Fine Asymptotic Geometry in the Heisenberg Group

MOON DUCHIN & CHRISTOPHER MOONEY

ABSTRACT. For every finite generating set on the integer Heisenberg group $H(\mathbb{Z})$, we know from a fundamental result of Pansu on nilpotent groups that the word metric has the large-scale structure of a Carnot-Carathéodory Finsler metric on the real Heisenberg group $H(\mathbb{R})$. We study the properties of those limit metrics and obtain results about the geometry of word metrics that reflect the dependence on generators, notably, asymptotic density results for geometric properties.

For example, stability of geodesics and distortion of subgroups can be made statistical. This contributes to a small literature on asymptotic density results that depend nontrivially on generators for nonfree groups. Our methods also allow us to a pursue a "geometry of numbers" for nilpotent groups.

1. INTRODUCTION

In this paper, we focus on the three-dimensional integer Heisenberg group $H(\mathbb{Z})$, a uniform lattice in the real Heisenberg group $H(\mathbb{R})$, with the word metric coming from a finite generating set. We study large-scale geometric properties that are sensitive to the choice of generators, such as the shape and stability of long geodesics, and asymptotic density of classes of elements.

We will take an approach based on work of Pansu and Breuillard [3, 11], exploiting the fact that all word metrics on $H(\mathbb{Z})$ have the large-scale structure of Carnot-Carathéodory Finsler metrics on $H(\mathbb{R})$; that is, the Lie group with its CC metric is the asymptotic cone of the lattice with the chosen word metric. However, we will study large-scale geometric properties that are lost when one passes

Indiana University Mathematics Journal ©, Vol. 63, No. 3 (2014)

to the asymptotic cone. (Note that studying the Heisenberg group up to quasi-isometry is even coarser than working with the asymptotic cone; quasi-isometries identify all word metrics as well as all left-invariant metrics on $H(\mathbb{R})$ with one another.) Thus, we develop a technology for working with these word metrics that uses the homogeneous dilation at finite scale together with combinatorial arguments, instead of the usual tools of coarse geometry. We think of this set of questions and techniques as belonging to "fine asymptotic geometry."

The geometry of nilpotent groups is closely tied to geometric minimax problems, and so Section 2 is devoted to the explicit solution of relevant isoperimetric problems in normed planes. In Section 3, we apply those findings, giving a very explicit description of the limit metric associated to an arbitrary word metric. The main result of this section is Theorem 3.1, which is summarized here.

Theorem (Structure Theorem). *For any finite symmetric generating set S of $H(\mathbb{Z})$, the limiting* CC *metric admits a complete description of its geodesics, classified into no more than $|S|^2 - 2|S|$ combinatorial types of* regular geodesics *and no more than $|S|$ types of* unstable geodesics. *The unit sphere S in the* CC *metric is a piecewise union of the graphs of finitely many quadratic polynomials.*

As a corollary of this description, we can analyze the uniqueness of geodesics in these limit metrics, describing exactly which points $x \in H(\mathbb{R})$ are reached by more than one geodesic segment based at 0.

We use this structure theorem in Section 4 to understand the geodesics in Cayley graphs for $H(\mathbb{Z})$, showing that although word geodesics may be very unruly, they are tracked to within a controlled distance by much better-behaved CC geodesics: the Hausdorff distance is of lower-order growth than the length of the word (see Lemma 4.3).

Lemma (Tracking Lemma). *Away from a certain unstable locus of endpoints, word geodesics are sublinearly tracked by* CC *geodesics.*

This opens the door to the study of geometric probability in the discrete Heisenberg groups, and we give some applications in Section 5–Section 7. For instance, in Theorem 5.3, we quantify the instability of word geodesics. A sequence of group elements may be said to be geodesically stable with respect to a generating set if, in the Cayley graph, the Hausdorff distance between the geodesic spellings of the word can only differ from each other sublinearly in the word length. When this is made precise, one can see that stable elements have zero density in free abelian groups, but full density in infinite hyperbolic groups. The situation is different in the Heisenberg group.

Theorem (Instability Theorem). *The asymptotic density of geodesically stable elements of $H(\mathbb{Z})$ is a rational number strictly between 0 and 1, depending on the generating set. For the standard generators, this density is precisely $\frac{19}{31}$.*

As a second application (Section 6), we give quantitative measures of subgroup distortion that are finer than the usual ones, illustrated by computations for abelian subgroups of $H(\mathbb{Z})$.

Finally, in Section 7, we apply the structure theorem to a generalized Gauss circle problem: Theorem 7.2 asserts that, in every CC limit metric, we can count lattice points to first order in the annular shell between balls. This gives fine quantitative data about group growth.

Theorem (Counting Theorem). *For any* CC *metric induced by a polygon with integer vertices, the number of lattice points between the spheres of radius* n *and* $n-1$ *is equal to* $4Vn^3 + O(n^2)$*, where* V *is the volume of the* CC *unit ball.*

Note that this is strictly better than what is logically implied by the best-known estimate of the number of lattice points in the ball of radius n, which is $Vn^4 + O(n^3)$ by [14]. (In fact, the full theorem is more precise than what is stated above: we can count points in the annulus in any radial direction.) Lattice point counting has many applications; in Euclidean space, counting points in dilates of polytopes and smooth figures is its own industry, with connections to algebraic geometry and number theory. (See [1] for an excellent introduction.) This result can be thought of as a step toward a "geometry of numbers" for nilpotent groups, which combines aspects of the circle problem with aspects of Ehrhart theory.

Finally, we include an appendix with more detailed computations for the standard generators specifically. We give a geometric/combinatorial description of the discrete sphere S_n for $S = \mathsf{std}$, and we show how to count spherical points in any cone. The counting measure on discrete spheres is thus shown to limit to a cone measure on the limit shape S, for which we have a concrete and finite description. This enables spherical averaging—sharper than asymptotic density calculations, which are averages over balls—with respect to the standard generators.

1.1. Nilpotent groups and the Heisenberg group. We review a few facts about the sub-Riemannian geometry of the Heisenberg group; see [7] for a comprehensive reference.

Our main object of study is the three-dimensional integer Heisenberg group $H(\mathbb{Z})$, the subgroup of matrices with integer entries in $H(\mathbb{R})$ (the real Heisenberg group). 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

$$X_0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad Y_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

There are *horizontal planes* at every point in $H(\mathbb{R})$, generated by a push-forward of X_0 and Y_0 by left multiplication in $H(\mathbb{R})$ to produce left-invariant vector fields X and Y. We say that a curve in $H(\mathbb{R})$ is *admissible* if all of its tangent vectors lie in these horizontal planes.

We will use the *exponential coordinates* on $H(\mathbb{R})$ given by the following representation:

$$(x, y, z) \leftrightarrow \begin{pmatrix} 1 & x & z + \frac{1}{2}xy \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}.$$

These coordinates have the property that $(x, y, z)^n = (nx, ny, nz)$. It is easily verified that the curve $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ is admissible if and only if

$$\gamma_3' = \frac{1}{2}(\gamma_1\gamma_2' - \gamma_2\gamma_1').$$

For integers x and y, define $\varepsilon(x, y)$ to be $\frac{1}{2}$ if x and y 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.

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 $(x, y, z) \mapsto (x, y)$. Recall that a path in a metric space is called *rectifiable* if it has finite length.

Lemma 1.1 (Standard fact from Heisenberg geometry). *For any rectifiable path $\gamma = (\gamma_1, \gamma_2)$ from $(0, 0)$ to (x, y), there is a unique admissible curve $\bar{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$. The lifted curve connects the origin to the point (x, y, z), where z is equal to the signed Euclidean area of R, the region in the plane enclosed by γ and a straight chord from $(0, 0)$ to (x, y).*

Proof. To construct the lift, we just need to supply the coordinate function γ_3, which can be done by integrating in the defining formula for admissibility. Note that $x(t)y'(t) - y(t)x'(t)$ is identically zero along a linear parametrization of the straight chord between the endpoints, so we can concatenate that chord with the path in the plane to obtain the closed curve ∂R. We get

$$z = \int_{\partial R} \gamma_1\gamma_2' - \gamma_2\gamma_1' = \int_R \mathrm{d}x \wedge \mathrm{d}y,$$

which is the area of R. □

From now on, the area of this region R will be called the *balayage area* associated to a curve γ.

Choose a centrally symmetric convex body $Q \subset \mathfrak{m}$ with boundary L, and let $\| \cdot \|_L$ denote the corresponding norm; this is the norm having Q as its unit ball and L as its unit sphere. We use the notation $\mathfrak{m}_L = (\mathfrak{m}, \| \cdot \|_L)$ for this normed plane. This induces a Carnot-Carathéodory Finsler metric (from now on, CC metric) on the Lie group: the distance between two points is the infimal length of an admissible path between them, measured by using the norm on the tangent vectors. (This is well defined for admissible curves because the tangent vectors lie in copies of $\mathfrak{m}$ pushed forward under multiplication in $H(\mathbb{R})$, and it is a classical fact that this gives $H(\mathbb{R})$ the structure of a geodesic space.) Note that measuring the length of an admissible curve in $H(\mathbb{R})$ with respect to this CC metric is equivalent to measuring the length of its projection to $\mathfrak{m}$ with respect to the norm $\| \cdot \|_L$. If L is a polygon, we call the induced metric a *polygonal* CC *metric* on $H(\mathbb{R})$. (In higher dimension, if L is a polytope, the associated norms are sometimes called *crystalline*.)

These CC metrics are equipped with a dilation $\delta_t(x, y, z) = (tx, ty, t^2 z)$ that is a metric similarity, scaling lengths and distances by t, areas in $\mathfrak{m}$ by t^2, and volumes by t^4. For any set $E \subset H(R)$, let ΔE denote its full cone under dilation, and let $\hat{E}$ denote its cone to the origin:

$$\Delta E = \{\delta_t(\mathsf{x}) \mid \mathsf{x} \in E,\ t \geq 0\}; \quad \hat{E} = \{\delta_t(\mathsf{x}) \mid \mathsf{x} \in E,\ 0 \leq t \leq 1\}.$$

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 $H(\mathbb{R})$ if and only if its L-length is minimal among all curves with the same endpoints and enclosing the same area.

Our motivating questions are about probability, or *asymptotic density*, in the Heisenberg groups. More generally, for a lattice Γ with finite generating set S in a Lie group G, to measure the density of subsets $U \subset \Gamma$ and $V \subset G$, let $B_n = B_n^S$ be the ball of radius n in the word metric, and let $B_r(0)$ be the ball of radius r about the identity in G. Then,

$$\mathrm{Prob}_\Gamma(U) := \lim_{n\to\infty} \frac{|B_n \cap U|}{|B_n|}; \quad \mathrm{Prob}_G(V) := \lim_{r\to\infty} \frac{\mathrm{vol}(B_r(0) \cap V)}{\mathrm{vol}(B_r(0))}.$$

Note that Lebesgue measure is left invariant; hence, it is a Haar measure on $H(\mathbb{R})$, and so we will be able to talk about ratios of volumes unambiguously.

1.2. Limit shapes and limit metrics. We have the following extremely general theorem, first proven for nilpotent groups by Pansu, and extended to all periodic pseudometrics on simply connected solvable Lie groups of polynomial growth by Breuillard [3]. Here, we just state the result for the very special case of word metrics on $H(\mathbb{Z}) \leq H(\mathbb{R})$. We will always assume that generating sets are symmetric ($S = -S$).

Theorem 1.2 (**Pansu** [11]). *Consider a word metric on $H(\mathbb{Z})$ given by a finite generating set S. Let $\pi(S)$ be the linear projection of the generators S to the horizontal subspace $\mathfrak{m}$, let Q be the convex hull of $\pi(S)$, and let L be its boundary polygon. Then, the limit $S = \lim_{n\to\infty} \delta_{1/n} S_n$ exists (as a Gromov-Hausdorff limit), and is equal to the unit sphere in the CC metric induced by the norm $\|\cdot\|_L$ on $\mathfrak{m}$.*

Equivalently,

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

Theorem 1.3 (**Krat** [10]). *Furthermore, there exists a constant $K = K(S)$ such that $d_{\mathrm{CC}}(\mathsf{x}, 0) - K \leq |\mathsf{x}|_S \leq d_{\mathrm{CC}}(\mathsf{x}, 0) + K$ for all $\mathsf{x} \in H(\mathbb{Z})$.*

Note that Krat's result finding a bounded difference between the word metric and the CC metric establishes something stronger for $H(\mathbb{Z})$ than is true in the general nilpotent case treated by Pansu, where one only has that the ratio goes to 1. In fact Breuillard showed in [3] that there exist 2-step nilpotent groups where bounded difference fails.

For the word metric $(H(\mathbb{Z}), S)$, we will call S and the induced CC metric the *limit shape* and *limit metric*, respectively. The volume of the unit ball in the CC metric will turn out to be an important number to attach to both the word metric and the CC metric itself; we will denote it by

$$V = V(S) = V(L) := \mathrm{vol}(\hat{S}).$$

The Hausdorff dimension of $H(\mathbb{R})$ is equal to 4, despite the topological dimension of 3, as one can see clearly by considering the growth of a cube under the dilation. Thus, the volume of the ball of radius r in the CC metric equals $V \cdot r^4$, and in $H(\mathbb{Z})$, we have $|B_n| = V \cdot n^4 + O(n^3)$.

1.3. Relationship to prior literature on $H(\mathbb{Z})$**.** Pansu's seminal paper from 1983 [11], based on his dissertation, shows that the ratio of limit metric to word metric goes to 1 for all virtually nilpotent groups with finite generating sets, and that these word metrics on $H(\mathbb{Z})$ satisfy $|B_n| = \alpha \cdot n^d + o(n^d)$. Breuillard's preprint from 2007 [3] extends this result (that the growth has a well-defined leading coefficient) to a larger class of groups, and computes the coefficient. Breuillard also gives extremely concrete geometric constructions in that paper, especially in the Appendix, where some explicit computations that inspired the current work are shown for $(H(\mathbb{Z}), \mathsf{std})$. Also, an interesting new work of Breuillard and Le Donne [4] quantifies the rate of convergence for the limit analogous to Theorem 1.2 for all word metrics in nilpotent groups.

In 1996, Stoll showed that 2-step nilpotent groups with infinite cyclic derived subgroup have a finite generating set for which the growth series is rational; on the other hand, higher Heisenberg groups (of dimension five and up) have a finite generating set for which the growth series is transcendental [13]. These results are recovered by Breuillard by geometric methods. Stoll 1998 [14] refines Pansu's asymptotics by showing that, for 2-step nilpotent groups with finite generating sets, $|B_n| = \alpha \cdot n^d + O(n^{d-1})$, and this error term is "often" sharp—compare our counting theorem below (Theorem 7.2), which gets a finer result, and also applies in any direction (i.e., it counts points in cones as well as the whole space).

M. Shapiro wrote a study of $H(\mathbb{Z})$ with respect to its standard generators in 1989 [12], giving a description of S_n, then computing an exact formula for $|S_n|$ as a quasi-polynomial function of n (a polynomial whose coefficients oscillate with finite period). In this remarkable formula, he not only gives a well-defined leading coefficient, but in fact computes all the coefficients, and finds that only the constant term oscillates (with period twelve, as it turns out). He uses this study to establish that $H(\mathbb{Z})$ has infinitely many cone types, and is almost convex in the sense of Cannon. Blachère 2003 [2] computes an exact word length formula with respect to the standard generators, using it to show "almost-connectedness" of discrete spheres S_n. We note that the calculations of Shapiro and Blachère for S_n with respect to $S = \mathsf{std}$ are somewhat combinatorial, and that they are essentially equivalent to the more geometric description of S_n given below in Section A (which also give directional information, as with the counting results).

The paper [9] of Duchin-Lelièvre-Mooney studied the limit shapes for $\mathbb{Z}^d$, considering not only the limit metric guaranteed by Pansu but also a *limit measure* for any finite generating set S. It was shown that counting measure on $(1/n)S_n$ converges to a measure on L called *cone measure*, which assigns to a measurable set $\sigma \subset L$ the measure $\mu_L(\sigma) = \mathrm{vol}(\hat{\sigma})/\mathrm{vol}(\hat{L})$. (That is, it is the proportion of the area of Q that is subtended by σ, in this case by Euclidean dilation.) Below, in Theorem A.1 of Appendix A, we will obtain the analogous result for $H(\mathbb{Z})$ with respect to the standard generators, which enables us to calculate spherical averages; conjecturally, cone measure is the limit measure for all finite generating sets.

In work based on her dissertation, Dani 2007 [8] has some striking results on asymptotic density: for any virtually nilpotent group, she gives an exact formula for the density of finite-order elements. Further, she shows that the values of these densities range over all of $\mathbb{Q} \cap [0, 1]$ as the groups vary.

Finally, after this paper was initially posted, we were alerted to the fact that many of these questions had previously been considered by Michael Stoll in written notes (including a structure theorem and counting results exactly like our Theorem 3.1 and Theorem 7.2). His intended applications were somewhat different from ours, and the notes were never published or made public. We hope that this gives an indication that this direction of inquiry is natural and interesting from more than one point of view.

2. Isoperimetry and Isoarea in Normed Planes

2.1. The classical problem. We describe the unit sphere in the CC metric as the set of endpoints of geodesics of length one based at $\mathsf{0}$. As seen above, understanding CC geodesics in $H(\mathbb{R})$ amounts to the problem of finding paths in $\mathfrak{m}$ of minimal L-length given their endpoints and balayage area. For closed loops, this is just the classical isoperimetric problem in the normed space $\mathfrak{m}_L$, which was elegantly solved by Busemann in the 1940s as follows.

Let the *polar dual* Q^* of a convex, centrally symmetric polygon Q be defined by

$$Q^* = \{\mathsf{x} \in \mathfrak{m} \mid \mathsf{x} \cdot \mathsf{y} \le 1,\ \forall\, \mathsf{y} \in Q\}.$$

Then if u and v are successive vertices of Q, the dual Q^* has a vertex a corresponding to the edge between them which is the solution to $\mathsf{a} \cdot \mathsf{u} = \mathsf{a} \cdot \mathsf{v} = 1$. (That is, if the side is called σ and the line through the origin perpendicular to σ is called ℓ, then the vertex dual to σ is a vector pointing in the direction of ℓ, and whose length is the reciprocal of the length of the projection of u or v to ℓ.) In particular, if Q is a polygon with $2N$ sides, then Q^* is also a $2N$-gon.

Theorem 2.1 (**Busemann** [6]). *Consider the norm on $\mathbb{R}^2$ induced by a convex centrally symmetric body Q. Then, the maximal ratio of enclosed area to perimeter is uniquely achieved (up to scale) by the* isoperimetrix *I, which is defined to be the boundary of the rotate by $\pi/2$ of the polar dual Q^* of Q.*

If Q is a polygon, then the sides of I are parallel to the vertex directions of Q. Here is the calculation: suppose a and b are successive vertices of Q^*, so that they are dual to successive edges of Q, sharing a vertex v. Then, a side of I is obtained by rotating $a - b$ by $\pi/2$, so that side of I is perpendicular to $a - b$. But since $a \cdot v = 1$ and $b \cdot v = 1$ (by construction of a and b), we have $(a - b) \cdot v = 0$, which shows that the side of I is parallel to v.

Most of the rest of this section is devoted to solving modifications of this problem that are relevant for the nilpotent geometry. We will define a *perimeter path* to be a curve in $\mathfrak{m}_L$ that follows the shape of an isoperimetrix, providing solutions to the Dido problem in $\mathfrak{m}_L$, and will show that these together with $\mathfrak{m}_L$-geodesics solve the dual isoarea problem in $\mathfrak{m}_L$.

2.2. Perimeter paths. Let I_1 be a scaled copy of I, scaled to have perimeter one in the L-norm. Fix a parametrized 1-periodic curve $\iota : \mathbb{R} \to \mathbb{R}^2$ so that the image of ι is I_1, the parametrization is by arclength relative to the norm, and ι traverses I_1 counterclockwise.

For any choice of parameters $t \in [0, 1]$ and $T \in (t, t + 1]$, let

$$\mathrm{Per}_{t,T}(s) = \frac{\iota(sT + t - st) - \iota(t)}{T - t}, \quad s \in [0, 1].$$

Each $\mathrm{Per}_{t,T}$ is a path of length one in the norm, whose shape follows along the perimeter of $(1/(T - t))I_1$ (a scaled copy of the isoperimetrix). These perimeter paths start at the origin and therefore must terminate inside the unit ball in the norm (namely, Q).

Figure 2.2 shows a classification of perimeter paths for two different choices of Q.

Definition 2.2. Given a normed plane $\mathfrak{m}_L$, let $A : Q \to \mathbb{R}$ be the *balayage function* giving the maximal balayage area among all paths of length one from $(0, 0)$ to (x, y).

Lemma 2.3 (Perimeter paths solve the Dido problem). *Given $(x, y) \in Q$, a path γ of L-length 1 connecting $(0, 0)$ and (x, y) has maximal balayage area among all such paths if and only if it is a perimeter path: $\gamma = \mathrm{Per}_{t,T}$ where we have $\mathrm{Per}_{t,T}(1) = (x, y)$.*

Proof. Consider a particular perimeter path with $\mathrm{Per}_{t,T}(1) = (x, y)$. Let α be the straight chord from $(0, 0)$ to (x, y), and let $\lambda = 1/(T - t)$. Let A_{Per} denote the area enclosed by $\mathrm{Per}_{t,T}$ and α; we want to show that this is maximal. Note that the perimeter path can be completed to λI_1 (a scaled copy of I) by extending the domain up to $s = \lambda$. Thus, α can be realized as a chord of λI_1 that cuts it into two pieces, and A_{Per} is the area of one of those two pieces. But now suppose there were a different path γ in $\mathbb{R}^2$, of length one in the L-norm, connecting $(0, 0)$ to (x, y) while enclosing maximal area $A(x, y) \geq A_{\mathrm{tr}}$ with the chord α. Maximality of area implies convexity, and so, in particular, the path γ always stays on one side of the line through α. But then γ can also be concatenated with

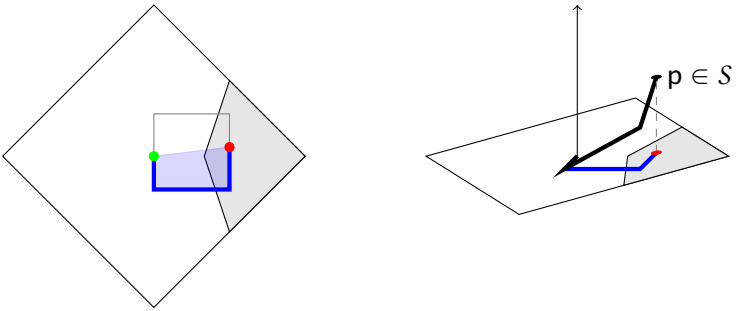

FIGURE 2.1. Here, $\mathfrak{m}_L$ has the ℓ^1 norm, and so the isoperimetrix is a coordinate square. The dark path on the left is a perimeter path of length one based at $\mathsf{0}$ that starts on the left side and ends on the right side of a full copy of I. It encloses as much area as possible for paths of length one with those endpoints. By varying the relative start and end points on those sides while maintaining unit length and basepoint $\mathsf{0}$, one obtains all the endpoints in the shaded quadrilateral. (Compare to Figure 2.2.) On the right is the CC geodesic obtained as the admissible lift of that perimeter path, with height equal to the balayage area. Its endpoint is on the CC unit sphere.

$\mathrm{Per}_{t,T}([1,\lambda])$, producing a figure which must be a copy of λI_1, by uniqueness of the isoperimetrix. Thus, $A_{\mathrm{Per}} = A(x,y)$, showing that perimeter paths enclose maximal area among all paths of length one, as desired. Indeed, this also means that the image of γ is uniquely determined, which forces it to be a perimeter path, though it is possible that $\gamma = \mathrm{Per}_{t,T} = \mathrm{Per}_{t',T'}$ for another choice of parameters with $t' = t + k$, $T' = T + k$. □

From here on, we will assume that Q is a $2N$-gon, in which case I_1 will also be a $2N$-gon, and ι will be piecewise affine: $\iota(s) = s\mathsf{a} + \mathsf{b}$ for fixed vectors a, b when s is in the subinterval of values corresponding to a side of I. Let ι be chosen so that $\iota(0)$ is a vertex in a copy of I_1 and let $\sigma_1, \sigma_2, \ldots, \sigma_{2N}$ be the sides listed counterclockwise. Let ℓ_i be the sidelength of σ_i for all i, so that $\sum \ell_i = 1$.

Define

$$Q_{ij} := \left\{ \frac{\iota(T) - \iota(t)}{T - t} \mid \iota(t) \in \sigma_i,\ \iota(T) \in \sigma_j \right\}.$$

This is the subset of Q consisting of endpoints $\mathrm{Per}_{t,T}(1)$ of those perimeter paths whose shape on I_1 begins on the ith side and ends on the jth side.

We say that two continuous, piecewise linear paths have the *same shape* if there exist vector directions $\mathsf{v}_1, \ldots, \mathsf{v}_k$ such that each path follows $a_1\mathsf{v}_1, a_2\mathsf{v}_2, \ldots, a_k\mathsf{v}_k$ in the same order, with $a_1, a_k \geq 0$, $a_2, \ldots, a_{k-1} > 0$, and $\mathsf{v}_i \neq \mathsf{v}_{i+1}$ for all i.

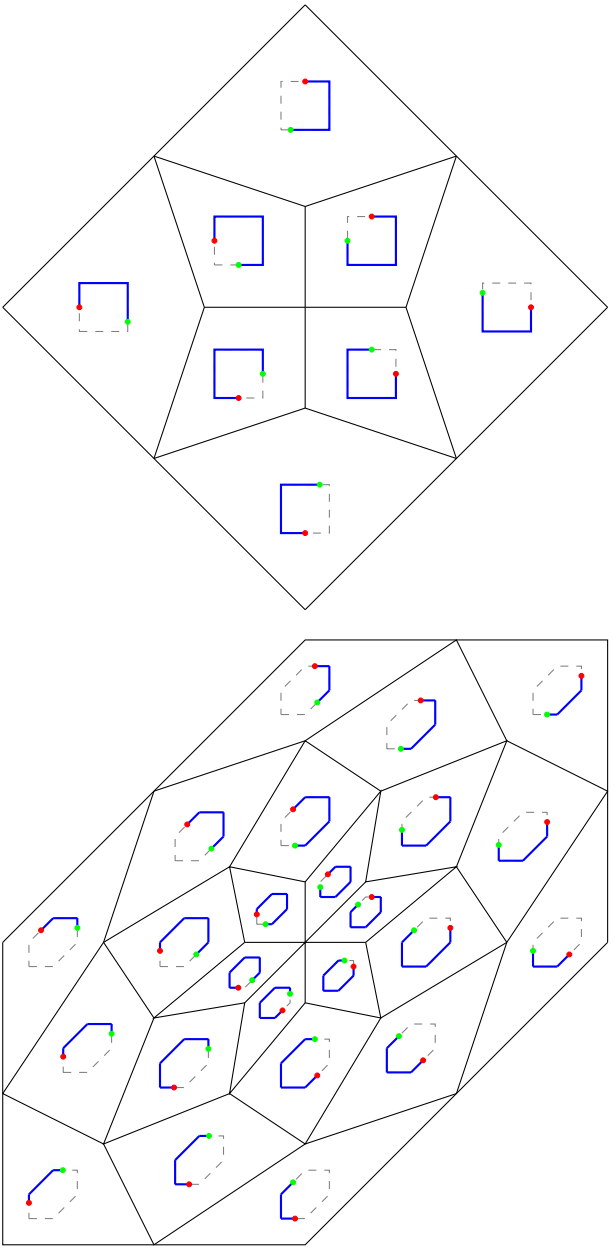

FIGURE 2.2. The combinatorics of perimeter paths for a square and a hexagon. There are 8 and 24 nondegenerate quadrilaterals, respectively, with constant combinatorics on each. Compare to Figure 2.1.

Theorem 2.4 (Combinatorics of perimeter paths). *If Q is a polygon with $2N$ sides, then there are $4N^2 - 4N$ quadrilaterals Q_{ij}, covering Q and overlapping only on their boundaries, such that the perimeter paths have "constant combinatorics" on each piece: that is, the perimeter paths ending at any two points in Q_{ij} have the same shape.*

Proof. We will show that the quadrilaterals Q_{ii} and $Q_{i,i+1}$ are degenerate (have no area in $\mathfrak{m}$). Discarding those, there remain $(2N)(2N-2)$ nondegenerate quadrilaterals, which will be seen to have the properties described in the statement above.

It will be helpful to note right away that, within a given Q_{ij}, each level set in t or T is a straight line. This is because ι is affine on each edge. Thus, within the quadrilaterals and for fixed T, the function Per is of the form $(t\mathsf{a} + \mathsf{b})/(c - t)$, and so its derivative is $(1/(c-t)^2)[c\mathsf{a} + \mathsf{b}]$, which has constant direction. This shows that t-level sets (and, similarly, T-level sets) are a family of straight segments sweeping out each quadrilateral, though not in general a parallel family. Let us also consider level sets of Per at $T - t = k$. In this case, we have

$$\mathrm{Per}_{t,t+k}(1) = \frac{\iota(t+k) - \iota(t)}{k},$$

which produces a parallel family of straight segments within Q_{ij} indexed by k.

We now consider paths starting and ending on the ith side. This quadrilateral Q_{ii} has two components: either the whole perimeter path is contained in the ith side, or the perimeter path traverses all the sides of I before returning. Recall that the side σ_i is parallel to a vertex direction from L, say, for vertex v_i. If the perimeter path is totally contained in σ_i, then the image $\mathrm{Per}_{t,T}(1)$ is the vertex v_i itself, because it points in that direction and has length one in the L norm. In the other case, we point in the $-\mathsf{v}_i$ direction. The magnitude of the vectors in this component of Q_{ii} range from 0 (achieved when $T = 1 + t$, and so the path closes up) to $\ell_i/(1-\ell_i)$ (achieved when t is at the end of σ_i and T has wrapped around to the beginning of σ_i). Thus, each Q_{ii} is a vertex of L together with a line segment in the interior of Q.

Now we will show that $Q_{i,i+1}$ is the edge of L with endpoints v_i, v_{i+1}. This follows simply from noticing that each of those perimeter paths is a two-sided path following direction v_i for time a and then direction v_{i+1} for time $1-a$, so that it terminates at the point $a\mathsf{v}_i + (1-a)\mathsf{v}_{i+1}$. Notice that these particular perimeter paths are geodesic in $\mathbb{R}^2$ with respect to the L-norm; and indeed, since all these perimeter paths have L-length one, they are $\mathfrak{m}_L$-geodesics if and only if they terminate on L itself (the sphere of radius one).

For the remaining quadrilaterals, there are four extreme points: we may start at either endpoint of σ_i and terminate at either endpoint of σ_j. Indeed, there is a quadrilateral traced out by holding the start point or end point on I_1 fixed at one extreme while moving the other from extreme to extreme, then alternating which

is fixed and which is moving, making a circuit of length four. (Its four subpaths are level sets for t or T, and so they are straight lines.)

Finally, observe that $\mathrm{Per}_{t,t+k}(1)$ is a convex, centrally symmetric polygon for every fixed $k \in (0,1]$. This polygon varies continuously in k, is identically zero when $k = 1$ and equals L itself for k sufficiently small. We can conclude that all points in Q are hit by perimeter paths. □

2.3. $\mathfrak{m}_L$***-geodesics.*** Note that perimeter paths are typically inefficient in $\mathfrak{m}_L$, but lift to geodesics of $(H(\mathbb{R}), \mathrm{cc})$. Arbitrary geodesics of $\mathfrak{m}_L$ also lift to geodesics of the CC metric. So, next, we consider geodesic segments in $\mathfrak{m}_L$ emanating from the origin. (One can check that a given path in $\mathfrak{m}$ from $\mathsf{0}$ to an interior point of an edge $\sigma \subset L$ is an $\mathfrak{m}_L$-geodesic by verifying that all of its tangent vectors point towards σ.)

Lemma 2.5 (Balayage area of geodesics). *Every point $(x, y) \in L$ is reached by a unique perimeter path $\mathrm{Per}_{t,T}$ of length 1, and this path encloses a nonnegative area $A = A(x, y)$. There are $\mathfrak{m}_L$-geodesics to (x, y) enclosing every signed area in the range $[-A, A]$.*

Proof. For each fixed $0 < a < 1$, consider the path that goes distance s in direction v_i, then $(1 - a)$ in direction v_{i+1}, and finally $(a - s)$ in direction v_i, as s varies from 0 to a. □

In general, the full set of $\mathfrak{m}_L$-geodesics to a side with extreme points v_i and v_{i+1} consists of parametrizations by arclength of arbitrary piecewise-differentiable paths whose tangent vectors point in the interval of directions between those two extremes.

Lemma 2.6 (Isoarea problem). *A curve in Q of L-length one has minimal length among all curves with the same endpoints, and balayage area if and only if it is a perimeter path or an $\mathfrak{m}_L$-geodesic.*

Proof. The reverse implication is easy: on the one hand, we have already shown that perimeter paths uniquely maximize area to their endpoints. But then, for $\mathfrak{m}_L$-geodesics, there is clearly no shorter path between their endpoints.

Now, for the forward implication: let β be a curve in Q of L-length one which has minimal length among all curves with the same endpoints and balayage area. If β ends on L, then it is an $\mathfrak{m}_L$-geodesic by definition. Otherwise, let us say β ends at $(x, y) \in Q^\circ$ and has balayage area A'. We know that there is a perimeter path $\gamma = \mathrm{Per}_{t,T}$ with $\mathrm{Per}_{t,T}(1) = (x, y)$ enclosing area $A(x, y) \geq A'$. If $A(x, y) = A'$, then β is a perimeter path by Lemma 2.3. If not, then we can modify γ to be an improvement on β: we know that the region enclosed by γ is convex, so we can find a chord of this body starting and ending on γ so that the modified curve γ', which follows γ except for a shortcut along this chord, has length < 1 and encloses area A'. This then contradicts the minimality of the length of β. □

3. Structure of Polygonal CC Metrics

3.1. Unit sphere and* CC *geodesics. Returning to the Heisenberg geometry, we find that we have identified the geodesics, and can thus map the shape of spheres.

Recall that Pansu's theorem tells us that the word metric $(H(\mathbb{Z}), S)$ is asymptotic to the CC norm on $H(\mathbb{R})$ induced by L, the boundary of the convex hull of $\pi(S)$. Note that the number of sides of L is at most $|S|$, but could be smaller if S has several elements with the same projection to $\mathfrak{m}$, or has "insignificant" elements which do not project to vertices of L. If L is a $2N$-gon, then there are $4N^2 - 4N$ quadrilateral regions Q_{ij} classifying the shapes of perimeter paths, while the $2N$ sides of L classify the $\mathfrak{m}_L$-geodesics.

Theorem 3.1 (Structure Theorem). *Consider the* CC *metric induced on $H(\mathbb{R})$ by a polygonal norm $\|\cdot\|_L$ on $\mathfrak{m}$. The balayage function $A(x, y)$ is a continuous function over Q whose restriction to each Q_{ij} is a quadratic polynomial in x, y.*

Let $\mathbf{z}(x, y)$ be the multivalued function given by

$$\mathbf{z} = \begin{cases} \{\pm A\}, & (x, y) \in Q^\circ, \\ [-A, A], & (x, y) \in L. \end{cases}$$

Then, the graph of $\mathbf{z}$ is precisely the unit sphere S in the CC *metric.*

Equivalently, the full set of CC *geodesics of length one based at the origin is identical to the set of admissible lifts of perimeter paths and $\mathfrak{m}_L$-geodesics of length one (and their reflections in the origin).*

Proof. For points $(x, y, z) \in H(\mathbb{R})$ with $z \geq 0$ and $(x, y) \in Q^\circ$, saying that an admissible path α from 0 to (x, y, z) is geodesic means that $\pi(\alpha)$ is the shortest path in $\mathfrak{m}$ from $(0, 0)$ to (x, y) enclosing (nonnegative) balayage area z. By central symmetry of the isoperimetrix, we can solve the problem for minimal signed area by traversing the isoperimetrix clockwise rather than counterclockwise. Therefore, the minimal signed area enclosed by a path of length one from the origin to (x, y) is given by $-A(-x, -y) = -A(x, y)$, which tells us that the graph of $\mathbf{z}$ is symmetric over the xy plane.

Continuity of $A(x, y)$ on Q is immediate from the continuity of $\mathrm{Per}(t, T)$ in its parameters. It only remains to show that the restriction of $A(x, y)$ to each Q_{ij} is a quadratic polynomial.

Observe that finding the perimeter path to a point $(x, y) \in Q_{ij}$ amounts to finding scalars s_1, s, s_2 (with $s_1, s_2 \leq s$) such that

$$(x, y) = s_1 \mathsf{w}_i + s(\mathsf{w}_{i+1} + \cdots + \mathsf{w}_{j-1}) + s_2 \mathsf{w}_j,$$

where w_i are the vectors defining the sides of I, as in Figure 3.1.

The area enclosed is equal to that of a polygon of fixed shape with $j - i$ sides of length proportional to s, so that its area is proportional to s^2, plus two triangles of area proportional to $s_1 \cdot s$ and $s_2 \cdot s$, respectively. There are three linear equations relating s_1, s_2, and s, given by the total arclength equaling one and the Euclidean

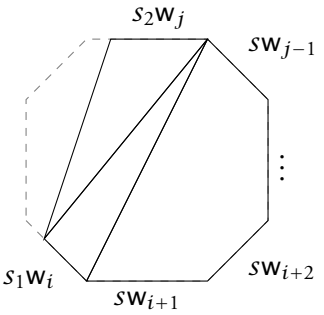

FIGURE 3.1. The balayage area of a path that follows along a scaled copy of the isoperimetrix is given by a quadratic expression in s_1, s_2, s.

displacement between endpoints equaling x and y in the horizontal and vertical direction, respectively. Thus, we can solve to get expressions for s_1, s_2, and s that are linear in x, y; this means that the area is quadratic in x, y, as required. □

By the definition of a CC metric, the sphere of radius t based at $\mathsf{0}$ is precisely $\delta_t S$ (so S is homeomorphic to a two-sphere), and the spheres with other centers are obtained from $\delta_t S$ by left-multiplication in $H(\mathbb{R})$.

We want to break down the sphere, which is the graph of $\mathbf{z}$ over Q, into the graphs over the various quadrilaterals $Q_{ij} \subset \mathfrak{m}$. We will call these the *panels* of the sphere, and define them by

$$\mathrm{Panel}_{ij} := \{(x, y, z) \mid (x, y) \in Q_{ij},\ z = \pm A(x, y)\}, \quad j \neq i + 1,$$

and

$$\mathrm{Panel}_{i,i+1} := \{(x, y, z) \mid (x, y) \in Q_{i,i+1},\ z \in [-A, A]\},$$

so that $S = \bigcup_{1 \le i,j \le 2N} \mathrm{Panel}_{ij}$. Note that we can leave out the Panel_{ii} without changing this equality, because they are contained in the union of the other panels. The $\mathrm{Panel}_{i,i+1}$ are flat and perpendicular to $\mathfrak{m}$ (projecting to edges of L), and the fact that $\mathbf{z}$ is quadratic over the interior implies that they are bounded above and below by parabolas in a plane through that edge that is perpendicular to $\mathfrak{m}$. We will call these the *side panels* of S (these are the panels cut away in Figure 3.2).

The combinatorics of perimeter paths on quadrilaterals dictates the shape of CC geodesics in the following straightforward way. Note that all geodesics from $\mathsf{0}$ to any point are just dilations of geodesics from $\mathsf{0}$ to S. These, in turn, are lifts of perimeter paths and $\mathfrak{m}_L$-geodesics, whose shape is determined by which quadrilateral contains the projection from S to Q. That is, the shapes of all possible geodesics from $\mathsf{0}$ to x are determined by which of the quadrilaterals Q_{ij} contains the point $\pi \circ \delta_{1/d}(\mathsf{x})$, where $d = d_{\mathrm{CC}}(\mathsf{x}, \mathsf{0})$. Let us call this point of Q the *footprint* of x.

Let the *regular points* in the unit sphere be the union of the panels over the quadrilaterals interior to Q, and let the *unstable points* be those in side panels:

$$S_{\mathsf{reg}} = \bigcup_{j \neq i+1} \mathrm{Panel}_{ij}; \quad S_{\mathsf{uns}} = \bigcup \mathrm{Panel}_{i,i+1},$$

so that $S = S_{\mathsf{reg}} \cup S_{\mathsf{uns}}$. Now, define $\mathsf{Reg} = \Delta S_{\mathsf{reg}}$ and $\mathsf{Uns} = \Delta S_{\mathsf{uns}}$, so that $H(\mathbb{R}) = \mathsf{Reg} \cup \mathsf{Uns}$. These are constructed so that points of Uns are reached by families of $\mathfrak{m}_L$-geodesics, while points of Reg are reached by perimeter paths. That is, we use the regular/unstable distinction to divide the space $H(\mathbb{R})$ according to which of the two types of geodesic reaches each point.

Let the *volume subtended by a panel* be the volume of the region obtained by coning off to the origin by dilation. We will see that the decomposition of the sphere into regular and unstable points gives us useful invariants coming from volume:

$$V_{\mathsf{reg}} = \mathrm{vol}(\hat{S}_{\mathsf{reg}}); \quad V_{\mathsf{uns}} = \mathrm{vol}(\hat{S}_{\mathsf{uns}}).$$

Recalling that V is the volume of the unit ball, we have $V = V_{\mathsf{reg}} + V_{\mathsf{uns}}$.

Remark 3.2 (Rationality). We note that Breuillard carried out a full description for the standard generators of the kind given in this section, and indicated key elements of such a description for general S, in [3, Proposition 9.1]. For instance, it is stated without argument there that the balayage function should be piecewise quadratic. Breuillard sketches an argument that V (the volume of the unit ball for any limit metric) is rational, noting that Q has integer vertices, and so its polar dual and therefore the isoperimetrix I must have rational vertices (since $Q^* = \{\mathsf{x} \in \mathfrak{m} \mid \mathsf{x} \cdot \mathsf{y} \leq 1, \ \forall\, \mathsf{y} \in Q\}$). Following this line in our language, we see that the vectors w_i have rational projections to each coordinate direction as well as rational length in the L-norm (though not in the Euclidean norm), and so the linear relations between x, y and s_1, s_2, s described in the proof of Theorem 3.1 are rational, which ensures that the balayage area over Q_{ij} is a rational quadratic polynomial in x and y. Furthermore, the vertices of the quadrilaterals Q_{ij} are rational (when a perimeter path begins and ends at a vertex, and has arclength 1, then each coordinate of the endpoint is a sum of rational numbers, scaled by a rational number). Thus, the volume subtended by each panel is given by an integral over a region bounded entirely by graphs of rational quadratic polynomials (both the balayage area and the tracks of the dilation) over rational polygonal domains, and so each such volume is in fact rational.

3.2. Nonuniqueness and regular points. The union of the degenerate quadrilaterals has a geometric significance. Recall that the quadrilateral Q_{ii} is a line segment based at the origin along with a vertex. Let v_i' denote the other endpoint of the line segment Q_{ii} emanating from $\mathsf{0}$. The other degenerate quadrilaterals cover the boundary L, whose vertices are called v_i. Then, with the convention that $Q_{2N,2N+1} = Q_{2N,1}$, let

$$\mathsf{N}_0 := \bigcup_{1 \leq i \leq 2N} Q_{ii} \setminus \{\mathsf{v}_i'\}; \quad L_0 := \bigcup_{1 \leq i \leq 2N} Q_{i,i+1} \setminus \{\mathsf{v}_i\} = L \setminus \{\mathsf{v}_i\}; \quad \mathsf{N} := \mathsf{N}_0 \cup L_0.$$

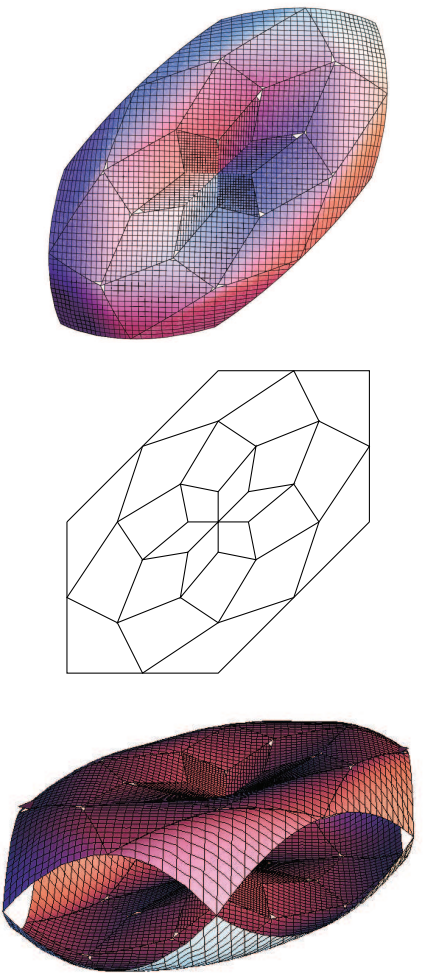

FIGURE 3.2. Two views of the limit shape S for the hexagonal generators, with the quadrilateral decomposition of the footprint Q reproduced for reference. Only the regular part, S_{reg}, is shown. The unstable part, S_{uns}, is cut away in the picture: it is made up of six flat panels perpendicular to the xy-plane that are bounded by parabolas.

Recall that the footprint of $\mathsf{x} \in H(\mathbb{R})$ is $\pi \circ \delta_{1/d}(\mathsf{x})$, for $d = d_{\mathrm{CC}}(\mathsf{x}, 0)$. Recall as well that $z(\mathsf{x})$ is the height (z-coordinate) of the point x and that, for $\mathsf{x} \in Q$, $A(\mathsf{x})$ is the most possible area enclosed by a unit-length path to x.

Proposition 3.3 (Uniqueness of* CC *geodesics). *A point* $\mathsf{x} \in S$ *is reached by more than one geodesic of length one based at the origin if and only if* $\pi(\mathsf{x}) \in \mathsf{N}$ *and* $\mathsf{x} \notin S_{\mathsf{reg}} \cap S_{\mathsf{uns}}$.

Thus, a point $\mathsf{x} \in H(\mathbb{R})$ *is reached by more than one geodesic based at the origin if and only if the footprint of* x *is in* N *and* $\mathsf{x} \notin \mathsf{Reg} \cap \mathsf{Uns}$.

That is, the points with footprints in N_0 are reached by multiple perimeter paths, and those with footprints in L_0 are reached by $\mathfrak{m}_L$-geodesics. Only if they have maximal height is the $\mathfrak{m}_L$-geodesic unique (in which case it coincides with a perimeter path).

Proof. First, we consider the case of perimeter paths. It is impossible for a convex polygon (in this case λI_1) to have two different interior chords which are parallel, have the same length, and subtend the same perimeter on each side. Thus, the only way this nonrigidity can occur is for the chord α to be contained in a side of λI_1. But this is precisely the case when the start and end points of Per fall on the same side, which was examined in the discussion of Q_{ii}. The vertices of L have perimeter paths enclosing zero area, and so the geodesics are unique in that case. Finally, if $\pi(\mathsf{x})$ is interior to a side of L, then geodesics are lifts of $\mathfrak{m}_L$-geodesics that are not straight lines. As long as these geodesics do not enclose maximal balayage area, they can always be perturbed to nearby curves with the same endpoints, length, and enclosed area. ◻

From now on, we call N the *nonuniqueness locus* in Q.

Theorem 3.4 (Probability of unique geodesics). *In* $H(\mathbb{R})$ *with the* CC *metric induced by a polygon* L*, the asymptotic density of unique geodesics equals* V_{reg}/V.

That is, choose x *uniformly in the ball of radius* r*, and consider the probability that there is only one geodesic from* $\mathsf{0}$ *to* x*. For every* r*, this probability equals* V_{reg}/V.

Proof. The only instances of nonuniqueness occur over N, as we have seen. But there is no volume over N_0, and none of the volumes of Uns is contributed by $\mathsf{Uns} \cap \mathsf{Reg}$. Thus, the probability of being reached by two or more distinct geodesics is precisely the proportion of the volume of the unit ball that is in the unstable part (i.e., subtended by the side panels). ◻

Example 3.5 (Volume calculations in* ℓ^1 *norm). Let $\|\cdot\|_L$ be the ℓ^1 norm on $\mathfrak{m}$, and give $H(\mathbb{R})$ the corresponding CC metric. We find that the volume subtended by panels with four-sided combinatorics is $\frac{13}{216}$, the volume subtended by three-sided panels is $\frac{11}{54}$, and the volume subtended by the side panels is $\frac{1}{6}$. Thus, $V_{\mathsf{reg}} = \frac{13}{216} + \frac{11}{54} = \frac{19}{72}$ and $V_{\mathsf{uns}} = \frac{1}{6}$. This adds up to give the total volume of the unit ball as $V = \frac{31}{72}$, which agrees with calculations by Breuillard [3] and Stoll [13] of the volume growth.

Thus, for this standard polygonal CC metric, the density of points reached by nonunique geodesics is $\frac{12}{31}$, or about 39%, versus $\frac{19}{31} \approx 61\%$ for unique geodesics.

4. GEODESICS IN THE WORD METRIC

Recall that the *Hausdorff distance* $d_{\mathsf{Haus}}(\alpha, \beta)$ between a pair of sets is the smallest $\varepsilon \geq 0$ such that each set lies in the ε-neighborhood of the other. Below, we discuss paths and the images of those paths interchangeably, to make sense of the Hausdorff distance between paths. The goal of this section is to make precise the following statement: "word geodesics can be approximated with CC geodesics."

Lemma 4.1 (Continuity of* CC *geodesics). *Fix an arbitrary $\rho > 0$. Suppose we are given $x \in S$ such that $\pi(x)$ does not lie in the ρ-tubular neighborhood $\mathcal{N}_\rho(N_0)$ of N_0 and a* CC *geodesic α from 0 to x. Then, for every $\varepsilon > 0$, there exists $\delta > 0$ such that whenever $y \in S$ with $d_{\mathrm{CC}}(x, y) < \delta$, there exists a* CC *geodesic β from 0 to y with $d_{\mathsf{Haus}}(\alpha, \beta) < \varepsilon$.*

Proof. This is clear for points interior to any regular panel. For points x in the boundaries of regular panels, one can see from the combinatorial description that a sequence of points approaching x from any panel must have geodesics approaching the unique geodesic to x (see Figure 2.2). Suppose x is in the interior of a side panel and y is close to x. If $\pi(x) = \pi(y)$, then we homotope $\pi(\alpha)$ to a path $\bar{\beta}$ with the same endpoints such that the tracks of individual points are small. As long as every tangent vector of every point on every intermediate path points towards the same side, this is a homotopy through $\mathfrak{m}_L$-geodesics. If $\pi(x) \neq \pi(y)$, then we can concatenate this homotopy with another similar homotopy ending at a path from the origin to $\pi(y)$ with only a small change to the area. Either way, lift $\bar{\beta}$ to a geodesic β ending at y that is close to α. □

This continuity statement is false if $\pi(x) \in N_0$. For instance, take x to be the point on S above 0 and $\{y_n\}$ to be a sequence of points interior to a panel, converging to x. Then, for each n, there is a unique geodesic γ_n from 0 to y_n; these converge to a particular one of the many geodesics from 0 to x, which include the lifts of $\mathrm{Per}_{t,t+1}$ for all $t \in [0, 1]$. These other geodesics to x are thus not closely approximated by geodesics to the y_n even as $y_n \to x$.

Let S be a finite generating set for $H(\mathbb{Z})$, and K be the maximum of $d_{\mathrm{CC}}(0, a)$ over all generators $a \in S$. Minimal-length spellings $w = a_1 \cdots a_n$ written in letters from S may be realized as *discrete geodesics* $x_0, x_1, \ldots, x_n$ in $H(\mathbb{R})$ where $x_0 = 0$, $x_i = a_1 \cdots a_i$, and the CC distance between successive points is no more than K. We say that the discrete geodesic $x_0, x_1, \ldots, x_n$ is *ε-linearly tracked* (or just ε-tracked) by an admissible path α if they have the same endpoints and if the set $\{x_i\}$ stays inside the εn-neighborhood of α.

Lemma 4.2 (Geodesic spellings vs. admissible paths). *For any $\varepsilon > 0$ and $x \in H(\mathbb{Z})$ with sufficiently large $n = |x|$, every discrete geodesic from 0 to x is ε-linearly tracked by an admissible path of length $n + O(\sqrt{n})$.*

Proof. Let $z(x)$ denote the height (i.e., the z-coordinate) of $x \in H(\mathbb{R})$, and $M := \max z(S)$. Let $x_0, \ldots, x_n$ be a discrete geodesic from 0 to $x \in H(\mathbb{R})$ in

letters from S. Finally, define a polygonal path γ in $\mathfrak{m}$ by connecting vertices $\pi(\mathsf{x}_0), \ldots, \pi(\mathsf{x}_n)$ with straight lines.

The curve γ has a unique admissible lift $\bar{\gamma}$; its endpoints are at $\mathsf{0}$ and at a point $\mathsf{y} \in H(\mathbb{R})$ which lies in the same vertical line as x. Since $\pi(S) \subset Q$, the curve γ lies in nQ, and since it is made up of straight segments of L-length ≤ 1, the length of γ (and hence $\bar{\gamma}$) is bounded above by n. We will establish the following equation:

$$z(\mathsf{x}) = z(\mathsf{y}) + \sum_{i=1}^{n} z(\mathsf{a}_i). \tag{$\dagger$}$$

This says that the height of x is the balayage area of γ plus the heights of all of the letters used in the spelling.

Equation ($\dagger$) can be proven by induction. The group law, in exponential coordinates, says that

$$(x, y, z)(x', y', z') = \left(x + x', y + y', z + z' + \frac{xy' - yx'}{2}\right).$$

In particular, if $\mathsf{x}_{n-1} = (x, y, z)$ and $\mathsf{a}_n = (x', y', z')$, then we see that the change in height between x_{n-1} and $\mathsf{x}_n = \mathsf{x}_{n-1}\mathsf{a}_n$ is equal to the height of a_n plus the area of an appropriate triangle in $\mathfrak{m}$. That means that the height of the spelling path differs from the height predicted by balayage area by precisely the height of the generators, as claimed in ($\dagger$).

Now set $C = d_{\mathrm{CC}}((0, 0, M), \mathsf{0})$, so that

$$d_{\mathrm{CC}}(\mathsf{0}, (0, 0, Mn)) = d_{\mathrm{CC}}(\mathsf{0}, \delta_{\sqrt{n}}(0, 0, M)) = C\sqrt{n}.$$

Let α be a path obtained by concatenating $\bar{\gamma}$ with a CC geodesic from y to x. From ($\dagger$), it follows that

$$|z(\mathsf{x}_i) - z(\mathsf{y}_i)| \leq Mi \leq Mn,$$

for all i, where y_i is the point on $\bar{\gamma}$ corresponding to the vertex x_i. Therefore, $d_{\mathrm{CC}}(\mathsf{x}_i, \mathsf{y}_i) \leq C\sqrt{n}$, and so the discrete geodesic lives in the $C\sqrt{n}$-neighborhood of α, which means that the path $Cn^{-1/2}$-tracks the spelling. Furthermore, the difference in lengths between α and $\bar{\gamma}$ is bounded by $C\sqrt{n}$. Thus, the length of α is bounded above by $n + C\sqrt{n}$. On the other hand, Theorem 1.3 provided a global constant C' such $d_{\mathrm{CC}}(\mathsf{0}, \mathsf{x}) \geq n - C'$. This proves that the length of α is $n + O(\sqrt{n})$. ◻

Lemma 4.3 (Tracking Lemma). *For every $\varepsilon > 0$, $\rho > 0$, and sufficiently large $\mathsf{x} \in H(\mathbb{Z})$ whose footprint is not in $\mathcal{N}_\rho(\mathsf{N}_0)$, every discrete geodesic from $\mathsf{0}$ to x is ε-linearly tracked by a* CC *geodesic.*

Proof. Suppose otherwise; then, there is an $\varepsilon > 0$ such that for every $n \geq 0$, there exist points $\mathsf{x}_n \in H(\mathbb{Z})$ with $|\mathsf{x}_n| \geq n$ whose footprints are not in $\mathcal{N}_\rho(\mathsf{N}_0)$, and there are choices of discrete geodesics from $\mathsf{0}$ to x_n which stay $\varepsilon|\mathsf{x}_n|$-far away from any CC geodesics to the same point. Let α_n be the admissible paths approximating these discrete geodesics, as described above.

Denote $\bar{\mathsf{x}}_n = \delta_{1/|\mathsf{x}_n|}\mathsf{x}_n$ and $\bar{\alpha}_n = \delta_{1/|\mathsf{x}_n|}\alpha_n$. The lengths of the $\bar{\alpha}_n$ are converging to 1 by Lemma 4.2. By passing to a subsequence, we may assume that the $\bar{\mathsf{x}}_n$ converge to a limit $\bar{\mathsf{x}} \in S$ with $\pi(\bar{\mathsf{x}}) \notin \mathcal{N}_\rho(\mathsf{N}_0)$, and that the paths $\bar{\alpha}_n$ converge in Hausdorff distance to a geodesic $\bar{\alpha}$ from $\mathsf{0}$ to $\bar{\mathsf{x}}$ by Arzelà-Ascoli. By Lemma 4.1, there is a large enough n such that there is a geodesic $\bar{\beta}_n$ within a Hausdorff distance of $\varepsilon/2$ from $\bar{\alpha}$, and for large enough n, we know that $\bar{\alpha}_n$ is within $\varepsilon/2$ of $\bar{\alpha}$. But then, for $\beta_n = \delta_{|\mathsf{x}_n|}\bar{\beta}_n$, we have $d_{\mathsf{Haus}}(\alpha_n, \beta_n) < \varepsilon|\mathsf{x}_n|$. Thus, the discrete geodesics are ε-tracked by the CC geodesics β_n, giving us a contradiction. ◻

A key idea in the above proof is that additive quasi-geodesics get close to true geodesics as the additive constant goes to zero. Again, this fails near N_0.

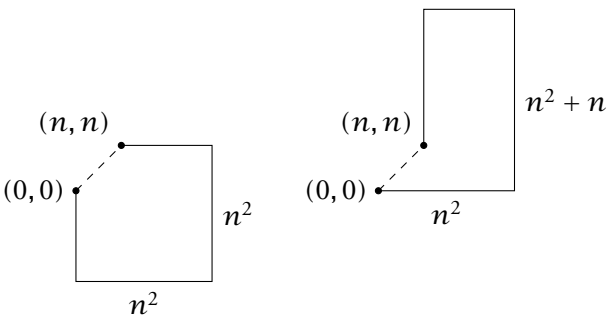

FIGURE 4.1. In these diagrams in $(\mathfrak{m}, \ell^1)$, we see a true geodesic α_n from $(0,0)$ to (n,n) on the left, and a path β_n with different combinatorics (enclosing the same area but with excess length) on the right. When these are rescaled so that their lifts terminate on the sphere of radius 1, the ratio of lengths goes to 1 while the Hausdorff distance is bounded away from zero.

Example 4.4. Let $\mathsf{x}_n = (n, n, n^4 - \frac{1}{2}n^2)$. There is a unique CC geodesic α_n from the origin to x_n, and its projection to the horizontal plane is shown in Figure 4.1 on the left. On the right is pictured a nongeodesic path β_n to x_n which has length $4n^2$ (while the true geodesic has length $L_n = 4n^2 - 2n$). Let $\mathsf{y}_n = \delta_{1/L_n}\mathsf{x}_n$ be the dilate of x_n that lies in the unit sphere S. Note that $\mathsf{y}_n \to (0, 0, \frac{1}{16})$ as $n \to \infty$. By scaling back the paths β_n by the same factor, we obtain paths whose lengths approach 1. However, the Hausdorff distance between α_n and β_n is $n^2 - n$; thus, after scaling they still have a Hausdorff distance of at

least $\frac{1}{4}$. In the limit as $n \to \infty$, both kinds of paths limit to valid closed geodesics to $(0, 0, \frac{1}{16})$. But for any finite value of n, the path with the wrong combinatorics is far from any geodesic.

5. Geodesic Stability

As a consequence of the tracking lemma, we can make a quantitative study of the stability of geodesics, measuring how much two geodesics with the same endpoints can differ. (Compare to the Morse Lemma for hyperbolic spaces.)

Definition 5.1. For a geodesic space X with basepoint x_0, let $G(x)$ be the set of all geodesics from x_0 to x. Then, the *instability function* $\mathbf{I}(x)$ measures how far apart they can be, relative to the size of x:

$$\mathbf{I}(x) := 2 \cdot \frac{\operatorname{diam} G(x)}{d(x, x_0)} = \frac{2}{d(x, x_0)} \cdot \sup_{\alpha,\beta \in G(x)} d_{\mathsf{Haus}}(\alpha, \beta),$$

where the metric on $G(x)$ is given by Hausdorff distance.

We say that a point x is *ε-stable* if $\mathbf{I}(x) \le \varepsilon$.

In general, $\operatorname{diam} G(x) \le d(x, x_0)/2$, because any point on a geodesic from x_0 to x is within $d(x, x_0)/2$ of one of the endpoints, and therefore within that distance of any other geodesic. Thus, the factor of 2 in the definition normalizes $\mathbf{I}$ so that $0 \le \mathbf{I}(x) \le 1$ for any X, x_0, x.

Note that if X has unique geodesics, then $\mathbf{I} \equiv 0$. If X is δ-hyperbolic, then $\operatorname{diam} G(x) \le \delta$ for all x, and so $\mathbf{I}(x) \to 0$ as $x \to \infty$.

The asymptotic behavior of this instability function is far from being quasi-isometry invariant. Obviously, $\mathbf{I} \equiv 0$ on the Euclidean plane $\mathbb{R}^2$, but the situation is quite different for $\mathbb{R}^2$ with a polygonal norm. For instance, in the ℓ^1 norm, the elements along the x and y axes have $\mathbf{I} = 0$, but the diagonal has $\mathbf{I}(a, a) = 1$. It is easily verified that, with any generating set, the limit shape L divides the plane into sectors, and the instability is 0 exactly in the vertex directions. In particular, one can recover the vertex directions of L from the function $\mathbf{I}$. Since the word metric on $(\mathbb{Z}^2, S)$ is asymptotic to $\mathfrak{m}_L = (\mathbb{R}^2, \|\cdot\|_L)$, it is also true for word metrics that $\mathbf{I}$ detects the directions of significant generators.

The goal of this section is to study instability in the Heisenberg group. As one would expect, $H(\mathbb{Z})$ is more geodesically stable than $\mathbb{Z}^2$ but less stable than a hyperbolic group. More intriguingly, the stability depends on the generating set in a different way than for those other groups.

We will call an element of $H(\mathbb{Z})$ *regular* if it is regular as an element of $H(\mathbb{R})$.

Lemma 5.2 (Stability of regular points). *For every $\varepsilon > 0$ and $\rho > 0$, all sufficiently large regular points $\mathsf{x} \in H(\mathbb{Z})$ with footprints not in $\mathcal{N}_\rho(\mathsf{N}_0)$ satisfy*

$$d_{\mathsf{Haus}}(g, h) < \varepsilon |\mathsf{x}|$$

for all word geodesics g, h from $\mathsf{0}$ to x.

Proof. Let S be a finite generating set for $H(\mathbb{Z})$, and give $H(\mathbb{R})$ the corresponding CC metric. Let $T \geq 0$ be large enough so that when $|\mathsf{x}| \geq T$, the footprint of x avoids $\mathcal{N}_\rho(\mathsf{N}_0)$, and g is a word geodesic from $\mathsf{0}$ to x, then there is a CC geodesic γ from $\mathsf{0}$ to x such that $d_{\mathsf{Haus}}(g, \gamma) < (\varepsilon/2)|g| = (\varepsilon/2)|\mathsf{x}|$. Since regular points have unique CC geodesics, we are done. □

Then, since the proportion of Reg lying over $\mathcal{N}_\rho(\mathsf{N}_0)$ goes to zero as $\rho \to 0$, we can conclude from the point of view of geometric probability that all of the regular part of $H(\mathbb{R})$ is ε-linearly stable for every ε, so that Hausdorff distance between word geodesics is bounded above by *every* linear function of word length. That is, for regular points x of $H(\mathbb{Z})$, all word geodesics $\overline{\mathsf{0x}}$ satisfy a sublinear fellow traveling property. This is much stronger control than one has in $\mathbb{Z}^d$, where only words that are powers of single generators can be ε-stable for every ε; therefore, almost all points are unstable. At the other extreme, all elements sufficiently far from the basepoint in a hyperbolic group G are stable. The Heisenberg group is intermediate between these in terms of geodesic stability: only the regular part is ε-stable for all ε, and its measure depends nontrivially on the generators.

To sum up this comparison, we have the following result:

Theorem 5.3 (Instability Theorem). *With respect to any finite generating set,*

$$\mathsf{Stab}(X) := \lim_{\varepsilon \to 0} \mathrm{Prob}(\mathbf{I}(\mathsf{x}) \leq \varepsilon) = \begin{cases} 0, & \text{if } X = \mathbb{Z}^d, \\ \dfrac{V_{\mathsf{reg}}}{V}, & \text{if } X = H(\mathbb{Z}), \\ 1, & \text{if } X \text{ is unbounded, } \delta\text{-hyperbolic.} \end{cases}$$

6. SUBGROUP DISTORTION

With these techniques, we can give finer measures for properties usually studied up to quasi-isometry equivalence. In this section, we consider the distortion of subgroups; in the following section, we consider counting problems related to group growth.

Recall that, for a subgroup $K \leq H$, its distortion is measured by comparing $|g|_K$ with $|g|_H$, where $K = \langle S' \rangle$ and $H = \langle S \rangle$. This quantifies the relative efficiency of traveling around K in its own Cayley graph rather than the Cayley graph for the ambient group. Traditionally, we say that a subgroup is *undistorted* if the ratio of the two quantities is bounded above and below by a constant, or equivalently if each quantity is linearly bounded with respect to the other. We likewise define other rates of distortion (not paying attention to the coefficients but only the rate), so that they are independent of the choice of generators. For instance, the central cyclic subgroup $\{(0, 0, *)\}$ is quadratically distorted in $H(\mathbb{Z})$ for any choices of generators, because $\mathsf{e}_3^{n^2} = [\mathsf{e}_1^n, \mathsf{e}_2^n]$. For the same reason, the abelian subgroup $K := \{(*, 0, *)\} \leq H$ is quadratically distorted as well.

For an arbitrary subgroup $K \leq H$, consider the distortion function $d(g) = |g|_K / |g|_H$, again with respect to fixed generating sets S' of K and S of H. If

there were some global bound $M > 0$ such that $d(g) < M$ for all $\mathsf{p} \in K$, then K would be undistorted in the usual sense. Instead, we can study the distribution of values of $d(g)$ over B_n^K as $n \to \infty$. If there is a limit shape for word metrics on K, one can hope that this converges to a distribution on the limit shape, which we will call the *distortion profile*. There are several invariants one could attach to the distortion profile; for instance, let us say that a subgroup is *strongly statistically undistorted* if the average value of d on B_n^K limits to a finite value. We can also compute a *distortion index*, defined even if the asymptotic average of d is infinite, by taking an ℓ^{-1} average (the reciprocal of the asymptotic average of $1/d$), so that a higher distortion index means more distortion. We say the subgroup has *statistical distortion* n^α if the average value of $(\ln \|\mathsf{p}\|_K)/(\ln \|\mathsf{p}\|_H)$ tends to α (and that the subgroup is (*weakly*) *statistically undistorted* if $\alpha = 1$).

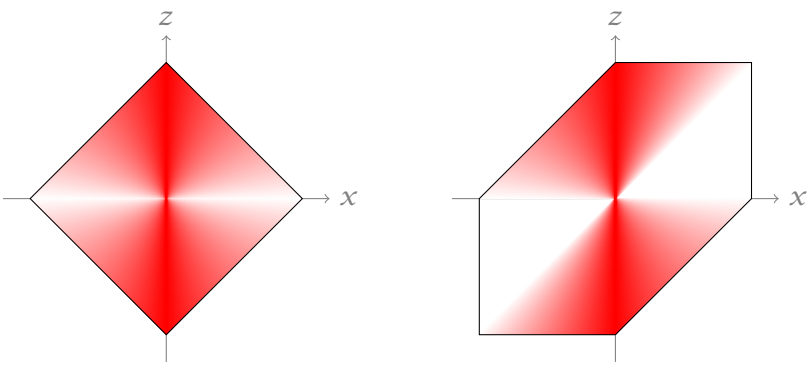

FIGURE 6.1. The distortion profiles for $K = \{(*, 0, *)\}$ in $H(\mathbb{Z})$ with two different generating sets. The value $d = 1$ is plotted as white, going red as $d \to \infty$.

Returning to the example of $\mathsf{p} = (x, 0, z) \in K = \{(*, 0, *)\} \leq H = H(\mathbb{Z})$, we have asymptotic formulas for the numerator and denominator in the distortion expression from using the limit shapes for $(\mathbb{Z}^2, S')$ and $(H(\mathbb{Z}), S)$, respectively. We find that the distortion function tends to a limiting distribution on a convex region in the xz-plane (see Figure 6.1). For instance, if we use the standard choices $S' = \pm\{\mathsf{e}_1, \mathsf{e}_3\}$ and $S = \pm\{\mathsf{e}_1, \mathsf{e}_2\}$ (or $S = \pm\{\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3\}$), we have $\|\mathsf{p}\|_K = x + z$ and $x \leq \|\mathsf{p}\|_H \leq x + 4\sqrt{z}$. (We take $x, z \geq 0$ without loss of generality.) We have

$$\lim_{t\to\infty} d(t\mathsf{p}) = \lim_{t\to\infty} \frac{tx + tz}{tx + 4\sqrt{t}\sqrt{z}} = 1 + \frac{z}{x}$$

for $x \neq 0$. This tells us about the asymptotic distribution of d over B_n^K: as $n \to \infty$, when suitably rescaled, it approaches the distribution of $1 + z/x$ over the ℓ^1 unit ball. (Again, this is stated for the first quadrant, for simplicity.) We can thus compute that this subgroup is weakly (but not strongly) statistically undistorted,

and that its distortion index is 2. That means that even though this K is globally quadratically distorted, a linear relationship between the two word metrics, with K-length twice as large as H-length, is nonetheless in this sense typical. It is easy to see that the distortion profile depends nontrivially on the choices of generators for K and H; for instance, if the generators of K are changed to $\pm\{\mathsf{e}_1, \mathsf{e}_3, \mathsf{e}_1\mathsf{e}_3\}$, the profile changes as in Figure 6.1. The new distortion index is $\frac{3}{2}$, meaning that with these generators the subgroup is less distorted.

7. LATTICE POINT COUNTING

Because the balls in the CC metric get "fat" (i.e., the volume of a 1-neighborhood of the sphere is lower-order than the volume of the ball), it is easy to see that the number of lattice points in a ball of large radius equals the volume of the ball (to first order). What is more delicate is to count lattice points in an annular region between a sphere of radius n and a sphere of radius $n-1$.

As before, the footprint of the limit shape, Q, is covered by finitely many quadrilaterals Q_{ij}. For a subset U of $H(\mathbb{R})$, let $\Delta_n U = \{\delta_t(U) \mid n-1 < t \leq n\}$ be the "annular" region between the $(n-1)$st and nth dilate of U.

Recall the function A and set-valued function $\mathbf{z}$ on $\mathfrak{m}$ described in Theorem 3.1. These functions are related by $A = \max \mathbf{z}$; that is, A is the height (maximum z-coordinate) of the unit sphere in the CC metric over points in Q, or equivalently, it is the maximum balayage area of curves of L-length one from the origin to (x, y). We can compute the height of the CC sphere of radius n above $(x, y) \in nQ$ using the formula $H_n(x, y) := n^2 \cdot A(x/n, y/n)$. (Note that for fixed (x, y) and large enough n, this is continuous and strictly increasing in n.) The unit ball is compact, and so A takes a maximum value $A_{\max}$ on Q (the largest height achieved by the unit sphere). Then, for fixed (x, y), the height function $H_n(x, y)$ is bounded above by $A_{\max} \cdot n^2$. But this is not by itself good enough to get control on the height difference between the n-sphere and the $(n-1)$-sphere over a point, $h_n := H_n - H_{n-1}$.

Lemma 7.1 (Control on height). *Take a polygonal* CC *metric on $H(\mathbb{R})$, and let $h_n(x, y)$ be the height difference between the spheres of radius n and $(n-1)$ over (x, y) for $(x, y) \in (n-1)Q$. Fix any nondegenerate quadrilateral Q_{ij}, and let $\Omega_n \subset nQ_{ij}$ denote the complement of the 1-neighborhood of the boundary. Then, $h_n = O(n)$ over Ω_n.*

Proof. First, for $\mathsf{x} \in \Omega_n$, one easily verifies that both $(1/n)\mathsf{x}$ and $(1/(n-1))\mathsf{x}$ lie in Q_{ij}. Recall from the structure of the CC sphere (Theorem 3.1) that the nonnegative part of S over Q_{ij} is the graph of a single quadratic polynomial in x, y; denote this polynomial by $f(x, y) = a_1x^2 + a_2xy + a_3y^2 + b_1x + b_2y + c$. For all $(x, y) \in \Omega_n$, we have

$$\left(x, y, n^2 \cdot f\left(\frac{x}{n}, \frac{y}{n}\right)\right) \in \delta_n(\mathrm{Panel}_{ij}),$$

and so

$$h_n(x,y) = n^2 \cdot f\left(\frac{x}{n}, \frac{y}{n}\right) - (n-1)^2 \cdot f\left(\frac{x}{n-1}, \frac{y}{n-1}\right) = b_1 x + b_2 y + 2cn - c.$$

Since $(x,y) \in \Omega_n \subset nQ_{ij} \subset nQ$, both x and y are $O(n)$, and therefore h_n is as well. □

Theorem 7.2 (Counting Theorem). *Consider the* CC *metric induced on* $H(\mathbb{R})$ *by a norm* $\|\cdot\|_L$ *on* $\mathfrak{m}$*, with* L *a polygon with integer vertices. Let* σ *be an arbitrary measurable subset of* S*, the* CC *unit sphere. Then,*

$$\#(H(\mathbb{Z}) \cap \Delta_n \sigma) = 4n^3 \operatorname{vol}(\hat{\sigma}) + O(n^2).$$

Proof. First, we consider $\sigma \subset S_{\mathsf{reg}}$. For this case, it suffices to treat small subsets σ which project to squares in $\mathfrak{m}$ whose closures are contained in the interior of a single quadrilateral, that is, σ projecting to squares

$$U := \pi(\sigma) = (u, u+\varepsilon) \times (v, v+\varepsilon)$$

such that $[u, u+\varepsilon] \times [v, v+\varepsilon] \subset Q_{ij}^{\circ}$. Since the dilation δ_t is just a homothety on $\mathfrak{m}$, the dilate tU is a square for every t. The condition on the closure of U guarantees that when n is large enough, nU does not intersect the 1-neighborhood of the boundary of nQ_{ij}.

Consider the rectangle $U_n := nU \cap (n-1)U$, let $A_n := \Delta_n \sigma \cap \pi^{-1}(U_n)$ be the part of the nth annular shell that is vertically over U_n, and let $B_n := \Delta_n \sigma \setminus A_n$ be the rest (see Figure 7.1).

We will make the chain of comparisons

$$\#(H(\mathbb{Z}) \cap \Delta_n \sigma) \sim \#(H(\mathbb{Z}) \cap A_n) \sim \operatorname{vol}(A_n) \sim \operatorname{vol}(\Delta_n \sigma) \sim 4n^3 \operatorname{vol}(\hat{\sigma}),$$

while keeping track of the error term.

First, we estimate the number of lattice points in B_n, and its volume. The projection $\pi(B_n)$ is contained in a square annulus in $\mathfrak{m}$ of width $O(1)$ and side-length $O(n)$. Thus, the projection contains at most $O(n)$ lattice points (x,y). By the previous lemma, the difference in heights between $\delta_n S$ and $\delta_{n-1} S$ over each of these points is at most $O(n)$, and so the total number of lattice points in B_n is at most $O(n^2)$. Thus,

$$\#(H(\mathbb{Z}) \cap \Delta_n \sigma) = \#(H(\mathbb{Z}) \cap A_n) + O(n^2).$$

Likewise, we can get an upper bound on the volume of B_n by integrating $h_n(x,y)$ over $(x,y) \in \pi(B_n)$, and so $\operatorname{vol}(B_n) = O(n^2)$. This then shows that $\operatorname{vol}(A_n) = \operatorname{vol}(\Delta_n \sigma) + O(n^2)$. For the comparison on the far right, we have

$$\operatorname{vol}(\Delta_n \sigma) = \operatorname{vol}(\delta_n \hat{\sigma}) - \operatorname{vol}(\delta_{n-1} \hat{\sigma}) = [4n^3 + O(n^2)] \operatorname{vol}(\hat{\sigma}),$$

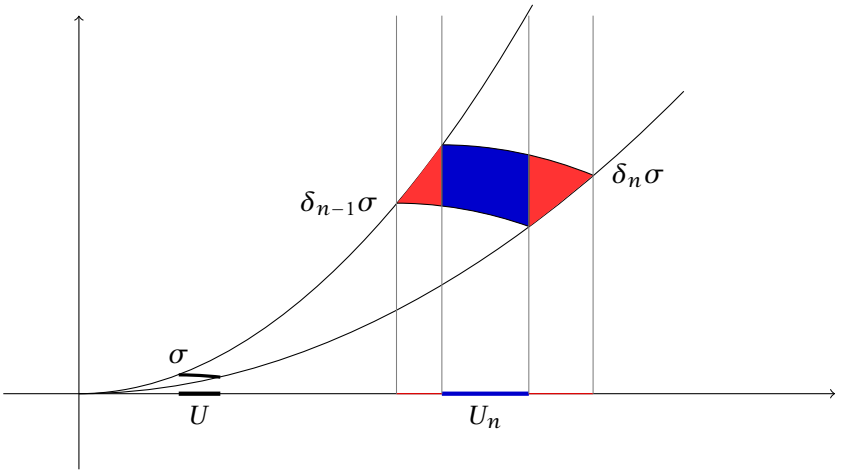

FIGURE 7.1. This shows an xz-plane slice, with U_n along the x-axis, A_n in blue, and B_n in red. We have $A_n \cup B_n = \Delta_n\sigma$; we want to count the lattice points in that entire region. The A_n lattice points are easier to count, and the number in B_n is lower order.

because δ_t expands volume by t^4.

All that remains is to compare $\#(H(\mathbb{Z}) \cap A_n)$ with $\mathrm{vol}(A_n)$. Fix attention on an integer point $(x, y) \in U_n$. For all the lattice points in the vertical line over (x, y), let us count the number in $\Delta_n\sigma$. We find $\#(H(\mathbb{Z}) \cap A_n) = (\sum_{U_n \cap \mathbb{Z}^2} h_n) + O(n^2)$, because $\#(U_n \cap \mathbb{Z}^2) = O(n^2)$ and there is bounded error on each line. On the other hand, $\mathrm{vol}(A_n) = \int_{U_n} h_n$. Regard the sum as a Riemann sum approximation to the integral over the integer squares. Since $h(n)$ is linear, the Riemann sum gives an equal error on each of the $O(n^2)$ squares, for total error of order n^2.

Next, we deal with the case $\sigma \subset S_{\mathsf{uns}}$, where it suffices for us to consider $\sigma = \gamma \times [h, h + \varepsilon]$ with γ a subinterval of an edge of L. Let $\Sigma = \Delta\gamma$ denote the sector of $\mathfrak{m}$ determined by γ, and let U_n be the trapezoid in Σ between $n\gamma$ and $(n-1)\gamma$. We see that $\Delta\sigma = \bigcup_{t \geq 0} t\gamma \times [ht^2, (h+\varepsilon)t^2]$. Let the height difference from top to bottom at (x, y) be denoted $T(x, y) = \varepsilon \cdot \|(x, y)\|_L^2$.

We make the comparison

$$\#(H(\mathbb{Z}) \cap \Delta_n\sigma) \sim \mathrm{vol}(\Delta_n\sigma)$$

by considering the Riemann sum approximation of $\int_{U_n} T$ by $\sum_{U_n \cap \mathbb{Z}^2} T$. (This is where the assumption that L has integer vertices is used; one verifies that the Riemann sum has lower-order error.) For any $\mathsf{u}, \mathsf{v} \in U_n$, the L-norm is between $n-1$ and n, and so $T(\mathsf{v}) - T(\mathsf{u}) = \varepsilon\|\mathsf{v}\|_L - \varepsilon\|\mathsf{u}\|_L \leq 2\varepsilon n$, which gives a bound on

the error per term in the Riemann sum. Since U_n has bounded width, its area and $\#\mathbb{Z}^2 \cap U_n$ are both $O(n)$. Putting these factors together, we get that the total error is $O(n^2)$, and so we have shown that $\#(H(\mathbb{Z}) \cap \Delta_n\sigma) = 4n^3 \operatorname{vol}(\hat{\sigma}) + O(n^2)$, as in the regular case. □

Appendix A. Limit Measure in the Standard Generators

Here, we focus on $H(\mathbb{Z})$ with the standard generating set $S = \mathsf{std} = \pm\{\mathsf{e}_1, \mathsf{e}_2\}$, and consider the word metric with respect to those generators. The goal for this section is to show for this special case that counting measure on discrete spheres in the word metric limits to cone measure (with respect to Heisenberg dilation) on the limit shape S.

Theorem A.1 (Limit measure for standard generators). *We consider the* CC *metric induced by the ℓ^1 norm on $\mathfrak{m}$, which is the limit metric for $(H(\mathbb{Z}), \mathsf{std})$. For any measurable set $\sigma \subset S$, we have $\#(S_n \cap \Delta\sigma) = n^3 \operatorname{vol}(\hat{\sigma}) + O(n^2)$, and therefore*

$$\lim_{n\to\infty} \frac{\#(S_n \cap \Delta\sigma)}{\#S_n} = \frac{\operatorname{vol}(\hat{\sigma})}{\operatorname{vol}(\hat{S})}.$$

Thus, for discrete spheres S_n in $(H(\mathbb{Z}), \mathsf{std})$, the counting measure on $\delta_{1/n}(S_n)$ converges to cone measure on the square $L \subset \mathfrak{m}$.

To prove this, we will study the distribution of points in S_n, finding that the set S_n has points on the line $\{(x, y, t) \mid t \geq 0\}$ if and only if x, y are integers and $x + y + n$ is even; in that case, it contains all lattice points between $\delta_n(S)$ and $\delta_{n-2}(S)$. This is illustrated in Figure A.1.

Lemma A.2 (*CC ***distance versus height). *For any polygonal* CC *metric, distance from the origin is an increasing function of height: regarded as a function of $t \geq 0$, the value $d_{\mathrm{CC}}((x, y, t), 0)$ is continuous, stays constant on an interval $[0, T(x, y)]$, and is strictly increasing thereafter.*

Proof. Let $\mathsf{x}_t = (x, y, t)$. The distance of a point x from $\mathsf{0}$ is the value d such that $\delta_{1/d}(\mathsf{x}) \in S$. If $\mathsf{x}_t \in \Delta S_{\mathsf{uns}}$, then the dilate of x_t that hits S will intersect a side panel, and so the distance of $\mathsf{x}_t = (x, y, t)$ to the origin depends only on x, y. Now, Uns is a closed set with interior which contains $(x, y, 0)$ for every $(x, y) \in \mathfrak{m}$. Thus, $(x, y, t) \in \mathsf{Uns}$ for a closed interval of times $[0, T(x, y)]$.

Now, consider x_t and x_s with $s > t$. These have the same projection to $\mathfrak{m}$, and so do $\delta_r(\mathsf{x}_t)$ and $\delta_r(\mathsf{x}_s)$ for any r. Thus, if $\delta_{1/d}(\mathsf{x}_t) \in S$, we must have $\delta_{1/d}(\mathsf{x}_s)$ vertically above S, meaning that $d_{\mathrm{CC}}(\mathsf{x}_s, \mathsf{0}) > d$. □

Note also that if α is a path in a Cayley graph for $H(\mathbb{Z})$ (i.e., a word written in terms of the generators), then there is a corresponding admissible path in $H(\mathbb{R})$ which ends at the same point and whose CC length is the same as the word length in that word metric. (See the proof of Lemma 4.2 for the construction, noting that the extra height created by generators is zero for $S = \mathsf{std}$, because the generators

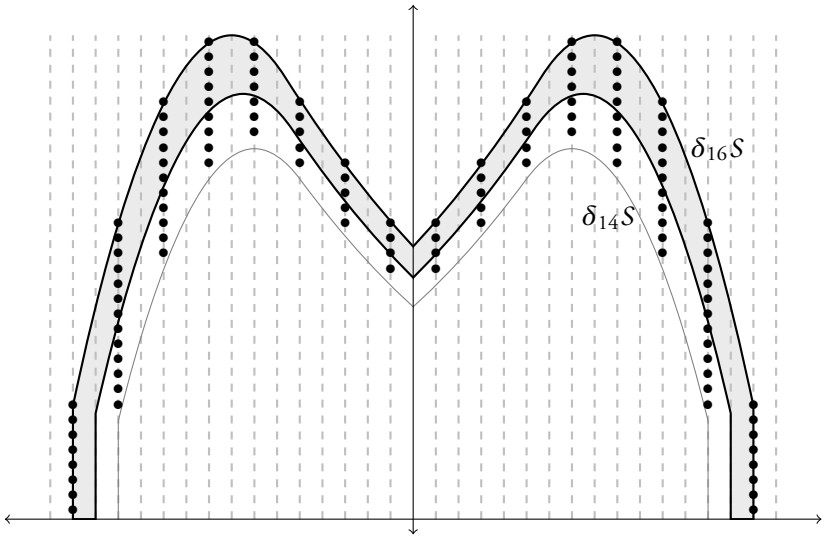

FIGURE A.1. This exactly depicts the cross-section $y = 1$ of $H(\mathbb{R})$, with the points of the sphere S_{16} in $(H(\mathbb{Z}), \mathsf{std})$ marked and three dilates of the CC sphere shown for comparison. The bottom-right-most lattice point is $(15, 1, \frac{1}{2}) = \mathsf{e}_1^8\mathsf{e}_2\mathsf{e}_1^7$. As we show here, S_n has very nearly as many points as the lattice points contained in $\Delta_n S$ (the shaded region): it has twice the points on half the lines.

lie in $\mathfrak{m}$.) That is, word paths are realizable as admissible paths in an obvious way, as in Figure A.2.

However, the CC geodesic in $H(\mathbb{R})$ between two points in $H(\mathbb{Z})$ is often shorter than the geodesic in the word metric, as seen in the figure: the CC geodesics need not have corners at integer points, even if they begin and end at integer points.

FIGURE A.2. Comparing geodesics: $\mathsf{x} = (4, 0, 3)$ is reached by a three-sided CC geodesic of length $5\frac{1}{2}$, while its word length in the standard generators is 6.

Lemma A.3 (Word length versus height). *For any $(x, y) \in \mathbb{Z}^2$, let $\varepsilon = \varepsilon(x, y) = \frac{1}{2}$ if x, y are odd and $\varepsilon = 0$ otherwise, so that $(x, y, \varepsilon + m) \in H(\mathbb{Z})$ for all $m \in \mathbb{Z}$. Then, $|(x, y, \varepsilon + m)|_{\mathsf{std}}$ is nondecreasing for $m = 0, 1, 2, \ldots$.*

Proof. Suppose without loss of generality that (x, y) is in the first quadrant of $\mathfrak{m}$. Let g be a geodesic in $(H(\mathbb{Z}), \mathsf{std})$ from the origin to a point $\mathsf{x} = (x, y, z)$ where $z > 0$. Then, z is equal to the signed area enclosed between g and the straight chord from $\mathsf{0}$ to $\pi(\mathsf{x})$. The path contains only e_1 moves and e_2 moves, and there must be some subword $\mathsf{e}_1\mathsf{e}_2$, or else the signed area would be nonpositive. But then let g' be the same path with some $\mathsf{e}_1\mathsf{e}_2$ subword replaced by $\mathsf{e}_2\mathsf{e}_1$. This has the same length, and its balayage area is one less; thus, it is a path in $H(\mathbb{Z})$ from $\mathsf{0}$ to $(x, y, z-1)$. This may no longer be geodesic, but its length provides an upper bound: $|(x, y, z-1)| \le |(x, y, z)|$. ◻

Next, it is quite easy to see that there is a parity condition on lengths of paths from the origin to (x, y, z).

Lemma A.4 (Parity). *For any path of length n in $H(\mathbb{Z})$ from 0 to (x, y, z), n has the same parity as $x + y$.*

Proof. All paths in $(H(\mathbb{Z}), \mathsf{std})$ are of the form $\Pi \mathsf{e}_1^{a_i}\mathsf{e}_2^{b_i}$, with $\sum a_i = x$, $\sum b_i = y$, and length $n = \sum |a_i| + |b_i|$. Since $r + |r|$ is even for all integers, it follows that $x + y + n \in 2\mathbb{Z}$.

Geometrically, this is just the observation that if a path in $\mathfrak{m}$ to (x, y) goes horizontally past x or vertically past y, it will have to backtrack by the same number of steps. ◻

For fixed $(x, y) \in \mathbb{Z}^2$ and $n \in \mathbb{N}$, let us count the points $(x, y, z) \in S_n$. We first note that $\pi(\delta_n S) = nQ$, so there are no points of S_n over (x, y) if $\|(x, y)\|_L > n$.

Theorem A.5 (Word length in terms of* CC *). *Fix $(x, y) \in \mathbb{Z}^2$. For $z \ge 0$ in $\mathbb{Z} + \varepsilon$, let n be the unique integer with the same parity as $x + y$ such that $n - 2 < d_{\mathrm{CC}}((x, y, z), \mathsf{0}) \le n$. Then, $|(x, y, z)|_{\mathsf{std}} = n$.*

Stated another way, this formula says that, for all $\mathsf{x} \in H(\mathbb{Z})$,

$$d_{\mathrm{CC}}(\mathsf{x}, \mathsf{0}) \le |\mathsf{x}|_{\mathsf{std}} < d_{\mathrm{CC}}(\mathsf{x}, \mathsf{0}) + 2,$$

and describes which of the two integers in that range is the correct value.

Proof. We will restrict ourselves to the half-space $z \ge 0$ without loss of generality; the situation on the other half-space is obtained by reflection.

The point $(x, y, \varepsilon + m)$ lies in the unstable part of $H(\mathbb{R})$ for $0 \le m \le T$, then in a sector with three-sided combinatorics for some $T \le m \le T'$, and finally in a sector with four-sided combinatorics for $m \ge T'$. (See Figure 2.2.) We consider those three cases separately.

Fix $n \ge 0$ of the same parity as $x + y$. By Lemma A.2, there is a unique real value $r \ge 0$ such that (x, y, r) is reached by a CC geodesic of length n. Take that r, and suppose that the geodesic γ is three-sided. Note that $\pi(\gamma)$ is a path in $\mathfrak{m}$ moving only horizontally and vertically; without loss of generality, γ has the form $\mathsf{e}_2^{-t}\mathsf{e}_1^{x}\mathsf{e}_2^{y+t}$, where $t > 0$ and $x + 2t + y = n$. But we know $x + y$ has the same

parity as n, and so it follows that t is an integer. Thus, y is the realization of a word geodesic, and so $|(x,y,r)| = d_{CC}((x,y,r),0)$.

Now assume that y is a four-sided geodesic from the origin to (x,y,r), with length n. Without loss of generality, y has the form $e_2^{y-x-t}e_1^{x+t}e_2^{x+t}e_1^{-t}$, where $t \geq 0$ and $3x - y + 4t = n$. Since $x + y + n$ is even, it follows that $4t$ is an integer with the same parity as $4x$, and so t is either an integer or a half-integer. If t is an integer, then y is the realization of a word geodesic and $|(x,y,r)| = d_{CC}((x,y,r),0)$. If not, then let $s = t - \frac{1}{2}$ and $u = t + \frac{1}{2}$ be the nearest integers. We can form an integer path $e_2^{y-x-u}e_1^{x+s}e_2^{x+u}e_1^{-s}$ of length n by shortening the horizontal sides and lengthening the vertical sides by $\frac{1}{2}$. This encloses area $z = (x+s)(x+u) - \frac{1}{2}xy$, which is in $\varepsilon + \mathbb{Z}$, and so the endpoint (x,y,z) of the new path is in $H(\mathbb{Z})$. This area is smaller than $r = (x+t)^2 - \frac{1}{2}xy$ by $\frac{1}{4}$, and so it must be the nearest lattice point. This new path may not be a geodesic, but it establishes the inequality $|(x,y,z)| \leq n$. On the other hand, shortening all four sides of y by $\frac{1}{4}$ produces a 4-sided CC geodesic to (x,y,r') of length $n-1$, and one easily checks that $r' < z < r$. This shows that $d_{CC}((x,y,z),0) > n-1$, and it follows that $|(x,y,z)| = n$.

In these cases, we have established that the highest lattice point over (x,y) inside the closed CC ball of radius n has word length exactly n. We now want to show that the lowest lattice point over (x,y) outside the closed CC ball of radius $n-2$ also has word length exactly n. Its CC norm is larger than $n-2$, and so it has word length larger than $n-2$, and by parity considerations the word length must be exactly n.

The last case is easy: for any unstable point $(x,y,z) \in H(\mathbb{R})$, its CC norm is $d_{CC}((x,y,z),0) = x + y$. We claim that these points satisfy $|(x,y,z)|_{\mathsf{std}} = x + y$ as well. To see this, just note that starting with the two-sided path $e_1^x e_2^y$, which encloses area $\frac{1}{2}xy$ with length $x + y$, one can shave off area one unit at a time with generator swaps (replacing some occurrence of e_1e_2 with e_2e_1) until the area enclosed is z. This produces a word path whose length matches the CC geodesic, and so it too is geodesic. □

We have shown that if σ is a subset of S_{reg}, then $|S_n \cap \Delta\sigma| \sim |H(\mathbb{Z}) \cap \Delta_n\sigma|$, because S_n has twice as many points as $\Delta_n\sigma$, but on half of the lines (see Figure A.1). On the other hand, $|S_n \cap \Delta\sigma| = |H(\mathbb{Z}) \cap \Delta_n\sigma|$ exactly for $\sigma \subset S_{\mathsf{uns}}$. Putting these together and then using the Counting Theorem (Theorem 7.2) to estimate the number of lattice points in $\Delta_n\sigma$, we obtain Theorem A.1.

We close by conjecturing that cone measure is the limit measure for any finite generating set on $H(\mathbb{Z})$. In particular, we conjecture that changes to a generating set that involve only the central letter do not affect the limit measure. Indeed, one can show that $\pm\{e_1, e_2, e_3\}$, which clearly has the same limit shape as std, has the same limit measure as well: the spheres differ by only one point in each vertical fiber.

Acknowledgments. Many thanks to Ralf Spatzier, Alex Eskin, and Emmanuel Breuillard for illuminating conversations on numerous aspects of the problems considered here. We thank Sebastian Hensel for suggesting the idea of measuring statistical subgroup distortion, and Larry Guth for a useful conversation about it. Thanks also to Laurent Bartholdi for sharing with us the unpublished notes of Michael Stoll mentioned above. Finally, thanks to the anonymous referee for useful comments.

The first and second author were partially supported by NSF DMS-0906086 and NSF RTG-0602191, respectively.

References

[1] M. Beck and S. Robins, *Computing the Continuous Discretely*: *Integer-point Enumeration in Polyhedra*, Undergraduate Texts in Mathematics, Springer, New York, 2007. MR2271992 (2007h:11119)

[2] S. Blachère, *Word distance on the discrete Heisenberg group*, Colloq. Math. **95** (2003), no. 1, 21–36. http://dx.doi.org/10.4064/cm95-1-2. MR1967551 (2003k:20058)

[3] E. Breuillard, *Geometry of groups of polynomial growth and shape of large balls*, preprint, available at http://arxiv.org/abs/arXiv:0704.0095.

[4] E. Breuillard and E. Le Donne, *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. http://dx.doi.org/10.1073/pnas.1203854109. MR3153949

[5] D. Yu. Burago, *Periodic metrics*, Representation Theory and Dynamical Systems, Adv. Soviet Math., vol. 9, Amer. Math. Soc., Providence, RI, 1992, pp. 205–210. MR1166203 (93c:53029)

[6] H. Busemann, *The isoperimetric problem in the Minkowski plane*, Amer. J. Math. **69** (1947), 863–871. MR0023552 (9,372h)

[7] L. Capogna, D. Danielli, S. D. Pauls, and J. T. Tyson, *An Introduction to the Heisenberg Group and the sub-Riemannian Isoperimetric Problem*, Progress in Mathematics, vol. 259, Birkhäuser Verlag, Basel, 2007. MR2312336 (2009a:53053)

[8] P. Dani, *The asymptotic density of finite-order elements in virtually nilpotent groups*, J. Algebra **316** (2007), no. 1, 54–78. http://dx.doi.org/10.1016/j.jalgebra.2007.06.023. MR2354853 (2008j:20124)

[9] M. Duchin, S. Lelièvre, and Ch. Mooney, *The geometry of spheres in free abelian groups*, Geom. Dedicata **161** (2012), 169–187. http://dx.doi.org/10.1007/s10711-012-9700-x. MR2994037

[10] S. A. Krat, *Asymptotic properties of the Heisenberg group*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **261** (1999), no. Geom. i Topol. 4, 125–154, 268 (Russian, with English and Russian summaries); English transl., J. Math. Sci. (New York) **110** (2002), no. 4, 2824–2840. http://dx.doi.org/10.1023/A:1015306413677. MR1758422 (2002b:20057)

[11] P. Pansu, *Croissance des boules et des géodésiques fermées dans les nilvariétés*, Ergodic Theory Dynam. Systems **3** (1983), no. 3, 415–445 (French, with English summary). http://dx.doi.org/10.1017/S0143385700002054. MR741395 (85m:53040)

[12] M. Shapiro, *A geometric approach to the almost convexity and growth of some nilpotent groups*, Math. Ann. **285** (1989), no. 4, 601–624. http://dx.doi.org/10.1007/BF01452050. MR1027762 (91d:20035)

[13] M. Stoll, *Rational and transcendental growth series for the higher Heisenberg groups*, Invent. Math. **126** (1996), no. 1, 85–109. http://dx.doi.org/10.1007/s002220050090. MR1408557 (98d:20033)

[14] ———, *On the asymptotics of the growth of 2-step nilpotent groups*, J. London Math. Soc. (2) **58** (1998), no. 1, 38–48. http://dx.doi.org/10.1112/S0024610798006371. MR1666070 (99k:20068)

MOON DUCHIN:
Department of Mathematics
Tufts University
Medford, MA 02155, USA
E-MAIL: moon.duchin@tufts.edu
URL: http://mduchin.math.tufts.edu

CHRISTOPHER MOONEY:
Department of Mathematics
Bradley University
Peoria, IL 61625, USA
E-MAIL: cpmooney@bradley.edu

KEY WORDS AND PHRASES: Geometric group theory, sub-Riemannian geometry, nilpotent groups, counting problems.

2010 MATHEMATICS SUBJECT CLASSIFICATION: 20F65, 20F18, 11N45.

Received: March 11, 2013.