are given by quadratic polynomials in n_1, n_2 on each set in the partition. Therefore there is a corresponding partition of $\mathfrak{m}$ so that these heights are quadratic on each piece on which $\mathcal{I}_{\mathsf{w}} \neq \emptyset$.

Proof. Fixing w, the height C_{w} can be seen as a function of (n_1, n_2) whose degree-two term equals $\frac{1}{2} n_1 n_2 (\mathsf{a}_{\mathrm{I}} \wedge \mathsf{a}_{\mathrm{I}+1})$, because $\mathsf{w}(n_1, n_2)$ fellow travels the two-sided figure $\mathsf{a}_{\mathrm{I}}^{n_1} \mathsf{a}_{\mathrm{I}+1}^{n_2}$. Fellow traveling ensures that the enclosed areas differ by at most an amount proportional to the length of the shape plus the boost provided by corner words, which are terms of degree one and zero.

W is the highest height of a minimal-length spelling path reaching the shadow of $\mathsf{w}(n_1, n_2)$. The simplification argument above shows that the spelling path realizing height W must also be boundedly close in projection to $a_{\mathrm{I}}^{n_1} a_{\mathrm{I}+1}^{n_2}$, so the difference $W - C_{\mathsf{w}}$ is a linear function as well. If it is positive, then $C'_{\mathsf{w}} = C_{\mathsf{w}}$. if it is negative, then C'_{w} are given by quadratic polynomials on each residue class of $C_{\mathsf{w}} \pmod N$.

The lowering process can take the pattern all the way down below height zero as long as n_2 is sufficiently large compared to N. If it is not, then the quadratic expression for C''_{w} in terms of n_1, n_2 is given by linear functions of n_1 for each small value of n_2.

Finally, the (a, b) are linearly related to (n_1, n_2) via $(a, b) = n_1 \mathsf{a}_{\mathrm{I}} + n_2 \mathsf{a}_{\mathrm{I}+1} + \bar{\mathsf{c}}$, where $\bar{\mathsf{c}}$ is the sum of the corner words, so a change of basis finishes the proof. □

7. Word geodesics vs. CC geodesics

Theorem 25 (Realization by shapes and patterns)**.** For every generating set S, the following two equivalent conditions hold:

- there is a $K = K(S)$ such that every group element has a geodesic spelling for which the shadow is K–close to the shadow of a CC geodesic;
- there is a $K = K(S)$ such that every group element has a geodesic spelling which is either the rearrangement of some pattern from Patt_K or the evaluation of some general shape from Shape_K.

Proof. Suppose γ is a geodesic spelling in $(H(\mathbb{Z}), S)$ evaluating to $(a, b, c) \in H(\mathbb{Z})$. Recall that N was defined as the least common multiple of the areas spanned by pairs of letters in the generating alphabet. Then any single neighboring generator-swap suffices, if performed enough times, to produce area changes of any multiple of N. The steps will be organized to ensure that, though the height may change, it stays in the same residue class modulo N. Throughout, we will be assuming $n = \ell(\gamma) \gg N$.

First we simplify γ to γ_1 (Lemma 18) by shuffling letters, cashing in insignificant generators, and balancing lengths. We know that γ_1 is K–almost balanced with respect to the induced norm on $\mathfrak{m}$. This means that it has the form $c_{\mathrm{I}-1} a_{\mathrm{I}}^{n_{\mathrm{I}}} c_{\mathrm{I}} \cdots a_{\mathrm{J}}^{n_{\mathrm{J}}} c_{\mathrm{J}}$ and that for $\mathrm{I} < i < j < \mathrm{J}$, the values $\frac{n_i}{\sigma_i}$ and $\frac{n_j}{\sigma_j}$ differ by at most a bounded amount and that the values $\frac{n_{\mathrm{I}}}{\sigma_{\mathrm{I}}}$ and $\frac{n_{\mathrm{J}}}{\sigma_{\mathrm{J}}}$ can exceed these by at most a bounded amount, though of course these values are not necessarily integral. We can rewrite such a spelling γ_1 as

$$\gamma_1 = c'_{\mathrm{I}-1} \widehat{a}_{\mathrm{I}}^{s^-} c'_{\mathrm{I}} \widehat{a}_{\mathrm{I}+1}^{s} \cdots \widehat{a}_{\mathrm{J}-1}^{s} c'_{\mathrm{J}-1} \widehat{a}_{\mathrm{J}}^{s^+} c'_{\mathrm{J}}.$$

In particular the projection $\pi(\gamma_1)$ fellow-travels a CC geodesic.

Depending on the number of sides, we next apply unsimplification for shapes or patterns (Lemma 20 or 24) to obtain γ_2. In the pattern case, note that (a, b, c) is geodesically spelled by some rearrangement of the pattern w, because the pattern was obtained in the first place by shuffling the original spelling.

To complete the proof of the Theorem for shapes, we observe that we are in one of three cases: either the height difference $z_2 - z_0 < (2SS)_0$ so that we can not apply unsimplifcation; the unsimplification process had at least one three-sided surgery; or unsimplification had only two-sided surgery. If any three-sided surgery was performed, then our new spelling γ_2 evaluates to the same word as γ but is shorter, contradicting geodesity of γ. If only two-sided surgery was needed, then a γ_2 of equal length to γ has been produced, but with lower eccentricity. Finally, if $z_2 - z_0$ is smaller than some fixed bound, then the steps in the proof only made minor changes to γ, and retracing the argument this implies that γ was boundedly close to isoperimetric at the beginning of the process. □

Example 26. We give an example to illustrate an eccentric word geodesic being improved by the shape algorithm above. Consider the standard generators, fix a value D and take $M \gg D$. Let γ be the closed rectangular path

$$\mathsf{e}_1^{M-D}\mathsf{e}_2^{M+D}\mathsf{e}_1^{-M+D}\mathsf{e}_2^{-M-D}.$$

This has length $4M$ and encloses area $M^2 - D^2$, so it evaluates to the group element $(0, 0, M^2 - D^2)$. The CC geodesic reaching the same element would have length $4\sqrt{M^2 - D^2}$, which is strictly greater than $4M - 1$ if M is large enough compared to D, and this means that γ is a geodesic. It is already cyclically ordered and has no out-of-place letters, so $\gamma_1 = \gamma$. Balancing the sides produces $\gamma_2 = \mathsf{e}_1^M\mathsf{e}_2^M\mathsf{e}_1^{-M}\mathsf{e}_2^{-M}$, which has area M^2. Now we perform a 2-sided surgery, replacing $\mathsf{e}_1^M\mathsf{e}_2^M$ with $\mathsf{e}_1^{M-1}\mathsf{e}_2^{D^2}\mathsf{e}_1\mathsf{e}_2^{M^2-D^2}$. This reduces the area by D^2 while preserving length, so creates a geodesic to $(0, 0, M^2 - D^2)$ that 1–fellow-travels the CC geodesic.

8. Competition among shapes and patterns

8.1. **Linear comparison for shapes.** We have seen that when ω is a shape (simple or general), the map $\mathrm{Dom}_K \to \mathbb{Z}^3$ induced by ω taking $X \mapsto (s^-, s, s^+) \mapsto (a, b, \ell)$ is injective and affine. Therefore, for a given shape ω, the inverse map $(a, b, \ell) \mapsto \mathsf{s} = (s^-, s, s^+)$ is an affine function on $\omega(\mathrm{Dom}_K) \subset \mathbb{Z}^3$.

First, we define the *domain of competition*, $\mathrm{DomComp}_K(\omega, \omega')$ for a pair of shapes ω and ω' to be the inputs for which they reach the same horizontal position at nearby lengths:

$$\{(X, X') \in \mathrm{Dom}_K(\omega) \times \mathrm{Dom}_K(\omega') : |\ell - \ell'| \leq K \text{ and } \pi(\omega(X)) = \pi(\omega'(X'))\}$$

Define a *competition function* $f_{\omega\omega'} : \mathrm{DomComp}(\omega, \omega') \to \mathbb{Z}$ to be the difference in heights, $z(\omega(X)) - z(\omega'(X'))$. We show that if two shapes ever compete, then the domain of competition decomposes into rational families where that height difference is given by a linear function.

Definition 27. Given a general shape ω of type (I, J) and data X with lengths $\mathsf{s} = (s^-, s, s^+)$, we define the *trace* $\tau = \tau(\omega(X))$ to be the corresponding $\mathbf{I}$–arc

$$\tau = \widehat{\mathsf{a}}_{\mathrm{I}}^{s^-}\widehat{\mathsf{a}}_{\mathrm{I}+1}^{s} \dots \widehat{\mathsf{a}}_{\mathrm{J}-1}^{s}\widehat{\mathsf{a}}_{\mathrm{J}}^{s^+}.$$

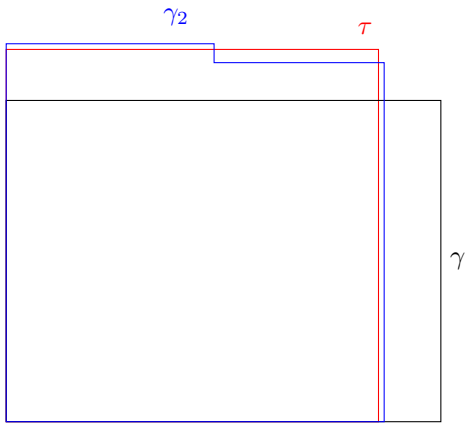


FIGURE 5. Here, a word geodesic γ with large eccentricity is shown compared to the corresponding CC geodesic τ, which can't be realized with integers. The algorithm *balances* γ and then *chips away* area to produce a geodesic γ_2 which evaluates to the same group element as the original γ but tracks close to τ.

That is, τ is equal to $\pi(\omega(X))$ with the break words deleted and the exponent differentials erased. Observe that by construction,

- the dependence of τ on X factors through (s^-, s, s^+);
- τ begins at $\mathsf{0} \in \mathfrak{m}$, and synchronously fellow-travels $\pi(\omega(X))$ with a fellow-traveller constant which depends only on ω and is independent of X; and
- for a given ω the difference between the projection of the endpoint of $\omega(X)$ and the endpoint of τ is independent of X, i.e., is constant on $\mathrm{Dom}_K(\omega)$. This is because this difference depends only on the b and c data from ω.

Lemma 28 (Linear comparison for shapes)**.** If $\mathrm{DomComp}_K(\omega, \omega')$ is nonempty, then there is a finite partition such that each piece $U_\delta \subset \mathrm{DomComp}_K(\omega, \omega')$ is defined by linear equations, linear inequalities, and congruences, and the comparison function $f_{\omega\omega'}|_{U_\delta}$ is linear.

Proof. Take K to be the bounded-difference constant from Corollary 17. We will partition the domain of competition into pieces for each $-K \le \delta \le K$ consisting of the subset of positions (a, b) reached by ω at some length ℓ and by ω' at length $\ell + \delta$. Call this subset U_δ. Let (s^-, s, s^+) and (t^-, t, t^+) denote the length data extracted from X and X' respectively.

Fixing this δ, we first consider the case where ω and ω' have the same combinatorial type, that is, $\mathrm{I} = \mathrm{I}'$, $\mathrm{J} = \mathrm{J}'$. Further, if $\mathrm{I} = \mathrm{J}$, recall that we have restricted the domain so that $s^+ = t^+ = 0$. We claim that the trace τ' fellow travels τ in projection and that the distance between corresponding sides is independent of a, b and ℓ. This is because the affine maps

$$(s^-, s, s^+) \mapsto (a, b, \ell)$$
$$(t^-, t, t^+) \mapsto (a, b, \ell + \delta)$$

have the same linear part, hence so do their inverses. Thus, $(s^- - t^-, s - t, s^+ - t^+)$ is constant on the domain of competition, which ensures fellow-traveling.

It follows that the area between the traces is linear on U_δ, as is the area between each of the shapes and its respective trace. (Area between two planar paths with different endpoints is measured by closing up with a straight chord.) Clearly the boost of each shape is also linear on DomComp. Thus $f_{\omega\omega'}$ is linear for each value of δ in the case where ω and ω' have the same combinatorial type. (Notice that in the case where τ and τ' might a priori differ as in the second case of Lemma 12, because of our restriction to Cone_0, they actually fellow-travel and the argument goes through.)

Next, consider the case where τ and τ' are almost the same combinatorial type. In this case $\omega(X)$ and $\omega'(X')$ are also of almost the same combinatorial type. For specificity let us consider the case where $\mathrm{I}+1=\mathrm{I}'$, $\mathrm{J}=\mathrm{J}'$, and s^- and $t-t^-$ are both bounded, so that there are only finitely many possible pairs $(s^-, t-t^-)$. For any such pair, the subset of U_δ realizing that pair is defined by linear equations. If we fix those values—i.e., treat $a_{\mathrm{I}}^{s^-}\, a_{\mathrm{I}+1}^{t-t^-}$ as a break word in $\omega(X)$—we can define new traces of the same combinatorial type and appeal to the case above.

Finally, we turn to the case where τ and τ' end close to the origin and have different types. Here, ω and ω' can only compete when τ and τ' are close to being the full polygon. But this implies that there are finitely many values (a,b) for which they compete. Furthermore, the set of (X, X') mapping to each of these finitely many (a,b) is determined by linear equalities and inequalities. For each such (a,b), the areas of $\omega(X)$ and $\omega(X')$ differ from a full isoperimetrix of scale s by amounts which are linear in X and X' respectively. Thus their areas differ from each other by amounts which are linear in X and X', and once again their respective boosts are also linear in X and X'. The result now follows. □

8.2. **Testing geodesity for shapes.** Consider the set

$$\{(a, b, w_n + j) : n \geq n_0(a,b), \quad 1 \leq j \leq w_{n+1} - w_n\},$$

containing the elements of the Heisenberg group in the (positive) regular range, i.e., the set of $(a,b,c) \in H(\mathbb{Z})$ with $c > W = w_{n_0}(a,b)$. By Bounded Difference (Cor 17), such an element $(a, b, w_n + j)$ has word length between $n+1$ and $n+K$.

Since Shape_K is a finite set, we can fix an arbitrary ordering of its elements.

Definition 29. For a general shape ω, let $G_\omega^\Delta(n)$ be the set of $(a,b,j) \in \mathbb{Z}^3$ such that $1 \leq j \leq w_{n+1} - w_n$ and ω is the first shape to geodesically realize $(a, b, w_n + j)$ at length $n + \Delta$.

Theorem 30 (Deciding geodesity for shapes)**.** For each shape ω and each $0 \leq \Delta \leq K$, the $G_\omega^\Delta(n)$ form a bounded rational family in $\mathbb{Z}^3$.

Proof. We will show that membership in $G_\omega^\Delta(n)$ is tested by finitely many linear equations, linear inequalities, and congruences.

For a shape ω, consider

$$A_\omega(n) = \{(a,b) : \omega \text{ produces a spelling of length } n \text{ over } (a,b)\}.$$

To see that $A_\omega(n)$ is a rational family, recall that $X \mapsto \ell(\omega(X)) = n$ is an affine map. Thus the sets $\{X \in \mathrm{Dom}_K(\omega) : \ell(\omega(X)) = n\}$ constitute a rational family. The map $X \mapsto (a,b)$ is also affine and thus the sets $A_\omega(n)$ are the affine push-forwards of a rational family and hence themselves rational.

For each shape ω consider

$$H_\omega(n) = \{(a,b) : \omega \text{ realizes the highest-height element } (a,b,w_n) \text{ at length } n\},$$

which is empty unless ω is a simple shape. We claim that for each simple shape ω, $H_\omega(n)$ is a rational family. For each $(a,b) \in A_\omega(n)$, ω fails to produce the highest-height element if there is ω' producing a higher element over (a,b) at length at most n. However, since $w_{n-2} < w_n$ this only needs to be tested for length n and $n-1$. Thus, for each potential competitor ω' only two inequalities need to be tested. But these are tested by the linear inequality $f_{\omega,\omega'}(X,X') \geq 0$ at $\delta = 0$ and $\delta = -1$. It follows that the sets $H_\omega(n)$ form a rational family as claimed.

Now for each pair of shapes α and β, note that

$$H_\alpha(n+1) \cap H_\beta(n) = \{(a,b) : \beta \text{ realizes } w_n(a,b) \text{ and } \alpha \text{ realizes } w_{n+1}(a,b)\}$$

is a rational family picking out positions at which α is highest-height at length $n+1$ and β is highest-height at length n. Given (a,b), we can search the finite list of shapes to find such a pair, and then (a,b,n) affinely determine s_α and $(a,b,n+1)$ determine s_β so that $\alpha(\mathsf{s}_\alpha)$ and $\beta(\mathsf{s}_\beta)$ are the highest-height paths. Thus, we can test the requirement that j satisfy $1 \leq j \leq w_{n+1}(a,b) - w_n(a,b)$ using equations which are linear in our data by seeing whether there exist shapes α, β for which $j \leq f_{\alpha\beta}(\mathsf{s}_\alpha, \mathsf{s}_\beta)$ at $\delta = 1$.

The requirement that ω realizes (a,b,w_n+j) at length $n+\Delta$ is similarly tested by $j = f_{\omega\beta}(\mathsf{s}, \mathsf{s}_\beta)$ at $\delta = \Delta$, i.e., by linear equalities and inequalities.

Finally, for any ω which realizes (a,b,w_n+j) at length $n+\Delta$, we must test whether this is geodesic, i.e., whether this length is shortest-possible. This is accomplished by testing all potential competitors ω' at lengths $\Delta' < \Delta$. This is finitely many competitors ω' and finitely many values Δ', and therefore determined by finitely many linear equalities and inequalities.

Finally, to see that ω is the lowest-numbered shape to produce such a geodesic, we simply check $\omega' < \omega$ at length Δ. □

8.3. **Linear comparison for patterns.** We will establish linear competition for patterns as we did for shapes above. For patterns w and w', define

$$\mathrm{DomComp}(\mathsf{w}, \mathsf{w}') = \{(n_1, n_2, n_1', n_2') : (a,b) = (a',b')\},$$

requiring that both paths end at the same horizontal position. Notice that on $\mathrm{DomComp}(\mathsf{w}, \mathsf{w}')$, the length difference $\ell(\mathsf{w}(n_1,n_2)) - \ell(\mathsf{w}'(n_1',n_2'))$ is constant.

Lemma 31 (Linear comparison for patterns)**.** The comparison functions $\max \mathcal{I}_\mathsf{w} - \max \mathcal{I}_{\mathsf{w}'}$ and $\min \mathcal{I}_\mathsf{w} - \min \mathcal{I}_{\mathsf{w}'}$ are affine on a finite partition of $\mathrm{DomComp}(\mathsf{w}, \mathsf{w}')$.

Equivalently, these can be regarded as affine functions on (n_1, n_2), or affine in (a,b) on those (a,b) whose fibers are reached by both w and w'.

Proof. The three statements are equivalent because on the appropriate sets, each of the three quantities, (n_1, n_2, n_1', n_2'), (n_1, n_2) and (a,b) determines the other two by an affine map. Linearity follows from the fact that the tops of the intervals are given piecewise by quadratic polynomials with the same leading coefficient, and the bottoms of the intervals are piecewise linear, over finitely many rational families that partition DomComp. □

8.4. **Testing geodesity for patterns.** Each pattern w is easily seen to determine maps from (n_1, n_2) to length, horizontal position, and $\ell(\mathsf{w}(n_1,n_2)) - n_0$. Notice that for each w, (a,b,ℓ) is an affine function of (n_1,n_2), and the map $(n_1,n_2) \mapsto (a,b)$ is injective.

Since Patt_K is a finite set, we can fix an arbitrary ordering of patterns as we did for shapes.

Definition 32. For a pattern w, let $G^\Delta_{\mathsf{w}}(n)$ be the set of $(a,b) \in \mathbb{Z}^2$ such that $n = n_0(a,b) = |(a,b)|_{\pi(S)}$ and w realizes some (a,b,c) at length $n+\Delta$.

Lemma 33 (Positions reached by patterns)**.** For each shape w and each $0 \le \Delta \le K$, the $G^\Delta_{\mathsf{w}}(n)$ form a bounded rational family in $\mathbb{Z}^2$.

Proof. The set of (a,b) reached by w is the push-forward under an affine map of the set of non-negative pairs (n_1,n_2). Now observe that in the Ith sector of the plane, the length $n_0 = n_0(a,b)$ is a periodic linear function in which the linear coefficient is independent of (a,b) and the constant term depends on the congruence class of (a,b) modulo the group generated by a_I and $\mathsf{a}_{\mathsf{I}+1}$. Note also that if $\mathsf{w}(n_1,n_2)$ ends over (a,b), then (n_1,n_2) and (a,b) are affine functions of each other. Of course then length of $\mathsf{w}(n_1,n_2)$ is an affine function of (n_1,n_2). Thus the difference $\ell(\mathsf{w}(n_1,n_2)) - n$, which gives Δ, is is a periodic function, and the result follows. □

Corollary 34 (Counting with patterns)**.** For each w and $0 \le \Delta \le K$ there are polynomials $p^\Delta_{\mathsf{w}}(a,b)$ of degree at most two such that for $(a,b) \in G^\Delta_{\mathsf{w}}(n)$ the number of group elements (a,b,c) with $c \ge 0$ geodesically spelled by w at length $n+\Delta$, and by no smaller-numbered pattern, is given by $p^\Delta_{\mathsf{w}}(a,b)$.

Proof. Clearly, the unstable elements of length $n_0+\Delta$ are those reached by some pattern w at length $n_0+\Delta$ but not by w' with length $n_0+\Delta'$ for any $\Delta' < \Delta$.

The p_{w} are defined by making the comparisons of the interval $\mathcal{I}_{\mathsf{w}}$ against competing intervals $\mathcal{I}_{\mathsf{w}'}$, and enumerating the points over (a,b) assigned to w as a finite sum/difference of the appropriate quadratic polynomials. □

9. The growth series

The growth series of (H,S) is now given as follows. The generators S determine a constant K so that the positive-height regular elements are enumerated by

$$\mathbb{S}^{\mathsf{reg}}(x) = \sum_{\omega} \sum_{n=0}^{\infty} \sum_{\Delta=0}^{K} \sum_{G^\Delta_\omega(n)} x^\Delta\, x^n,$$

where $\omega \in \mathrm{Shape}_K$ are the shapes described above.

The series enumerating unstable elements with $c \ge 0$ is

$$\mathbb{S}^{\mathsf{uns}}(x) = \sum_{\mathsf{w}} \sum_{n=0}^{\infty} \sum_{\Delta=0}^{K} \sum_{G^\Delta_{\mathsf{w}}(n)} p^\Delta_{\mathsf{w}}(a,b)\, x^\Delta\, x^n,$$

where $\mathsf{w} \in \mathrm{Patt}_K$ are the patterns described above. The difference in appearance between the two expressions corresponds to the fact that regular CC geodesics of a certain length only hit each fiber in a single point, while unstable CC geodesics may hit in an interval of size that is quadratic in the length.

Both series are rational by Theorem 4, because Shape_K and Patt_K are finite sets, the $G(n)$ are bounded rational families, and the p are polynomial. We then appeal to the height-reversing bijection $g \mapsto g^{-1}$ to similarly count the elements of non-positive height. This double-counts the elements at height zero.

Lemma 35 (Zero-height elements)**.** Let $\sigma^0(n) = \#\{(a,b,0) : |(a,b,0)|_S = n\}$ be the spherical growth function of height-zero elements. Then $\mathbb{S}^0(x) = \sum \sigma^0(n)x^n$ is rational.

Proof. The fiber over (a,b) has an element with $c = 0$ if and only if ab is even. Thus, our problem reduces to counting the set of such $(a,b) \in \mathbb{Z}^2$ with respect to the generating set $\pi(S)$. It is well-known that the set of lex-least geodesics in an abelian group is a regular language. Those ending at an element (a,b) with ab even is a regular subset of these. The set in question therefore has rational growth. □

Finally, we have

$$\mathbb{S}(x) = 2{\cdot}\mathbb{S}^{\mathsf{reg}}(x) + 2{\cdot}\mathbb{S}^{\mathsf{uns}}(x) - \mathbb{S}^{\mathsf{0}}(x).$$

This establishes that the spherical growth series $\mathbb{S}(x)$ and thus also the growth series $\mathbb{B}(x)$ is rational for any finite generating set of $H(\mathbb{Z})$, finishing the proof of Theorem 1.

10. Applications, remarks, and questions

10.1. **Languages.** Each shape defines a language $\mathcal{L}(\omega)$. For $\textsc{j} > \textsc{i} + 1$, these languages are not regular. For $\textsc{j} > \textsc{i} + 2$, they are not context-free.

This is attributable to non-commutativity: what could be accomplished with a bounded counter if the group were abelian is a non-regular language otherwise. For instance, $\{a^n b^n\}$ is non-regular, even though $\{(ab)^*\}$ enumerates words with the same letters. The words represented by our shapes of geodesics need to be nearly balanced, and this breaks regularity.

It was pointed out to us by Cyril Banderier that a recursion with positive integer coefficients implies the existence of *some* regular language enumerated by the function, though not necessarily the language of geodesics for (G,S). This holds in the special case of (H, std), which is extremely intriguing.

10.2. **Cone types.** We recall the definition of *cone type* from [8].

Definition 36. Consider the Cayley graph $\mathsf{Cay}(G,S)$ of group G with generating set S. Given $g \in G$, the *cone at* G, denoted $C(g)$, consists of all paths σ based at g with the property that word length $|\sigma(t)|$ is strictly increasing along σ. The *cone type* of g consists of the cone of g translated to the origin, i.e., $g^{-1}(C(g))$.

For $\mathsf{Cay}(G,S)$ to have finitely many cone types is almost exactly the same thing as having the language of geodesics in $\mathsf{Cay}(G,S)$ be a regular language. If $\mathsf{Cay}(G,S)$ has finitely many cone types, these cone types can be used as the states of a finite state automaton which accepts the language of geodesics. This is because the cone type of G tells us which generators are outbound at g. However, the cone type of g encodes additional information, namely which edges are "half outbound": if an edge e of $\mathsf{Cay}(G,S)$ connects two elements g and g' with $|g| = |g'| = n$, then the midpoint of this edge is at distance $n + \frac{1}{2}$ from the origin. We believe that there is no known example of a group presentation for which the language of geodesics is regular, but which has infinitely many cone types.

¿From the shape theorem we easily recover the (already known) fact that H has infinitely many cone types in every generating set. In particular, it has no generating set where the language of geodesics is regular.

To see this, just note that there are infinitely many possibilities for how long a geodesic continues in a particular significant direction before turning to the successive direction, depending on what shape has reached the point $g = (a, b, c)$ at what scale.

Brian Rushton has pointed out to us that the presence of infinitely many cone types implies that there is no associated subdivision rule. (See [23].)

10.3. **Almost convexity.** A metric space is called *almost convex* (k) or $AC(k)$ if there exists a constant $N = N(k)$ such that for any two elements x, y in any metric sphere $S_n(x_0)$ with $d(x, y) \leq k$, there is a path of length at most N connecting x and y in $B_n(x_0)$. That is, convexity would require that for two points on a sphere of any radius, connecting them inside the ball is efficient; almost-convexity is the existence of an additive bound on the inefficiency. This was defined by Cannon in [9], where he also showed that for Cayley graphs of finitely generated groups, $AC(2) \Rightarrow AC(k) \quad \forall k$. The importance of this property is that it gives a fast algorithm for constructing the Cayley graph. Almost-convexity is known for hyperbolic groups and virtually abelian groups with any finite set of generators, and for Coxeter groups and certain 3–manifold groups with standard generators. Several weakenings and strengthenings of the property have been proposed and studied by various authors. It was established for $H(\mathbb{Z})$ with standard generators in [24], but to our knowledge has not been extended to arbitrary generators, which we settle here by using once again the comparison of the CC and word metrics.

Intriguingly, the dissertation of Carsten Thiel [27] establishes that higher Heisenberg groups are *not* AC in their standard generators, which corresponds remarkably to Stoll's finding of non-rational growth for the same examples.

Lemma 37. The CC metric on $H(\mathbb{R})$ induced by any rational polygonal norm is almost convex. (That is, it is $AC(k)$ for all k.)

Proof. Consider $x, y \in \mathcal{S}_n$ with $d_{\text{CC}}(x, y) \leq 2$. First we show that if there exist geodesics $\overline{0x}$ and $\overline{0y}$ that K–fellow-travel in projection, then there exists a connecting path from x to y of bounded length inside the ball. To construct this path, begin with a constant $m \gg 1$. We will build a path from $\pi(x)$ to $\pi(y)$ as follows: backtrack distance mK along $\pi(\overline{0x})$. Connect geodesically to the point w that is $n - mK$ from the origin along $\pi(\overline{0y})$ and finish by connecting w to $\pi(y)$ along $\pi(\overline{0y})$. This path has length at most $(2m + 1)K$. Its lift connects x not to y but to something else in the same fiber over $\pi(y)$, differing in height by at most mK^2 because that is the most area that can be contained in the "rectangular" strip enclosed by the path we have built. To correct this, we can splice a loop into our planar path at the point w. This loop follows a parallelogram with sides $t\mathsf{u}$ and $t\mathsf{v}$ for some successive significant generators, where t is chosen so that the area of the parallelogram, $t^2(\mathsf{u} \wedge \mathsf{v})$, is the height differential to be made up. This has length at most $4\sqrt{\frac{m}{\mathsf{u}\wedge\mathsf{v}}}K$. Since m was chosen to be large, this length is less than mK and so the lift of the concatenated path stays inside $\mathcal{B}_n$. Thus we have connected x to y by a path inside the ball, of length bounded independent of x, y, n.

To complete the proof, we must reduce to this case. By possibly inserting one extra point z and separately considering the two pairs x, z and z, y, we will cover all possibilities with the following cases.

Case 1: x, y both unstable and in the same sector.

Then there are fellow-traveling geodesics as required: if the sector is between significant directions a_I and $\mathsf{a}_{\mathrm{I}+1}$, then x is reached in exactly one way by a geodesic whose shadow is of the form $\mathsf{a}_\mathrm{I}^i\mathsf{a}_{\mathrm{I}+1}^j\mathsf{a}_\mathrm{I}^k$. Likewise y has a unique such geodesic, and they must fellow-travel to reach nearby endpoints.

Case 2: x, y both regular and of the same combinatorial type.

In this case, the geodesics from the origin are unique, and both project to P–arcs for the defining polygon P of the norm with the same combinatorial type. ¿From the fellow-traveling lemma for P–arcs (Lemma 12) we know that these fellow-travel in projection and we proceed as before.

Case 3: One of x, y projects to the origin (say $y = (0, 0, c)$).

In this case we fix any geodesic from the origin to x. There are many geodesics reaching y (corresponding to choosing any starting position on P), and we can take one of the same combinatorial type as the path chosen for x. These then fellow-travel in projection.

This concludes the proof of $AC(2)$. To deduce $AC(k)$ for any other k, one can simply apply $\delta_{2/k}$ to send the points x, y to a sphere on which they are at most distance 2 apart, then apply $\delta_{k/2}$ to the path constructed above. In this way, we get $N(k) = \frac{k}{2}N(2)$. □

Theorem 38. The Heisenberg group is almost convex with any word metric.

Proof. Start with g_1, g_2 with $|g_1| = |g_2| = n$ and $|g_1 g_2^{-1}| \leq 2$, and let K be the constant bounding the difference between the word and CC metrics, as in Cor 17. Then if B_n is the ball of radius n in the word metric and $\mathcal{B}_n$ is the ball of radius n in the associated CC metric, we have $B_n \subset \mathcal{B}_{n+K}$ and $\mathcal{B}_{n-K} \cap H(\mathbb{Z}) \subseteq B_n$.

Fix any $p \gg 2K$. Let h_1 be a group element obtained by backtracking p steps along a geodesic spelling of g_1, so that $|h_1| = n - p$, and define h_2 similarly. The distance $|h_1 h_2^{-1}|$ is at most $2 + 2p$, and since the (continuous) group is $AC(2 + 2p)$, there is a constant $N(2+2p)$ so that a CC path γ exists between h_1 and h_2 of length at most N and contained totally inside $\mathcal{B}_{n-p+K}$. As γ is traversed from h_1 to h_2, construct an ordered set of integer points by choosing a nearest point at each time. Since the diameter of a fundamental domain for $H(\mathbb{Z})$ is bounded, say by Δ, each of these points is contained in the Δ–neighborhood of γ and therefore each is within 2Δ CC distance of the previous and next point in the sequence. These round-off points all lie in $\mathcal{B}_{n-p+K+\Delta}$. Two successive points can be connected by a word path of length at most $2\Delta + K$, and the word path from h_1 to h_2 built by concatenating these must lie inside $\mathcal{B}_{n-p+2K+2\Delta}$. There are at most $N/2\Delta$ round-off points, so the total length of the word path from h_1 to h_2 is bounded by $(N/2\Delta)(2\Delta + K)$. Since p was chosen to ensure that $n - p + K + 2\Delta < n - K$, this path lies inside $\mathcal{B}_{n-K} \cap H(\mathbb{Z}) \subseteq B_n$. Piecing this together we obtain a path from g_1 to g_2 inside B_n of length at most $(N/2\Delta)(2\Delta + K) + 2p$. □

10.4. **Open questions.**

10.4.1. *Scope of rational growth in the nilpotent class.* Our argument should carry through with small modifications for groups that are virtually $H(\mathbb{Z}) \times \mathbb{Z}^d$. We know from Stoll's result that not all two-step groups have rational growth, even with respect to their standard generators. However it is possible (for instance) that free nilpotent groups do.

Question 39. Which nilpotent groups have rational growth in all generating sets?

On the other hand, one could try to mimic and extend the Stoll construction.

Question 40. Does every nilpotent group have rational growth with respect to at least one generating set? In the other direction, for which nilpotent groups is the fundamental volume transcendental for standard generators (which would rule out rationality by Theorem 2)?

10.4.2. *Period and coefficients.* In the polynomial range (i.e., $f(n) \leq An^d$ for some A, d), rational growth is equivalent to the property that $f(n)$ is *eventually quasi-polynomial*, i.e., there are a finite period N, polynomials $f_1, \ldots, f_N$, and a threshold T such that

$$n \geq T, \quad n = kN + i \implies f(n) = f_i(n).$$

For example, Shapiro's computation of the spherical growth for the Heisenberg group with standard generators showed it to be eventually quasipolynomial of period twelve, and in fact only the constant term oscillates:

$$\sigma(n) = \frac{1}{18}\left(31n^3 - 57n^2 + 105n + c_n\right),$$

where $c_n = -7, -14, 9, -16, -23, 18, -7, 32, 9, 2, -23, 0$, and then repeats mod 12, for $n \geq 1$. (So that $\sigma(1) = 4$, $\sigma(2) = 12$, and so on.)

It follows that the (ball) growth function $\beta(n) = \sum_{k=0}^{n} \sigma(k)$ is also quasipolynomial of period twelve, with only its constant term oscillating. We note that this implies that the growth function for standard generators is within bounded distance of a true polynomial in n.

Preliminary calculations indicate that several other generating sets also have the property that only the constant terms oscillate; in these examples, the periods relate both to the sidedness of the fundamental polygon and to the index of the sublattice of $\mathbb{Z}^2$ generated by its extreme points.

Question 41. How does the generating set S determine the period of quasipolynomiality of the growth function? Which coefficients oscillate? We know that the top coefficient of $\beta(n)$ is the volume of the CC ball; is the second coefficient well-defined, and if so is it a "surface area"? Are all growth functions bounded distance from polynomials?

References

[1] H. Bass, The degree of polynomial growth of finitely generated nilpotent groups. Proc. London Math. Soc. (3) 25 (1972), 603–614.

[2] M. Benson, Growth series of finite extensions of $\mathbb{Z}^n$ are rational, Invent. Math. 73 (1983), no. 2, 251–269.

[3] Max Benson, On the rational growth of virtually nilpotent groups, Ann. Math. Stud 111 (1987), 185–196.

[4] E. Breuillard, Geometry of groups of polynomial growth and shape of large balls. arXiv:0704.0095

[5] E. Breuillard and E. LeDonne, On the rate of convergence to the asymptotic cone for nilpotent groups and subFinsler geometry. Proc. Natl. Acad. Sci. USA 110 (2013), no. 48, 19220–19226.

[6] H. Busemann, The isoperimetric problem in the Minkowski plane. AJM 69 (1947), 863–871.

[7] J. Cannon, The growth of the closed surface groups and compact hyperbolic Coxeter groups. Circulated typescript, Univ. Wisconsin, 1980.

[8] J. Cannon, The combinatorial structure of cocompact discrete hyperbolic groups. Geom. Dedicata 16 (1984), no. 2, 123–148.

[9] J. Cannon, Almost convex groups, Geom. Dedicata 22 (1987), no. 2, 197–210.

[10] L. Capogna, D. Danielli, S. Pauls and J. Tyson, *An Introduction to the Heisenberg Group and to the Sub-Riemannian Isoperimetric Problem*. Birkhauser, Progress in Mathematics, 2007.
[11] M. Duchin and C.P. Mooney, *Fine asymptotic geometry of the Heisenberg group*, Indiana University Math Journal 63 No. 3 (2014), 885–916.
[12] D.B.A. Epstein, J.W. Cannon, D.F. Holt, S.V.F. Levy, M.S. Paterson, and W.P. Thurston, Word processing in groups. Jones and Bartlett, 1992.
[13] R. Grigorchuk and P. de la Harpe, On problems related to growth, entropy, and spectrum in group theory, J. Dynam. Control Systems 3 (1997), no. 1, 51–89.
[14] M. Gromov, Groups of polynomial growth and expanding maps. Inst. Hautes tudes Sci. Publ. Math. No. 53 (1981), 53–73.
[15] M. Gromov, Hyperbolic groups. Essays in group theory, 75–263, Math. Sci. Res. Inst. Publ., 8, Springer, New York, 1987.
[16] Y. Guivarc'h, Groupes de Lie à croissance polynomiale. (French) C. R. Acad. Sci. Paris Sér. A-B 271 1970 A237–A239.
[17] Y. Guivarc'h, Croissance polynomiale et priodes des fonctions harmoniques. (French) Bull. Soc. Math. France 101 (1973), 333–379.
[18] P. de la Harpe, Topics in geometric group theory. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 2000.
[19] S.A. Krat, Asymptotic properties of the Heisenberg group. Journal of Mathematical Sciences, Vol. 110, No. 4 (2002) 2824–2840.
[20] A. Mann, How groups grow. London Mathematical Society Lecture Note Series, 395. Cambridge University Press, Cambridge, 2012.
[21] W. Neumann and M. Shapiro, Automatic structures, rational growth, and geometrically finite hyperbolic groups. Invent. Math. 120 (1995), no. 2, 259–287.
[22] P. Pansu, Croissance des boules et des géodésiques fermées dans les nilvariétés. Ergodic Theory Dynam. Systems 3 (1983), no. 3, 415–445.
[23] B. Rushton, Classification of subdivision rules for geometric groups of low dimension. Conform. Geom. Dyn. 18 (2014), 171–191.
[24] M. Shapiro, A geometric approach to the almost convexity and growth of some nilpotent groups. Math. Ann. 285, 601–624 (1989).
[25] M. Stoll, Rational and transcendental growth series for the higher Heisenberg groups. Invent. math. 126, 85–109 (1996).
[26] M. Stoll, On the asymptotics of the growth of 2–step nilpotent groups. J. London Math. Soc. (2) 58 (1998) 38–48.
[27] C. Thiel, Zur fast-Konvexität einiger nilpotenter Gruppen. (German) [On the almost convexity of some nilpotent groups] Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 1991.

Department of Mathematics, Tufts University

Department of Biology and Biochemistry, University of Bath