

Dynamics and Equidistribution on Homogeneous Spaces

Author address:

DEPARTEMENT MATHEMATIK, ETH ZÜRICH, RÄMISTRASSE 10, 8092 ZÜRICH,
SWITZERLAND
E-mail address: `hartnick@math.ethz.ch`

Contents

Preface	vii
Part 1. Arithmetic counting problems	1
Chapter 1. Counting integral binary quadratic forms	3
1.1. Binary quadratic forms and discriminant varieties	3
1.2. Integral points and the counting problem	4
1.3. Lattices and the abstract counting problem	9
1.4. Equidistribution implies the good counting property	14
1.5. The group $SL_2(\mathbb{R})$ and the hyperbolic plane	17
1.6. Ergodicity, mixing and the Howe-Moore property	19
1.7. The Howe-Moore theorem for $SL_2(\mathbb{R})$	21
1.8. The wavefront lemma in the hyperbolic plane	22
1.9. Counting definite quadratic forms	23
Chapter 2. Equidistribution for symmetric pairs	27
2.1. Motivation and overview	27
2.2. CAT(0) spaces and symmetric spaces	28
2.3. The space of positive definite matrices	31
2.4. Reductive groups and their symmetric spaces	34
2.5. Riemannian symmetric pairs	36
2.6. Flats and Weyl chambers	40
2.7. Parabolic subgroups and horospheres	42
2.8. The wavefront lemma for Riemannian symmetric pairs	46
2.9. Non-Riemannian symmetric pairs and relative decompositions	47
2.10. The wavefront lemma for general symmetric pairs	49
2.11. The Howe-Moore theorem for $SL_d(\mathbb{R})$	49
2.12. Howe-Moore theorems for semisimple groups	50
2.13. Equidistribution for symmetric pairs	51
Part 2.	53
Chapter 3. Unipotent flows	55
3.1. More examples of counting problems	55
3.2. Counting beyond strict equidistribution	56
3.3. Ergodic measures and Property (T)	57
3.4. The Eskin-Mozes-Shah theorem	58
3.5. Ratner's measure classification theorem	58
Bibliography	63

Preface

These are notes for the seminar on *Dynamics and Equidistribution on Homogeneous Spaces* organized by Marc Burger and Manfred Einsiedler at ETH Zürich in the spring semester of 2010. They are not proofread and meant for internal use only. Please send corrections and remarks to hartnick@math.ethz.ch.

Tobias Hartnick

Part 1

Arithmetic counting problems

Counting integral binary quadratic forms

1.1. Binary quadratic forms and discriminant varieties

A *binary quadratic form* q over an integral domain R is a homogeneous polynomial of degree 2 over R in two variables, i.e. q is of the form

$$(1.1) \quad q(X, Y) = aX^2 + bXY + cY^2$$

for some $a, b, c \in R$. We denote the set of all binary quadratic forms over R by $\mathcal{P}(R)$. The group $SL_2(R)$ acts on $\mathcal{P}(R)$ via

$$g.q(X, Y) = q(g^{-1}(X, Y)).$$

Explicitly, if

$$g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

and q is given as in (1.1), then

$$\begin{aligned} g.q(X, Y) &= (\delta^2 a - \gamma\delta b + \gamma^2 c)X^2 + (-2\beta\delta a + (\alpha\delta + \beta\gamma)b - 2\alpha\gamma c)XY \\ &\quad + (\beta^2 a - \alpha\beta b + \alpha^2 c)Y^2. \end{aligned}$$

An important invariant of this action is given by the *discriminant*. If q is given as in (1.1), then the latter is given by

$$(1.2) \quad d := d(q) := b^2 - 4ac.$$

Invariance of d can be checked directly from the above formula; it implies that $SL_2(R)$ preserves the *discriminant varieties*

$$V_d(R) := \{q \in \mathcal{P}(R) \mid d(q) = d\}.$$

The behaviour of the family V_d as d changes over R depends very much on R . If R is a field, then the sets $V_d(R)$ are affine varieties over R . Note that $V_0(R)$ is singular, while $V_d(R)$ is non-singular for $d \neq 0$.

For $R = \mathbb{C}$ all the $V_d(\mathbb{C})$ for $d \neq 0$ are isomorphic; moreover, $SL_2(\mathbb{C})$ is transitive on all $V_d(\mathbb{C})$, and the point-stabilizers for different d are conjugate. For $d = 0$ there are two $SL_2(\mathbb{C})$ -orbits, given by 0 and its complement respectively.

The picture becomes slightly more interesting for $R = \mathbb{R}$: Again, the case $d = 0$ is special and corresponds to a singular double cone (with two orbits, one of which is a fixed point). The varieties $V_d(\mathbb{R})$ for $d \neq 0$ are all non-singular, however, their shape depends crucially on the sign of d : Namely, $V_d(\mathbb{R})$ is a one-sheeted hyperboloid for $d > 0$, and a two-sheeted hyperboloid for $d < 0$. Arithmetically this corresponds to the fact that forms of positive discriminant are *indefinite*, while forms of negative discriminant are *definite*; the two sheets of the hyperboloid then correspond to

positive definite and *negative definite* forms respectively, and we denote them by $V_d^+(\mathbb{R})$ and $V_d^-(\mathbb{R})$ accordingly. With this notation the classification of quadratic forms over \mathbb{R} can be summarized as follows:

Proposition 1.1.1. *The action of $SL_2(\mathbb{R})$ is transitive on $V_d(\mathbb{R})$ if $d > 0$ and transitive on $V_d^\pm(\mathbb{R})$ if $d < 0$. The stabilizer of a point in $V_d(\mathbb{R})$ is conjugate to $SO(1,1)$ if $d > 0$ and conjugate to $SO(2)$ if $d < 0$. In particular,*

$$V_d(\mathbb{R}) \cong SL_2(\mathbb{R})/SO(1,1) \quad (d > 0)$$

and

$$V_d^\pm(\mathbb{R}) \cong SL_2(\mathbb{R})/SO(2) \quad (d < 0).$$

Remark 1.1.2. A more familiar model for the homogeneous space $SL_2(\mathbb{R})/SO(2)$ is given by the upper halfplane

$$\mathbb{H}^2 := \{z \in \mathbb{C} \mid \Im(z) > 0\}.$$

For any $d < 0$ we can give an explicit equivariant identification of $V_d^\pm(\mathbb{R})$ with \mathbb{H}^2 as follows: Given $q \in V_d^\pm(\mathbb{R})$, let X_q denote the set of complex zeros of q . Since q is homogeneous, X_q can be considered as a subset of \mathbb{CP}^1 . Since $\deg q = 2$, the set X_q consists of two points, and since q is real (and has no real zeros), these two points are complex conjugates of each other. In particular, X_q intersects $\mathbb{H}^2 \subset \mathbb{CP}^1$ in a single point. Mapping q to this zero then defines an $SL_2(\mathbb{R})$ -equivariant bijection between $V_d^\pm(\mathbb{R})$ and \mathbb{H}^2 . This follows from the fact that a quadratic form of fixed discriminant is uniquely determined by its zeros.

1.2. Integral points and the counting problem

Our main interest here lies in the set $\mathcal{P}(\mathbb{Z})$ of *integral* binary quadratic forms and its subsets $V_d(\mathbb{Z})$. We will think of $V_d(\mathbb{Z})$ as integral points on $V_d(\mathbb{R})$; unlike the real or complex case, the sets $V_d(\mathbb{Z})$ may well be empty. In fact we have:

Proposition 1.2.1. *Let $d \in \mathbb{Z}$. Then $V_d(\mathbb{Z}) \neq \emptyset$ if and only if $d \equiv 0 \pmod{4}$ or $d \equiv 1 \pmod{4}$.*

PROOF. Let q be given as in (1.1). Then

$$d(q) = b^2 - 4ac \equiv b^2 \pmod{4},$$

so $d(q)$ is necessarily a square modulo 4, and the only such squares are 0 and 1. Conversely, if $d = 4k + 1$, then $q_d(X, Y) := X^2 + XY - kY^2 \in V_d(\mathbb{Z})$; similarly, if $d = 4k$, then $q_d(X, Y) := X^2 - kY^2 \in V_d(\mathbb{Z})$. \square

In view of the proposition we will call $d \in \mathbb{Z}$ a *discriminant* if it is congruent to 0 or 1 modulo 4. For the rest of this section we assume that d is a non-zero discriminant. Then $V_d(\mathbb{Z}) \subset V_d(\mathbb{R})$ is a discrete subset, whence we can count the number of integral points in a given compact subset of $V_d(\mathbb{R})$. The asymptotics of this number as the compact subset grows is then a measure for the size of $V_d(\mathbb{Z})$ inside $V_d(\mathbb{R})$. It turns out a posteriori that the precise shape of the compact subsets involved in this asymptotic counting is of little relevance as long as the compact subsets behave sufficiently like balls. For simplicity, let us fix a norm on $V_d(\mathbb{C})$ by setting

$$\|aX^2 + bXY + cY^2\| = |a|^2 + \frac{|b|^2}{2} + |c|^2.$$

We then denote by $B(0, T)$ the ball of radius T around 0 in $V_d(\mathbb{R})$ with respect to $\|\cdot\|$. Now the *counting problem* for integral binary quadratic forms of discriminant d can be formulated as follows:

Problem 1.2.2. *Determine the asymptotic of the numbers*

$$N(V_d, T) := |V_d(\mathbb{Z}) \cap B(0, T)|$$

as $T \rightarrow \infty$.

A first step towards the solution of this counting problem is the following classical result of Gauß:

Theorem 1.2.3. *Let d be a non-zero discriminant. Then there are finitely many $SL_2(\mathbb{Z})$ -orbits in $V_d(\mathbb{Z})$.*

To see that the content of the theorem is non-trivial, we consider the case $d = 0$. In this case the greatest common divisor of the coefficients is invariant under $SL_2(\mathbb{Z})$. This shows in particular, that the number of $SL_2(\mathbb{Z})$ -orbits in $V_0(\mathbb{Z})$ is infinite.

PROOF OF THEOREM 1.2.3. The idea is to show that every integral binary quadratic form can be brought into a certain normal form by applying elements of $SL_2(\mathbb{Z})$, and to show that there are only finitely many quadratic forms in this normal form for a given discriminant. Classically this goes by the name of *reduction theory* of quadratic forms; the precise reduction algorithm is different for definite and indefinite forms. We will only deal with the definite case here and refer the reader to [BV07, Chapter 6] for the indefinite case. In fact, we will only deal with positive definite forms, but the negative definite case is easily reduced to the positive one. Thus let $q(X, Y) = aX^2 + bXY + cY^2$ be a positive definite form. We identify q with its coefficients thus writing $q = (a, b, c)$. We then call q *normal* if $-a < b \leq a$ and *reduced* if in addition $a < c$ or $a = c$ and $b \geq 0$. If $d = d(q)$ and q is reduced then

$$-d = 4ac - b^2 \geq 4a^2 - a^2 = 3a^2,$$

hence

$$|b| \leq a \leq \sqrt{-d/3},$$

which shows that there are only finitely many reduced forms of discriminant d . It thus remains to show that every $SL_2(\mathbb{Z})$ -orbit of positive definite forms contains a reduced form. We first claim that every positive definite form can be turned into a normal one by multiplication with a suitable power of the matrix

$$T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Indeed, since

$$T^s \cdot (a, b, c) = (a, b + 2sa, c)$$

we may choose $s = -[(a - b)/2a]$. Actually, this s is unique and we may thus refer to $N(a, b, c) := T^s \cdot (a, b, c)$ as the normalization of (a, b, c) . On the other hand, applying the matrix

$$S := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

transforms (a, b, c) into $(c, -b, a)$. Starting with (a, b, c) we now run the following algorithm:

- (1) Replace (a, b, c) by $N(a, b, c)$. If the result is reduced, then return $N(a, b, c)$
- (2) Otherwise apply S . Apply Step (1) to the resulting form.

By construction, the algorithm produces a reduced form equivalent to (a, b, c) if it terminates. On the other hand, if (a_i, b_i, c_i) is the form produced after i iterations of Step (2), then the sequence (a_i) is a strictly decreasing sequence of non-negative integers. Thus the algorithm terminates, and the theorem follows. \square

According to Theorem 1.2.3, the counting problem 1.2.2 for $V_d(\mathbb{Z})$ can be reduced into two separate steps:

- (1) Determine a finite set $X \subset V_d(\mathbb{Z})$ of representatives for the various $SL_2(\mathbb{Z})$ -orbits in $V_d(\mathbb{Z})$.
- (2) For each $q \in X$ estimate the asymptotic of the numbers

$$N(q, T) := |SL_2(\mathbb{Z}) \cdot q \cap B(0, T)|$$

as $T \rightarrow \infty$.

These two problems are actually very different in nature: The first step is specific to quadratic forms and can be solved by determining all reduced quadratic forms of a given discriminant. A naive way to do this, say in the positive definite case, is to list all pairs (a, b) with

$$|b| \leq a \leq \sqrt{-d/3},$$

and to check whether $c := \frac{b^2 - d}{4a}$ is an integer. Then one selects those triples (a, b, c) with $c \in \mathbb{Z}$ and (a, b, c) reduced. Thus the first step can be solved in principle. (For more efficient algorithms see [BV07].)

For the second step, there is no obvious algorithm. Note, however, that this step is not at all specific to quadratic forms. Rather, it is a prototype of a very general counting problem, which will be at the core of this book: Let $G := SL_2(\mathbb{R})$, $H < G$ a subgroup conjugate to $SO(1, 1)$ or $SO(2)$ (depending on whether $d > 0$ and $d < 0$) and $\Gamma := SL_2(\mathbb{Z})$. Then the second step can be reformulated as follows:

Problem 1.2.4. *Given $q \in G/H$, compute the asymptotics of*

$$N(q, T) := |\Gamma \cdot q \cap B(0, T)|$$

as $n \rightarrow \infty$.

In our specific setting, we can solve this problem explicitly. Let us denote by $\mu(T)$ the volume of the ball $B(0, T)$. Then we will compare $\mu(T)$ to $N(q, T)$. Heuristically, the proportion $N(q, T)/\mu(T)$, i.e. the number of points per volume inside $B(0, T)$ should converge to the ratio

$$\frac{\text{Vol}((H \cap \Gamma) \backslash H)}{\text{Vol}(\Gamma \backslash G)},$$

where H denotes the stabilizer of q in $SL_2(\mathbb{R})$. One would thus expect the asymptotic of $N(q, T)$ to be

$$N(q, T) \sim \frac{\text{Vol}((H \cap \Gamma) \backslash H)}{\text{Vol}(\Gamma \backslash G)} \cdot \mu(T).$$

However, to even make sense of this statement, one needs to overcome two problems: Firstly, one has to make precise what one means by volume; secondly, one has to make sure that the volumes involved in the formula are *finite*. As far as the first

problem is concerned, we will use suitably normalized Haar measures to define the above volumes. The precise definition requires a somewhat technical discussion concerning volumes on homogeneous spaces, and we defer this discussion to Section 1.3 below. On the other hand, the second problem is a real problem: Our heuristic makes sense only if the discrete subgroups $\Gamma < G$ and $H \cap \Gamma < H$ have finite covolume. In general, discrete subgroups of finite covolume are called *lattices*, and we will discuss such subgroups in more detail in Section 1.3. For the present discussion it suffices to know that $SL_2(\mathbb{Z}) < SL_2(\mathbb{R})$ and every cocompact discrete subgroup are lattices. Our problem thus concerns the subgroup $H \cap \Gamma$, which may or may not be a lattice in H . This is not an issue in the case of definite quadratic forms; here $H \sim SO(2)$ is compact, and thus every discrete subgroup is a lattice. However, in the indefinite case, the group H is conjugate to $SO(1,1)$, hence non-compact. To see that $H \cap \Gamma$ may fail to be a lattice in this case, we consider the following example:

Example 1.2.5. Let $d := 4$ and consider $q(X, Y) = X^2 + 2XY \in V_4(\mathbb{Z})$. Let $H := \text{Stab}_G(q)$. Now assume

$$g = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in H \cap \Gamma;$$

then

$$\delta^2 - 2\gamma\delta = 1, \quad -2\beta\delta + 2\alpha\delta + 2\beta\gamma = 2, \quad \beta^2 - 2\alpha\beta = 0.$$

If $\beta = 0$, then $2\alpha\delta = 2$, whence $\alpha = \delta =: \epsilon \in \{\pm 1\}$. Consequently, $\delta = 0$ and thus $g = \pm \mathbf{1}$. Otherwise $\beta = 2\alpha$, and the second equation becomes $\alpha(2\gamma - \delta) = 1$. Now the first equation implies $\delta - 2\gamma = \delta = \epsilon \in \{\pm 1\}$, so $\alpha = -\epsilon$. In any case, there are only finitely many possibilities for g . We deduce that $H \cap \Gamma$ is finite, and thus cannot be a lattice in H .

These kinds of problems occur for all square numbers $d = f^2$. Actually, it should be expected that squares can cause problems: If $d = f^2 d'$ for some integer $f > 1$ and $q(X, Y) = a'X^2 + b'XY + c'Y^2$ is a binary quadratic form of discriminant d' , then $f a'X^2 + f b'XY + f c'Y^2$ is a form of discriminant d . Thus $V_d(\mathbb{Z})$ contains a copy of $V_{d'}(\mathbb{Z})$, which has to be taken special care of in any sort of counting problem. In order to avoid this problem we define:

Definition 1.2.6. A non-zero discriminant d is called a *fundamental discriminant* if there does not exist an integer $f > 1$ such that d/f^2 is a discriminant.

For fundamental discriminants, the aforementioned problems do not arise:

Proposition 1.2.7. *Suppose $q(X, Y) = aX^2 + bXY + cY^2$ is an integral binary quadratic form of fundamental discriminant d . If $H = \text{Stab}_G(q)$, then*

$$H \cap \Gamma = \left\{ \left(\begin{array}{cc} \frac{x-by}{2} & -cy \\ ay & \frac{x+by}{2} \end{array} \right) \mid (x, y) \in \mathbb{Z}^2, x^2 - dy^2 = 4 \right\}.$$

Moreover, $H \cap \Gamma$ is a co-compact lattice in H .

In fact, $H \cap \Gamma$ is co-compact in H as soon as d is not a square, but the above explicit description of $H \cap \Gamma$ is not available in this more general case. The equation

$$x^2 - dy^2 = 4$$

used for the parametrization of $H \cap \Gamma$ is the famous *Pell equation* of discriminant d .

PROOF OF PROPOSITION 1.2.7. It is easy to check that the matrices of the above form act as automorphisms. For the converse let us assume $a, b, c \neq 0$ (the remaining cases can be dealt with separately). Since d is fundamental we have $\gcd(a, b, c) = 1$. Now assume

$$U := \begin{pmatrix} s & t \\ u & v \end{pmatrix} \in SL_2(\mathbb{Z})$$

stabilizes q . Then writing out the matrix equation one obtains

$$at = -cu, \quad bu = a(v - s), \quad bt = c(s - v).$$

Then $\gcd(a, b, c) = 1$ implies $a|u$, whence we can solve the above equation uniquely in \mathbb{Z} . If we set $x := s + v$ and $y := u/a$ then we obtain $x^2 - dy^2 = 4$ and

$$U = \begin{pmatrix} \frac{x-by}{2} & -cy \\ ay & \frac{x+by}{2} \end{pmatrix},$$

which yields the above description of $H \cap \Gamma$. It remains to deduce that $H \cap \Gamma$ is a lattice in H . Since H is one-dimensional it suffices to show that U is infinite, i.e. that the Pell equation has infinitely many solutions for every fundamental discriminant d . The latter is indeed the case. For this consider the sequence of convergents $\frac{h_n}{k_n}$ of the continued fraction expansion of d . Then it follows from the theory of continued fractions [RS92, Thm. IV.2.2] that for some n , the pair $x = h_n$, $y = k_n$ provides a solution to the Pell equation, which is different from the trivial solution $(2, 0)$. Once we have a single non-trivial solution (x_1, y_1) we can produce infinitely many others by the recurrence relation

$$x_{n+1} := x_1 x_n + d y_1 y_n, \quad y_{n+1} := x_1 y_n + y_1 x_n.$$

This shows that the Pell equation has infinitely many solution for a fundamental discriminant d and finishes the proof. \square

By the proposition, our above heuristic for the counting problem is meaningful, if we restrict attention to fundamental discriminants. Even better, in this case, it provides the correct result:

Theorem 1.2.8. *Let d be a fundamental discriminant. Then for every integral binary quadratic form q of discriminant d with stabilizer $H = \text{Stab}_G(q)$ we have*

$$(1.3) \quad N(q, T) \sim \frac{\text{Vol}((H \cap \Gamma) \backslash H)}{\text{Vol}(\Gamma \backslash G)} \cdot \mu(T)$$

We will derive Theorem 1.2.8 from our solution of a more general abstract counting problem below. Before we turn to this more general setting, let us describe the consequences of Theorem 1.2.8 for our initial problem 1.2.2: The function $\mu(T)$ is well-understood at least since the work of Siegel; it grows linearly in T []. In particular, $N(q, T)$ grows linearly in T for every q of fundamental discriminant. Combining this with Theorem 1.2.3 we obtain:

Corollary 1.2.9. *For every fundamental discriminant d there exists $c_d > 0$ such that*

$$N(V_d, T) \sim c_d \cdot T.$$

Note that Theorem 1.2.8 is much stronger: It allows us to actually compute the coefficients c_d (at least in theory - in practice, they remain rather mysterious). In fact, Corollary 1.2.9, unlike the theorem, can be established by elementary means and has been known since the 19th century. Moreover, linear growth is already suggested by a much simpler heuristic: Consider a homogeneous polynomial $p : \mathbb{R}^n \rightarrow \mathbb{R}$ of degree k . Then p is expected to take roughly $c \cdot T^k$ values on $\mathbb{Z}^n \cap B(0, T)$ (between $-\frac{c}{2}T^k$ and $\frac{c}{2}T^k$), while there are approximately T^n integral points. Thus the levelsets of p should on average contain $\frac{1}{c}T^{n-k}$ points. In the case of the discriminant function we have $n = 3$, $d = 2$, which predicts linear growth. This naive heuristic already shows that Corollary 1.2.9 is not too surprising; however, unlike the heuristic underlying Theorem 1.2.8, it cannot be made precise (and cannot explain, why fundamental discriminants behave different from non-fundamental ones).

To round off our discussion on non-fundamental discriminants, it is interesting to note that in those cases, where $H \cap \Gamma$ is not a lattice in H , the asymptotic behaviour of $N(V_d, T)$ is *not* linear; instead, it grows with a rate of $T \log T$ \square .

1.3. Lattices and the abstract counting problem

We have seen in the last section that the key to the solution of the counting problem for integral binary quadratic forms was provided by Theorem 1.2.8. We now aim to provide a more general formulation of the situation in this theorem. Our first formulation of the problem will be as general as possible; we will investigate later, which additional assumptions are needed for the general theorem to be correct.

To start with, the theorem concerns a topological group G together with two closed subgroups H and Γ , where the latter was assumed discrete. Problem 1.2.4 can already be defined in this generality. However, the point of our heuristic was to compare the growth of $N(q, n)$ with the volume growth of the sequence B_n of balls used to define $N(q, n)$. For this we need a good notion of volume on (homogeneous spaces of) G . Such a notion exists for any locally compact group; however, in view of the applications we have in mind, we will only need the case, where G is a Lie group. In this case, we can describe an invariant measure explicitly:

Denote by \mathfrak{g} the Lie algebra of G . Since any left-invariant differential form on G is uniquely determined by its value at the identity $e \in G$ and conversely, every form on \mathfrak{g} can be extended to a left-invariant form we have

$$\Omega^{\dim G}(G)^G \cong \bigwedge^{\dim \mathfrak{g}} \mathfrak{g}^* \cong \mathbb{R};$$

thus there is - up to scalar multiple - a unique non-zero left-invariant volume form $\omega \in \Omega^{\dim G}(G)^G$. We then find a corresponding left-invariant Radon measure on G given by

$$\mu_G(A) := \left| \int_A \omega \right|$$

for any bounded open subset $A \subset G$. Again, μ_G is unique up to a positive factor and any choice for μ_G will be referred to as a *left-Haar measure* on G . Similarly, we can use right-invariant forms on G to define a *right-Haar measure* on G . For the next few paragraphs we will always work with right-Haar measures, and this distinction is important. Later we will see that for our groups of interest, left-

and right-Haar measures coincide, but to establish this fact, we have to carefully distinguish left and right.

In general, there is no preferred normalization for μ_G . If G is compact, then μ_G is finite and may be normalized to total mass 1. More importantly, if G is discrete, then μ_G assigns a constant mass c to every singleton $\{g\} \subset G$. In this case we normalize $c := 1$. With this normalization we refer to μ_G as the *counting measure* on G .

For the formulation of our counting problem we are interested in invariant measures on homogeneous spaces $H \backslash G$. We can push-forward the measure μ_G along the canonical projection $G \rightarrow H \backslash G$, but the result will in general not be invariant for the G -action. To formulate a criterion we need to introduce a function, which measures the difference between right- and left-Haar measure:

Lemma 1.3.1. *Let G be a Lie group and fix a right-Haar measure μ_G . Then there exists a continuous homomorphism $\Delta_G : G \rightarrow (\mathbb{R}^{>0}, \cdot)$ such that for all $f \in C_c(G)$ and all $h \in G$*

$$(1.4) \quad \int_G f(hg) d\mu_G(g) = \Delta_G(h) \cdot \int_G f(g) d\mu_G(g).$$

PROOF. Existence of a function Δ_G satisfying (1.4) follows from the uniqueness of the right-Haar measure. Applying (1.4) to the product hk for $h, k \in G$ we see that Δ_G is necessarily a homomorphism. In particular, it suffices to establish continuity of Δ_G at the identity $e \in G$. By (1.4) we have for any $h \in G$, $f \in C_c(G)$ with $\int f d\mu_G = 1$ the inequality

$$|\Delta_G(e) - \Delta_G(h)| \leq \int_G |f(g) - f(hg)| d\mu_G(x).$$

Now using uniform continuity of compactly supported functions the right hand side can be made smaller than any given ϵ for h sufficiently close to e . \square

The homomorphism Δ_G is called the *modular function*¹ on G and the group G is called *unimodular* if its modular function is trivial; equivalently, the measure μ_G is bi-invariant under G . Evidently, the counting measure on a discrete group is bi-invariant, therefore discrete groups are unimodular. Similarly, connected simple Lie groups are unimodular, since they do not admit continuous homomorphisms into G ; this argument can be extended to semisimple Lie groups. On the other hand, there also exist many nilpotent unimodular Lie groups. The relation between the modular function and measures on homogeneous spaces is provided by the following lemma:

Lemma 1.3.2 (Weil). *Let G be a Lie group and $H < G$ be a closed subgroup. Then there exists a G -invariant measure μ on $H \backslash G$ if and only if $\Delta_G|_H = \Delta_H$. In this case the measure μ is unique up to constant and can be normalized in such a way that*

$$(1.5) \quad \int_{H \backslash G} \int_H f(hg) d\mu_H(h) d\mu(Hg) = \int_G f(g) d\mu_G(g).$$

¹Actually, *right-modular function* would be a more adequate term, but we will not use left-modular functions here.

PROOF. Assume first that a G -invariant measure μ exists. It is then easy to check that the left hand side of 1.5 defines a right-Haar measure on G . (This yields in particular the uniqueness statement of the lemma.) By normalizing μ_G accordingly, we may thus assume that (1.5) holds. Now let $h' \in H$ and denote by $l_{h'}$ the left-multiplication by h' . Then

$$\begin{aligned} \Delta_G(h') \cdot \int_G f(g) d\mu_G(g) &= \int_G (f \circ l_{h'})(g) d\mu_G(g) \\ &= \int_{H \backslash G} \int_H f(h'hg) d\mu_H(h) d\mu(Hg) \\ &= \Delta_H(h') \cdot \int_H f(hg) d\mu_H(h) d\mu(Hg) \\ &= \Delta_H(h') \cdot \int_G f(g) d\mu_G(g), \end{aligned}$$

which yields $\Delta_G|_H = \Delta_H$. Conversely, assume $\Delta_G|_H = \Delta_H$. Denote by I_G the integral with respect to μ_G and by $T_H : C_c(G) \rightarrow C_c(H \backslash G)$ the functional

$$(T_H f)(Hg) := \int_H f(hg) d\mu_H(h).$$

We then want to establish the existence of a functional I making the diagram

$$\begin{array}{ccc} C_c(G) & \xrightarrow{I_G} & \mathbb{R} \\ T_H \downarrow & \nearrow I & \\ C_c(H \backslash G) & & \end{array}$$

commute, i.e. we want to show that $\ker(T_H) \subset \ker(I_G)$. Thus let $f \in \ker T_H$, i.e. assume

$$0 = \int_H f(hg) d\mu_H(h) = 0$$

for all $g \in G$. Now let $k \in C_c(G)$ be an auxiliary function with the property that $T_H(k) \equiv 1$ on the image of $\text{supp} f$ in $H \backslash G$. Then

$$\int_H k(hg) d\mu_H(h) \cdot f(g) = f(g)$$

and thus Fubini's theorem yields

$$\begin{aligned}
0 &= \int_G \int_H f(hg) d\mu_H(h) k(g) d\mu_G(g) \\
&= \int_H \int_G f(hg) k(g) d\mu_H(h) d\mu_G(g) \\
&= \int_H \int_G f(g) k(h^{-1}g) \Delta_G(h^{-1}) d\mu_G(g) d\mu_H(h) \\
&= \int_H \int_G f(g) k(h^{-1}g) \Delta_H(h^{-1}) d\mu_G(g) d\mu_H(h) \\
&= \int_G f(g) \int_H k(h^{-1}g) \Delta_H(h^{-1}) d\mu_H(h) d\mu_G(g) \\
&= \int_G f(g) \int_H k(hg) d\mu_H(h) d\mu_G(g) \\
&= \int_G f(g) d\mu_G(g).
\end{aligned}$$

□

Recall that a *lattice* in G is a discrete subgroup, such that $\Gamma \backslash G$ admits a finite G -invariant measure. Then the lemma implies:

Corollary 1.3.3. *Let G be a Lie group. If there exists a lattice Γ in G , then G is unimodular.*

PROOF. By Lemma 1.3.2, the subgroup Γ is contained in $\ker \Delta_G$, whence there is a surjection $\Gamma \backslash G \rightarrow \ker \Delta_G \backslash G =: H$. Pushing forward the finite G -invariant measure along this surjection, we obtain a finite Haar measure on H . Thus H is compact; however, $H \cong \text{Im}(\Delta_G)$, which is a subgroup of $\mathbb{R}^{>0}$. Consequently, H must be trivial, and thus $\Delta_G \equiv 1$. □

Thus in order to give meaning to the various objects occurring in Theorem 1.2.8 we have to make the following assumptions on G , H and Γ :

- (1) G is a connected unimodular Lie group.
- (2) H is a closed unimodular subgroup of G .
- (3) Γ is a lattice in G .
- (4) $H \cap \Gamma$ is a lattice in H .

For ease of reference, let us refer to a triple (G, H, Γ) satisfying (1)-(4) above an *admissible triple*. Given such a triple we choose Haar measures μ_G on G and μ_H on H and a G -invariant measure μ on $H \backslash G$ such that (1.5) holds. The same formula then determines measures m_G on $\Gamma \backslash G$ and m_H on $\Gamma \backslash H$ by taking the counting measure on Γ .

Given an admissible triple (G, H, Γ) we want to count the asymptotic growth of the orbit $v\Gamma$, where $v = He$ is the canonical base point of $H \backslash G$. This makes sense only if the orbit $v\Gamma$ is discrete in $H \backslash G$. Somewhat surprisingly, this property comes for free:

Proposition 1.3.4. *If (G, H, Γ) is an admissible triple, then the inclusion*

$$\iota: (H \cap \Gamma) \backslash H \rightarrow \Gamma \backslash G, \quad [h] \rightarrow [\Gamma h]$$

is a homeomorphism onto its image. In particular, if $v = He$, then $v\Gamma$ is discrete in $H\backslash G$.

PROOF. For the purpose of the proof we fix a right-invariant metric on G . We then denote by B_r^G and B_r^H respectively the open balls of radius r with respect to this metric around the identity in G , respectively, H . To see that ι is a homeomorphism onto its image, we have to show that it preserves convergence of sequences. Thus let (x_n) be a sequence in $(H \cap \Gamma)\backslash H$ and assume that $\iota(x_n)$ converges to $y \in \Gamma\backslash G$. We then have to show that x_n converges to some $x \in (H \cap \Gamma)\backslash H$. Let us abbreviate $\Lambda := H \cap \Gamma$, so that $x_n = \Lambda h_n$ for some $h_n \in H$. Also choose $g \in G$ with $y = \Gamma g$. Then

$$\Gamma h_n \rightarrow \Gamma g.$$

We have to show that $g \in \Gamma H$. Assume otherwise; then $\Gamma h_n \rightarrow \Gamma g$ in a direction transverse to the H -orbits in $\Gamma\backslash G$. Thus we have $\Gamma h_n \notin \Gamma g B_1^H$ and may assume $\Gamma h_n \notin \Gamma h_m \overline{B_1^H}$ for $n \neq m$ (by passing to a subsequence if necessary). Recall that $r > 0$ is called an *injectivity radius* at $y \in \Gamma\backslash G$ if the natural projection $G \rightarrow \Gamma\backslash G$ restricts to an isometry on $\pi^{-1}(y)B_r^G$, where $\pi : G \rightarrow \Gamma\backslash G$ is the projection map. Since Γ is discrete, every $y \in \Gamma\backslash G$ admits a positive injectivity radius. We now choose an injectivity radius $r < 1$ for $y = \Gamma g$. Since $\Gamma h_n \rightarrow \Gamma g$ we then find n_0 with $h_n \in g B_{r/2}^H$ for all $n \geq n_0$. By definition of the injectivity radius it follows that the sets $\Gamma h_n B_{r/2}^H = \Gamma B_{r/2}^H h_n$ are all disjoint, hence so are the sets $\Lambda B_{r/2}^H h_n$. Now H is unimodular and $m_H(\Lambda B_{r/2}^H) > 0$, which yields to the contradiction

$$m_H(\Lambda\backslash H) \geq m_H\left(\bigcup_{n=n_0}^{\infty} \Lambda B_{r/2}^H h_n\right) = \sum_{n=n_0}^{\infty} m_H(\Lambda B_{r/2}^H h_n) = \sum_{n=n_0}^{\infty} m_H(\Lambda B_{r/2}^H) = \infty.$$

□

The present proof is based on [Ein06, Lemma 4.1]; for an entirely different proof of the proposition see [Rag72, Thm. 1.13].

As a last step in the formulation of the abstract counting problem, we need to choose a sequence $B_n \subset H\backslash G$ which replaces the balls $B(0, T)$ used in the counting problem for binary quadratic forms. Clearly, we want the sets B_n to exhaust all of $H\backslash G$, so we should clearly demand $\mu(B_n) \rightarrow \infty$. On the other hand, this assumption is not quite sufficient to make any non-trivial statement, since one easily constructs a family of sets B_n with $\mu(B_n) \rightarrow \infty$, which avoids Γ altogether. The following condition makes sure that this kind of degenerate behaviour does not happen:

Definition 1.3.5. A sequence B_n of subsets of $H\backslash G$ is called *well-rounded* if for every $\epsilon > 0$ there exists an identity neighbourhood U in G such that for all $n \in \mathbb{N}$,

$$\frac{\mu(\partial B_n \cdot U)}{\mu(B_n)} < \epsilon$$

Now we can finally formulate our abstract counting problem::

Problem 1.3.6. Let (G, H, Γ) be an admissible triple and B_n be a well-rounded sequence of subsets of $H\backslash G$ with $\mu(B_n) \rightarrow \infty$. Given $v = He \in H\backslash G$, estimate the asymptotic behaviour of

$$N(v, n) := |v\Gamma \cap B_n|.$$

Following our heuristic we now define:

Definition 1.3.7. An admissible triple (G, H, Γ) has the *good counting property* if for any well-rounded sequence B_n of subsets in $H \backslash G$ and $v = He$ we have

$$N(v, n) \sim \frac{m_H((H \cap \Gamma) \backslash H)}{m_G(\Gamma \backslash G)} \cdot \mu(B_n).$$

With this language, Theorem 1.2.8 can be reformulated by saying that if H denotes the stabilizer of an integral binary quadratic form of fundamental discriminant in $SL_2(\mathbb{R})$, then the triple $(SL_2(\mathbb{R}), H, SL_2(\mathbb{Z}))$ has the good counting property. Since the whole problem is conjugation-invariant we can replace H by $SO(2)$ or $SO(1, 1)$, depending on the sign of the discriminant in question. Then we have to replace $SL_2(\mathbb{Z})$ by a conjugate Γ . Actually, we do not need to know Γ precisely; all we need is that

- Γ is still a lattice in G and intersects $SO(1, 1)$ in a lattice;
- $\Gamma \backslash G$ is non-compact. (We say that Γ is a *non-uniform lattice* in G .)

In this generality we shall prove.

Theorem 1.3.8. (i) *If Γ is a non-uniform lattice in $SL_2(\mathbb{R})$, then $(SL_2(\mathbb{R}), SO(2), \Gamma)$ has the good counting property.*
(ii) *If Γ is a non-uniform lattice in $SL_2(\mathbb{R})$, which intersects $SO(1, 1)$ in a lattice, then $(SL_2(\mathbb{R}), SO(1, 1), \Gamma)$ has the good counting property.*

Clearly Theorem 1.3.8 implies Theorem 1.2.8. In the next section we will provide a dynamical condition which implies the good counting property. We will then prove Theorem 1.2.8 by establishing this dynamical condition for the triples in question.

1.4. Equidistribution implies the good counting property

Let (G, H, Γ) be an admissible triple. Then the measures m_H on $(H \cap \Gamma) \backslash H$ and m_G on $\Gamma \backslash G$ are finite, and we denote by \widehat{m}_H and \widehat{m}_G the corresponding probability measures given by

$$\widehat{m}_H := \frac{1}{m_H((H \cap \Gamma) \backslash H)} \cdot m_H, \quad \widehat{m}_G := \frac{1}{m_G(\Gamma \backslash G)} \cdot m_G.$$

Moreover, we denote by $\iota : (H \cap \Gamma) \backslash H \rightarrow \Gamma \backslash G$ the natural inclusion. By abuse of notation we will denote the push-forward $\iota_* \widehat{m}_H$ by the same letter \widehat{m}_H^* . Finally, denote by ρ_g the right translation by g on $\Gamma \backslash G$. Then for $g \in G$ the measure $(\rho_g)_* \widehat{m}_H^*$ is supported on $Y_g := (\iota((H \cap \Gamma) \backslash H))g$. Now we define:

Definition 1.4.1. We say that the triple (G, H, Γ) has the *equidistribution property* or that the translates Y_g of $\iota((H \cap \Gamma) \backslash H)$ *equidistribute* inside $\Gamma \backslash G$ if

$$(\rho_{g_n})_* \widehat{m}_H^* \xrightarrow{w^*} \widehat{m}_G$$

as $\Gamma g_n \rightarrow \infty$ in $\Gamma \backslash G$.

Here and in the sequel, given a locally compact space X and a sequence $x_n \in X$ we use the notation $x_n \rightarrow \infty$ to indicate that x_n leaves every compact subset of X . Note that the equidistribution property can only be defined if the lattice Γ is non-uniform. This is the reason why we can prove Theorem 1.3.8 only for non-uniform lattices.

It was observed by Eskin and McMullen in [EM93] that equidistribution can be used in order to establish the good counting property:

Theorem 1.4.2. *Let (G, H, Γ) be an admissible triple. If (G, H, Γ) has the equidistribution property, then it has the good counting property.*

The proof is based on the following lemma:

Lemma 1.4.3. *Let (G, H, Γ) be an admissible triple satisfying the equidistribution property and $A_n \subset \Gamma \backslash G$ be a (not necessarily well-rounded) sequence of bounded sets with $\mu(A_n) \rightarrow \infty$. Then we have weak- $*$ -convergence*

$$(1.6) \quad \frac{|A_n \cap v\Gamma g|}{\mu(A_n)} d\widehat{m}_G(g) \rightarrow \frac{m_H((H \cap \Gamma) \backslash H)}{m_G(\Gamma \backslash G)} d\widehat{m}_G(g).$$

PROOF. Given any bounded subset $A \subset H \backslash G$ we define a function $F_A : \Gamma \backslash G \rightarrow \mathbb{Z}$ by $F_A(g) := |A \cap v\Gamma g|$. We have to show that for every $\alpha \in C_c(\Gamma \backslash G)$,

$$(1.7) \quad \int_{\Gamma \backslash G} \alpha(g) \frac{F_{A_n}(g)}{\mu(A_n)} d\widehat{m}_G(g) \rightarrow \frac{m_H((H \cap \Gamma) \backslash H)}{m_G(\Gamma \backslash G)} \int_{\Gamma \backslash G} \alpha(g) d\widehat{m}_G(g).$$

The idea of the proof is to bring into play the double fibration

$$\begin{array}{ccc} & (\Gamma \cap H) \backslash G & \\ & \swarrow \quad \searrow & \\ \Gamma \backslash G & & H \backslash G \end{array}$$

to pass from integrals over $\Gamma \backslash G$ to integrals over $H \backslash G$. This is a two-step process given by combining an *unfolding* from functions on $\Gamma \backslash G$ to functions on $(\Gamma \cap H) \backslash G$ with a *fiber integration* which allows one to pass down to $H \backslash G$. (This kind of two-step passage is well-known in the context of the Rankin-Selberg method.) Note that the measures involved are related according to the Weil relation (1.5). Besides the measures $\widehat{m}_G, \widehat{m}_H, m_G, m_H$ introduced above we will also use a measure m_0 on $(\Gamma \cap H) \backslash G$, which we normalize such that (1.5) holds with respect to μ_G and the counting measure on $\Gamma \cap H$. Now we start our argument by rewriting the function F_{A_n} as

$$F_{A_n}(g) = |A_n \cap v\Gamma g| = \sum_{\gamma \in (\Gamma \cap H) \backslash \Gamma} \chi_{A_n}(v\gamma g).$$

Let us introduce the notation $f_{A_n} d\mu := \frac{1}{\mu(A_n)} \int_{A_n} d\mu$ for the average over A_n . Then for any $\alpha \in C_c(\Gamma \backslash G)$ we compute

$$\begin{aligned}
& \int_{\Gamma \backslash G} \alpha(g) \frac{F_{A_n}(g)}{\mu(A_n)} d\widehat{m}_G(g) \\
&= \frac{1}{\mu(A_n) \cdot m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} \alpha(g) \left(\sum_{\gamma \in (\Gamma \cap H) \backslash \Gamma} \chi_{A_n}(v\gamma g) \right) dm_G(g) \\
&= \frac{1}{\mu(A_n) \cdot m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} \left(\sum_{\gamma \in (\Gamma \cap H) \backslash \Gamma} \alpha(\gamma g) \chi_{A_n}(v\gamma g) \right) dm_G(g) \\
&= \frac{1}{\mu(A_n) \cdot m_G(\Gamma \backslash G)} \cdot \int_{(\Gamma \cap H) \backslash G} \alpha(g) \chi_{A_n}(vg) dm_0(g) \\
&= \frac{1}{\mu(A_n) \cdot m_G(\Gamma \backslash G)} \cdot \int_{H \backslash G} \chi_{A_n}(vg) \int_{(\Gamma \cap H) \backslash H} \alpha(\Gamma hg) dm_H(h) d\mu(g) \\
&= \frac{m_H((\Gamma \cap H) \backslash H)}{m_G(\Gamma \backslash G)} \cdot \int_{A_n} \int_{(\Gamma \cap H) \backslash H} \alpha(\Gamma hg) d\widehat{m}_H(h) d\mu(g).
\end{aligned}$$

As far as the inner integral is concerned, the equidistribution property yields

$$\int_{(\Gamma \cap H) \backslash H} \alpha(\Gamma hg) d\widehat{m}_H(h) \rightarrow \int_{\Gamma \backslash G} \alpha(x) d\widehat{m}_G(x).$$

If n is large then for most $g \in A_n$ the left integral is ϵ -close to the right one. One deduces that also

$$\int_{A_n} \int_{(\Gamma \cap H) \backslash H} \alpha(\Gamma hg) d\widehat{m}_H(h) d\mu(g) \rightarrow \int_{\Gamma \backslash G} \alpha(x) d\widehat{m}_G(x),$$

which establishes (1.7) and hence (1.6). \square

Now we can prove the theorem:

PROOF OF THEOREM 1.4.2. Let B_n be a well-rounded sequence in $\Gamma \backslash G$ with $\mu(B_n) \rightarrow \infty$. We will use the following reformulation of well-roundedness: If we fix $\epsilon > 0$, then we can find an identity neighbourhood $U = U^{-1}$ in G such that

$$(1.8) \quad (1 - \epsilon) \cdot \mu \left(\bigcup_{u \in U} B_n u \right) \leq \mu(B_n) \leq (1 + \epsilon) \cdot \mu \left(\bigcap_{u \in U} B_n u \right).$$

Let us abbreviate

$$B'_n := \bigcup_{u \in U} B_n u, \quad B''_n := \bigcap_{u \in U} B_n u.$$

These sets may no longer be well-rounded, but they still satisfy $\mu(B'_n) \rightarrow \infty$, $\mu(B''_n) \rightarrow \infty$. Now, for all $u \in U$ we have

$$F_{B'_n}(u) = |B'_n \cap v\Gamma u| = |B'_n u^{-1} \cap v\Gamma| \geq |B_n \cap v\Gamma| = F_{B_n}(e),$$

and

$$F_{B''_n}(u) = |B''_n \cap v\Gamma u| = |B''_n u^{-1} \cap v\Gamma| \leq |B_n \cap v\Gamma| = F_{B_n}(e).$$

Thus

$$(1.9) \quad F_{B''_n}(u) \leq F_{B_n}(e) \leq F_{B'_n}(u) \quad (u \in U).$$

Now choose $\alpha \in C_c(\Gamma \backslash G)$ with $\alpha \geq 0$, $\text{supp}(\alpha) \subset \Gamma \cdot U$ and $\int_{\Gamma \backslash G} \alpha d\widehat{m}_G = 1$. Then applying (1.8) and (1.9) we obtain

$$\int_{\Gamma \backslash G} \alpha(g) \cdot \frac{F_{B'_n}(g)}{\mu(B'_n)} d\widehat{m}_G(g) \geq \inf_{u \in U} \frac{F_{B'_n}(u)}{\mu(B'_n)} \geq \frac{F_{B_n}(e)}{(1-\epsilon)^{-1} \cdot \mu(B_n)}$$

and

$$\int_{\Gamma \backslash G} \alpha(g) \cdot \frac{F_{B''_n}(g)}{\mu(B''_n)} d\widehat{m}_G(g) \leq \sup_{u \in U} \frac{F_{B''_n}(u)}{\mu(B''_n)} \leq \frac{F_{B_n}(e)}{(1+\epsilon)^{-1} \cdot \mu(B_n)}.$$

Now applying Lemma 1.4.3 to the sequences B'_n and B''_n , we see that both integrals converge to $\frac{m_H((\Gamma \cap H) \backslash H)}{m_G(\Gamma \backslash G)}$. We thus get

$$(1-\epsilon) \cdot \overline{\lim} \frac{|B_n \cap v\Gamma|}{\mu(B_n)} \leq \frac{m_H((\Gamma \cap H) \backslash H)}{m_G(\Gamma \backslash G)} \leq (1+\epsilon) \cdot \underline{\lim} \frac{|B_n \cap v\Gamma|}{\mu(B_n)},$$

which for $\epsilon \rightarrow 0$ finally yields the good counting property. \square

Remark 1.4.4. We see from the proof of Theorem 1.4.2 that we do not need the full strength of our assumption. Rather, it suffices to assume the weaker property (1.6).

1.5. The group $SL_2(\mathbb{R})$ and the hyperbolic plane

Our next goal is to establish the equidistribution property (and hence the good counting property) for admissible triples of the form $(SL_2(\mathbb{R}), SO(2), \Gamma)$. To get a geometric understanding of these triples, we consider the isometric action of $G := SL_2(\mathbb{R})$ on the Poincaré upper halfplane

$$\mathbb{H}^2 := \{z \in \mathbb{C} \mid \Im(z) > 0\},$$

which we always think of as equipped with the hyperbolic metric. If x and y denote the standard (real and imaginary) coordinates on \mathbb{H}^2 then the associated volume form is given by $\frac{dx \wedge dy}{y^2}$. We assume that the reader is familiar with elementary hyperbolic geometry on the level of [BP92, Chapter A]. Nevertheless, let us recall some basic facts in order to fix our notations: First of all, the action of G on \mathbb{H}^2 (in fact, on all of \mathbb{CP}^1) is given by fractional linear transformations, i.e.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z := \frac{az + b}{cz + d}.$$

It follows from this formula that the kernel of this action is given by $\{\pm 1\}$ and the stabilizer of $i \in \mathbb{H}^2$ is precisely the rotation group $K := SO(2)$, whose Lie algebra \mathfrak{k} consists of the antisymmetric matrices in the Lie algebra \mathfrak{g} of G . If we denote by \mathfrak{p} the space of symmetric matrices in \mathfrak{g} (which is not a Lie algebra!), then $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ as vector spaces (in fact, as $\text{Ad}(K)$ -modules). Now, $\exp(\mathfrak{p})$ consists of positive definite symmetric matrices in G , and since every invertible matrix can be written as a product of a rotation and a positive definite symmetric matrix we have a diffeomorphism

$$\mathfrak{p} \times K \rightarrow G, \quad (X, k) \mapsto \exp(X) \cdot k,$$

called the *polar decomposition*. Since G acts transitively on \mathbb{H}^2 , this implies

$$\mathbb{H}^2 \cong G/K \cong \exp(\mathfrak{p}).$$

Now every symmetric matrix can be diagonalized using a rotation; thus if A denotes the subgroup of diagonal matrices in G , then $\exp(\mathfrak{p}) = \bigcup_{k \in K} kAk^{-1}$ and thus

$$(1.10) \quad G = KAK, \quad \mathbb{H}^2 \cong KA.$$

Let us interpret the latter decomposition geometrically: Let $a_t = \text{diag}(e^{t/2}, e^{-t/2})$ so that $A = \{a_t \mid t \in \mathbb{R}\}$. Then $a_t.i = e^t i$, i.e. the A -orbit through i is precisely the geodesic $(0, \infty)$. Then the decomposition $\mathbb{H}^2 \cong KA$ means that we can reach every point in \mathbb{H}^2 by first moving along this geodesic and then turning along a circle with hyperbolic midpoint i . Since each circle (of positive radius) intersects the geodesic $(0, \infty)$ twice, the map $K \times A \rightarrow \mathbb{H}^2$ is $2 : 1$ away from i . To obtain proper coordinates we define $A^\pm := \{a_t \mid \pm t > 0\}$. Then

$$(1.11) \quad G = K\overline{A^+}K, \quad \mathbb{H}^2 \cong K\overline{A^+}$$

and the map $K \times A^+ \rightarrow \mathbb{H}^2 \setminus \{i\}$ is a bijection. Its inverse map is called *standard hyperbolic polar coordinates* centered at i . Of course, we can replace A by a K -conjugate in this decomposition. Geometrically, this corresponds to replacing $(0, \infty)$ by an arbitrary geodesic through i . We can also move the whole situation around in the hyperbolic plane by left-multiplying an element of G to change the basepoint i . We then obtain general *hyperbolic polar coordinates* which consist of a first parameter, describing a movement along a geodesic, and a second parameter describing rotation around a point on this geodesic. If we move this point out to infinity, then we obtain *horospherical coordinates*. For simplicity, consider again the geodesic $(0, \infty)$ and the corresponding group A . The stabilizers of 0 and ∞ are respectively given by the groups

$$N^- := \left\{ \begin{pmatrix} 1 & 0 \\ x & 1 \end{pmatrix} \mid x \in \mathbb{R} \right\}, \quad N^+ := \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \mid x \in \mathbb{R} \right\},$$

thus the orbits $N^\pm.x$ are horospheres through x based at 0 and ∞ respectively. Since every point in \mathbb{H}^2 can be described by first moving along $(0, \infty)$ and then rotating along such a horosphere we obtain decompositions

$$(1.12) \quad G = N^\pm AK, \quad \mathbb{H}^2 \cong N^\pm A.$$

Here the maps $N^\pm \times A \rightarrow \mathbb{H}^2$ are honest bijections, and their inverse maps are called (positive and negative) *standard horospherical coordinates*. Again we can replace A by its conjugate and N^\pm by the stabilizers of the points of infinity of the corresponding geodesic to obtain general horospherical coordinates.

Our geometric interpretation of the groups A, N^\pm has an important consequence: Suppose a_n is a sequence in A leaving any compact subset; let us assume $a_n \in A^+$ for all $n \gg 0$. Then the horosphere $N^+ a_n.i$ at $a_n.i$ has smaller and smaller radius as n growth. This shows that

$$(1.13) \quad a_n^{-1} N^+ a_n \rightarrow \{e\},$$

i.e. N^+ is contracted by the conjugation action of A^+ . Similarly, the horosphere N^- is contracted by the conjugation action of A^- . More generally, if we go out to infinity along a geodesic ray, then the horospheres based at the endpoint of this ray get contracted (and those of the opposite endpoint get expanded).

1.6. Ergodicity, mixing and the Howe-Moore property

In order to establish the equidistribution property for the triple $(SL_2(\mathbb{R}), SO(2), \Gamma)$ we need some more tools from ergodic theory. In view of later applications we prefer to introduce these tools in a general setup:

Let (X, \mathcal{B}, μ) be a probability space and G a locally compact group. An action of G on X is called *measure-preserving* if the action map $G \times X \rightarrow X$ is measurable and $g_*\mu = \mu$ for all $g \in G$. A measurable subset $A \in \mathcal{B}$ is called *G -invariant up to nullsets* if for all $g \in G$ we have $\mu(g^{-1}A \Delta A) = 0$. (Here $A \Delta B := A \setminus B \cup B \setminus A$ denotes the symmetric difference of two sets A, B .) Equivalently, $\mu(g^{-1}A \cap A) = \mu(A)$ for all $g \in G$. Now we define:

Definition 1.6.1. A measure-preserving action of a locally compact group G on a probability space (X, \mathcal{B}, μ) is called

- *ergodic*, if $\mu(A) \in \{0, 1\}$ for every $A \in \mathcal{B}$, which is G -invariant up to nullsets.
- *mixing*, if G is non-compact and for all $A, B \in \mathcal{B}$ and all sequences $g_n \in G$ with $g_n \rightarrow \infty$ we have $\mu(g_n^{-1}A \cap B) \rightarrow \mu(A)\mu(B)$.

The probabilistic interpretation of mixing is as follows: Recall that two events modelled by sets $A, B \in \mathcal{B}$ with $\mu(A \cap B) = \mu(A)\mu(B)$ are called *independent* in the probabilistic language. Since $\mu(g_n B) = \mu(B)$ for all $g_n \in G$, mixing implies that the events A and $g_n B$ become more and more independent as g_n growth, even if A and B were very much dependent at the beginning. The most obvious relation between mixing and ergodicity is provided by the following observation:

Lemma 1.6.2. *Every mixing action is ergodic.*

PROOF. With notation as above, let G act mixing on (X, \mathcal{B}, μ) and assume A is G -invariant up to nullsets. Pick a sequence $g_n \in G$ with $g_n \rightarrow \infty$; then mixing implies

$$\mu(A) = \mu(g_n^{-1}A \cap A) \rightarrow \mu(A)^2,$$

i.e. $\mu(A) = \mu(A)^2$, whence $\mu(A) \in \{0, 1\}$. □

The precise relation between ergodicity and mixing is rather subtle. For example, if the G -action on (X, \mathcal{B}, μ) is mixing, then even the G -action on $(X \times X, \mathcal{B} \otimes \mathcal{B}, \mu \times \mu)$ is ergodic. For $G = \mathbb{Z}$ this property (called *weak mixing*) is equivalent to mixing, but for general groups G this need not be true. To see that ergodicity itself does not imply mixing, we consider the following example:

Example 1.6.3. Let $G = \mathbb{Z}$, $(X, \mathcal{B}, \mu) = (S^1, \mathcal{B}, \lambda)$, where \mathcal{B} is the Borel algebra of S^1 and λ is the push-forward of the Lebesgue measure via $\pi : \mathbb{R} \rightarrow S^1 = \mathbb{R}/\mathbb{Z}$. For every $\alpha \in \mathbb{R}$ there is a measure-preserving G -action on S^2 via $n.[x] := [x + n\alpha]$. This action is ergodic iff $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. We claim that it is never mixing. Indeed, let A, B be two disjoint small segments, say $A = \pi([0, \frac{1}{10}])$, $B = \pi([\frac{5}{10}, \frac{6}{10}])$. Then for ever $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ we find an unbounded sequence n_k with $(\pi^{-1}(A) + n_k\alpha) \cap \pi^{-1}(B) = \emptyset$, so the action cannot be mixing.

We are now looking for a property which allows us to deduce mixing (even for subgroups) from ergodicity. In order to formulate this property, it is convenient to pass from statements about measure preserving actions to statements about

certain representations. Recall that a *unitary representation* (\mathcal{H}, π) of G consists of a Hilbert space \mathcal{H} and a homomorphism $\pi : G \rightarrow \mathcal{U}(\mathcal{H})$, where $\mathcal{U}(\mathcal{H})$ denotes the group of unitary operators on \mathcal{H} ; such a representation is called *strongly continuous*, if the map $G \times \mathcal{H} \rightarrow \mathcal{H}$ is continuous (with respect to the product topology on the left hand side). Every measure-preserving action of G on (X, \mathcal{B}, μ) induces a strongly continuous unitary representation of G by setting $\mathcal{H} := L^2(X, \mu)$ and defining $(\pi(g)f)(x) = f(g^{-1}x)$. (In the case of a right-action one defines $(\pi(g)f)(x) = f(xg)$ instead.) We can now translate ergodicity and mixing into this language:

Lemma 1.6.4. *Let G act measure-preservingly on (X, \mathcal{B}, μ) and let $(L^2(X, \mu), \pi)$ be the associated strongly continuous unitary representation. Then:*

- (i) *The G -action is ergodic iff the only G -invariant vectors in $(L^2(X, \mu), \pi)$ are the constants.*
- (ii) *The G -action is mixing iff for all $f, h \in L^2(X, \mu)$ and for every sequence $g_n \in G$ with $g_n \rightarrow \infty$ we have*

$$\int \pi(g_n)(f) \cdot h d\mu \rightarrow \int f d\mu \int h d\mu.$$

PROOF. The implications \Leftarrow follow by choosing the functions in question to be characteristic functions. The implications \Rightarrow then follow by approximating general L^2 -functions by characteristic ones. \square

Given a strongly continuous unitary representation (\mathcal{H}, π) and vectors $\xi, \eta \in \mathcal{H}$ we can define a continuous function $m_{\xi, \eta}$ on G by the formula

$$m_{\xi, \eta}(g) := \langle g \cdot \xi, \eta \rangle.$$

In view of their interpretation in the finite-dimensional case, these functions are called the *matrix coefficients* of the representation (\mathcal{H}, π) , and it is easy to see that they determine (\mathcal{H}, π) uniquely. We use them to define:

Definition 1.6.5. A locally compact group G has the *Howe-Moore property* if $m(g_n) \rightarrow 0$ as $g_n \rightarrow \infty$ for all matrix coefficients of strongly continuous unitary representation (\mathcal{H}, π) without non-zero fixed vectors.

Clearly, if (\mathcal{H}, π) is a strongly continuous unitary representation and $\xi \in \mathcal{H}$ is a non-zero fixed vector, then $m_{\xi, \xi} \equiv \|\xi\|^2$ does not vanish at infinity, so the Howe-Moore property asks for vanishing of all possible matrix coefficients at infinity. This is a very strong property, and it allows us to deduce mixing from ergodicity:

Proposition 1.6.6. *Let G be a locally compact group with the Howe-Moore property. Let (X, \mathcal{B}, μ) be an ergodic probability G -space. Then the action of any non-compact subgroup $H < G$ on X is mixing. In particular, every ergodic G -action is mixing.*

PROOF. Denote by $L_0^2(X, \mu)$ the orthogonal complement of $\mathbb{C} \cdot 1_X$ in $L^2(X, \mu)$. Since G is ergodic, there is no non-zero invariant vector in $L_0^2(X)$; then the Howe-Moore property implies that all matrix coefficients of $L_0^2(X)$ vanish. In particular, if $f, h \in L^2(X, \mu)$ and $\pi : L^2(X, \mu) \rightarrow L_0^2(X, \mu)$ denotes the orthogonal projection then for all $h_n \in H$ with $h_n \rightarrow \infty$ we have

$$\begin{aligned} 0 &= \lim_{n \rightarrow \infty} m_{\pi(f), \pi(h)}(h_n) = \lim_{n \rightarrow \infty} \int_X (\pi(h_n)(f - \int_X f d\mu)) \cdot (h - \int_X h d\mu) d\mu \\ &= \lim_{n \rightarrow \infty} \int \pi(h_n)(f) \cdot h d\mu - \int f d\mu \int h d\mu, \end{aligned}$$

which is the desired mixing statement. \square

Corollary 1.6.7. *Let G be a locally compact group, $\Gamma < G$ a lattice and assume that G has the Howe-Moore property. Then for every non-compact closed subgroup $H < G$ the H -action on $\Gamma \backslash G$ is mixing, in particular ergodic.*

In view of Example 1.6.3, this shows that \mathbb{Z} does not have the Howe-Moore property. Actually, discrete groups hardly ever have the Howe-Moore property:

Lemma 1.6.8. *Let G be a locally-compact group with the Howe-Moore property and H an open subgroup of infinite index. Then H is compact. In particular, if a discrete subgroup has an infinite subgroup of infinite index, then it does not have the Howe-Moore property.*

PROOF. We only provide a brief sketch: Consider the space $\mathcal{H}_0 := L^2(H \backslash G, \mu)$, where μ is a suitable G -quasiinvariant measure on $H \backslash G$. Then one can make G act in a strongly continuous unitary way on \mathcal{H}_0 . Denote by \mathcal{H} the orthogonal complement of the constant functions in \mathcal{H} . Then the action of G on \mathcal{H} does not have fixed vectors. Now assume H is non-compact let $h_n \in H$ with $h_n \rightarrow \infty$ and denote by ξ the characteristic function on H . Since H is open we have $\xi \neq 0$; on the other hand, the matrix coefficient $m_{\xi, \xi}$ is constant along h_n and does not vanish at infinity. This contradicts the Howe-Moore property. \square

Even for non-discrete groups, the Howe-Moore property is rather strong. For example, nilpotent Lie groups never have it. It is thus remarkable, that all simple real Lie groups have the Howe-Moore property. In the next section we provide a proof in the $SL_2(\mathbb{R})$ case. The general case will be treated in ?? below.

1.7. The Howe-Moore theorem for $SL_2(\mathbb{R})$

The purpose of this section is to provide a proof of the following important result:

Theorem 1.7.1. *The group $G = SL_2(\mathbb{R})$ has the Howe-Moore property.*

For the proof we will use the notation introduced in Section 1.5 regarding special subgroups of $SL_2(\mathbb{R})$. We start by an easy reduction: Let $g_n \in G$ with $g_n \rightarrow \infty$. According to (1.11) we then find sequence $k_n, h_n \in K$ and $a_n \in \overline{A}^+$ with $g_n = k_n a_n h_n$. Since K is compact we then have $g_n \rightarrow \infty$ iff $a_n \rightarrow \infty$. We claim that moreover, for any strongly continuous unitary representation (\mathcal{H}, π) we have

$$(1.14) \quad \forall \xi, \eta \in \mathcal{H} : m_{\xi, \eta}(g_n) \rightarrow 0 \quad \Leftrightarrow \quad \forall \xi', \eta' \in \mathcal{H} : m_{\xi', \eta'}(a_n) \rightarrow 0.$$

Indeed, the implication \Rightarrow is trivial; for the converse let $\xi, \eta \in \mathcal{H}$. Choose a sequence n_k such that $h_{n_k} \xi \rightarrow \xi'$ and $k_{n_k}^{-1} \eta \rightarrow \eta'$ converge. Then

$$\lim_{k \rightarrow \infty} m_{\xi, \eta}(g_{n_k}) = \lim_{k \rightarrow \infty} \langle a_{n_k} h_{n_k} \xi, k_{n_k} \eta \rangle = \lim_{k \rightarrow \infty} \langle a_{n_k} \xi', \eta' \rangle = \lim_{k \rightarrow \infty} m_{\xi', \eta'}(a_{n_k}) = 0.$$

By varying the sequence n_k we then obtain (1.14). It therefore suffices to establish the Howe-Moore property for $g_n = a_n \in \overline{A}^+$. Now we come to the key ingredient in the proof of Theorem 1.7.1:

Lemma 1.7.2 (Mautner). *Let G be a locally compact group, (\mathcal{H}, π) a strongly continuous unitary representation of G , $(a_n) \subset G$ and $\xi \in \mathcal{H}$. Let ξ_0 be an accumulation point of $(\pi(a_n)\xi)$ with respect to the weak topology. Then*

$$U(a_n) := \{g \in G \mid a_n^{-1} g a_n \rightarrow e\}$$

fixes ξ_0 .

PROOF. By passing to a subsequence we may assume $\pi(a_n)\xi \rightarrow \xi_0$ weakly. Then for all $g \in U(a_n)$ and $\eta \in \mathcal{H}$ we have

$$\begin{aligned} |\langle \pi(g)\xi_0 - \xi_0, \eta \rangle| &= \lim_{n \rightarrow \infty} |\langle \pi(g)\pi(a_n)\xi - \pi(a_n)\xi, \eta \rangle| \\ &= \lim_{n \rightarrow \infty} |\langle \pi(a_n)^{-1}\pi(g)\pi(a_n)\xi - \xi, \pi(a_n^{-1})\eta \rangle| \\ &\leq \lim_{n \rightarrow \infty} \|\pi(a_n^{-1}ga_n)\xi - \xi\| \cdot \|\pi(a_n^{-1})\eta\| \\ &= \|\eta\| \cdot \lim_{n \rightarrow \infty} \|\pi(a_n^{-1}ga_n)\xi - \xi\| = 0. \end{aligned}$$

Since $\eta \in \mathcal{H}$ was arbitrary we deduce that $\pi(g)\xi_0 - \xi_0 = 0$ as claimed. \square

The other main ingredient of the proof is the following lemma:

Lemma 1.7.3 (Gelfand trick). *Let $G = SL_2(\mathbb{R})$ and let (\mathcal{H}, π) be a strongly continuous unitary representation. If $\xi_0 \in \mathcal{H}$ is fixed by N^+ , then it is also fixed by A .*

PROOF. If $\xi_0 = 0$, then nothing is to be proved. We may thus assume $\xi_0 \neq 0$. We consider the continuous function $\phi : G \rightarrow \mathbb{R}$ given by $\phi(g) := \langle \pi(g)\xi_0, \xi_0 \rangle$. We then have

$$\phi(g) = \|\xi_0\|^2 \iff \pi(g)\xi_0 = \xi_0 \quad (g \in G)$$

by the equality case of Cauchy-Schwartz. By assumption the function ϕ is N^+ -biinvariant and we have to show that it is constant on A . By right-invariance the function ϕ descends to a function $\hat{\phi}$ on $G/N^+ \cong \mathbb{R}^2 \setminus \{0\}$; here the identification is given by $g \mapsto g \cdot (1, 0)^\top$. In this picture the N^+ orbits are of two types: The first type is given by horizontal lines with non-zero y -coordinate, the complement of these orbits is given by the pointed x -axis and consists entirely of fixed points. Since $\hat{\phi}$ is continuous and constant along N^+ -orbits, it must also be constant along the punctured x -axis, which is precisely the A -orbit in $\mathbb{R}^2 \setminus \{0\}$. Thus ϕ is A -invariant, and the lemma follows. \square

Now we can establish the Howe-Moore property for $G = SL_2(\mathbb{R})$:

PROOF OF THEOREM 1.7.1. Let (\mathcal{H}, π) be a strongly continuous unitary representation without non-zero fixed vectors and let $\xi, \eta \in \mathcal{H}$. We have to show the vanishing of $m_{\xi, \eta}(a_n)$ as $n \rightarrow \infty$, where $a_n \in \bar{A}^+$ leaves is a sequence leaving every compact subset. By (1.13) the sequence a_n contracts N^+ . Now let ξ_0 be any accumulation point of $\pi(a_n)\xi$. By Mautner's lemma, N^+ stabilizes ξ_0 and by Gelfand's trick the same is true for A . In particular, a_n^{-1} stabilizes ξ_0 and contracts N^- , hence N^- stabilizes ξ_0 using Mautner's lemma again. But then ξ_0 is stabilized by N^-AN^+ , which is open in G and thus generates G . This shows that ξ_0 is a fixed point of G , whence $\xi_0 = 0$. Since ξ_0 was an arbitrary accumulation point of $\pi(a_n)\xi$ we have $\pi(a_n)\xi \rightarrow 0$ weakly and thus $m_{\xi, \eta}(a_n) \rightarrow 0$, which finishes the proof. \square

1.8. The wavefront lemma in the hyperbolic plane

Before we establish the good counting property for definite quadratic forms, we highlight a trivial consequence of the hyperbolicity of the geodesic flow on \mathbb{H} which will play a crucial role in our proof. The groups A and N^+ are defined as before.

Lemma 1.8.1 (Wave front lemma). *Let U be an open identity neighbourhood in G . Then there exists an open identity neighbourhoods V in AN^+ such that $V \subset U$ and for all $a \in \overline{A^+}$,*

$$(1.15) \quad V \cdot a \subset a \cdot V.$$

PROOF. Since multiplication is continuous we find an open identity neighbourhood U' in G with $(U')^2 \subset U$. By (1.13) we find $U_{N^+} \subset U' \cap N^+$ such that $a^{-1}U_{N^+}a \subset U_{N^+}$ for all $a \in \overline{A^+}$. Now let $U_A := U' \cap A$. Then

$$U_A U_{N^+} \subset (U')^2 \subset U;$$

moreover, we have for every $a \in \overline{A^+}$ the inclusion

$$U_A U_{N^+} \cdot a \subset U_A \cdot a \cdot (a^{-1}U_{N^+}a) \subset a \cdot U_A U_{N^+}.$$

We may thus choose $V := U_A U_{N^+}$. \square

This type of statement turns out to be much harder to establish in more general situations, where hyperbolicity of the geodesic flow is not available. Its generalizations are the key to establish equidistribution for more general admissible triples.

1.9. Counting definite quadratic forms

We finally return to the problem of counting integral binary quadratic forms. Our goal is to solve the problem in the case, where the discriminant is negative, i.e. the quadratic forms in question are (positive or negative) definite. By means of Theorem 1.4.2 this problem can be considered as solved (modulo some explicit volume computations), once the following equidistribution result is established:

Theorem 1.9.1. *If Γ is a non-uniform lattice in $SL_2(\mathbb{R})$, then $(SL_2(\mathbb{R}), SO(2), \Gamma)$ has the equidistribution property.*

The two main ingredients in the proof of Theorem 1.9.1 are the Howe-Moore property of $SL_2(\mathbb{R})$ (i.e. Theorem 1.7.1) and the wavefront lemma (Lemma 1.8.1). Besides we will use some standard decomposition properties of the Haar measure with respect to subgroups. The following lemma summarizes what we need:

Lemma 1.9.2. *Let G be a unimodular Lie group and G_1, G_2 two closed subgroups with $G = G_1 G_2$ and $G_1 \cap G_2 = \{e\}$. Assume that G_1 is unimodular and denote by $\mu_{G_2}^{(R)}$ the right Haar measure on G_2 , which is normalized in such a way that*

$$\mu_G(A \times B) = \mu_{G_1}(A) \cdot \mu_{G_2}(B)$$

for some bounded $A \subset G_1$, $B \subset G_2$ of positive measure. Then for all $f \in C_c(G)$,

$$\int_G f(g) d\mu_G(g) = \int_{G_1} \int_{G_2} f(g_1 g_2) d\mu_{G_2}^{(R)}(g_2) d\mu_{G_1}(g_1)$$

PROOF. See [Kna02, Chapter VIII]. \square

We now turn to the proof of the main result:

PROOF OF THEOREM 1.9.1. We fix a Haar measure μ_G on $G := SL_2(\mathbb{R})$ and denote by μ_K the Haar measure on $K := SO(2)$ with $\mu_K(K) = 1$. The group Γ will always be equipped with the counting measure. With this convention, the measures μ_G and μ_K determine measures m_G and m_K on $\Gamma \backslash G$ and $((K \cap \Gamma) \backslash K)$

respectively. We have to show that for every sequence g_n in G which leaves every compact set and every $f \in C_c(\Gamma \backslash G)$ we have

$$\begin{aligned} & \frac{1}{m_K((K \cap \Gamma) \backslash K)} \cdot \int_{((K \cap \Gamma) \backslash K)} f(\Gamma k g_n) dm_K((K \cap \Gamma)k) \\ & \rightarrow \frac{1}{m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} f(\Gamma g) dm_G(\Gamma g). \end{aligned}$$

Two simplifications suggest themselves: First of all, since $K \cap \Gamma$ is finite, we can rewrite the left hand side as

$$(1.16) \quad \int_K f(\Gamma k g_n) d\mu_K(k).$$

The second simplification concerns the sequence g_n . According to (1.11) we can write $g_n = k_n a_n k'_n$ with $a_n \in \overline{A^+}$ and $k_n, k'_n \in K$. By right-invariance of μ_K we may replace g_n by $a_n k'_n$ without changing the integral (1.16). We claim that it is enough to establish convergence for $g_n = a_n \in \overline{A^+}$ and all $f \in C_c(\Gamma \backslash G)$. Indeed, assume first that k'_n converges to some $k \in K$. Then applying the assumed convergence result for a_n to the function $\Gamma g \mapsto f(\Gamma g k)$ we see that

$$\begin{aligned} \int_K f(\Gamma k g_n) d\mu_K(k) & \rightarrow \frac{1}{m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} f(\Gamma g k) dm_G(\Gamma g) \\ & = \frac{1}{m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} f(\Gamma g) dm_G(\Gamma g). \end{aligned}$$

Since this limit does not depend on k , we can group the k'_n into convergent subsequence to establish our claim. To summarize our considerations so far, we have reduced ourselves to proving that for $g_n \rightarrow \infty$ with $g_n \in \overline{A^+}$ and all $f \in C_c(\Gamma \backslash G)$

$$(1.17) \quad \int_K f(\Gamma k g_n) d\mu_K(k) \rightarrow \frac{1}{m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} f(\Gamma g) dm_G(\Gamma g)$$

Now we fix $\epsilon > 0$. Since f is compactly supported, it is in particular uniformly continuous, i.e. there is an identity neighbourhood U in G such that

$$(1.18) \quad |f(\Gamma g u) - f(\Gamma g)| < \epsilon \quad (g \in G, u \in U).$$

We fix such a U and choose V as in Lemma 1.8.1. Now let $v \in V$ and all $k \in K$. By the defining property of V we find for each $v \in V$ and every $n \in \mathbb{N}$ an element $v'_n \in V$ such that $v \cdot g_n = g_n \cdot v'_n$. Thus (1.18) applied to $g := k g_n$ and $u := v'_n$ yields

$$|f(\Gamma k v g_n) - f(\Gamma k g_n)| = |f(\Gamma k g_n v'_n) - f(\Gamma k g_n)| < \epsilon.$$

Now denote by $\mu_{AN^+}^{(R)}$ a *right* Haar measure on AN^+ . Then

$$\left| \frac{1}{\mu_{AN^+}^{(R)}(V)} \cdot \int_V \int_K f(\Gamma k v g_t) d\mu_k(k) d\mu_{AN^+}^{(R)}(v) - \int_K f(\Gamma k g_t) d\mu_k(k) \right| < \epsilon.$$

Instead of (1.17) we may thus show for some normalization of $\mu_{AN^+}^{(R)}$ that for all $v \in V$, $k \in K$ and $g_n \in \overline{A^+}$ with $g_n \rightarrow \infty$ we have

$$(1.19) \quad \begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{\mu_{AN^+}^{(R)}(V)} \cdot \int_V \int_K f(\Gamma k v g_n) d\mu_k(k) d\mu_{AN^+}^{(R)}(v) \\ &= \frac{1}{m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} f(\Gamma g) dm_G(\Gamma g). \end{aligned}$$

By Lemma 1.9.2 we can choose the normalization of $\mu_{AN^+}^{(R)}$ in such a way that

$$\begin{aligned} & \frac{1}{\mu_{AN^+}^{(R)}(V)} \cdot \int_V \int_K f(\Gamma k v g_n) d\mu_k(k) d\mu_{AN^+}^{(R)}(v) \\ &= \frac{1}{\mu_{AN^+}^{(R)}(V)} \cdot \int_{AN} \int_K f(\Gamma k v g_n) d\mu_k(k) \chi_V(v) d\mu_{AN^+}^{(R)}(v) \\ &= \frac{1}{\mu_G(KV)} \cdot \int_G f(\Gamma g g_n) d\mu_G(g) \\ &= \frac{\int_{\Gamma \backslash G} f(\Gamma g g_n) \left(\sum_{\gamma \in \Gamma} \chi_{KV}(\gamma g) \right) dm_G(\Gamma g)}{\int_{\Gamma \backslash G} \left(\sum_{\gamma \in \Gamma} \chi_{KV}(\gamma g) \right) dm_G(\Gamma g)} \end{aligned}$$

Now by Theorem 1.7.1 and Corollary 1.6.7 the action of A on $\Gamma \backslash G$ is mixing. Thus the last expression converges to

$$\frac{1}{m_G(\Gamma \backslash G)} \cdot \int_{\Gamma \backslash G} f(\Gamma g) dm_G(\Gamma g),$$

which establishes (1.19) and thus finishes the proof. \square

At this point we have established Part (i) of Theorem 1.3.8 and, in particular, Theorem 1.2.8 for definite forms. The indefinite case requires some more machinery. We prefer to explain this machinery in the more general setup of the next chapter. We close this chapter with some remarks concerning the structure of the proof of Theorem 1.2.8 in the definite case. As we have seen, the proof is based on three major steps:

- Establish a suitable wavefront lemma.
- Combine this with the Howe-Moore property of G to establish equidistribution.
- Deduce the good counting property from equidistribution.

We have seen that the third step works in complete generality, while the first step was nearly trivial in the present case. We will see in the next chapter, that the first step will provide major difficulties in more general situations. In the present setup we could use the interplay between geodesics and horospheres in the hyperbolic plane, and the geometry of the hyperbolic plane played a major role for this. Conversely, the equidistribution result above can be reinterpreted as a statement on the geometry of the hyperbolic plane and its quotients:

Corollary 1.9.3 (Equidistribution of spheres). *Let $x \in \mathbb{H}^2$ and let $S_r(x) \subset \mathbb{H}^2$ be the sphere of hyperbolic radius r with hyperbolic midpoint x . Let Γ be a non-uniform lattice in $SL_2(\mathbb{R})$ and denote by $\pi : \mathbb{H}^2 \rightarrow \Gamma \backslash \mathbb{H}^2$ the canonical projections. Then the images $\pi(S_r)$ of the spheres equidistribute in $\Gamma \backslash \mathbb{H}^2$ as $r \rightarrow \infty$.*

PROOF. We may assume $x = i$; then $S_r(i) = Kg_r$, where

$$g_r = \begin{pmatrix} e^{\frac{r}{2}} & 0 \\ 0 & e^{-\frac{r}{2}} \end{pmatrix} \in \overline{A}^+$$

and $g_r \rightarrow \infty$. Now denote by m the push forward of the hyperbolic volume form along π and let $f \in C_c(\Gamma \backslash \mathbb{H}^2)$. Using the same notation as in the proof of Theorem 1.9.1 we then have to show

$$(1.20) \quad \lim_{r \rightarrow \infty} \int_K f(\Gamma kg_r i) d\mu_K(k) = \frac{1}{m(\Gamma \backslash \mathbb{H}^2)} \int_{\Gamma \backslash \mathbb{H}^2} f(x) dm(x).$$

Now denote by $p : \Gamma \backslash G \rightarrow \Gamma \backslash \mathbb{H}^2$ the projection along G ; since p is proper, the function $\tilde{f} := f \circ \pi$ is compactly supported. Now

$$\int_K f(\Gamma kg_r i) d\mu_K(k) = \int_K \tilde{f}(\Gamma kg_r) d\mu_K(k);$$

on the other hand

$$\frac{1}{m(\Gamma \backslash \mathbb{H}^2)} \int_{\Gamma \backslash \mathbb{H}^2} f(x) dm(x) = \frac{1}{m_{\Gamma \backslash G}(\Gamma \backslash G)} \int_{\Gamma \backslash G} \tilde{f}(\Gamma g) dm_G(\Gamma g),$$

so that (1.20) becomes a special case of (1.17). \square

Equidistribution for symmetric pairs

2.1. Motivation and overview

The aim of this section is to provide a far reaching generalization of Theorem 1.9.1. Let G be a Lie group; then G is called *simple* if its Lie algebra \mathfrak{g} is simple in the sense that it does not possess any non-trivial ideals. Equivalently, G has a discrete center, and the quotient of G by this center is simple as an abstract group. (Thus, for example, $SL_n(\mathbb{R})$ is simple in the above sense, although it has a non-trivial center and thus cannot be simple as an abstract group.) An *involution* σ of a Lie group G is an automorphism of order 2. We use the notation

$$G^\sigma := \{g \in G \mid \sigma(g) = g\}$$

to denote the corresponding fixed point group. A pair (G, H) of Lie groups is called a *symmetric pair*¹ if there exists an involution σ of G such that $H = G^\sigma$. For example, the pair $(SL_2(\mathbb{R}), SO(2))$ is symmetric (with associated involution $\theta(g) = (g^\top)^{-1}$). This example shows that the following theorem generalizes Theorem 1.9.1

Theorem 2.1.1. *If (G, H) is a symmetric pair with G simple and non-compact, and Γ is a non-uniform lattice in G , which intersects H in a lattice, then (G, H, Γ) has the equidistribution property, hence the good counting property.*

Note that the theorem also applies to admissible triples involving the symmetric pair $(SL_2(\mathbb{R}), SO(1,1))$, thereby finishing the proof of Theorem 1.2.8. We will ultimately prove a more general version of the theorem, which applies to *semisimple* Lie groups (i.e. Lie groups whose Lie algebra is a direct sum of simple Lie algebras) and even to the slightly larger class of so-called *reductive* Lie groups. However, in this more general setup one will have to restrict the class of admissible lattices. For example, if $G = G_1 \times G_2$ is a direct product and $H = G_2$, then equidistribution fails for product lattices of the form $\Gamma_1 \times \Gamma_2$. This is closely related to the fact that product groups do not have the Howe-Moore property in the strong sense of Definition 1.6.5. This defect can be repaired by requiring the lattice Γ in question to be *irreducible*. The precise meaning of this will be discussed below. Ultimately, we will prove:

Theorem 2.1.2. *If (G, H) is a symmetric pair with G reductive and non-compact, and Γ is a non-uniform, irreducible lattice in G , which intersects H in a lattice, then (G, H, Γ) has the equidistribution property, hence the good counting property.*

The proof of Theorem 2.1.2 will follow the same three steps as the proof of Theorem 1.9.1 above: We will first prove a wavefront lemma, then deduce equidistribution

¹Most authors, including [Hel01], merely require $(G^\sigma)^0 \subset H \subset G^\sigma$. The additional flexibility gained by this is useful for a systematic study of such pairs, but to keep our exposition simple, we prefer our more restrictive definition.

using a suitable version of the Howe-Moore theorem and finally conclude by means of Theorem 1.4.2. Besides the fact that we have to take into account the technicalities about irreducibility of lattices explained above, establishing the Howe-Moore theorem is not much harder than in (and in fact reduced to) the $SL_2(\mathbb{R})$ case. The main difficulty thus lies in establishing a suitable wavefront lemma. This will occupy the larger part of this chapter.

In the case of $SL_2(\mathbb{R})$ we used the geometry of the hyperbolic plane - more precisely the interplay between horospheres and geodesic rays. In the present setup we will replace \mathbb{H}^2 by a certain metric space associated with the group G , which is called the associated *symmetric space*. The geometry of these spaces is discussed exhaustively in [Hel01]. If G is non-compact (as we will assume throughout), then this space happens to be non-positively curved in the sense of Bruhat and Tits. Such non-positively curved spaces are also called CAT(0) spaces, and we will start this chapter by investigating their basic properties (see also [BH99] for a more exhaustive treatment). We will then discuss the space of positive-definite matrices, i.e. the symmetric space of $GL_n(\mathbb{R})$, in some detail, following the beautiful exposition in [Lan99]. We then turn to general reductive groups and their associated symmetric spaces. The general definition of a reductive group is rather technical (see [Kna02, Chapter 7]); however, any such group can be embedded in a certain nice way into $GL_n(\mathbb{R})$. In order to keep the necessary background in Lie theory to the minimum, we will thus *define* reductive groups as certain subgroups of $GL_n(\mathbb{R})$. Correspondingly, their symmetric spaces will be spaces of positive definite matrices. We hope that this concrete approach will make the material accessible to readers without a strong background in Lie theory. Once the geometry of the symmetric space is understood, we can start to establish the wavefront lemma. *Horospheres* can still be defined in symmetric spaces, and will play a similar role as before. However, the role of geodesic rays will now be taken by certain convex cones called *Weyl chambers*. The interplay between such Weyl chambers and horospheres will ultimately yield the desired wavefront lemma. After proving this lemma, we briefly discuss the necessary modifications in the statement and proof of the Howe-Moore theorem. We then deduce Theorem 2.1.2.

2.2. CAT(0) spaces and symmetric spaces

Throughout this section we denote by (X, d) a complete metric space. Given $x \in X$ and $r > 0$ we denote by $B_r(x)$ the open and by $\overline{B}_r(x)$ the closed ball of radius r centered at x -

Definition 2.2.1. A point $m \in X$ is called a *midpoint* of $x, y \in X$, if

$$d(x, m) = d(y, m) = \frac{1}{2}d(x, y).$$

In a general metric space midpoints need not exist, and need not be unique if they exist. For example, on the n -sphere S^n the north pole and the south pole have the whole equator as midpoint set. On the contrary, in Hilbert spaces, trees or hyperbolic spaces, any two points admit a unique midpoint. In [BT72] Bruhat and Tits established the following condition, which guarantees unique midpoints:

Proposition 2.2.2. *Let (X, d) be a complete metric space and $x, y, m \in X$. Assume*

$$(2.1) \quad \forall z \in X : d(x, y)^2 + 4d(m, z)^2 \leq 2d(x, z)^2 + 2d(y, z)^2.$$

Then m is the unique midpoint of x and y .

PROOF. Specializing (2.1) to $z := x$ and $z := y$ respectively we obtain

$$(2d(m, x))^2 \leq d(x, y)^2, \quad (2d(m, y))^2 \leq d(x, y)^2,$$

whence

$$(2.2) \quad 2d(m, x) \leq d(x, y), \quad 2d(m, y) \leq d(x, y).$$

This implies

$$d(m, x) + d(m, y) \leq d(x, y)$$

and thus

$$(2.3) \quad d(m, x) + d(m, y) = d(x, y)$$

by the triangle inequality. Combining (2.2) and (2.3) we see that m is a midpoint; if m' is another midpoint then applying (2.1) to $z := m'$ we obtain

$$\forall z \in X : d(x, y)^2 + 4d(m, m')^2 \leq 2d(x, m')^2 + 2d(y, m')^2 = d(x, y)^2,$$

whence $d(m, m') = 0$ and thus $m = m'$. \square

Definition 2.2.3. A complete² metric space (X, d) with the property that for all $x, y \in X$ there exists $m \in X$ such that (2.1) holds, is called a *CAT(0) space*. Inequality (2.1) is called the *Bruhat-Tits inequality*. If X is a CAT(0) space and $p, q \in X$, then we denote by $\text{mid}(p, q)$ the unique midpoint of p and q .

A complete metric space is CAT(0) in the above sense if and only its curvature in the sense of E. Cartan, Alexandroff and Toponogov is ≤ 0 , hence the name. For the definition of this curvature and the corresponding definition of a CAT(κ)-space, see [BH99].

Exercise 2.2.4. Show that every Hilbert space is CAT(0) and that in this case we have equality in the Bruhat-Tits inequality.

Definition 2.2.5. Let (X, d) be a complete metric space and $I \subset (\mathbb{R}, d)$ be connected. Then an isometry $c : I \rightarrow X$ is called a *geodesic segment*. If $I = [0, \infty)$, then c is called a *geodesic ray* at $c(0)$ and if $I = \mathbb{R}$ then it is called a *geodesic*. The space (X, d) is called *uniquely geodesic* if any two $x, y \in X$ can be joined by a geodesic segment.

Corollary 2.2.6. *A CAT(0) space is uniquely geodesic.*

PROOF. Geodesic segments can be constructed by taking iterated midpoints and their limits. Uniqueness follows from uniqueness of these midpoints. \square

Not every space with unique midpoints is necessarily CAT(0). In fact, to be a CAT(0) space requires existence and uniqueness of generalized midpoints (called *circumcenters*) in the sense of the following definition:

²Many authors do not require a CAT(0) space to be complete; our CAT(0) spaces are then called *Hadamard spaces* by these authors. Since we will only deal with complete CAT(0) spaces here, we prefer to include this hypothesis in our definition.

Definition 2.2.7. Let $S \subset (X, d)$ be bounded. Then the *circumradius* of S is defined as

$$r_S := \inf\{r > 0 \mid \exists x \in X : S \subset \overline{B_r}(x)\}$$

and $x_0 \in X$ is called a *circumcenter* of S if $S \subset \overline{B_{r_S}}(x_0)$.

Note that a circumcenter for $S = \{x, y\}$ is nothing but a midpoint for the pair x, y , so circumcenters generalize midpoints. Now we have:

Proposition 2.2.8 (Serre). *Every bounded subset of a CAT(0) space has a unique circumcenter.*

The proof of Proposition 2.2.8 is based on the following estimate:

Lemma 2.2.9. *Let X be a CAT(0) space, $x_1, x_2 \in X$, $r_1, r_2 > 0$ and $S \subset \overline{B_{r_1}}(x_1) \cap \overline{B_{r_2}}(x_2)$. Then*

$$(2.4) \quad d(x_1, x_2)^2 \leq 2(r_1^2 - r_S^2) + 2(r_2^2 - r_S^2).$$

PROOF. Denote by m the midpoint of x_1 and x_2 . For every $z \in S$ we have $d(x_j, z) \leq r_j$ and thus the Bruhat-Tits inequality yields

$$\begin{aligned} d(x_1, x_2)^2 + 4d(m, z)^2 &\leq 2d(x_1, z)^2 + 2d(x_2, z)^2 \\ &\leq 2r_1^2 + 2r_2^2. \end{aligned}$$

Since $\inf_{z \in S} d(m, z)^2 \geq r_S^2$ we obtain (2.4). \square

PROOF OF PROPOSITION 2.2.8. For the existence, let r_n be a decreasing sequence converging to r_S . By definition we then find $x_n \in X$ such that $S \subset \overline{B_{r_n}}(x_n)$. By (2.4) the sequence x_n is a Cauchy sequence, hence converges to some $x_0 \in X$. By construction, x_0 is a circumcenter for S . Uniqueness is immediate from (2.4). \square

We now deduce an important fixed point theorem; we denote by $\text{Is}(X, d)$ the isometry group of (X, d) .

Theorem 2.2.10 (Bruhat-Tits fixed point theorem). *Let (X, d) be a CAT(0) space, $K \subset \text{Is}(X, d)$. Assume that K has a bounded orbit in X . Then K has a fixed point in X .*

PROOF. Let S be a bounded orbit of K and x_0 its circumcenter. Then for all $k \in K$ we have

$$S \subset \overline{B_{r_S}}(x_0) \Rightarrow S = kS \subset k\overline{B_{r_S}}(x_0) = \overline{B_{r_S}}(kx_0),$$

hence $kx_0 = x_0$ by uniqueness of circumcenters. This shows that x_0 is a fixed point and finishes the proof. \square

With this fundamental result at hand, we end our brief exposition of CAT(0) spaces. Much more could be said at this point, but since the exposition in [BH99] is hard to improve, we refrain from doing so. Instead we now focus on the class of CAT(0) spaces, in which we will ultimately be interested in:

Definition 2.2.11. A locally compact CAT(0) space (X, d) is called *symmetric* if

- (Sym1) For all $x, m \in X$ there is $y \in X$ such that m is a midpoint of x and y .
- (Sym2) For all $m \in X$ there exists $\sigma_m \in \text{Is}(X, d)$ such that for all $x, y \in X$ with midpoint m we have $\sigma_m(x) = y$.

It is easy to see that the isometry σ_m from (Sym2) is unique, if it exists. It is called the *point involution* at $m \in X$.

We say that a CAT(0) space is *geodesically complete* if every geodesic segment is part of a geodesic. Every CAT(0) space satisfying (Sym1) is clearly geodesically complete.

Example 2.2.12. Consider \mathbb{R}^n with the l^2 -metric. As a Hilbert space, this is in particular CAT(0). Given $v \in \mathbb{R}^n$ we define an isometry σ_v of \mathbb{R}^n by $\sigma_v(x) = -x + 2v$. Then \mathbb{R}^n is a symmetric space with point involutions σ_v .

We will see another example of a symmetric space in the next section. More general (in fact, all) examples can be obtained from this by the following lemma. Recall that a subset Y of a uniquely geodesic metric space X is called *totally geodesic* if every geodesic segment in X which intersects Y in at least two points is actually contained in Y . Note that this property is stronger than just geodesic convexity of Y : If a geodesic intersects Y in two points y_1, y_2 , then we require not only the geodesic segment joining y_1 and y_2 , but the whole geodesic through y_1 and y_2 to be contained in Y . Now we have:

Lemma 2.2.13. *A subset Y of a symmetric space X is a symmetric space with respect to the induced metric iff it is totally geodesic. In this case, the point involutions σ_y for $y \in Y$ preserve Y , and the point involutions of Y are just the restrictions of these.*

PROOF. A subset Y of a CAT(0) space X is totally geodesic if and only if it contains with any two points also their midpoint. From this it follows easily that Y is CAT(0) if and only if it is geodesically convex. However, if Y is symmetric, then it must be geodesically complete, and a geodesically convex and geodesically complete subspace is totally geodesic. It thus remains only to show that the point involution at $y \in Y$ maps Y to itself if $Y \subset X$ is totally geodesic. Now if $z \in Y$ is different from y then Y contains the geodesic through y and z , and $\sigma_y(z)$ lies on this geodesic. This shows $\sigma_y(Y) \subset Y$ and finishes the proof. \square

Before we close this section, let us mention that there is a more general definition of symmetric spaces, which can be found in [Hel01]. These spaces need not be CAT(0); if we restrict to irreducible ones, than they fall into three classes, given by compact symmetric spaces, non-compact non-flat symmetric spaces and the flat space \mathbb{R} respectively. Any symmetric space X in the sense of [Hel01] splits canonically as a product $X = X_{nc} \times \mathbb{R}^n \times X_c$, with X_{nc} non-compact non-flat and X_c compact. Then X is CAT(0) if and only if X_c is reduced to a point. The above definition is a well-known characterization of these CAT(0) symmetric spaces. Characterizing CAT(0) symmetric spaces among CAT(0) spaces is actually an area of active research. For the state of the art see [CM09].

2.3. The space of positive definite matrices

The purpose of this section is to introduce a key example of a symmetric space. We denote by $\text{Sym}_n(\mathbb{R})$ the set of real symmetric n -by- n matrices. A matrix $p \in \text{Sym}_n(\mathbb{R})$ is called *positive definite* if all its eigenvalues are positive, and we denote by $Pd_n(\mathbb{R})$ the set of all positive definite symmetric matrices. Note that $Pd_n(\mathbb{R})$ is open in $\text{Sym}_n(\mathbb{R})$, hence a submanifold. It is easy to see that the matrix exponential

function provides a diffeomorphism

$$(2.5) \quad \exp : \text{Sym}_n(\mathbb{R}) \rightarrow Pd_n(\mathbb{R}), \quad A \mapsto \sum_{n=0}^{\infty} \frac{A^n}{n!}.$$

Indeed, using the fact that every symmetric matrix is diagonalizable, it is easy to see that the power series for the logarithm converges on $Pd_n(\mathbb{R})$ and provides a two-sided inverse of \exp . From this fact one easily obtains the polar decomposition of $GL_n(\mathbb{R})$. The space of all real n -by- n matrices is the Lie algebra of $GL_n(\mathbb{R})$, hence denoted $\mathfrak{gl}_n(\mathbb{R})$, and similarly the space of all antisymmetric such matrices is denoted $\mathfrak{o}_n(\mathbb{R})$. Then the fact that every matrix can be written as a sum of an antisymmetric matrix and a symmetric matrix becomes

$$\mathfrak{gl}_n(\mathbb{R}) = \mathfrak{o}_n(\mathbb{R}) \oplus \text{Sym}_n(\mathbb{R}).$$

We now recall from linear algebra that this decomposition has a global analogon:

Proposition 2.3.1 (Polar decomposition). *The multiplication map of $GL_n(\mathbb{R})$ induces a diffeomorphism*

$$O_n(\mathbb{R}) \times Pd_n(\mathbb{R}) \rightarrow GL_n(\mathbb{R}).$$

Using diagonalization it is easy to see that the $GL_n(\mathbb{R})$ -action on $Pd_n(\mathbb{R})$ given by

$$g \cdot p := gpg^\top$$

is transitive. Its kernel is given by $\pm \mathbf{1}$ and the stabilizer of $\mathbf{1}$ is by definition the orthogonal group $O_n(\mathbb{R})$. We thus have an identification

$$GL_n(\mathbb{R})/O_n(\mathbb{R}) \rightarrow Pd_n(\mathbb{R}), \quad gO_n(\mathbb{R}) \mapsto gg^\top.$$

We now use the inclusion $Pd_n(\mathbb{R}) \hookrightarrow \text{Sym}_n(\mathbb{R})$ to identify the tangents space of $Pd_n(\mathbb{R})$ at $\mathbf{1}$ with $\text{Sym}_n(\mathbb{R})$. On the latter we have a natural scalar product given by $\langle A, B \rangle := \text{tr}(AB)$. Since this scalar product is invariant under the stabilizer action of $O_n(\mathbb{R})$ we deduce:

Lemma 2.3.2. *There exists a unique $GL_n(\mathbb{R})$ -invariant Riemannian metric $\langle \cdot, \cdot \rangle$ on $Pd_n(\mathbb{R})$ with*

$$\langle A, B \rangle_{\mathbf{1}} = \text{tr}(AB).$$

PROOF. We recall that if a manifold X is homogeneous under a Lie group G and $x \in X$ is a base point, then a scalar product on $T_x X$ extends to a G -invariant Riemannian metric on X if this scalar product is invariant under the action of $\text{Stab}_G(x)$ on $T_x X$. Thus the lemma follows from the fact that $(A, B) \text{tr}(AB)$ is invariant under the action of $O_n(\mathbb{R})$. \square

In the sequel we will always consider $Pd_n(\mathbb{R})$ as a Riemannian manifold with the above Riemannian metric. In particular, we will consider $Pd_n(\mathbb{R})$ as a metric space with respect to the metric induced by the Riemannian metric above. On the other hand, the scalar product on $\text{Sym}_n(\mathbb{R})$ turns the latter into a Hilbert space with a well-defined distance function given by $d(A, B) = \|B - A\|$.

Lemma 2.3.3. *The exponential function (2.5) is distance-increasing, i.e. for all $A, B \in \text{Sym}_n(\mathbb{R})$ we have $d(A, B) \leq d(\exp(A), \exp(B))$.*

PROOF. [BH99, Lemma 10.40].³ \square

³They prove much more than we need here; one should find a better reference for our purpose.

Now we can describe geodesics in $Pd_n(\mathbb{R})$:

Corollary 2.3.4. (i) For all $g \in GL_n(\mathbb{R})$, $A \in \text{Sym}_n(\mathbb{R})$ with $\|A\| = 1$ the map $\gamma(t) := g \exp(tA)g^\top$ is a geodesic.
(ii) For every $A \in \text{Sym}_n(\mathbb{R})$ we have $d(\mathbf{1}, \exp(A)) = \|A\|$.

PROOF. (i) Since geodesics are invariant under isometries, we may assume $g = e$. Then, denoting by $L(\cdot)$ the Riemannian length of a curve, we have for all $s < t$

$$\begin{aligned} d(\gamma(s), \gamma(t)) &\leq L(\gamma|_{[s,t]}) = \int_s^t \|\dot{\gamma}(t)\|_{\gamma(t)} dt = \int_s^t \|\gamma(t)^{-\frac{1}{2}} A \gamma(t) \gamma(t)^{-\frac{1}{2}}\| dt \\ &= \int_s^t \|A\| dt = t - s = d(sA, tA) \leq d(\gamma(s), \gamma(t)), \end{aligned}$$

showing that $d(\gamma(s), \gamma(t)) = t - s$. (ii) is immediate from (i). \square

Now we can prove:

Theorem 2.3.5. $Pd_n(\mathbb{R})$ is a symmetric CAT(0) space.

PROOF. Let $p, q \in Pd_n(\mathbb{R})$. We claim that there exists an isometry g of $Pd_n(\mathbb{R})$ such that $g.p = (g.q)^{-1}$. Indeed, let $p = e^A$, $q = e^B$ with $A, B \in \text{Sym}_n(\mathbb{R})$ and define $h_1 := e^{-\frac{B}{2}}$. We then find C with $h_1.p = e^C$ and define $h_2 := e^{-\frac{C}{4}}$. Then

$$h_2.h_1.p = e^{-\frac{C}{4}}.e^C = e^{\frac{C}{2}}, \quad h_2.h_1.q = h_2.e^{-\frac{B}{2}}.e^B = e^{-\frac{C}{4}}.\mathbf{1} = e^{-\frac{C}{2}},$$

so we may choose $g := h_2 h_1$. It thus suffices to establish the Bruhat-Tits inequality for $x = e^A$, $y = e^{-A}$ and $m = \mathbf{1}$. Thus let $B \in \text{Sym}_n(\mathbb{R})$ and $z = e^B$. Now the Hilbert space $\text{Sym}_n(\mathbb{R})$ is CAT(0), whence

$$d(A, -A)^2 + 4d(0, B)^2 \leq 2d(A, B)^2 + 2d(-A, B)^2.$$

Moreover, $d(A, -A) = d(x, y)$ and $d(0, B) = d(\mathbf{1}, z)$ by Corollary 2.3.4. Now Lemma 2.3.3 applies and yields

$$d(x, y)^2 + 4d(\mathbf{1}, z)^2 \leq 2d(A, B)^2 + 2d(-A, B)^2 \leq 2d(x, z)^2 + 2d(y, z)^2,$$

which is the desired Bruhat-Tits inequality. We have thus established that $Pd_n(\mathbb{R})$ is CAT(0). To establish (Sym1) we may assume $x = \mathbf{1}$; then m is the midpoint of x and $y := m^2$ by Lemma 2.3.4. As far as (Sym2) is concerned we claim that the point involution at $p \in Pd_n(\mathbb{R})$ is given by

$$(2.6) \quad \sigma_p(q) := pq^{-1}p.$$

Again, it suffices to check this for $p = \mathbf{1}$, where it follows from Corollary 2.3.4. \square

Combining this with Lemma 2.2.13 we obtain:

Corollary 2.3.6. Every totally geodesic subspace of $Pd_n(\mathbb{R})$ is a symmetric CAT(0) space.

Actually, the examples from Corollary 2.3.6 already exhaust *all* examples of symmetric CAT(0) spaces. This classification result is, however, beyond the scope of our introduction; see [Hel01]. As another consequence of Theorem 2.3.5 we can now finish our description of geodesics in $Pd_n(\mathbb{R})$:

Corollary 2.3.7. Every geodesic in $Pd_n(\mathbb{R})$ is of the form $\gamma(t) := g \exp(tA)g^\top$ for some $g \in GL_n(\mathbb{R})$ and $A \in \text{Sym}_n(\mathbb{R})$ with $\|A\| = 1$. Every geodesic segment in $Pd_n(\mathbb{R})$ is part of a geodesic.

PROOF. Segments of the above family of geodesics can be used to join $\mathbf{1}$ to any other point in $Pd_n(\mathbb{R})$, hence any two points in $Pd_n(\mathbb{R})$. Since a CAT(0) space is uniquely geodesic, there are no other geodesic segments. \square

In preparation of the discussions in the next section, a small technicality concerning isometries of $Pd_n(\mathbb{R})$ deserves attention: We should emphasize that $GL_n(\mathbb{R})$ is *not* the isometry group of $Pd_n(\mathbb{R})$. First of all, the subgroup $\{\pm \mathbf{1}\}$ of $GL_n(\mathbb{R})$ acts trivially on $Pd_n(\mathbb{R})$, and thus the natural map $\iota : GL_n(\mathbb{R}) \rightarrow \text{Is}(Pd_n(\mathbb{R}))$ has a kernel of order 2. More importantly, this map is *not* onto. Even worse, it does not contain the involution $\sigma_{\mathbf{1}}$. On the positive side, $\iota(GL_n(\mathbb{R}))$ does contain the identity component of $\text{Is}(Pd_n(\mathbb{R}))$ and has finite index in the isometry group (in fact, index 2). We will not prove this fact here (see again [Hel01] for details) but confine ourselves with the following easier observation:

Lemma 2.3.8. *Let $p, q \in Pd_n(\mathbb{R})$. Then $\sigma_p \circ \sigma_q \in \iota(GL_n(\mathbb{R}))$.*

PROOF. We have

$$(\sigma_p \circ \sigma_q)(x) = p(qx^{-1}q)^{-1}p = pq^{-1}.x.$$

\square

2.4. Reductive groups and their symmetric spaces

We have seen in Corollary 2.3.6 that totally geodesic subspace of $Pd_n(\mathbb{R})$ are symmetric spaces. To each of these spaces we will associate a subgroup of $GL_n(\mathbb{R})$, which acts transitively and by isometries. These groups are closely related to the corresponding isometry groups, but not quite the same (similarly as $GL_n(\mathbb{R})$ is not quite the isometry group of $Pd_n(\mathbb{R})$). We will use the following terminology:

Definition 2.4.1. Let $G \subset GL_n(\mathbb{R})$ be a closed subgroup. Then G is called *reductive* if

$$G^\top := \{g \in GL_n(\mathbb{R}) \mid g^\top \in G\} \subset G.$$

It is called a *nice* subgroup if

$$\forall A \in \text{Sym}_n(\mathbb{R}) : \exp(A) \in G \Rightarrow A \in \mathfrak{g},$$

where $\mathfrak{g} \subset \mathfrak{gl}_n(\mathbb{R})$ is the Lie algebra of G .

Since we are only dealing with groups of matrices here, we do not need any differential geometric machinery. In particular, the Lie algebra \mathfrak{g} is simply given by

$$\mathfrak{g} := \{A \in M_n(\mathbb{R}) \mid \forall t > 0 : e^{tA} \in G\}$$

and the *adjoint representation* $\text{Ad} : G \rightarrow GL(\mathfrak{g})$ is just matrix conjugation

$$\text{Ad}(g)(A) := gAg^{-1}.$$

The definition of a nice reductive group is made in such a way that the following holds:

Proposition 2.4.2. *Let G be a nice reductive group and*

$$X_G := G \cdot \mathbf{1} \subset Pd_n(\mathbb{R}).$$

Then

- (i) $X_G = G \cap Pd_n(\mathbb{R})$.
- (ii) $\mathbf{1} \in X_G$ and X_G is a totally geodesic subspace of $Pd_n(\mathbb{R})$, hence a CAT(0) symmetric space.

PROOF. (i) Since G is reductive we have $X_G \subset GG^\top \subset G$, which yields \subset . Conversely, let $v \in G \cap Pd_n(\mathbb{R})$. Since $v \in Pd_n(\mathbb{R})$ we find $A \in Sym_n(\mathbb{R})$ such that $v = \exp(A)$. Since G is nice we have $A \in \mathfrak{g}$ and thus $\exp\left(\frac{A}{2}\right) \in G$. This implies

$$v = \exp\left(\frac{A}{2}\right) \cdot \mathbf{1} \in G \cdot \mathbf{1}.$$

(ii) Given $v \in X_G$, we have $v = \exp(A)$ with $A \in \mathfrak{g}$. In particular,

$$\text{mid}(\mathbf{1}, v) = \exp\left(\frac{A}{2}\right) = \exp\left(\frac{A}{4}\right) \cdot \mathbf{1} \in X_G.$$

Since X_G is homogeneous, this implies that $\text{mid}(p, q) \in X_G$ for all $p, q \in X_G$. This implies that X_G is totally geodesic. \square

It is thus not surprising, that we have the following partial converse of the proposition:

Theorem 2.4.3. *Let X be a totally geodesic subspace of $Pd_n(\mathbb{R})$ containing $\mathbf{1}$. Then*

$$G := \{g \in GL_n(\mathbb{R}) \mid g \cdot X \subset X\}$$

is a nice reductive group and $X = X_G$. Moreover, if $K := G \cap O_n(\mathbb{R})$ and $\mathfrak{p} := \mathfrak{g} \cap Sym_n(\mathbb{R})$, then multiplication induces a diffeomorphism $K \times \exp(\mathfrak{p}) \rightarrow G$ and $X \cong G/K$.

PROOF. Denote by $\iota : GL_n(\mathbb{R}) \rightarrow \text{Is}(Pd_n(\mathbb{R}))$ the canonical map. Then by Lemma 2.3.8 we have for all $g \in GL_n(\mathbb{R})$ the implication

$$(2.7) \quad \forall p, q \in X : \iota^{-1}(\sigma_p \circ \sigma_q) \subset G.$$

Now let $p, q \in X$, $m := \text{mid}(p, q)$. Since X is totally geodesic we have $m \in X$. Let $g \in \iota^{-1}(\sigma_m \circ \sigma_p) \subset G$. Then

$$q = \sigma_m(p) = (\sigma_m \circ \sigma_p)(p) = g \cdot p,$$

showing that G is transitive on X and hence $X = G \cdot \mathbf{1}$. We now claim that G is nice. For this let $A \in Sym_n(\mathbb{R})$ with $\exp(A) \in G$. Then for all $n \in \mathbb{Z}$ we have $\exp(2nA) = \exp(A)^n \cdot \mathbf{1} \in X$. Since X is totally geodesic, it thus contains the geodesic segment joining $\mathbf{1}$ to $\exp(2nA)$, hence the full geodesic $\{\exp(tA)\}$. This shows $A \in \mathfrak{g}$, whence G is nice. Now let $p \in X$; then for all $q \in Pd_n(\mathbb{R})$ we have

$$\iota(p)(q) = p \cdot q = pq^{-1}p = (\sigma_{\mathbf{1}} \circ \sigma_{p^{-1}})(q),$$

thus $p \in \iota^{-1}((\sigma_{\mathbf{1}} \circ \sigma_{p^{-1}}) \subset G$. Since $p \in X$ was arbitrary we obtain

$$(2.8) \quad X \subset G.$$

Now let $g \in G$. By Proposition 2.3.1 we can write g as $g = e^A k$ for some $A \in Sym_n(\mathbb{R})$ and $k \in O_n(\mathbb{R})$. Now $e^{2A} = g \cdot \mathbf{1} \in X$, and hence $e^{2A} \in G$ by (2.8). Since G is nice this implies $A \in \mathfrak{g}$, hence $e^A \in G$ and thus also $k \in G$. By definition, $A \in \mathfrak{p}$ and $k \in K$, hence $G = \exp(\mathfrak{p})K = K \exp(\mathfrak{p})$. It remains to show that G is reductive. For this let $g \in G$ and write $g = kp$ with $k \in K$, $p \in \exp(\mathfrak{p})$. Then $p, k \in G$ and thus $g^\top = pk^{-1} \in G$ as was to be shown. \square

The decomposition $G \cong K \times \exp(\mathfrak{p})$ is called the *Cartan decomposition* of G . In the next section we study this decomposition more systematically.

Exercise 2.4.4. A subgroup $G \subset GL_n(\mathbb{R})$ is called *algebraic* if there exist polynomials p_1, \dots, p_m in n^2 variables over \mathbb{R} such that

$$G = \{(a_{ij}) \in GL_n(\mathbb{R}) \mid \forall 1 \leq k \leq m; p_k(a_{ij}) = 0\}.$$

Show that every algebraic subgroup of $GL_n(\mathbb{R})$ is nice.

All the classical subgroups of $GL_n(\mathbb{R})$ including $SL_n(\mathbb{R})$, $SO(p, q)$, $\text{Sym}(2n)$, \dots are algebraic. Thus the exercise shows that there are many examples of reductive groups (and hence symmetric spaces). However, we would like to stress that there are also classical examples of nice reductive subgroups which are *not* algebraic. The simplest example is $\mathbb{R}^{>0} \subset GL_1(\mathbb{R})$; more general examples of non-algebraic nice reductive subgroups arise from automorphism groups of symmetric cones. As for $\mathbb{R}^{>0}$ these groups are algebraic in the sense that they are isomorphic to the real points of an algebraic group over \mathbb{R} , but their standard embedding into some $GL_n(\mathbb{R})$ is not algebraic.

In the sequel we will restrict attention to *non-compact* nice reductive groups G . This is hardly a restriction: If $X \subset Pd_n(\mathbb{R})$ is totally geodesic and not just a point, then it contains a geodesic. Since all geodesics in $Pd_n(\mathbb{R})$ leave every compact set, this implies that X is non-compact, whence the corresponding group G is non-compact (acting transitively on X). Conversely, if G is a compact nice reductive group, then X_G is just a point. This shows in particular, that given a totally geodesic subspace $X \subset Pd_n(\mathbb{R})$ the group G with $X = X_G$ is not unique. (In fact, it is unique up to compact factors, but we will not need this fact here.)

2.5. Riemannian symmetric pairs

Throughout this section we denote by G a non-compact nice reductive Lie group and by X the associated symmetric space. Note that by definition G comes together with a fixed embedding into $GL_n(\mathbb{R})$. Various constructions depend on this embedding.

We recall that (G, H) is called a *symmetric pair* if H is the fixed point set of an involution σ of G . In this case the Lie algebra \mathfrak{h} of H has a canonical complement \mathfrak{q} in the Lie algebra \mathfrak{g} of G . Indeed, let us denote by the same letter σ the differential of σ at the identity. Then $\sigma : \mathfrak{g} \rightarrow \mathfrak{g}$ is a linear automorphism with $\sigma^2 = \text{Id}$, hence $\text{Spec}(\sigma) \subset \{\pm 1\}$. By definition, \mathfrak{h} is the $+1$ eigenspace of σ ; if we denote by \mathfrak{q} the -1 eigenspace, then

$$(2.9) \quad \mathfrak{g} = \mathfrak{h} \oplus \mathfrak{q};$$

this decomposition is called the decomposition of \mathfrak{g} *induced by* σ .

The Cartan decomposition constructed in the last section fits nicely into this framework: Define the *standard Cartan involution* on G by $\theta : g \mapsto (g^\top)^{-1}$. (This depends on the embedding of G into $GL_n(\mathbb{R})$.) Then in the notation of the last section we have $K = G^\theta$, hence (G, K) is a symmetric pair called the *standard Riemannian symmetric pair* of G . (The terminology comes from the fact, that the associated homogeneous space G/K is in a natural way a Riemannian manifold.) The *standard infinitesimal Cartan involution* $\theta : \mathfrak{g} \rightarrow \mathfrak{g}$ is given by $A \mapsto -A^\top$, and thus the induced decomposition of \mathfrak{g} is given by $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$, where \mathfrak{p} is defined as in

the last section. We refer to this decomposition as the *standard polar decomposition*.

Definition 2.5.1. A symmetric pair (G, K') is called a *Riemannian symmetric pair* if K' is conjugate to the group $K = G \cap O(n)$ in G . An involution θ' is called a *Cartan involution* (and its derivative an *infinitesimal Cartan involution*) if it is conjugate to the standard Cartan involution by an inner automorphism of G . The associated splitting of \mathfrak{g} is then called a *polar decomposition*.

The way we have defined these terms they depend a priori on the embedding of G into $GL_n(\mathbb{R})$. We will show that they are acutally independent of this embedding by providing an abstract characterization. The key observation towards this characterization is the following classification result:

Proposition 2.5.2 (Classification of maximal compact subgroups). *Let G be a nice reductive group and let $K := O(n) \cap G$. Then for every compact subgroup $K' \subset G$ there exists $g \in G$ such that $gK'g^{-1} \subset K$. In particular, K is a maximal compact subgroup of G and any other maximal compact subgroup of G is conjugate to K .*

PROOF. If K' is compact, then the action of K' on X has a bounded orbit, hence a fixed point $p \in X$ by Theorem 2.2.10. Thus $K' \subset \text{Stab}_G(p)$. Now the proposition follows from the fact that $K = \text{Stab}_G(\mathbf{1})$ and all stabilizers of X are conjugate by homogeneity of X . \square

From this we obtain our desired abstract characterization of Riemannian symmetric pairs.

Corollary 2.5.3. *Let G be a non-compact nice reductive group and $K < G$ a closed subgroup. Then (G, K) is a Riemannian symmetric pair iff K is a maximal compact subgroup of G .*

Now let us turn to Cartan involutions:

Lemma 2.5.4. *For every maximal compact subgroup $K' \subset G$ there exists a unique Cartan involution θ' with $K' = G^{\theta'}$.*

PROOF. Let $K := G \cap O(n)$ and θ be the standard Cartan involution. Then there is $g \in G$ with $gKg^{-1} = K'$. Let $\theta' := c_g\theta c_g^{-1}$, where $c_g(h) := ghg^{-1}$; then $K' = G^{\theta'}$, which establishes existence. For the uniqueness it suffices to show that $\text{Stab}_G(\theta) = N_G(K)$. Thus let $g \in G$ and assume $\theta' := c_g\theta c_g^{-1} = \theta$. Then for all $h \in G$ we have

$$\theta(g)h\theta(g)^{-1} = ghg^{-1},$$

hence $\text{Ad}(g) = \text{Ad}(\theta(g))$, or equivalently, $g\theta(g)^{-1} = c \in Z(G)$. Then for all $k \in K$,

$$\theta(gkg^{-1}) = c^{-1}gkg^{-1}c = gkg^{-1},$$

showing $gkg^{-1} \in G^\theta = K$ and thus $g \in N_G(K)$, which finishes the proof. \square

A Cartan involution θ' induces an infinitesimal Cartan involution, hence a polar decomposition of \mathfrak{g} . If θ' is conjugate to θ by c_g for some $g \in G$, then the associated polar decomposition $\mathfrak{g} = \mathfrak{k}' \oplus \mathfrak{p}'$ is related to the standard polar decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ by $\mathfrak{k}' = \text{Ad}(g)(\mathfrak{k})$, $\mathfrak{p}' = \text{Ad}(g)(\mathfrak{p})$. For our further study of polar decompositions we introduce the *trace form*

$$B : \mathfrak{g} \otimes \mathfrak{g} \rightarrow \mathbb{R}, \quad (X, Y) \mapsto \text{tr}(XY).$$

Evidently, B is a symmetric bilinear form. It is also Ad-invariant by conjugation-invariance of the trace. Finally, we claim that B is non-degenerate. For this it suffices to observe that if θ denotes the infinitesimal standard Cartan involution, then

$$B(X, \theta X) = -\operatorname{tr}(XX^\top),$$

and the latter is negative unless $X = 0$. We record for later use that by this formula the pairing

$$(2.10) \quad \langle X, Y \rangle := -B(X, \theta(Y))$$

is actually a scalar product on \mathfrak{g} . Now we have:

Lemma 2.5.5. *Let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be a polar decomposition of \mathfrak{g} . Then \mathfrak{p} is the orthogonal complement of \mathfrak{k} in \mathfrak{g} with respect to B .*

PROOF. Since \mathfrak{p} is a complement of \mathfrak{k} we have only to show orthogonality. Since B is Ad-invariant, we may assume that the given polar decomposition is standard. Thus the lemma boils down to the fact that the pairing of a symmetric matrix $A = (a_{ij})$ with an antisymmetric matrix $A' = (a'_{ij})$ is 0. The latter follows from

$$\operatorname{tr}(AA') = \sum_i \sum_j a_{ij} a'_{ji} = -\sum_i \sum_j a_{ji} a'_{ij} = \sum_j \sum_i a_{ji} a'_{ij} = -\operatorname{tr}(AA').$$

□

By the lemma, every polar decomposition of \mathfrak{g} is given by the Lie algebra of a maximal compact subgroup of G and its B -orthogonal complement. In particular, it is uniquely determined by the compact subgroup in question. Summarizing, we have obtained bijections between each of the following collections:

- the collection of maximal compact subgroups $K \subset G$;
- the collection of Riemannian symmetric pairs (G, K) ;
- the collection of Cartan involutions on G ;
- the collection of polar decompositions $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ of \mathfrak{g} .

By Proposition 2.5.2, each of this collections can be identified with $G/N_G(K)$, where $N_G(K)$ denotes the normalizer of K in G .

In view of the last remark it seems desirable to obtain a better understanding of $N_G(K)$. This is our next goal. For technical reasons it will be convenient to restrict attention first to a subclass of nice reductive groups:

Definition 2.5.6. A nice reductive group G is called *semisimple* if it has finite center.

Recall that the derivative of the adjoint representation $\operatorname{Ad} : G \rightarrow GL(\mathfrak{g})$ is given by

$$\operatorname{ad} : \mathfrak{g} \rightarrow \operatorname{End}(\mathfrak{g}), \quad \operatorname{ad}(A)(B) := [A, B].$$

In particular, the kernel of ad is precisely the center $\mathfrak{z}(\mathfrak{g})$ of \mathfrak{g} . If G is semisimple, then $\mathfrak{z}(\mathfrak{g})$ is trivial, hence ad is injective. We may thus think of \mathfrak{g} as embedded into $\operatorname{End}(\mathfrak{g})$. On \mathfrak{g} we have the non-degenerate trace form B and the associated scalar product $\langle \cdot, \cdot \rangle$ given by (2.10). We may thus define the *adjoint* A^* of an endomorphism $A \in \operatorname{End}(\mathfrak{g})$ by the formula

$$\langle A(v), w \rangle = \langle v, A^*(w) \rangle.$$

Then we have:

Lemma 2.5.7. *Denote by θ the standard Cartan involution on \mathfrak{g} . Then for every $Z \in \mathfrak{g}$ we have*

$$(2.11) \quad (\text{ad}(Z))^* = -\text{ad}(\theta(Z))$$

PROOF. Since B is invariant under the adjoint action of G we can take derivatives to obtain

$$B(\text{ad}(Z)v, w) + B(v, \text{ad}(Z)(w)) = 0.$$

Thus, since θ is a Lie algebra homomorphism we compute

$$\begin{aligned} \langle v, -\text{ad}(\theta(Z))(w) \rangle &= -B(v, \theta([\theta(Z), w])) = -B(v, [Z, \theta(w)]) \\ &= -B(v, \text{ad}(Z)(\theta(w))) = B(\text{ad}(Z)(v), \theta(w)) \\ &= \langle \text{ad}(Z)(v), w \rangle. \end{aligned}$$

□

Corollary 2.5.8. *If G is semisimple, then the Killing form*

$$\kappa : \mathfrak{g} \otimes \mathfrak{g} \rightarrow \mathbb{R}, \quad (X, Y) \mapsto \text{tr}(\text{ad}(X)\text{ad}(Y))$$

is non-degenerate. Moreover, if $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ is a Cartan decomposition then κ is positive definite on \mathfrak{p} and negative definite on \mathfrak{k} and \mathfrak{k} and \mathfrak{p} are orthogonal with respect to κ .

PROOF. With notation as in the last lemma we have

$$\kappa(X, \theta(X)) = \text{tr}(\text{ad}(X)\text{ad}(\theta(X))) = -\text{tr}(\text{ad}(X)(\text{ad}(X))^*),$$

which is strictly negative for $X \neq 0$. The various claims can now be read off directly from this formula. □

Corollary 2.5.9. *If G is semisimple and K a maximal compact subgroup of G then $N_G(K) = K$.*

PROOF. Let $H := N_G(K)$ and denote by θ the Cartan involution corresponding to K . We claim that H is θ -stable. Indeed, if $h \in H$ and $k \in K$, then

$$\theta(h)k\theta(h)^{-1} = \theta(h\theta(k)h^{-1}) = \theta(hkh^{-1}) = hkh^{-1} \in K.$$

Now let $h \in H$ and write $h = k \exp(A)$ with $k \in K$, $A \in \mathfrak{p}$. Then

$$\exp(2A) = \theta(h)^{-1}h \in H.$$

Now consider the automorphism $\alpha := \text{Ad}(\exp(2A))$ of \mathfrak{g} . Since $\exp(2A) \in H$ we see that α preserves \mathfrak{k} ; but since B is invariant under the adjoint action, this implies that α also preserves the B -orthogonal complement of \mathfrak{k} , which is \mathfrak{p} . We may thus consider α as an element of $GL(\mathfrak{k}) \times GL(\mathfrak{p})$. Since α preserves the Killing form it is actually contained in the subgroup $O(\mathfrak{k}, \kappa|_{\mathfrak{k}}) \times O(\mathfrak{p}, \kappa|_{\mathfrak{p}})$. By Corollary 2.5.8 this group is compact. Thus α generates a bounded subgroup of $GL(\mathfrak{k}) \times GL(\mathfrak{p})$, hence of $GL(\mathfrak{g})$. This means that the geodesic generated by $2A$ in the symmetric space of G does not go out to infinity; thus $A = 0$ and $h = k \in K$. Since $h \in H$ was arbitrary, this finishes the proof. □

We have now arrived at the main result of this section:

Theorem 2.5.10. *Let G be semisimple and Φ a finite group of automorphisms of G . Then there exists a maximal compact subgroup K of G which is fixed by all elements of Φ .*

PROOF. Denote by $\mathcal{K}(G)$ the collection of all maximal compact subgroups of G . We have already observed that $\mathcal{K}(G) \cong G/N_G(K) = G/K$. Via this isomorphism, $\mathcal{K}(G)$ inherits the structure of a CAT(0) space and it can be checked that the action of $\text{Aut}(G)$ on $\mathcal{K}(G)$ is by isometries. Then the theorem follows from Theorem 2.2.10. \square

Here we will only need the case where Φ is the two-element group generated by an involution of G . Here we obtain the following special case:

Corollary 2.5.11. *Let G be a semisimple group and σ an involution of G . Denote by (G, H) the corresponding symmetric pair and by $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{q}$ the induced decomposition of \mathfrak{g} . Then the following hold:*

- (i) *There exists a maximal compact subgroup K of G , which is fixed by σ .*
- (ii) *There exists a Cartan involution θ commuting with σ .*
- (iii) *There exists a polar decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ such that*

$$\begin{aligned}\mathfrak{h} &= (\mathfrak{h} \cap \mathfrak{k}) \oplus (\mathfrak{h} \cap \mathfrak{p}); \\ \mathfrak{q} &= (\mathfrak{q} \cap \mathfrak{k}) \oplus (\mathfrak{q} \cap \mathfrak{p}).\end{aligned}$$

PROOF. (i) is the case $\Phi = \{e, \sigma\}$ of Corollary ???. To deduce (ii) and (iii) we use the fact that for a semisimple group a polar decomposition is also orthogonal with respect to the Killing form, and that every Lie algebra automorphism preserves the latter. Thus, if σ preserves \mathfrak{k} , then it also preserves \mathfrak{p} and thus commutes with θ . \square

The corollary can be extended to reductive groups. The key observation is that every reductive group is a central extension of a semisimple group. One then uses the fact that the center is characteristic and has a unique maximal compact subgroup. For details see [Kna02, Chapter VII]. In the situation of the corollary we say that the group K respectively the Cartan involution θ and the Cartan decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ are *adapted* to the symmetric pair (G, H) . The existence of an adapted Cartan decomposition will be crucial for our proof of the wavefront lemma for non-Riemannian symmetric pairs.

2.6. Flats and Weyl chambers

In the proof of the equidistribution theorem for $SL_2(\mathbb{R})$ a prominent role was played by certain subgroups which we denoted A and N^\pm at that point. We aim to introduce similar subgroups for arbitrary nice reductive groups; understanding their geometric meaning will be crucial for the proof of the wavefront lemma. In this section we will focus on generalizations of the group A and the subset A^+ defined in Section ??. Recall that in the $SL_2(\mathbb{R})$ case the group A parametrized a geodesic in \mathbb{H}^2 , while the subset A^+ parametrized a geodesic ray contained in that geodesic. The analogous concepts in the present situation are given as follows;

Definition 2.6.1. Let X be a geodesically complete locally compact CAT(0) space. A subset $F \subset X$ is called a *flat* if there exists an isometry $i : (\mathbb{R}^n, d_{l_2}) \rightarrow X$ with $i(\mathbb{R}^n) = F$. Since X is assumed locally compact, the dimension of flats in X is bounded, and its maximum is called the *rank* of X . A flat F is called a *maximal flat* if it is maximal with respect to inclusion. A geodesic ray r or a geodesic σ are called *regular* if they are contained in a unique maximal flat; otherwise they are

called *singular*. Given $p, q \in X$ we denote by $\sigma_{p,q}$ the unique geodesic through p and q . The *regular set* of a point p in a maximal flat F containing p is defined as

$$\text{Reg}(p, F) := \{q \in F \setminus \{p\} \mid \sigma_{p,q} \text{ is regular}\}$$

The connected components of $\text{Reg}(p, F)$ are called *Weyl chambers* in F based at p .

In this language, the hyperbolic plane \mathbb{H}^2 has rank 1. Maximal flats are geodesics, and Weyl chambers are geodesic rays. In general symmetric spaces flats and regular geodesics can be described as follows:

Proposition 2.6.2. *Let G be a nice reductive group and let X_G be the associated symmetric space, $K = \text{Stab}_G(\mathbf{1})$ and $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ the corresponding Cartan decomposition.*

- (i) *Flats containing $\mathbf{1}$ in X_G are precisely the sets of the form $F = \exp(\mathfrak{a})\mathbf{1}$, where $\mathfrak{a} \subset \mathfrak{p}$ is an abelian subalgebra.*
- (ii) *Let r be a geodesic ray through $\mathbf{1}$ in X_G . Then $r(t) = \exp(tA)$ for some $A \in \mathfrak{p}$ of norm 1, and r is regular if and only if the centralizer $\mathfrak{z}_{\mathfrak{p}}(A)$ is an abelian Lie algebra. In this case, $\mathfrak{z}_{\mathfrak{p}}(A)$ is a maximal abelian subalgebra of \mathfrak{p} .*
- (iii) *Let \mathfrak{a} is a maximal abelian subalgebra of \mathfrak{p} and $F = \exp(\mathfrak{a})\mathbf{1}$ the corresponding flat and $p = \exp(A)\mathbf{1}$ with $A \in \mathfrak{a}$. Then $p \in \text{Reg}(\mathbf{1}, F)$ if and only if $\mathfrak{z}_{\mathfrak{p}}(A)$ is an abelian Lie algebra; in this case $\mathfrak{z}_{\mathfrak{p}}(A) = \mathfrak{a}$.*

PROOF. (i) is [BH99, Prop. 10.45]. (ii) The description of geodesic rays is a consequence of Corollary 2.3.7. We deduce from (i) that $\exp(tA)$ and $\exp(tB)$ are in a common maximal flat if and only if $[A, B]$ commute. From this observation (ii) and (iii) follow readily. \square

Here we have the slight notational problem that positive definite matrices can be considered as both elements of G and of X_G . (We recall that for $p \in Pd_n(\mathbb{R})$ we have $p\mathbf{1} = p^2$.) As a set a maximal abelian subgroup A of $G \cap Pd_n(\mathbb{R})$ is the same as the corresponding flat $F = A\mathbf{1}$ in X_G . We still want to distinguish between the two notationally; we reserve the letter F for a maximal flat in X_G and the letter A for the subgroup G with the same underlying set. Similarly, a positive Weyl chamber $W^+ \subset F$ at $\mathbf{1}$ will be denoted A^+ when considered as a subset of A .

Example 2.6.3. Let us consider the case of $G = GL_n(\mathbb{R})$. Flats in X_G are abelian subgroups of $Pd_n(\mathbb{R})$, and a maximal such subgroup is given by the group

$$A := \{\text{diag}(\lambda_1, \dots, \lambda_n) \mid \lambda_j \geq 0\}.$$

Any element of A is of the form $g(t_1, \dots, t_n) = \text{diag}(e^{t_1}, \dots, e^{t_n})$ with $t_j \in \mathbb{R}$. The centralizer of $g(t_1, \dots, t_n)$ coincides with A if and only if $t_i \neq t_j$ for $i \neq j$. Thus

$$\text{Reg}(\mathbf{1}, A) = \{g(t_1, \dots, t_n) \mid \forall i \neq j : t_i \neq t_j\}.$$

We see from this description that there are $n!$ Weyl chambers in A based at $\mathbf{1}$. We can parametrize these by $\sigma \in \mathfrak{S}_n$, setting

$$W_\sigma := \{g(t_1, \dots, t_n) \mid t_{\sigma(1)} > \dots > t_{\sigma(n)}\}.$$

We call $W^+ := W_{\text{id}}$ the *standard positive Weyl chamber* of G and denote the corresponding subgroup of A by A^+ . By definition,

$$A^+ = \{\text{diag}(\lambda_1, \dots, \lambda_n) \mid \lambda_1 > \dots > \lambda_n > 0\}.$$

In a general CAT(0) space it may happen that maximal flats have different dimensions. This is not the case in symmetric spaces. Indeed, let F be a maximal flat through $\mathbf{1}$ and A the associated subgroup of G . Let \mathfrak{a} be its Lie algebra; then for every $k \in K$ the algebra $\mathfrak{a}' := \text{Ad}(k)(\mathfrak{a})$ is again abelian, contained in \mathfrak{p} (since $[\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p}$) and maximal. Thus, K acts on the maximal flats through $\mathbf{1}$.

Proposition 2.6.4. *Let G be nice reductive and $K = \text{Stab}_G(\mathbf{1})$. Then the action of K on flats through $\mathbf{1}$ is transitive.*

PROOF. Let $\mathfrak{a}, \mathfrak{a}'$ be maximal abelian subalgebras of \mathfrak{p} corresponding to flats $F, F' \subset X_G$ and let $H \in \mathfrak{a}, H' \in \mathfrak{a}'$ be generators of regular geodesics. Now for $k \in K$ the flat $k.F$ contains the regular geodesic generated by $\text{Ad}(k)(H)$. The angle α_k between this geodesic and the one generated by H' is given by

$$\cos(\alpha_k) = B(\text{Ad}(k)(H), H').$$

Now since K is compact, the function $k \mapsto \cos(\alpha_k)$ takes its minimum in some $k_0 \in K$. We claim that $k_0.F = F'$. Indeed, we have

$$\left. \frac{d}{dt} \right|_{t=0} B(\text{Ad}(\exp(tZ))\text{Ad}(k_0)(H), H') = 0 \quad (Z \in \mathfrak{k}),$$

which leads to the equation

$$B([H', \text{Ad}(k_0)(H)], Z) = 0 \quad (Z \in \mathfrak{k}).$$

Since B is definite on \mathfrak{k} this implies

$$[H', \text{Ad}(k_0)(H)] = 0,$$

hence $\text{Ad}(k_0)(H) \in \mathfrak{z}_{\mathfrak{p}}(H') = \mathfrak{a}'$. We conclude that the geodesic generated by $\text{Ad}(k_0)(H)$ is contained in both F' and $k_0.F$. Since it is regular, these two flats must coincide. \square

This has a number of important consequences:

Corollary 2.6.5. *G acts transitively on maximal flats in X_G . In particular, all maximal flats have the same dimension given by the rank of X_G .*

Corollary 2.6.6. *If G is nice reductive, $K = \text{Stab}_G(\mathbf{1})$ and A the group corresponding to a maximal flat through $\mathbf{1}$ in X_G . Then*

$$G = KAK.$$

As in the $SL_2(\mathbb{R})$ -case one can establish the stronger decomposition $G = K\overline{A^+}K$. This requires to prove that $N_K(A)$ acts transitively on Weyl chambers. We will not need this fact here, and thus refer the reader to [Kna02, Chapter VII] for a proof.

2.7. Parabolic subgroups and horospheres

In our study of $SL_2(\mathbb{R})$ we have made crucial use of the horospherical subgroups N^\pm . These were closely related to the boundary of the hyperbolic plane. Indeed, if we consider \mathbb{RP}^1 as the boundary of \mathbb{H}^2 then the stabilizer of any boundary point was conjugate to AN^+ . Thus N^+ can be seen as the radial part of the stabilizer of a point at infinity. We will generalize this intuition to general symmetric spaces. We start by defining the boundary of a symmetric space:

Definition 2.7.1. Let X be a CAT(0) space. Two geodesic rays c_1, c_2 in X are called *asymptotic* if

$$\exists C > 0 \forall t > 0 : d(c_1(t), c_2(t)) < C.$$

This is an equivalence relation, and the equivalence class of a geodesic ray c is denoted $c(\infty)$. The collection of all these equivalence classes is called the *geodesic boundary* of X and denoted ∂X .

We observe that $\text{Is}(X, d)$ acts on the set of geodesic rays, and since asymptoticity is preserved under this action, it descends to an action on ∂X . Here we only consider ∂X as a $\text{Is}(X, d)$ -set without any further structure. However, let us at least mention that it is possible to define a topology and even a metric on ∂X . Moreover, there is a compactification \bar{X} of X (i.e. a compact space containing a homeomorphic copy of X as a dense open subset) such that $\partial X = \bar{X} \setminus X$. This justifies the term *boundary*. We could also define the geodesic boundary as geodesic rays emanating from a fixed base point. This is equivalent, since in a CAT(0) space X there exists for every $x \in X$ and every geodesic ray c a unique geodesic ray c' emanating from x and asymptotic to c [BH99, Prop. II.8.2]. In the case of a symmetric space, this yields a particularly nice parametrization of ∂X :

Proposition 2.7.2. *Let G be a nice reductive group, $\mathfrak{p} = \mathfrak{g} \cap \text{Sym}_n(\mathbb{R})$ and denote by $S^1(\mathfrak{p})$ the unit sphere in \mathfrak{p} . Given $A \in S^1(\mathfrak{p})$ denote by $r_A(t) := \exp(tA)$ the associated geodesic ray. Then the map*

$$(2.12) \quad S^1(\mathfrak{p}) \rightarrow \partial X_G, \quad A \mapsto r_A(\infty)$$

is a bijection.

PROOF. As discussed above we may identify the geodesic boundary with geodesic rays emanating from $\mathbf{1}$ and these are of the described form. \square

Let us warn the reader, that *as a G -space* the boundary can be much more complicated than suggested by the example of the hyperbolic plane:

Proposition 2.7.3. *Let G be a nice, reductive group and $X = X_G$. Then G acts transitively on ∂X if and only if $\text{rank}(X) = 1$.*

PROOF. If $\text{rank}(X) = 1$, then G acts transitively on geodesics by Corollary 2.6.5, and in fact on geodesic rays. (One can pass from a geodesic ray to its opposite by a suitable combination of point involutions.) If $\text{rank}(X) > 1$, then there exist singular geodesic rays, and the G -action preserves the proper subset of regular geodesic rays. \square

As we will see below, the stabilizers of different boundary points can vary quite a lot. Because of their importance, these stabilizers have a special name:

Definition 2.7.4. Let X a CAT(0) space, $G < \text{Is}(X, d)$ and $\xi \in \partial X$. Then

$$P_\xi := P_\xi(G) := \text{Stab}_G(\xi)$$

is called a *parabolic subgroup* of G .

Now we want to compute the parabolic subgroups of a given nice reductive group G . Given $H \in S^1(\mathfrak{p})$ we use the notation P_H (or $P_H(G)$) to denote the parabolic subgroup $P_{r_H(\infty)}$. Since

$$(2.13) \quad P_H(G) = P_H(GL_n(\mathbb{R})) \cap G$$

we can reduce our computation to the case $G = GL_n(\mathbb{R})$. Since H is a symmetric matrix we can find $k \in SO_n(\mathbb{R})$ such that

$$(2.14) \quad H' := kHk^{-1} = \text{diag}(\lambda_1, \dots, \lambda_n) \quad (\lambda_1 \geq \dots \geq \lambda_n).$$

Now we have:

Lemma 2.7.5. *Let H, k and H' as above and suppose H' has $k \leq n$ distinct eigenvalues with corresponding multiplicities r_1, \dots, r_k . Then*

$$\begin{aligned} P_{H'} &= \{g \in GL_n(\mathbb{R}) \mid \exists h \in G : \lim_{t \rightarrow \infty} e^{-tA'} g e^{tA'} = h\} \\ &= \left\{ \left(\begin{array}{ccc} A_{11} & \dots & A_{1k} \\ 0 & \ddots & \vdots \\ 0 & 0 & A_{kk} \end{array} \right) \in GL_n(\mathbb{R}) \mid A_{ij} \in M_{r_i, r_j}(\mathbb{R}) \right\}, \end{aligned}$$

and $P_H = kP_{H'}k^{-1}$.

PROOF. This is a straight forward computation; see [BH99, Prop. II.10.64]. \square

With other words, parabolic subgroups in G are intersections of conjugates of block upper triangular matrices in $GL_n(\mathbb{R})$ with G . Note, however, that the conjugation takes place within $GL_n(\mathbb{R})$, not G . From the explicit description of parabolic subgroups above we deduce:

Corollary 2.7.6. *If $H \in S^1(\mathfrak{p})$ is regular, then P_H depends only on the Weyl chamber of $\exp(H)$ based at $\mathbf{1}$ in the flat containing r_H .*

In view of the corollary we introduce the following notation: Let F be a flat at $\mathbf{1}$ and W^+ a Weyl chamber in F at $\mathbf{1}$. As before we use the notation A for the associated subgroup and A^+ for W^+ considered as a subset of A . Now write \mathfrak{a} for the Lie algebra of A and denote by \mathfrak{a}^+ the inverse image of A^+ under the exponential function. Then for $H \in \mathfrak{a}^+ \cap S(\mathfrak{p})$ we set $P_{W^+} := P_H$ and refer to P_{W^+} as the parabolic subgroup associated with the Weyl chamber W^+ .

Example 2.7.7. We continue the discussion from Example 2.7.7 and use the same notation. In particular, $G = GL_n(\mathbb{R})$ and W^+ is the Weyl chamber corresponding to

$$A^+ = \{\text{diag}(\lambda_1, \dots, \lambda_n) \mid \lambda_1 > \dots > \lambda_n > 0\}.$$

Then P_{W^+} is precisely the group of upper triangular matrices. If W^- denotes the opposite Weyl chamber (corresponding to $(A^+)^{-1}$), then P_{W^-} is the group of lower triangular matrices.

We now aim to define the radial (or horospherical) part of a parabolic subgroup. We remind the reader of the following terminology:

Definition 2.7.8. A matrix $M \in M_n(\mathbb{R})$ is called *unipotent* if $M - \mathbf{1}$ is nilpotent. A subgroup of $GL_n(\mathbb{R})$ is called *unipotent* if all its elements are unipotent.

Every Lie group G contains a unique maximal connected unipotent normal subgroup denoted $R_u G$ and called the *unipotent radical* of G . Now we define:

Definition 2.7.9. Let G be a nice reductive group, W^+ a fixed Weyl chamber in X_G and W^- the Weyl chamber opposite W^+ (i.e. the set of inverses of W^+). Then the *horospherical group* N_{W^+} of W^+ is defined as the unipotent radical of P_{W^+} . If W^+ is clear from the context then we write N^\pm for N_{W^\pm} .

We compute the horospherical groups for $G = GL_n(\mathbb{R})$: ...

In the general case ...

Our next goal is to determine the Lie algebras of \mathfrak{n}^\pm given a fixed positive Weyl chamber W^+ in X_G . We denote again by A the group corresponding to the flat containing W^+ and by \mathfrak{a} its Lie algebra; as before, W^+ determines a convex cone \mathfrak{a}^+ in \mathfrak{a} . We call such a cone an *infinitesimal Weyl chamber* in \mathfrak{a} . Now we define:

Definition 2.7.10. A non-zero (real-valued) linear functional $\lambda \in \mathfrak{a}^*$ is called a *root* of \mathfrak{a} in \mathfrak{g} if the associated *root space*

$$g_\lambda := \{X \in \mathfrak{g} \mid \forall H \in \mathfrak{a} : \text{ad}(H)(X) = \lambda(H)(X)\}$$

is non-zero. The sets of roots of \mathfrak{a} in \mathfrak{g} is denoted $\Sigma(\mathfrak{g}, \mathfrak{a})$ (or Σ for short).

In accordance with this notation we also define

$$g_0 := \{X \in \mathfrak{g} \mid \forall H \in \mathfrak{a} : \text{ad}(H)(X) = 0\}.$$

By definition, this is just the centralizer of \mathfrak{a} in \mathfrak{g} . We then have the following *root space decomposition*:

Lemma 2.7.11.

$$\mathfrak{g} = g_0 \oplus \bigoplus_{\lambda \in \Sigma(\mathfrak{g}, \mathfrak{a})} g_\lambda.$$

PROOF. □

Since \mathfrak{g} is finite-dimensional, the set $\{\ker \lambda \mid \lambda \in \Sigma(\mathfrak{g}, \mathfrak{a})\}$ is a finite set, of hyperplanes in \mathfrak{a} . In particular,

$$\mathfrak{a}_{reg} := \mathfrak{a} \setminus \bigcup_{\lambda \in \Sigma(\mathfrak{g}, \mathfrak{a})} \ker \lambda$$

is a union of finitely many connected components, which are convex cones. Then the following is not surprising:

Lemma 2.7.12. (i) *Let $H \in \mathfrak{a}$ and $\|H\| = 1$. Then r_H is regular if and only if $H \in \mathfrak{a}_{reg}$.*

(ii) *The connected components of \mathfrak{a}_{reg} are precisely the infinitesimal Weyl chambers.*

PROOF. □

In particular, \mathfrak{a}^+ is a connected component of \mathfrak{a}_{reg} and thus every root is either positive or negative on \mathfrak{a}^+ . Thus if we define

$$\Sigma^\pm(\mathfrak{g}, \mathfrak{a}, \mathfrak{a}^+) := \{\lambda \in \Sigma(\mathfrak{g}, \mathfrak{a}) \mid \pm \lambda|_{\mathfrak{a}^+} > 0\},$$

then

$$\Sigma(\mathfrak{g}, \mathfrak{a}) = \Sigma^+(\mathfrak{g}, \mathfrak{a}, \mathfrak{a}^+) \cup \Sigma^-(\mathfrak{g}, \mathfrak{a}, \mathfrak{a}^+).$$

If the positive Weyl chamber is clear from the context we simply write Σ^\pm . Now the root space decomposition implies

$$(2.15) \quad \mathfrak{g} = \mathfrak{n}^- \oplus g_0 \oplus \mathfrak{n}^+,$$

where

$$\mathfrak{n}^\pm := \bigoplus_{\lambda \in \Sigma^\pm(\mathfrak{g}, \mathfrak{a}, \mathfrak{a}^+)} g_\lambda.$$

This decomposition, called the *triangular decomposition* of \mathfrak{g} , depends of course on our choice of positive Weyl chamber. Now we have:

Proposition 2.7.13. \mathfrak{n}^\pm is the Lie algebra of the horospherical group N^\pm associated with the same positive Weyl chamber.

PROOF. □

2.8. The wavefront lemma for Riemannian symmetric pairs

We have not collected enough results to establish the wavefront lemma for Riemannian symmetric pairs. The general case will only require minor modifications; however, for sake of clarity of ideas we deal with the simpler Riemannian case first. Our proof is based on the following *infinitesimal Iwasawa decomposition*:

Proposition 2.8.1 (Iwasawa). *Let G be a nice reductive group with Lie algebra \mathfrak{g} . Let $K = \text{Stab}_G(\mathbf{1})$, $A \subset G$ a subgroup corresponding to a maximal flat in X_G and N^\pm the horospherical subgroups corresponding to a choice of positive Weyl chamber A^+ in A . Then the corresponding Lie algebras satisfy*

$$\mathfrak{g} = \mathfrak{k} + \mathfrak{a} + \mathfrak{n}^+ = \mathfrak{k} + \mathfrak{a} + \mathfrak{n}^-.$$

In particular, the maps

$$K \times A \times N^\pm \rightarrow G, \quad (k, a, n) \mapsto kan$$

are open.

In fact, the decompositions $\mathfrak{g} = \mathfrak{k} + \mathfrak{a} + \mathfrak{n}^\pm$ are direct, and the corresponding global maps are diffeomorphisms. However, we will not need these facts here. (For $G = GL_n(\mathbb{R})$ they follow immediately from the Gram-Schmidt orthogonalization procedure.) In the present (weak) form the proposition generalizes to non-Riemannian symmetric pairs (while the stronger form of the proposition does not).

PROOF OF PROPOSITION 2.8.1. We start from the decomposition (2.15), writing $\mathfrak{g} = \mathfrak{n}^- \oplus \mathfrak{g}_0 \oplus \mathfrak{n}^+$. Now let $\lambda \in \Sigma \cup \{0\}$, $X \in \mathfrak{g}_\lambda$ and $Z \in \mathfrak{a}$, and denote by θ the infinitesimal standard Cartan involution. Now, since $Z \in \mathfrak{p}$ we have $\theta(Z) = -Z$ and thus

$$\text{ad}(Z)(\theta(X)) = -\text{ad}(\theta(Z))(\theta(X)) = -[\theta(Z), \theta(X)] = -\theta([Z, X]) = -\lambda(Z)\theta(X).$$

This shows that $\theta(X) \in \mathfrak{g}_{-\lambda}$. Since $X \in \mathfrak{g}_\lambda$ was arbitrary we deduce $\theta(\mathfrak{g}_\lambda) = \mathfrak{g}_{-\lambda}$. In particular,

$$(2.16) \quad \theta(\mathfrak{n}^\pm) = \mathfrak{n}^\mp, \quad \theta(\mathfrak{g}_0) = \mathfrak{g}_0.$$

Now every $X \in \mathfrak{n}^\pm$ can be written as $X = (X + \theta(X)) - \theta(X)$. We have $X + \theta(X) \in \mathfrak{k}$ (since it is θ -stable) and $\theta(X) \in \mathfrak{n}^\mp$, hence

$$(2.17) \quad \mathfrak{n}^\pm \subset \mathfrak{k} + \mathfrak{n}^\mp.$$

Now let $X \in L\mathfrak{g}_0$ and write $X = \frac{1}{2}(X + \theta(X)) + \frac{1}{2}(X - \theta(X))$ as above. Then, again, we have $\frac{1}{2}(X + \theta(X)) \in \mathfrak{k}$, while $\frac{1}{2}(X - \theta(X)) \in \mathfrak{p} \cap \mathfrak{g}_0$. The latter is just the centralizer of \mathfrak{a} in \mathfrak{p} , which is \mathfrak{a} itself by maximality. Thus

$$(2.18) \quad \mathfrak{g}_0 \subset \mathfrak{k} + \mathfrak{a}.$$

Now the proposition follows by combining (2.15), (2.17) and (2.18). □

Now we can finally prove:

Theorem 2.8.2 (Wavefront lemma, Riemannian symmetric case). *Let G be a nice reductive group with Lie algebra \mathfrak{g} . Let $K = \text{Stab}_G(\mathbf{1})$ and $A \subset G$ a subgroup corresponding to a maximal flat in X_G . For any identity neighbourhood U in G there is an identity neighbourhood V in G such that*

$$\forall g \in AK : KVg \subset KgU.$$

PROOF. The proof proceeds in three steps:

Step 1. Fix a positive Weyl chamber A^+ and consider the corresponding Iwasawa decomposition $G = KAN^+$. Since multiplication is continuous we can find identity neighbourhoods W_A in A and W_{N^+} in N^+ such that $W_A W_{N^+} \subset U$. Moreover, by shrinking W_{N^+} we can ensure that $a^{-1}W_{N^+}a \subset W_{N^+}$ for all $a \in \overline{A^+}$. We now define $V_{A^+} := KW_A W_{N^+}$. By Proposition 2.8.1 this is open in G . Moreover, for all $g \in \overline{A^+}$ we have

$$\begin{aligned} KV_{A^+}g &= KW_A W_{N^+}g = K(W_A g)(g^{-1}W_{N^+}ag) \subset K(gW_A)W_{N^+} \\ &\subset KgU. \end{aligned}$$

Thus we have established the desired property for $g \in \overline{A^+}$.

Step 2. Construct V_{A^+} for every Weyl chamber A^+ in A at o . Since there are only finitely many such Weyl chambers, the intersection

$$V_A := \bigcap V_{A^+}$$

is open. By Step 1 we then have $KV_A g \subset KgU$ for all $g \in A$.

Step 3. Since K is compact we find an identity neighbourhood $U' \subset U$ with $k^{-1}U'k \subset U$ for all $k \in K$. By Step 2 we find an identity neighbourhood V such that $KVa \subset KaU'$ for all $a \in A$. Now let $g = ak \in AK$ be arbitrary. Then

$$\begin{aligned} KVg &= KVak \subset KaU'k = Kak(k^{-1}U'k) \\ &\subset KgU, \end{aligned}$$

which finishes the proof. \square

2.9. Non-Riemannian symmetric pairs and relative decompositions

We now want to establish a wavefront lemma for more general symmetric pairs. For simplicity we restrict attention to the case where G is semisimple. Let $\sigma : G \rightarrow G$ be an involution and $H = G^\sigma$. Recall from ?? that σ induces an infinitesimal decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{q}$. By Corollary ?? there exists a Cartan involution θ commuting with σ . If \mathfrak{k} and \mathfrak{p} denote the corresponding $+1$ and -1 -eigenspace respectively, then we have

$$(2.19) \quad \mathfrak{g} = (\mathfrak{k} \cap \mathfrak{h}) \oplus (\mathfrak{k} \cap \mathfrak{q}) \oplus (\mathfrak{p} \cap \mathfrak{h}) \oplus (\mathfrak{p} \cap \mathfrak{q}).$$

We can group the four summands in three different ways to obtain Lie subalgebras of \mathfrak{g} : Namely, we have

$$\mathfrak{k} = (\mathfrak{k} \cap \mathfrak{h}) \oplus (\mathfrak{k} \cap \mathfrak{q}), \quad \mathfrak{h} = (\mathfrak{k} \cap \mathfrak{h}) \oplus (\mathfrak{p} \cap \mathfrak{h})$$

and

$$\mathfrak{g}_0 := (\mathfrak{k} \cap \mathfrak{h}) \oplus (\mathfrak{p} \cap \mathfrak{q}).$$

The latter is the fixed point algebra of the involution $\theta\sigma = \sigma\theta$, hence the Lie algebra of the group $G_0 := G^{\theta\sigma}$. Any two of the three groups K, H and G_0 intersect in $K_0 = G_0^\theta$, which has Lie algebra $\mathfrak{k}_0 = (\mathfrak{k} \cap \mathfrak{h})$. Note that G_0 is reductive with maximal compact subgroup K_0 . Indeed, if $g \in G_0$ then

$$\theta\sigma(\theta g) = \theta(\theta\sigma g) = \theta g,$$

so $\theta g \in G_0$. We now aim to reproduce the various decomposition theorems (polar, KAK , Iwasawa decomposition) in versions relative H . The key idea is to replace the maximal abelian subalgebra \mathfrak{a} of \mathfrak{p} by a maximal abelian subalgebra \mathfrak{a}_0 of $\mathfrak{p} \cap \mathfrak{q}$, the non-compact part of G_0 . Geometrically, such a subalgebra corresponds to a flat in X_G , which is orthogonal to the H -orbit at the origin and maximal with this property. We will then consider a root space decomposition of \mathfrak{g} with respect to this smaller abelian subalgebra \mathfrak{a}_0 . We denote by $\Sigma = \Sigma(\mathfrak{g}, \mathfrak{a}_0)$ the roots of \mathfrak{a}_0 in \mathfrak{g} so that

$$(2.20) \quad \mathfrak{g} = \mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0) \oplus \bigoplus_{\lambda \in \Sigma} \mathfrak{g}_{\lambda}.$$

Formally as before we can now define \mathfrak{a}_0^{reg} as the complement of the union of zero loci of the roots, and its connected components will still be called *Weyl chambers*. As before, a choice of positive Weyl chamber \mathfrak{a}_0^+ determines a decomposition of Σ into positive and negative roots and we define

$$\mathfrak{n}_0^{\pm} := \bigoplus_{\lambda \in \Sigma^{\pm}} \mathfrak{g}_{\lambda}$$

The corresponding analytic subgroups of G are denoted N_0^{\pm} . Note that these are nilpotent, hence $N^{\pm} = \exp(\mathfrak{n}^{\pm})$. Similarly, we define $A_0 := \exp(\mathfrak{a}_0)$. One new ingredient, which did not feature so prominently in the Riemannian case is the centralizer $M := Z_K(\mathfrak{a}_0)$ of \mathfrak{a}_0 in K . With these notations at hand we can now formulate the relative decompositions of G as follows:

Proposition 2.9.1. *With notations as above the following hold:*

- (i) *The map $(\mathfrak{p} \cap \mathfrak{h}) \times (\mathfrak{p} \cap \mathfrak{q}) \times K \rightarrow G$ given by $(X, Y, k) \mapsto e^X e^Y k$ is surjective (generalized polar decomposition).*
- (ii) *The map $H \times A_0 \times K \rightarrow G$ induced by multiplication is surjective (generalized KAK decomposition).*
- (iii) *The map $H \times M \times A_0 \times N_0^+ \rightarrow G$ is open onto a neighbourhood of the identity in G (generalized Iwasawa decomposition).*

PROOF. (i) *was not presented in class and is rather tedious...* (ii) Since G_0 is reductive with maximal compact subgroup K_0 we have $\mathfrak{p} \cap \mathfrak{q} = \bigcup_{k \in K_0} \text{Ad}(k)(\mathfrak{a}_0)$ (see Proposition ??). Now let $g \in G$. According to (i) we may write $g = e^X e^Y k$ with $X \in \mathfrak{p} \cap \mathfrak{h}, Y \in \mathfrak{p} \cap \mathfrak{q}$ and $k \in K$. Now $Y = hZh^{-1}$ for some $Z \in \mathfrak{a}_0$ and $h \in K_0$. Thus,

$$g = (e^X h) e^Z (h^{-1} k) \in H A_0 K.$$

(iii) It is enough to show that $\mathfrak{g} = \mathfrak{h} + \mathfrak{m} + \mathfrak{a}_0 + \mathfrak{n}_0^+$. As in the Riemannian case we show that σ and θ interchange \mathfrak{n}_0^+ and \mathfrak{n}_0^- and preserve $\mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0)$. Let us give the details for σ : Assume $X \in \mathfrak{g}_{\lambda}$ for $\lambda \in \Sigma \cup \{0\}$ and let $Z \in \mathfrak{a}_0$. Then $\sigma Z = -Z$ and thus

$$\text{ad}(Z)(\sigma X) = [Z, \sigma X] = -\sigma([Z, X]) = -\lambda(Z)\sigma(X),$$

showing $\sigma g_\lambda \subset g_{-\lambda}$. We deduce two things: Firstly, given $X \in \mathfrak{n}_0^-$ we have

$$X = (X + \sigma X) - \sigma(X) \in \mathfrak{h} + \mathfrak{n}_0^+,$$

hence $\mathfrak{n}_0^- \subset \mathfrak{h} + \mathfrak{n}_0^+$. Secondly, given $X \in \mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0)$ we have

$$X = \frac{1}{2}(X + \theta X) + \frac{1}{2}(X - \theta X) \in (\mathfrak{k} \cap \mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0)) + \mathfrak{p} = \mathfrak{m} + \mathfrak{p},$$

and since every $Y \in \mathfrak{p}$ can be written as

$$Y = \frac{1}{2}(Y + \sigma Y) + \frac{1}{2}(Y - \sigma Y) \in \mathfrak{h} + (\mathfrak{p} \cap \mathfrak{q} \cap \mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0)) = \mathfrak{h} + \mathfrak{a}_0$$

we have $\mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0) \subset \mathfrak{m} + \mathfrak{h} + \mathfrak{a}_0$. Now the root space decomposition of \mathfrak{g} implies

$$\mathfrak{g} = \mathfrak{n}_0^- + \mathfrak{z}_{\mathfrak{g}}(\mathfrak{a}_0) + \mathfrak{n}_0^+ \subset (\mathfrak{h} + \mathfrak{n}_0^+) + (\mathfrak{m} + \mathfrak{h} + \mathfrak{a}_0) + \mathfrak{n}_0^+.$$

□

2.10. The wavefront lemma for general symmetric pairs

Now we can finally prove:

Theorem 2.10.1 (Wavefront lemma, general symmetric case). *Let G be a semisimple Lie group with Lie algebra \mathfrak{g} . Let σ be an involution of G , $H = G^\sigma$, K an adapted maximal compact subgroup and $A_0 < G$ a subgroup corresponding to a flat in X_G which is orthogonal to the H -orbit as $\mathbf{1}$ and maximal with this property. For any identity neighbourhood U in G there is an identity neighbourhood V in G such that*

$$\forall g \in A_0 K : HVg \subset HgU.$$

PROOF. Compared to the proof of Theorem ?? only the first step needs a modification: Let U be an identity neighbourhood. Pick a Weyl chamber \mathfrak{a}_0^+ in \mathfrak{a}_0 . By continuity of multiplication we find identity neighbourhoods $W_M \subset M$, $W_{A_0} \subset A_0$ and $W_{N_0^+}$ such that $W_M W_{A_0} W_{N_0^+} \subset U$. Moreover, we may assume $a^{-1} W_{N_0^+} a \subset W_{N_0^+}$ for all $a \in \overline{A_0^+}$, where $A_0^+ := \exp(\mathfrak{a}_0^+)$. Set

$$V := V_{A_0^+} := HW_M W_{A_0} W_{N_0^+}.$$

By Proposition ??, V is an identity neighbourhood in G . Moreover, for all $a \in \overline{A_0^+}$ we have

$$HV a = H(W_M W_{A_0} a)(a^{-1} W_{N_0^+} a) \subset H(a W_M W_{A_0}) W_{N_0^+} = HaU.$$

The extension to $g \in AK$ is as in the Riemannian case. □

2.11. The Howe-Moore theorem for $SL_d(\mathbb{R})$

Besides the wavefront lemma, the second main ingredient in our proof of the equidistribution theorem for symmetric pairs is a suitable version for the Howe-Moore theorem. The extension of the theorem from $SL_2(\mathbb{R})$ to other *simple* Lie groups provides little difficulties. Let us explain the proof in the special case of $G = SL_d(\mathbb{R})$:

Theorem 2.11.1. *The group $SL_d(\mathbb{R})$ has the Howe-Moore property for all $d \geq 2$.*

PROOF. Let $G := SL_d(\mathbb{R})$. A maximal compact subgroup K of G is given by $SO_d(\mathbb{R})$. The standard maximal flat in G/K is given by the subgroup A of diagonal matrices. The standard positive Weyl chamber in this flat corresponds to the subset $A^+ \subset A$ given by

$$A^+ = \{\text{diag}(\lambda_1, \dots, \lambda_d) \mid \prod \lambda_j = 1, \lambda_1 > \dots > \lambda_d\}$$

We want to show that the matrix coefficients of a sequence g_n vanish as $g_n \rightarrow \infty$. As in the $SL_2(\mathbb{R})$ case we may assume $g_n \in \overline{A^+}$. Indeed, this follows from the decomposition $G = K\overline{A^+}K$ (which in the present case can be seen by elementary linear algebra). Now, if $g_n = \text{diag}(\lambda_1^{(n)}, \dots, \lambda_d^{(n)})$ leaves every compact set, then $\lambda_1^{(n)} \rightarrow \infty$ and $\lambda_d^{(n)} \rightarrow 0$. In the notation of Lemma ?? this implies

$$U(g_n) \supset \left\{ \begin{pmatrix} 1 & \dots & x \\ 0 & \ddots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \in SL_d(\mathbb{R}) \mid x \in \mathbb{R} \right\} =: N_{11}.$$

In particular, if ξ_0 is any accumulation point of $\pi(g_n)\xi$ for some $\xi \in \mathcal{H}$, where (\mathcal{H}, π) is a unitary representation of G , then N_{11} fixes ξ . Now we define an embedding of $\iota : SL_2(\mathbb{R}) \rightarrow SL_d(\mathbb{R})$ by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} a & \dots & b \\ \vdots & & \vdots \\ c & \dots & d \end{pmatrix}.$$

Then $(\mathcal{H}, \pi \circ \iota)$ is a unitary representation of $SL_2(\mathbb{R})$, and we can apply the Gelfand trick (Proposition ??) to show that ξ_0 is fixed by $\pi(\text{diag}(\lambda, 1, \dots, 1, \lambda^{-1}))$ for all $\lambda > 0$. Now we apply the Mautner lemma once more, using sequences of the form $a_n := \text{diag}(\lambda_n, 1, \dots, 1, \lambda_n^{-1})$. Depending on whether $\lambda_n \rightarrow \infty$ or $\lambda_n \rightarrow 0$ we see that ξ_0 is stabilized by elementary matrices with entries in the first/last row/column, but these matrices generate all of $SL_d(\mathbb{R})$ (as can be seen e.g. by considering the corresponding Lie algebras); hence $SL_d(\mathbb{R})$ stabilizes ξ_0 and thus $\xi_0 = 0$. We deduce $\pi(g_n)\xi \rightarrow 0$ for all $\xi \in \mathcal{H}$, hence all matrix coefficients vanish at g_n as $n \rightarrow \infty$. \square

For arbitrary simple groups the argument is only slightly more technically. However, in the semisimple case new problems arise. We will discuss them in the next section.

2.12. Howe-Moore theorems for semisimple groups

It is clear that non-simple semisimple groups cannot have the Howe-Moore property in the sense of Definition ?.?. Indeed, suppose $G = G_1 \times G_2$ and let $\Gamma = \Gamma_1 \times \Gamma_2$, where Γ_j are lattices in G_j . Then the G_1 -orbits in G/Γ preserve the second coordinate, hence cannot equidistribute. Thus in the formulation of the Howe-Moore theorem for semisimple Lie groups one has to be slightly more carefully.

In order to overcome this difficulty, we will use some structural facts about semisimple groups without further proofs (see [Kna02, Chapter VI]). Any such group G can be written as

$$G = G_1 \cdots G_k,$$

where the groups G_j are simple Lie groups with mutually finite intersection, which commute pairwise. We say that G is the *almost direct product* of the groups G_j ,

which are called the *simple factors* of G . Now the correct formulation of the Howe-Moore theorem for semisimple Lie groups is as follows:

Theorem 2.12.1. *Let (π, \mathcal{H}) be a unitary representation of a semisimple Lie group G . Assume that for every simple factor G_j of G the fixed point set is given by $\mathcal{H}^{G_j} = \{0\}$. Then all matrix coefficients of π vanish at infinity.*

PROOF. □

This Howe-Moore property does *not* imply that every subgroup of G acts mixingly on $\Gamma \backslash G$ for arbitrary lattices Γ , but it does imply this if Γ is far enough from being a product. Let us make this precise:

Definition 2.12.2. A lattice $\Gamma < G$ is called *irreducible* if the projection of Γ onto each simple factor is dense.

Now we have:

Lemma 2.12.3. *If Γ is an irreducible lattice in G , then the action of every non-compact subgroup $H < G$ on $\Gamma \backslash G$ is mixing.*

2.13. Equidistribution for symmetric pairs

UNDER CONSTRUCTION

Part 2

Unipotent flows

3.1. More examples of counting problems

In the previous two chapters we have established the good counting property for triples (G, H, Γ) , where G is a semisimple Lie group, Γ a lattice in G intersecting H in a lattice and H is a symmetric subgroup of G . The latter condition allows for many interesting examples, but even for many classical arithmetic counting problems it is too restrictive. Here are three examples of classical counting problems, where H is not symmetric:

Example 3.1.1. Let $G = SL_2(\mathbb{R})$, $\Gamma = SL_2(\mathbb{Z})$ and $H = N^+$, the group of upper triangular unipotent matrices. This is an admissible triple since

$$\Gamma \cap H = \left\langle \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right\rangle$$

is a cocompact lattice in H . As we have seen in the proof of the Gelfand trick we have $H \backslash G \cong \mathbb{R}^2 \setminus \{0\}$ via $Hg \mapsto (0, 1) \cdot g$. In this case the Γ -orbit of $v = (0, 1)$ is given by the primitive elements in \mathbb{Z}^2 . Thus the counting problem for (G, H, Γ) is equivalent to the problem of asymptotically counting primitive elements in \mathbb{Z}^2 . This is an important classical counting problem, whose answer is well-known: We have

$$|v\Gamma \cap B| \sim \frac{1}{\zeta(2)} \cdot \text{area}(B).$$

Here H is not even reductive (hence cannot be symmetric), so we cannot apply the arguments of the last chapter. Even worse, the equidistribution property does not hold: Let Y be the H -orbit through the basepoint in $X = \Gamma \backslash G$ and denote by m_Y the H -invariant probability measure on Y . If we project Y down to $\Gamma \backslash \mathbb{H}^2$ and identify the latter (measurably) with the standard fundamental domain of Γ on \mathbb{H}^2 , then Y is just a vertical line. If we move this line upwards, then it disappears to the cusp. Consequently, if we define

$$a_t := \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$$

then $\rho(a_t)_* m_Y \rightarrow 0$ in the weak-* topology. This shows that (G, H, Γ) does not have the equidistribution property in the strong sense of Definition ??.

Example 3.1.2. Let $G = SL_d(\mathbb{R})$, $\Gamma = SL_d(\mathbb{Z})$ and

$$H = \begin{pmatrix} SL_l(\mathbb{R}) & \mathbb{R}^{l \times m} \\ 0 & SL_m(\mathbb{R}) \end{pmatrix}$$

with $d = l + m$. Then H fixes $v := e_{l+1} \wedge \cdots \wedge e_{l+m} \in \bigwedge^m \mathbb{R}^d$. The orbit $v\Gamma$ consists of those $v_1 \wedge \cdots \wedge v_m$, which can be extended to a basis v_1, \dots, v_d of \mathbb{Z}^d . We can

identify these points with *rational d-planes*. In this case, a good norm leading to well-rounded balls is given by

$$\|v_1 \wedge \cdots \wedge v_m\| := \text{covol}(\langle v_1, \dots, v_m \rangle_{\mathbb{Z}}, \langle v_1, \dots, v_m \rangle_{\mathbb{R}}).$$

The corresponding counting problem is thus *counting of rational d-planes of bounded covolume*. Again, strict equidistribution fails, e.g. for

$$a_t := \text{diag}(e^{t/l}, \dots, e^{t/l}, e^{-t/m}, \dots, e^{-t/m}).$$

Example 3.1.3. Let $G = SL_d(\mathbb{R})$, $\Gamma = SL_d(\mathbb{Z})$ and $H = SL_l(\mathbb{R}) \times SL_m(\mathbb{R})$. Again, equidistribution fails.

3.2. Counting beyond strict equidistribution

In order to deal with some of the examples given above, it is worthwhile to return to our considerations in Section 1.4, and in particular to Remark 1.4.4. Namely, in order to deduce the good counting property, we do not need the full strength of the equidistribution property; it is enough if we have weak-* convergence

$$\frac{|A_n \cap v\Gamma g|}{\mu(A_n)} d\widehat{m}_G(g) \rightarrow \frac{m_H((H \cap \Gamma) \backslash H)}{m_G(\Gamma \backslash G)} d\widehat{m}_G(g).$$

for every sequence of bounded sets $A_n \subset \Gamma \backslash G$ with $\mu(A_n) \rightarrow \infty$. According to the proof of Lemma 1.4.3 this amounts to showing that for every $\alpha \in C_c(\Gamma \backslash G)$ we have

$$(3.1) \quad \int_{A_n} \int_{(\Gamma \cap H) \backslash H} \alpha(\Gamma hg) d\widehat{m}_H(h) d\mu(g) \rightarrow \int_{\Gamma \backslash G} \alpha(x) d\widehat{m}_G(x),$$

while equidistribution requires

$$\int_{(\Gamma \cap H) \backslash H} \alpha(\Gamma hg) d\widehat{m}_H(h) \rightarrow \int_{\Gamma \backslash G} \alpha(x) d\widehat{m}_G(x),$$

With other words, the *weak equidistribution property* (3.1) required to deduce the good counting property requires only equidistribution *on average over larger and larger balls*. This weaker property can still be established in many cases, where the strict equidistribution fails. We give two examples of this:

Example 3.2.1. Consider again the situation of Example 3.1.1. For the sequence a_t above, equidistribution fails, but for a_{-t} equidistribution holds by a result of Sarnak. Now we observe that

$$va_t k = (0, e^{-t/2})k \quad (k \in SO_2(\mathbb{R}));$$

this means that the equidistribution works for all points outside the unit ball and fails for all point inside the unit ball. But for equidistribution on average a compact ball does not matter! Therefore we do have the weak equidistribution property (3.1), which yields counting.

Example 3.2.2. The situation is similar for Example 3.1.2. Here we can use a one-sided version of the wavefront lemma to establish weak equidistribution, hence counting.

On the contrary, we cannot deal with Example 3.1.3 by such a trick, since the wavefront lemma fails completely in this case. Still, our examples indicate that there are more general counting results than those arising from symmetric pairs. We will present one such result (due to Eskin, Mozes and Shah) below.

3.3. Ergodic measures and Property (T)

Here we present a completely different motivation for the Eskin-Mozes-Shah theorem. Namely, we ask, whether the set of ergodic probability measures for a given group action is closed in the set of all probability measures. The following example shows that the answer is negative in general:

Example 3.3.1. Let $X := \mathbb{R}/\mathbb{Z}$ and $T_2(x) := 2x \pmod{1}$. For every sequence $\epsilon_1, \dots, \epsilon_n$ of digits in $\{0, 1\}$ we define x_n to be the rational number with periodic binary digit expansion

$$x_n = 0.\underbrace{0 \cdots 0}_{n} \epsilon_1 \cdots \epsilon_n.$$

Clearly the T_2 -orbit of x_n is periodic, hence carries a T_2 -ergodic probability measure μ_n . Now, if the sequence ϵ_k is chosen i.i.d., then the measures μ_n will converge in the weak-* topology to the measure $\mu := \frac{1}{2}(\delta_0 + m_X)$, where $m_{[0,1]}$ denotes Haar measure on X . Since both δ_0 and m_X are T_2 -invariant, μ is *not* ergodic.

Interestingly enough, in many cases related to *semisimple Lie groups*, the set of ergodic probability measures is closed. To define the property, which is responsible for this behaviour, we recall that given a unitary representation (π, \mathcal{H}) of a locally compact group G and a function $\chi \in C_c(G)$ the operator $\pi(\chi)$ on \mathcal{H} is defined by the formula

$$\langle v, \pi(\chi)w \rangle := \int_G \chi(g) \langle v, \pi(g)w \rangle d\mu_G(g).$$

Now we can define:

Definition 3.3.2. A locally compact group G has *Kazhdan's property (T)* if there exists $\theta > 0$ such that for all $\chi \in C_c(G)$ with $\chi(g) = \overline{\chi(g^{-1})}$ and $\int \chi d\mu_G = 1$ and for every unitary representation (π, \mathcal{H}) of G without G -invariant vectors

$$\|\pi(\chi)\| \leq 1 - \theta.$$

To explain the notation (T), we remark that the above definition says precisely that the trivial representation \mathbb{T} is isolated (indicated by the brackets (\cdot)) in the space of all unitary G -representations (with respect to the Fell topology). Every semisimple real Lie group of rank ≥ 2 has property (T). Moreover, if a locally compact group G has property (T), then the same is true for every lattice Γ in G . A non-trivial reformulation of property (T) as a fixed point property can be given as follows: Recall that the isometry group of a Hilbert space \mathcal{H} is given by the semidirect product of $U(\mathcal{H})$ with \mathcal{H} . A continuous homomorphism into this group is called an *isometric action* on \mathcal{H} . Then we have:

Proposition 3.3.3 (Delorme-Serre). *A locally compact group G has property (T) if and only if every isometric action on a Hilbert space has a fixed point.*

For groups with property (T) the answer to our initial question is positive:

Proposition 3.3.4 (E. Glasner-B. Weiss). *Let G be a locally compact group with property (T). If G acts continuously on a compact metric space X , then the set of ergodic measures is closed.*

PROOF. Suppose μ_n is a family of ergodic measure converging to μ and let $f \in C(X)$. Then for $\chi \in C_c(G)$ we can define the convolution

$$(\chi * f)(x) := \int f(g^{-1}x) \chi(g) d\mu_G(g) \in C(X).$$

Now assume that χ is as in the definition of property (T). Then we have in particular $\|\chi\|_1 = 1$ and thus

$$\|\chi * f - \int f d\mu_n\|_2 = \|\chi * (f - \int f d\mu_n)\|_2 \leq (1 - \theta) \cdot \|f - \int f d\mu_n\|_2.$$

Taking weak-* limits this yields

$$\|\chi * f - \int f d\mu\|_2 \leq (1 - \theta) \cdot \|f - \int f d\mu\|_2.$$

Now if f is G -invariant then $\chi * f = f$; hence the above inequality implies $\|f - \int f d\mu\|_2 = 0$. This shows that f is constant μ -almost everywhere, hence μ is ergodic. \square

3.4. The Eskin-Mozes-Shah theorem

Somehow the versions from the two talks differ... We have to clarify which version we want to state here.

3.5. Ratner's measure classification theorem

Recall that the good case of the Eskin-Mozes-Shah theorem produces for us a measure, which is invariant under a unipotent one-parameter subgroup. This raises the question on how to classify such subgroups. The answer to this question is provided by the famous measure classification theorem of M. Ratner, which we can state as follows:

Theorem 3.5.1 (Ratner's measure classification theorem). *Let G be a semisimple Lie group, $H < G$ a closed subgroup generated by unipotent one-parameter subgroups and μ an H -invariant ergodic probability measure on $\Gamma \backslash G$. Then there exists a closed connected subgroup L of G , which contains H , and $x_0 \in \Gamma \backslash G$ such that μ is the unique L -invariant probability measure supported on the orbit $x_0 L$.*

Given a subgroup L of G and a point $x_0 \in \Gamma \backslash G$, we will denote the unique L -invariant probability measure supported on $x_0 L$ by $\mu_{x_0 L}$. A probability measure of this form will be called *algebraic* for L .

The proof of Ratner's theorem in the general case is rather involved. However, many ideas can already be seen in the case where $H = SL_2(\mathbb{R})$. We will thus provide a proof in the $SL_2(\mathbb{R})$ -case, referring to [] for the general case. In the language introduced above we thus have to show:

Theorem 3.5.2 (Ratner's measure classification theorem, $SL_2(\mathbb{R})$ -case). *Let G be a semisimple Lie group, $H < G$ a closed subgroup isomorphic to $SL_2(\mathbb{R})$ and μ an H -invariant ergodic probability measure on $\Gamma \backslash G$. Then μ is algebraic for some group L containing H .*

Note that as a consequence of the Howe-Moore theorem, every subgroup of $SL_2(\mathbb{R})$ acts ergodically on $\Gamma \backslash G$ with respect to μ (see Proposition ??). This applies in particular to the unipotent subgroup N^+ , which consists of the elements

$$(3.2) \quad u_t := \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \quad (t \in \mathbb{R}).$$

The action of this subgroup will be at the core of the proof of Theorem 3.5.2. We now present the general strategy of the proof. Since G is finite-dimensional, it clearly suffices to establish the following:

Lemma 3.5.3. *In the situation of Theorem 3.5.2 assume that μ is invariant under some subgroup $L \supset H$. Then either μ is algebraic for L or there exists some closed connected group L' of larger dimension than L such that μ is L' -invariant.*

This suggests a strategy for the proof: Assume that μ is invariant under L , but not algebraic for L . Then we have to find a one-parameter group in G , which lies not in L and preserves μ . Since G is semisimple, the Killing form on \mathfrak{g} is non-degenerate; this suggests to look for the generator of the one-parameter subgroup in question in the orthogonal complement \mathfrak{l}^\perp of the Lie algebra \mathfrak{l} of L with respect to the Killing form. Now, how do we find this generator?

The key idea of Ratner is to make use of the unipotent flow given by the $\{u_t\}$ -action, where u_t is as in (3.2). More precisely, we will study the adjoint action of this one-parameter group on \mathfrak{g} . Thanks to the explicit representation theory of $SL_2(\mathbb{R})$, it is easy to analyze this action. The way this adjoint action comes into play is as follows: Assume we start with to point $x, x' \in X$, where here and in the sequel $X := \Gamma \backslash G$. Let us further assume that these two points are close together; then there exists $V \in \mathfrak{g}$ with

$$(3.3) \quad x' = x \exp(V).$$

Now let both points flow along the given unipotent flow for some time t and compare the endpoints $u_t.x$ and $u_t.x'$. We then have

$$u_t.x' = x'u_{-t} = x \exp(V)u_{-t} = xu_t \exp(\text{Ad}(u_t)(V)) = (u_t.x) \exp(\text{Ad}(u_t)(V))$$

Thus, if the point started out at logarithmic distance V , they end up at logarithmic distance $\text{Ad}(u_t)(V)$. In particular, if V is contained in the Lie algebra \mathfrak{c} of the centralizer $C_G(\{u_t\})$, then this distance is constant, i.e. the points x and x' do not diverge.

To make use of the unipotent flow, we will use the following uniform version of the Birkhoff ergodic theorem:

Proposition 3.5.4 (Birkhoff ergodic theorem, uniform version). *Let X be a locally compact, second countable space and let μ be a probability space, which is ergodic under a one-parameter group u_t . Then there exists a subset $\text{gen}(\mu) \subset X$ with $\mu(\text{gen}(\mu)) = 1$ such that for all $x \in \text{gen}(\mu)$ and for every $f \in C_c(X)$*

$$\frac{1}{T} \int_0^T f(xu_t) dt \rightarrow \int f d\mu$$

The elements of $\text{gen}(\mu)$ are called μ -generic points. It turns out that every pair of non-diverging generic points gives rise to an invariance of the measure μ :

Proposition 3.5.5 (Furstenberg argument). *Let $x', x \in X$ as in (3.3). Assume that $V \in \mathfrak{c}$, i.e. x' and x do not diverge, and that x and x' are moreover μ -generic. Then μ is invariant under $c := \exp(V)$.*

PROOF. Let $f \in C_c(\Gamma \backslash G)$ and $c := \exp(V)$. Then we compute

$$\begin{aligned} \int f(yc)d\mu(y) &= \int c^{-1}.f(y)d\mu(y) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T c^{-1}.f(xu_t)dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(xu_t c)dt = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(xcu_t)dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(x'u_t)dt = \int f(y)d\mu(y). \end{aligned}$$

□

We have thus reduced the proof of Theorem 3.5.2 to the following proposition:

Proposition 3.5.6. *In the situation of Theorem 3.5.2 assume that μ is invariant under some subgroup $L \supset H$, but not algebraic for L . Then there exists a ray $\mathbb{R} \cdot V \subset \mathfrak{l}^\perp \cap \mathfrak{c}$ such that for every $W \in \mathbb{R} \cdot V$ there exist $x, x' \in \text{gen}(\mu)$ with $x' = x \exp(W)$.*

Indeed, if the proposition holds, then μ is invariant under every $\exp(W)$ for $W \in \mathbb{R} \cdot V$, and thus invariant under $L' = \langle L, \exp(\mathbb{R} \cdot V) \rangle$; since $V \in \mathfrak{l}^\perp$, L' has larger dimension than L , and hence Lemma 3.5.3 and Theorem 3.5.2 will follow. We are thus left with proving Proposition 3.5.6.

3.5.1. Ratner's (H)-principle. For the proof of Proposition 3.5.6 we need to produce many pairs of non-diverging generic points whose distance is orthogonal to the $\{u_t\}$ -orbits. Again, these can be found using properties of the unipotent flow. It is tempting to start with two μ -generic point x_1, x_2 such that $x_2 = x_1 \exp(V_\epsilon)$ for some $V_\epsilon \in B_\epsilon^{\mathfrak{l}^\perp}(0)$ (i.e. the ball of radius epsilon around 0 in \mathfrak{l}^\perp) and just to let them flow. By general principles about unipotent flows the fastest growing direction of divergence will be invariant under the flow. This means that the logarithmic distance between the two points after time t can be written as

$$\text{Ad}(u_t)(V_\epsilon) = W(t, \epsilon) + R(t, \epsilon),$$

where $W(t, \epsilon)$ is the fastest growing direction of divergence and $R(t, \epsilon)$ is some remainder term. Note that being U -invariant is the same as being in \mathfrak{c} by definition. Thus the W part is the good direction that we want to extract, while the R part is some junk. Now the problem arises to separate the two parts. While W will become dominating for large t , we cannot simply let t go to ∞ , since $\text{Ad}(u_t)(V_\epsilon)$ will diverge. The situation will be better controlled if we choose ϵ smaller, but again we cannot just let $\epsilon \rightarrow 0$, because then $\text{Ad}(u_t)(V_\epsilon) \rightarrow 0$ and all information is lost. Instead we need to find for every fixed positive ϵ a time interval $[t_0(\epsilon), t_1(\epsilon)]$ such that for all $t \in [t_0(\epsilon), t_1(\epsilon)]$ the quantity $W(t, \epsilon)$ is large, while $R(t, \epsilon)$ is small. for this $t = t(\epsilon)$ we can then let $\epsilon \rightarrow 0$.

The existence of such a good time scale is far from obvious, but true in quite some generality. This is Ratner's famous (H)-principle. One way to formalize this in the present setting is the following:

Lemma 3.5.7 (Ratner's (H)-principle). *Let $n := \dim \mathfrak{l}^\perp$. Then for all $\eta > 0, \epsilon > 0$ there exist $y, y' \in X$, $t > 0$ and $V \in \mathfrak{l}^\perp$ with the following properties:*

- (i) $\|V\| < \epsilon$.

- (ii) $\{u_t \cdot y, u_t \cdot y'\} \subset \text{gen}(\mu)$.
- (iii) *There exists $W \in \mathfrak{g}$ such that*

$$\|\text{Ad}(u_t)(V) - W\| = O(\epsilon^{\frac{1}{n}})$$

and

$$c\epsilon \leq \|W\| \leq C\epsilon,$$

where c and C are universal constants depending only on \mathfrak{l}^\perp .

The hope is that $u_{t(\epsilon)}y_\epsilon$ and $u_{t(\epsilon)}y'_\epsilon$ can be chose in such a way that they converge as $\epsilon \rightarrow 0$ to some generic x_η and x'_η . If this is the case and $x'_\eta = x_\eta \exp(W_\eta)$, then (iii) will guarantee $W_\eta \neq 0$, and W_η is our good direction. Of course, there are many issues with this argument, in particular, the set of generic points need not be closed. Thus we need to formulate and prove a slight strengthening of Ratner's (H)-principle in order to be able to deduce Proposition 3.5.6. Namely, we want to make sure that certain pairs of points spent a big proportion of their time within the generic set of μ . To this end we establish:

Lemma 3.5.8. *There exists a compact subset $K \subset \text{gen}(\mu)$ and $T_0 > 0$ with the following properties;*

- (i) $\mu(K) > 0.9$;
- (ii) *If*

$$X_1 := \{y \in X \mid \forall T > T_0 : \frac{1}{T} \int_0^T \mathbf{1}_K(yu_t) dt > 0.8\},$$

then $\mu(X_1) > 0.99$.

- (iii) *If B_1^L denotes the unit ball in L and μ_L the Haar measure of L , then*

$$X_2 := \{z \in X \mid \frac{1}{m(B_1^L)} \int_{B_1^L} \mathbf{1}_{X_1}(l \cdot z) d\mu_L(l) > 0.9\},$$

then $\mu(X_2) > 0.9$.

Bibliography

- [BH99] Martin R. Bridson and André Haefliger, *Metric spaces of non-positive curvature*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 319, Springer-Verlag, Berlin, 1999.
- [BP92] R. Benedetti and C. Petronio, *Lectures on hyperbolic geometry*, Universitext, Springer-Verlag, Berlin, 1992. MR MR1219310 (94e:57015)
- [BT72] F. Bruhat and J. Tits, *Groupes réductifs sur un corps local*, Inst. Hautes Études Sci. Publ. Math. (1972), no. 41, 5–251.
- [BV07] Johannes Buchmann and Ulrich Vollmer, *Binary quadratic forms - an algorithmic approach*, Algorithms and Computation in Mathematics, vol. 20, Springer, Berlin, 2007.
- [CM09] P.-E. Caprace and N. Monod, *Isometry groups of non-positively curved spaces: structure theory*, J. Topol. **2** (2009), no. 4, 661–700.
- [Ein06] Manfred Einsiedler, *Ratner's theorem on $SL(2, \mathbb{R})$ -invariant measures*, Jahresber. Deutsch. Math.-Verein. **108** (2006), no. 3, 143–164.
- [EM93] Alex Eskin and Curt McMullen, *Mixing, counting, and equidistribution in Lie groups*, Duke Math. J. **71** (1993), no. 1, 181–209.
- [Hel01] Sigurdur Helgason, *Differential geometry, Lie groups, and symmetric spaces*, Graduate Studies in Mathematics, vol. 34, American Mathematical Society, Providence, RI, 2001, Corrected reprint of the 1978 original.
- [Kna02] A. W. Knaapp, *Lie groups beyond an introduction*, second ed., Progress in Mathematics, vol. 140, Birkhäuser Boston Inc., Boston, MA, 2002.
- [Lan99] Serge Lang, *Math talks for undergraduates*, Springer-Verlag, New York, 1999.
- [Rag72] M. S. Raghunathan, *Discrete subgroups of Lie groups*, Springer-Verlag, New York, 1972.
- [RS92] Andrew M. Rockett and Peter Szűsz, *Continued fractions*, World Scientific Publishing Co. Inc., River Edge, NJ, 1992.