Homogeneous Dynamics

Davide Ravotti

October 25, 2023

Welcome to the *Homogeneous Dynamics* course!

These lectures are intended to be an introduction to homogeneous dynamics, which nowadays is a very active subject of research. Homogeneous dynamics lies at the intersection of many areas in pure mathematics: of course, dynamics and ergodic theory, but also geometry, Lie group theory, representation theory, and more. There are also remarkable connections to several problems in number theory, some of which will be explored during the course.

The literature in the subject is vast and it would be impossible to cover it all. The choice I made to select the specific topics which will be discussed during these lectures was motivated mainly by two reasons. In part, of course, there are my personal preferences; more importantly, I wanted to focus on concrete examples (where computations can be carried out explicitly) which can help to build the intuition and provide insights on more general and abstract situations. It is my hope that this introduction can sparkle the curiosity in students to pursue this line of research.

One final disclaimer before starting: these lecture notes are a work-in-progress, and as such they need to be read with critical thinking. I tried to minimize the number of errors, but it would be widely optimistic of me to believe that there are none. If you spot mistakes, or have any comment in general, please let me know by sending me an email to `davide.ravotti@gmail.com`.

Davide Ravotti

Chapter 1

A quick recap: the case of linear flows on tori

In this first chapter, we will quickly review some basic notions in dynamics and ergodic theory, which the reader is assumed to be already familiar with. An exhaustive treatment of these topics can be found, for example, in [4, Chapters 2, 4.3].

In parallel, we will look at linear flows on tori. Very roughly speaking, the course consists in studying their non-Abelian analogues, as we will see later. Thus, focusing on this simple case can be a nice "warm-up" exercise.

1.1 Smooth flows on manifolds

The subject of this course is a special class of smooth flows. Let us recall the general definition.

Definition 1.1. *Let M be a smooth manifold, and let* $\mathrm{Diff}(M)$ *be the group of its diffeomorphisms. A* smooth flow $\varphi\colon \mathbb{R}\times M \to M$ *is a smooth map which satisfies*

$$\varphi_0 = \mathrm{Id}, \quad \textit{and} \quad \varphi_{t+s} = \varphi_t \circ \varphi_s = \varphi_s \circ \varphi_t, \quad \textit{for all } t,s \in \mathbb{R},$$

where $\varphi_t := \varphi(t,\cdot) \in \mathrm{Diff}(M)$.

In particular, Definition 1.1 implies that the continuous curve $t \mapsto \varphi_t$ is a group homomorphism between $\mathbb{R}$ and $\mathrm{Diff}(M)$, and $\{\varphi_t\}_{t\in\mathbb{R}}$ is said to be a *1-parameter group* of diffeomorphisms. We will often identify φ with $\{\varphi_t\}_{t\in\mathbb{R}}$.

Given a smooth flow φ, we can define a vector field X on M by

$$Xf(p) := \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f\circ\varphi_t(p), \quad \text{for all } f \in \mathscr{C}^\infty(M) \text{ and } p \in M.$$

The vector field X is called the *infinitesimal generator* of φ. Vice-versa, one can prove that, at least when M is compact, for any given smooth vector field X, there exists a unique smooth flow φ with infinitesimal generator X.

From here onward, M always denotes a smooth manifold, not necessarily compact, and φ is a smooth flow on M.

Let us turn to a very concrete example. Let $\mathbb{T}^n$ be the n-dimensional torus $\mathbb{T}^n := \mathbb{R}^n/\mathbb{Z}^n$. We will denote points in $\mathbb{T}^n$ using the symbol $[\![\cdot]\!]$, namely $[\![\mathbf{x}]\!] := \mathbf{x}+\mathbb{Z}^n$. For any $\mathbf{v}\in\mathbb{R}^n\setminus\{0\}$, we define the *linear flow in direction* $\mathbf{v}$ to be the smooth flow $\varphi^{\mathbf{v}}$ on $\mathbb{T}^n$ given by

$$\varphi^{\mathbf{v}}_t([\![\mathbf{x}]\!]) = [\![\mathbf{x}+t\mathbf{v}]\!], \quad \text{for } t\in\mathbb{R}.$$

It is easy to check that indeed $\varphi^{\mathbf{v}}$ is a well-defined smooth flow according to Definition 1.1. The associated infinitesimal generator X is the derivative in direction $\mathbf{v}$: for any $p = [\![\mathbf{x}]\!] \in \mathbb{T}^n$,

$$Xf(p) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f([\![\mathbf{x}+t\mathbf{v}]\!]) = \mathbf{v}\cdot\nabla_{[\![x]\!]} f = \sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}([\![x]\!]).$$

In other words, under the usual identification of the tangent space $T_p\mathbb{T}^n$ at p with $\mathbb{R}^n$, we have $X = \mathbf{v}$. The associated 1-parameter subgroup consists of the translations $\varphi_t^{\mathbf{v}}\colon [\![\mathbf{x}]\!] \mapsto [\![\mathbf{x}+t\mathbf{v}]\!]$ in direction $\mathbf{v}$.

Let us rephrase the example above in more algebraic terms. Our setting was the following. We considered the Abelian group $(\mathbb{R}^n,+)$, and we fixed a 1-dimensional subgroup $V = \{t\mathbf{v} : t \in \mathbb{R}\} < \mathbb{R}^n$. This subgroup V is everywhere tangent to the constant vector field $\mathbf{v} \in \mathbb{R}^n$, where we identified $\mathbb{R}^n = T_{\mathbf{x}}\mathbb{R}^n$ for all $\mathbf{x} \in \mathbb{R}^n$. In turn, V is identified with the 1-parameter group of translations

$$\{(\mathbf{x} \mapsto \mathbf{x}+t\mathbf{v}) : t \in \mathbb{R}\} \subset \mathrm{Diff}(\mathbb{R}^n).$$

We then fixed the discrete subgroup $\mathbb{Z}^n < \mathbb{R}^n$ and we considered the quotient space $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$. The key observation is that the 1-parameter group of translations $\mathbf{x} \mapsto \mathbf{x}+t\mathbf{v}$ associated to $\mathbf{v}$ *descends to the quotient*, which means that they commute with the canonical projection $\mathbf{x} \mapsto [\![\mathbf{x}]\!] = \mathbf{x}+\mathbb{Z}^n$. This tells us that, under the projection, we obtain a well-defined 1-parameter group of diffeomorphisms of $\mathbb{T}^n$, and hence a smooth flow $\varphi^{\mathbf{v}}$.

Homogeneous flows, which are the subject of this course, are a "non-Abelian" generalization of this simple example. Namely, we will replace

- $\mathbb{R}^n$ with a Lie group G (the Heisenberg group in Chapter 3 and $\mathrm{SL}(2,\mathbb{R})$ in Chapters 4–7),
- $\mathbb{Z}^n$ with a *lattice* Γ (a discrete subgroup of G with some additional properties that we will discuss in §2.4),
- $\mathbb{T}^n$ with the *left*[1] quotient $\Gamma\backslash G = \{\Gamma g : g \in G\}$,
- $V = \{t\mathbf{v} : t \in \mathbb{R}\}$ with a 1-parameter subgroup $\{g_t : t \in \mathbb{R}\}$ of G (generated by a "constant" vector field, which in the case above was $\mathbf{v} \in \mathbb{R}^n \setminus \{0\}$),
- $\varphi_t^{\mathbf{v}}\colon [\![\mathbf{x}]\!] \mapsto [\![\mathbf{x}+t\mathbf{v}]\!]$ with the multiplication *on the right*[2] $\Gamma g \mapsto \Gamma g \cdot g_t$.

We will make this analogy precise in the next chapters.

1.2 The topology of orbits

Let $\varphi\colon \mathbb{R}\times M \to M$ be a smooth flow, and let $p \in M$. The *orbit* of p is the set

$$\mathcal{O}_\varphi(p) = \{\varphi_t(p) : t \in \mathbb{R}\} \subset M.$$

Note that the orbit of any point $p \in M$ is an *immersed* smooth curve in M.

In dynamics, one is interested in the behaviour of orbits: do they "close up"? Do they accumulate in some regions? Do they visit all parts of the space? From the topological point of view, it is particularly important to try to understand their accumulation points and closure $\overline{\mathcal{O}_\varphi(p)} \subseteq M$.

[1] Note that, in the Abelian case, left and right cosets coincide.

[2] Taking left quotients and multiplying on the right is the conventional choice, but of course one could do the opposite (taking right quotients and multiplying on the left). Note that, again, multiplying on the right and projecting on the left quotient $\Gamma\backslash G$ commute.

Definition 1.2. *A point p is a* fixed point *if $\mathcal{O}_\varphi(p) = \{p\}$. A point p is* periodic *if there exists $T > 0$ such that*

$$\varphi_T(p) = p. \tag{1.1}$$

If p is periodic but not a fixed point, its period *is the smallest $T > 0$ for which* (1.1) *holds.*

Exercise 1.3. (a) *Show that the set of $T \in \mathbb{R}$ for which* (1.1) *holds is a subgroup of $\mathbb{R}$, in particular if p is periodic but not a fixed point, its period is well-defined.*

(b) *Show that, if p is a periodic point of period T, then its orbit is an embedded closed curve and*

$$\mathcal{O}_\varphi(p) = \{\varphi_t(p) : t \in [0,T]\}.$$

Periodic and fixed points have the smallest possible orbit closures, since their orbits are themselves closed. On the opposite, we may have points with *dense* orbits, that is, points whose orbit closure is the largest possible.

Definition 1.4. *A smooth flow φ is* minimal *is all orbits are dense, namely if*

$$\overline{\mathcal{O}_\varphi(p)} = M, \quad \textit{for all } p \in M.$$

Let us look at our motivating example. In the case of linear flows on the two dimensional torus, we have a pleasant dichotomy.

Theorem 1.5. *Let $\varphi^{\mathbf{v}} \colon \mathbb{T}^2 \to \mathbb{T}^2$ be a linear flow in direction $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2 \setminus \{0\}$. If v_1 and v_2 are* rationally dependent*, then every orbit is periodic; otherwise, if v_1 and v_2 are* rationally independent*, the flow $\varphi^{\mathbf{v}}$ is minimal.*

We will say that $\varphi^{\mathbf{v}}$ is a *rational* linear flow if we are in the first case, and it is an *irrational* linear flow if we are in the second one.

Before diving into the proof of Theorem 1.5, let us make a couple of simple observations. First, note that a rescaling $a\mathbf{v}$ of $\mathbf{v}$ for some $a > 0$ does not change the behaviour of the orbits of the flow. If $v_2 = 0$, then $v_1 \neq 0$. It is clear that all orbits of $\varphi^{\mathbf{v}}$ are periodic of period $1/v_1$ and consist of horizontal cirlces of the form $\mathbb{T}^1 \times \{p_2\}$, with $p_2 \in \mathbb{T}^1$, hence the result is proved in this case. If $v_2 \neq 0$, then, without loss of generality, we can assume that $\mathbf{v} = (v, 1)$. We divide the proof of Theorem 1.5 into two cases: when $v \in \mathbb{Q}$ (the rational case) and when $v \notin \mathbb{Q}$ (the irrational case).

Proof of Theorem 1.5 - *Case $v \in \mathbb{Q}$.* Let us write $v = a/b$ in reduced terms. Then, we claim that all orbits are periodic of period b. Indeed, let $p = [\![x_1, x_2]\!] \in \mathbb{T}^2$. Then,

$$\varphi^{\mathbf{v}}_b(p) = [\![x_1 + b \cdot a/b, x_2 + b]\!] = [\![x_1, x_2]\!] + [\![a, b]\!] = p.$$

If $T > 0$ is such that $\varphi^{\mathbf{v}}_T(p) = p$, then, looking at its second coordinate, we see that $x_2 + T + \mathbb{Z} = x_2 + \mathbb{Z}$. Hence $T \in \mathbb{N}$, and, looking at the first coordinate, $x_1 + T \cdot a/b + \mathbb{Z} = x_1 + \mathbb{Z}$. This implies that $(Ta)/b \in \mathbb{Z}$. Since a and b are coprime by assumption, b divides T. This proves the claim and hence the theorem in the rational case. □

Proof of Theorem 1.5 - *Case $v \notin \mathbb{Q}$.* We first claim that it is enough to prove the following statement.

($\star$) The *circle rotation* $R_v \colon \mathbb{T}^1 \to \mathbb{T}^1$ defined by $R_v([\![x]\!]) = [\![x + v]\!]$ is *minimal* (where, here, $[\![x]\!] = x + \mathbb{Z}$).

We leave as an exercise to the reader to check that indeed it is sufficient to prove $(\star)$. The idea is that the orbit of a point $p = [\![x_1, x_2]\!]$ under the flow $\varphi^{\mathbf{v}}$ is dense in $\mathbb{T}^2$ if and only if its intersection with the horizontal circle $\mathbb{T}^1 \times \{[\![x_2]\!]\}$ is dense in $\mathbb{T}^1 \times \{[\![x_2]\!]\}$. Indeed, the projection on the first coordinate of the intersection of the orbit of p with the circle $\mathbb{T}^1 \times \{[\![x_2]\!]\}$ is precisely the orbit of $[\![x_1]\!]$ under the rotation R_v.

We now focus on proving $(\star)$. Let $p = [\![x]\!] = x + \mathbb{Z} \in \mathbb{T}^1$ and $\varepsilon > 0$ be fixed; choose a natural number $N \geq \varepsilon^{-1}$ and partition $\mathbb{T}^1 \approx [0,1)$ into N intervals $I_k = [(k-1)N^{-1}, kN^{-1})$ for $k = 1, \dots, N$. We need to show that the orbit of p visits all intervals I_k.

Let us consider the set $O_N = \{p, R_v(p), \dots, R_v^N(p)\}$. Since $|O_N| = N+1$, by the Pigeonhole Principle, there exists a $\bar{k} \in \{1, \dots, N\}$ such that the interval $I_{\bar{k}}$ contains at least two distinct elements of O_N, say $R_v^n(p)$ and $R_v^m(p)$, with $n < m$. Let us call w the fractional part of $(m-n)v$. For any $y \in [0,1)$, we have

$$R_v^{m-n}([\![y]\!]) = [\![y + (m-n)v]\!] = [\![y + w]\!] = R_w([\![y]\!]),$$

namely, the map R_v^{m-n} is again a rotation of angle $w \in (0,1)$. Since we showed that the points $p' = R_v^n(p)$ and $R_w(p') = R_v^m(p)$ are both in the same interval $I_{\bar{k}}$, they are at distance less than N^{-1}. It follows that $0 < w < N^{-1} \leq \varepsilon$. Thus, the orbit of p under R_v contains the orbit of p under $R_v^{m-n} = R_w$, which is a rotation of angle less than ε. Since this latter set clearly intersects all intervals I_k, the proof is complete. □

In general, it is a hopeless task to try to understand all orbit closures. They can be quite complicated objects, with "fractal-like" structures and non-integer dimensions. However, in the particular case of linear flows on $\mathbb{T}^2$, orbit closures are well-behaved and we managed to classify all possibilities: we showed that all orbit closures are either the whole space $\mathbb{T}^2$ or circles isomorphic to $\mathbb{T}^1$. In higher dimensions, a similar phenomenon occurs: orbit closures of any linear flow on $\mathbb{T}^n$ are sub-tori isomorphic to $\mathbb{T}^k$, for some $k = 1, \dots, n$ (see Section 1.3.4 below).

1.3 Elements of Ergodic Theory

Ergodic theory is the study of dynamical systems from the point of view of measure theory. The measures on the phase space M that will be relevant for us are Borel invariant measures.

1.3.1 Invariant measures

Definition 1.6. *Let φ be a smooth flow on M. A Borel measure μ on M is an* invariant measure *for φ if for all Borel measurable sets $A \subset M$ and for all $t \in \mathbb{R}$,*

$$\mu(\varphi_t(A)) = \mu(A).$$

If $\mu(M) = 1$, then μ is a probability invariant measure. *The triple (M, φ, μ) is called a* probability preserving flow *(*ppf*, for short).*

The previous definition extends to all functions in $L^1(M) = L^1(M, \mu)$: if (M, φ, μ) is a ppf, then, for every function $f \in L^1(M)$ and for all $t \in \mathbb{R}$, the function $f \circ \varphi_t$ is in $L^1(M)$ and

$$\int_M f \circ \varphi_t \, \mathrm{d}\mu = \int_M f \, \mathrm{d}\mu.$$

Similarly, if $f \in L^2(M)$, then $f \circ \varphi_t \in L^2(M)$ for all $t \in \mathbb{R}$ and

$$\|f \circ \varphi_t\|_2 = \|f\|_2. \tag{1.2}$$

Let us see some examples of invariant measures. Clearly, the Lebesgue measure on the torus $\mathbb{T}^2$ is an invariant measure for all linear flows $\varphi^{\mathbf{v}}$. If the flow is irrational, we will see in Section 1.3.3 that there are no other invariant probability measures. However, if $\varphi^{\mathbf{v}}$ is rational, then we have uncountably many invariant probability measures supported on periodic orbits. This is a general fact: for any periodic orbit, there is an invariant probability measure supported on such orbit.

Exercise 1.7. (a) *Let $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2 \setminus \{0\}$, with v_1, v_2 rationally dependent. Let T be the period of all orbits of $\varphi^{\mathbf{v}}$. Show that*

$$T = \frac{\min\{\|\mathbf{w}\|_2 : \mathbf{w} \in \mathbb{R}\mathbf{v} \cap \mathbb{Z}^2,\ \mathbf{w} \neq 0\}}{\|\mathbf{v}\|_2}.$$

(b) *For any $p \in M$, let μ_p be the Borel measure defined by*

$$\mu_p(A) := \frac{1}{T}\,\mathrm{Leb}\{t \in [0,T] : \varphi_t^{\mathbf{v}}(p) \in A\}.$$

Show that μ_p is a probability invariant measure for $\varphi^{\mathbf{v}}$.

(c) *Prove that $\mu_p = \mu_q$ if and only if $\mathcal{O}_{\varphi^{\mathbf{v}}}(p) = \mathcal{O}_{\varphi^{\mathbf{v}}}(q)$. Deduce that $\varphi^{\mathbf{v}}$ has uncountably many probability invariant measures.*

It is actually easy to see that if there is more than one probability invariant measure, then there are uncountably many. Indeed, any convex combination of (probability) invariant measures is a (probability) invariant measure. In other words, probability invariant measures form a *simplex* in the space of probability measures on M.

The reader might wonder whether we are sure to find, in general, at least one probability invariant measure. When M is compact, the following result answers this question affirmatively.

Theorem 1.8 (Krylov-Bogolyubov)**.** *Let φ be a smooth flow on the compact manifold M. There exists one invariant probability measure.*

Proof. Recall that, when M is compact, the set of Borel (signed) measures coincides with $\mathscr{C}(M)^*$, the weak-$*$ dual of $\mathscr{C}(M)$. Recall also that, by Banach-Alaoglu's Theorem, the unit ball in $\mathscr{C}(M)^*$, which contains all (positive) probability measures, is weakly-$*$ compact. Fix any $p \in M$, and consider the family of (positive) probability measures $\{\mu_T\}_{T\in\mathbb{R}}$ given by

$$\mu_T(f) := \frac{1}{T}\int_0^T f \circ \varphi_t(p)\,\mathrm{d}t, \qquad \text{for } f \in \mathscr{C}(M).$$

By compactness, there exists an increasing sequence $T_n \to \infty$ such that μ_{T_n} weakly-$*$ converges to a (positive) probability measure μ. We claim that μ is invariant. Let $f \in \mathscr{C}(M)$ and $r \in \mathbb{R}$; then,

$$\begin{aligned}
|\mu_{T_n}(f \circ \varphi_r) - \mu_{T_n}(f)| &= \frac{1}{T_n}\left|\int_0^{T_n} f \circ \varphi_{t+r}(p)\,\mathrm{d}t - \int_0^{T_n} f \circ \varphi_t(p)\,\mathrm{d}t\right| \\
&= \frac{1}{T_n}\left|\int_{T_n}^{T_n+r} f \circ \varphi_t(p)\,\mathrm{d}t - \int_0^{r} f \circ \varphi_t(p)\,\mathrm{d}t\right| \\
&\leq \frac{2r\,\|f\|_{\mathscr{C}(M)}}{T_n} \to 0.
\end{aligned}$$

Therefore,

$$0 = \lim_{n\to\infty} |\mu_{T_n}(f \circ \varphi_r) - \mu_{T_n}(f)| = |\mu(f \circ \varphi_r) - \mu(f)|,$$

which shows that μ is an invariant measure for φ. □

We will mostly be concerned with *smooth* invariant measures, namely measures given by integrating a volume form on M. In this case, we can check whether a smooth measure is invariant by computing its Lie derivative with respect to the infinitesimal generator of the flow.

Proposition 1.9. *Let φ be a smooth flow with infinitesimal generator X, and let μ be a smooth probability measure given by a volume form ω on M. Then μ is invariant if and only if $\mathscr{L}_X(\omega) = 0$, where $\mathscr{L}_X(\omega) = \mathrm{d}(i_X\omega)$ is the Lie derivative of ω with respect to X and i is the contraction operator.*

Proof. Let $(\varphi_t)^*$ denote the pull-back by φ_t. By definition of the Lie derivative,

$$\mathscr{L}_X(\omega) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} (\varphi_t)^*(\omega),$$

hence $(\varphi_t)^*(\omega) = \omega$ if and only if $\mathscr{L}_X(\omega) = 0$. By Cartan's formula,

$$\mathscr{L}_X(\omega) = \mathrm{d}(i_X\omega) + i_X(\mathrm{d}\omega) = \mathrm{d}(i_X\omega),$$

which follows from the fact that $\mathrm{d}\omega = 0$ since ω is a n-form, where $n = \dim(M)$. □

Exercise 1.10 (Invariant measures of time-changes)**.** *Let φ be a smooth flow on M with infinitesimal generator X, and let μ be a smooth probability invariant measure. Show that, for any smooth positive function $\alpha\colon M \to \mathbb{R}_{>0}$, the flow*[3] *generated by the vector field αX preserves the measure equivalent to μ with density $1/\alpha$.*

Once we have chosen a probability invariant measure, we can ask about the properties of *typical* points, in other words the properties that are satisfied up to exceptional sets of measure zero. A fundamental result is the recurrence theorem by Poincaré, which, roughly speaking, says that typical orbits will come back close to their initial point infinitely often.

Theorem 1.11 (Poincaré Recurrence Theorem)**.** *Let (M, φ, μ) be a ppf. If $A \subset M$ is a measurable (Borel) set, for* almost every *$p \in A$ there exists an increasing sequence $T_n \to \infty$ such that $\varphi_{T_n}(p) \in A$.*

1.3.2 Ergodicity and the Ergodic Theorems

Given a flow φ on M, we say that a measurable set $A \subset M$ is *invariant* if $\varphi_t(A) = A$ for all $t \in \mathbb{R}$. If (M, φ, μ) is a ppf and $A \subset M$ is an invariant set of positive measure, then we can consider the subsystem (A, φ, μ_A) given by the restriction of the flow φ to A with the conditional probability invariant measure defined by

$$\mu_A(B) := \mu(B \cap A)/\mu(A), \qquad \text{for any measurable set } B.$$

When we have an invariant set of positive measure, we can then reduce ourselves to study a "simpler" system. Intuitivley, the notion of ergodicity plays the role of "indecomposability" in the context of ppf's. That is to say, an ergodic ppf cannot be decomposed into non-trivial invariant subsystems.

Definition 1.12. *Let (M, φ, μ) be a ppf. We say that μ is* ergodic*, or that (M, φ, μ) is an ergodic flow*[4] *if for every invariant measurable set $A \subset M$ we have $\mu(A) = 0$ or $\mu(A) = 1$.*

We recall the following characterization of ergodicity.

[3] This flow is called the *time-change* generated by α.

[4] Sometimes, by a little abuse of notation, when the reference measure μ is clear from the context, we will say that φ is ergodic.

Proposition 1.13. *Let (M,φ,μ) be a ppf. The following are equivalent:*

1. *μ is ergodic,*
2. *for every measurable set $A \subset M$ such that $\mu(\varphi_t(A)\triangle A) = 0$ for all $t \in \mathbb{R}$, then $\mu(A) = 0$ or $\mu(A) = 1$,*
3. *if $f\colon M \to \mathbb{C}$ is a measurable function such that $f \circ \varphi_t = f$ almost everywhere for all $t \in \mathbb{R}$, then there exists $c \in \mathbb{C}$ such that $f = c$ almost everywhere,*
4. *if $f \in L^2(M)$ is an* invariant function*, namely if $f \circ \varphi_t = f$ in L^2 for all $t \in \mathbb{R}$, then there exists $c \in \mathbb{C}$ such that $f = c$ in L^2.*

Let us go back once more to the case of linear flows on tori and let us consider the ppf $(\mathbb{T}^2, \varphi^{\mathbf{v}}, \mathrm{Leb})$. It is easy to see that, if the flow $\varphi^{\mathbf{v}}$ is rational, then it is *not* ergodic. Indeed, any set of the form

$$A_r = \bigcup \{\mathcal{O}_{\varphi^{\mathbf{v}}}(p) : p = [\![x_1, 0]\!] \in \mathbb{T}^2 \text{ with } 0 \le x_1 \le r\}$$

is an invariant set of with $\mathrm{Leb}(A_r) = r$. Choosing $r \in (0,1)$ appropriately gives an example of a non-trivial invariant set, thus disproving ergodicity.

Exercise 1.14. (a) *Show that the measures μ_p of Exercise* 1.7 *are ergodic.*

(b) *Show that any non-trivial convex combination of μ_p and μ_q, for p and q on different orbits, is not ergodic.*

(c*) *Finally, show that if μ is an ergodic invariant probability measure, then $\mu = \mu_p$ for some $p \in \mathbb{T}^2$.*

On the other hand, the Lebesgue measure is ergodic when the flow $\varphi^{\mathbf{v}}$ is irrational. There are several ways of proving this fact, here we see a proof that uses Fourier analysis.

Theorem 1.15. *Let $\varphi^{\mathbf{v}}$ be an irrational linear flow on $\mathbb{T}^2$. Then, the Lebesgue measure* Leb *is ergodic.*

Proof. We denote by $\cdot$ the scalar product in $\mathbb{R}^2$. For any $f \in L^2(\mathbb{T}^2)$, we can write a Fourier expansion

$$f([\![\mathbf{x}]\!]) = \sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i \mathbf{n}\cdot\mathbf{x}}, \quad \text{with} \sum_{\mathbf{n}\in\mathbb{Z}^2} |f_{\mathbf{n}}|^2 = \|f\|_2^2.$$

Assume that f is an invariant function, that is assume that $f \circ \varphi_t^{\mathbf{v}} = f$ for all $t \in \mathbb{R}$, where the equality holds in $L^2(\mathbb{T}^2)$. We want to show it is constant in L^2. For all $\mathbf{x} \in \mathbb{R}^2$ and $t \in \mathbb{R}$ we have

$$\sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i \mathbf{n}\cdot\mathbf{x}} = f([\![\mathbf{x}]\!]) = f([\![\mathbf{x}+t\mathbf{v}]\!]) = \sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i \mathbf{n}\cdot(\mathbf{x}+t\mathbf{v})} = \sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i t \mathbf{n}\cdot\mathbf{v}} e^{2\pi i \mathbf{n}\cdot\mathbf{x}}.$$

By uniqueness of the coefficients, we must have

$$f_{\mathbf{n}} = f_{\mathbf{n}}\, e^{2\pi i t \mathbf{n}\cdot\mathbf{v}} \quad \text{for all } \mathbf{n} \in \mathbb{Z}^2.$$

If $\mathbf{n} \neq 0$, then either $f_{\mathbf{n}} = 0$ or $e^{2\pi i t \mathbf{n}\cdot\mathbf{v}} = 1$ for all $t \in \mathbb{R}$, and this latter condition is verified if and only if $\mathbf{n}\cdot\mathbf{v} = 0$. Since $\mathbf{v}$ has rationally independent coordinates, this second possibility cannot occur; hence we deduce $f_{\mathbf{n}} = 0$ for all $\mathbf{n} \in \mathbb{Z}^2 \setminus \{0\}$. This proves that $f = f_0$ is equal to a constant in $L^2(\mathbb{T}^2)$, and thus completes the proof. □

Let (M,φ,μ) be an ergodic ppf. The ergodic theorems of Von Neumann and Birkhoff relate the *time averages* $\frac{1}{T}\int_0^T f \circ \varphi_t \,\mathrm{d}t$ of a measurable function $f \in L^2(M)$ (or $L^1(M)$) to the *space average* $\mu(f) = \int_M f \,\mathrm{d}\mu$.

Theorem 1.16 (Von Neumann Ergodic Theorem)**.** *Let (M,φ,μ) be a ppf. For every $f \in L^2(M)$, let $\mathrm{P} f \in L^2(M)$ be the projection of f onto the closed subspace of invariant functions. Then, the ergodic averages of f converge in $L^2(M)$ to $\mathrm{P} f$, namely*

$$\left\| \frac{1}{T}\int_0^T f\circ\varphi_t(p)\,\mathrm{d}t - \mathrm{P} f(p) \right\|_2 \to 0.$$

In particular, if (M,φ,μ) is ergodic, $\mathrm{P} f = \mu(f)$ and hence

$$\frac{1}{T}\int_0^T f\circ\varphi_t\,\mathrm{d}t \to \mu(f) \quad \text{in } L^2(M).$$

Theorem 1.17 (Birkhoff Ergodic Theorem)**.** *Let (M,φ,μ) be a ppf. For every $f \in L^1(M)$, there exists $f^* \in L^1(M)$ with*

$$\mu(f) = \mu(f^*), \quad \text{and} \quad f^*\circ\varphi_t = f^* \quad \text{for all } t \in \mathbb{R},$$

where the latter equality holds in $L^1(M)$, such that

$$\frac{1}{T}\int_0^T f\circ\varphi_t(p)\,\mathrm{d}t \to f^*(p),$$

for almost every *$p \in M$. If (M,φ,μ) is ergodic, then $f^*(p) = \mu(f)$ almost everywhere.*

1.3.3 Unique ergodicity

In Theorem 1.8, we saw that a smooth flow on a compact manifold M always has an invariant probability measure, and we also noticed that, if there is more than one, then there are uncountably many. The former case deserves a special name.

Definition 1.18. *Let φ be a smooth flow on a compact manifold M. If there exists only one invariant probability measure μ, the system (M,φ,μ) (or simply φ) is said to be* uniquely ergodic.

The reader might be wondering what the uniqueness of the invariant measure has to do with ergodicity. The following proposition shows that, in the case of a single invariant measure, ergodicity is automatically guaranteed.

Proposition 1.19. *Let φ be a smooth flow on a compact manifold M. The set of ergodic probability measures for φ coincides with the set of* extremal points[5] *of the simplex of invariant probability measures. In particular, if there exists a unique invariant probability measure μ, then it is ergodic.*

If (M,φ,μ) is uniquely ergodic, then, from the Ergodic Theorem, Theorem 1.17, we know that the ergodic averages of any L^1-function converge almost everywhere to its space average. On the other hand, one can show that, if the function is *continuous*, then the convergence is *uniform*.

Proposition 1.20. *Let φ be a smooth flow on a compact manifold M. The following are equivalent:*

1. *φ is uniquely ergodic,*
2. *there exists a unique ergodic invariant probability measure,*
3. *for every $f \in \mathscr{C}(M)$ there exists a constant C_f such that, for all $p \in M$,*

$$\frac{1}{T}\int_0^T f\circ\varphi_t(p)\,\mathrm{d}t \to C_f, \tag{1.3}$$

[5] A point in a simplex is extremal if it cannot be expressed as a non-trivial convex combination of two other points.

4. *for every $f \in \mathscr{C}(M)$, the convergence in* (1.3) *is uniform over M.*

Under any of the assumptions above, the constant C_f in (1.3) *equals $\mu(f)$, where μ is the unique invariant probability measure.*

We have seen already that for rational linear flows $\varphi^{\mathbf{v}}$ on $\mathbb{T}^2$ there exist uncountably many invariant measures. Let us now see that in the other case, when the coordinates of $\mathbf{v}$ are rationally independent, the flow is uniquely ergodic.

Theorem 1.21. *Let $\varphi^{\mathbf{v}}$ be an irrational linear flow on $\mathbb{T}^2$. Then, $(\mathbb{T}^2, \varphi^{\mathbf{v}}, \mathrm{Leb})$ is uniquely ergodic.*

Proof. Let $f \in \mathscr{C}(M)$ be fixed, and let us prove that the ergodic averages

$$A_T f(p) := \frac{1}{T}\int_0^T f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t$$

converge uniformly to $\mathrm{Leb}(f) = \int_{\mathbb{T}^2} f\,\mathrm{d}\,\mathrm{Leb}$. We claim that the family

$$\mathscr{A} := \{A_T f\}_{T>0} \subset \mathscr{C}(M).$$

is pre-compact in $\mathscr{C}(M)$, i.e., $\mathscr{A}$ has a compact closure. In order to do this, we check the assumptions of the Ascoli-Arzelà Theorem.

It is easy to see that $\mathscr{A}$ is equibounded: since $\|f \circ \varphi_t^{\mathbf{v}}\|_\infty = \|f\|_\infty$ for all $t \in \mathbb{R}$, it follows that, for any $T > 0$ and for all $p \in \mathbb{T}^2$, we have

$$|A_T f(p)| \leq \frac{1}{T}\int_0^T \|f\|_\infty\,\mathrm{d}t = \|f\|_\infty.$$

Let us verify that $\mathscr{A}$ is equicontinuous. We will use the fact that $\varphi_t^{\mathbf{v}}$ is an isometry for all $t \in \mathbb{R}$: if we denote by d the Euclidean distance on $\mathbb{T}^2$, we have that $d(\varphi_t^{\mathbf{v}}(p), \varphi_t^{\mathbf{v}}(q)) = d(p,q)$ for all $t \in \mathbb{R}$. With this in mind, let us fix $\varepsilon > 0$. Since f is uniformly continuous, there exists $\delta > 0$ such that $|f(p) - f(q)| < \varepsilon$ whenever $d(p,q) < \delta$. Then, for any $T > 0$, if $p, q \in \mathbb{T}^2$ are such that $d(p,q) < \delta$, we get

$$|A_T(p) - A_T(q)| \leq \frac{1}{T}\int_0^T |f \circ \varphi_t^{\mathbf{v}}(p) - f \circ \varphi_t^{\mathbf{v}}(q)|\,\mathrm{d}t < \frac{1}{T}\int_0^T \varepsilon\,\mathrm{d}t = \varepsilon.$$

By the Ascoli-Arzelà Theorem, the closure of $\mathscr{A}$ is compact in $\mathscr{C}(M)$, in particular $\mathscr{A}$ has limit points. Let $T_n \to \infty$ and $g \in \mathscr{C}(M)$ be such that

$$A_{T_n} f \to g \quad \text{in } \mathscr{C}(M).$$

By Birkhoff Ergodic Theorem, Theorem 1.17, for almost every point p we have

$$A_{T_n} f(p) \to \int_{\mathbb{T}^2} f\,\mathrm{d}\,\mathrm{Leb},$$

therefore $g = \mathrm{Leb}(f)$ almost everywhere. Since g is continuous, the equality must hold everywhere.

We have showed that all limit points of $\mathscr{A}$ are the constant function $\mathrm{Leb}(f)$. Therefore, the limit point is unique and we conclude that the whole family converges in $\mathscr{C}(M)$, namely

$$A_T f = \frac{1}{T}\int_0^T f \circ \varphi_t^{\mathbf{v}}\,\mathrm{d}t \to \int_{\mathbb{T}^2} f\,\mathrm{d}\,\mathrm{Leb}$$

uniformly on $\mathbb{T}^2$, which concludes the proof. □

We remark that the proof of Theorem 1.21 works in a greater generality: any isometry of a compact space which has an ergodic measure with full support is uniquely ergodic.

Exercise 1.22. *Let* $(\mathbb{T}^2, \varphi^{\mathbf{v}}, \mathrm{Leb})$ *be an irrational linear flow.*

(a) *Show that for any set* $A \subset \mathbb{T}^2$ with non-empty interior *there exists* $T_A > 0$ *such that* for all points $p \in M$ *there exists* $t \in [0, T_A]$ *such that* $\varphi_t^{\mathbf{v}}(p) \in A$ *(all points enter A before time* T_A*).*

(b*) *Provide a counterexample to* (a) *when we drop the assumption on A, namely give an example of a set* $A \subset \mathbb{T}^2$ with positive measure and empty interior *such that*

1. *almost every point enters A,*
2. *at least one point* $p \in M$ never *enters A,*
3. *for every* $T > 0$ *there exists a set* $B_T \subset \mathbb{T}^2$ of positive measure *such that all points in* B_T *do not enter A in the interval* $[0, T]$.

1.3.4 A glimpse at Ratner's Theorems

Let us summarize what we proved in the case of linear flows on the 2 dimensional torus:

- If the generator $\mathbf{v}$ has rationally independent coordinates, then
  1. the orbit closure of any point is the whole space $\mathbb{T}^2$ (Theorem 1.5),
  2. the orbit of any point equidistributes in $\mathbb{T}^2$ (Theorem 1.15),
  3. Leb is the only ergodic probability measure for $\varphi^{\mathbf{v}}$ (Theorem 1.21).
- If the generator $\mathbf{v}$ has rationally dependent coordinates, then
  1. all orbits are periodic, hence closed (Theorem 1.5),
  2. all orbits are not equidistributed in $\mathbb{T}^2$ (but, clearly, they equidistribute in their closure),
  3. any ergodic measure is the normalized Lebesgue measure on a periodic orbit (Exercise 1.14).

It is possible to generalize these results to linear flows on higher dimensional tori. Let us first recall some definitions.

A subspace $V < \mathbb{R}^n$ is called *rational* if the discrete Abelian group $V \cap \mathbb{Z}^n$ has rank precisely equal to $k := \dim(V)$. It is easy to see that this happens exactly when we can find a basis of V consisting of vectors in $\mathbb{Z}^n$. The subspace V carries a smooth measure, that we call Leb_V, given by the Lebesgue measure on V normalized so that the discrete subgroup $V \cap \mathbb{Z}^n$ has *covolume* 1 (see, e.g., Exercise 1.7). This measure descends to a measure on the k-dimensional torus $V/(V \cap \mathbb{Z}^n)$, as well as on its affine translates $\mathbf{x} + V/(V \cap \mathbb{Z}^n)$ for all $\mathbf{x} \in \mathbb{R}^n$. By a little abuse of notation, we will still call Leb_V any of these affine measures.

Theorem 1.23. *Let* $\varphi^{\mathbf{v}} \colon \mathbb{R} \times \mathbb{T}^n \to \mathbb{T}^n$ *be a linear flow on* $\mathbb{T}^n$*, with* $\mathbf{v} \in \mathbb{R}^n \setminus \{0\}$*. There exists a rational subspace* $V < \mathbb{R}^n$ *of dimension* $k \in \{1, \dots, n\}$ *which contains the line* $\mathbb{R}\mathbf{v}$ *for which the following holds.*

1. *(Orbit closure classification) The orbit closure of any point is an affine k-dimensional torus, namely for all* $p = [\![\mathbf{x}]\!] \in \mathbb{T}^n$ *we have*

$$\overline{\mathcal{O}_{\varphi^{\mathbf{v}}}(p)} = \mathbf{x} + V/(V \cap \mathbb{Z}^n).$$

2. *(Equidistribution) The orbit of any point $p \in \mathbb{T}^n$ equidistributes in its closure with respect to the affine measure* Leb_V.

3. *(Measure classification) Any ergodic measure for $\varphi^{\mathbf{v}}$ is an affine measure* Leb_V *on the affine torus $\mathbf{x}+V/(V\cap\mathbb{Z}^n)$ for some $\mathbf{x}\in\mathbb{R}^n$.*

Theorem 1.23 can be seen as a very simple case of a series of profound and general theorems by Marina Ratner which classify all possible orbit closures for unipotent actions, show that all orbits equidistribute in their closure, and prove that any ergodic measure is the affine translate of the Lebesgue (Haar) measure on a intermediate subgroup. The purpose of this complicated comment is only to whet you appetite for what will come in the rest of the course.

1.4 Further chaotic properties

1.4.1 Weak mixing

Let (M,φ,μ) be a ppf. As we have seen in (1.2), for every $t\in\mathbb{R}$ the *Koopman operator*

$$U_t\colon L^2(M)\to L^2(M),\quad U_t f = f\circ\varphi_t$$

is *unitary*. By Proposition 1.13, the flow φ is ergodic if and only if the eigenspace corresponding to the eigenvalue 1 has dimension 1, and consists of constant functions. Since U_t is unitary, if there are other eigenvalues, they must have modulus 1.

Definition 1.24. *We say that the ppf (M,φ,μ) is* weak mixing *if the only solutions to*

$$U_t f = e^{2\pi i t\alpha} f \quad \text{in } L^2(M) \text{ for all } t\in\mathbb{R}$$

are given by $\alpha=0$ and $f=c$ for some $c\in\mathbb{C}$.

As usual, when the reference measure μ is clear from the context, we will often simply say that φ is weak mixing when the condition in Definition 1.24 is satisfied.

Clearly, a weak mixing ppf is also ergodic. The converse, however, is not true, and a family of counterexamples is given precisely by our irrational linear flows.

Lemma 1.25. *Any linear flow $(\mathbb{T}^2,\varphi^{\mathbf{v}},\mathrm{Leb})$ is not weak mixing.*

Proof. It is sufficient to consider the irrational case, since we already know that rational linear flows are not ergodic and hence cannot be weak mixing. We claim that for any $\mathbf{n}\in\mathbb{Z}^2\setminus\{0\}$, the function

$$f_{\mathbf{n}}([\![\mathbf{x}]\!]) = e^{2\pi i\mathbf{n}\cdot\mathbf{x}} \in L^\infty(\mathbb{T}^2)\subset L^2(\mathbb{T}^2)$$

is a non-constant eigenfunction, and the α as in Definition 1.24 is $\alpha=\mathbf{n}\cdot\mathbf{v}\neq 0$. Indeed, for any $t\in\mathbb{R}$, we have

$$U_t f_{\mathbf{n}}([\![\mathbf{x}]\!]) = f_{\mathbf{n}}([\![\mathbf{x}+t\mathbf{v}]\!]) = e^{2\pi i\mathbf{n}\cdot(\mathbf{x}+t\mathbf{v})} = e^{2\pi i t\mathbf{n}\cdot\mathbf{v}}e^{2\pi i\mathbf{n}\cdot\mathbf{x}} = e^{2\pi i t\alpha} f_{\mathbf{n}}([\![\mathbf{x}]\!]).$$

Thus, irrational linear flows are ergodic but not weak-mixing. □

Weak-mixing is a *spectral* property, in the sense that it concerns the spectrum of the Koopman operators U_t of the system. If they have no pure point component (no eigenvalues), the flow is weak mixing. There are other equivalent characterizations of weak-mixing, which have a more "dynamical flavour"; we summarize them in Proposition 1.26 below.

Proposition 1.26. *Let (M,φ,μ) be a ppf. The following are equivalent.*

1. *(M,φ,μ) is weak mixing.*
2. *For any $f,g \in L^2(M)$,*
$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \left| \int_M f\circ\varphi_t \cdot \bar{g}\,\mathrm{d}\mu - \mu(f)\,\mu(\bar{g}) \right| \mathrm{d}t = 0.$$
3. *For any $f,g \in L^2(M)$,*
$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \left| \int_M f\circ\varphi_t \cdot \bar{g}\,\mathrm{d}\mu - \mu(f)\,\mu(\bar{g}) \right|^2 \mathrm{d}t = 0.$$
4. *For any $f,g \in L^2(M)$, there exists a set $J = J_{f,g} \subset \mathbb{R}$ of zero density such that*
$$\lim_{T\to\infty, T\notin J} \int_M f\circ\varphi_T \cdot \bar{g}\,\mathrm{d}\mu = \mu(f)\,\mu(\bar{g}).$$
5. *The product measure $\mu\times\mu$ is ergodic for the flow $\varphi\times\varphi$ on $M\times M$.*
6. *The product measure $\mu\times\mu$ is weak mixing for the flow $\varphi\times\varphi$ on $M\times M$.*
7. *For any ergodic ppf (N,ψ,ν), the system $(M\times N, \varphi\times\psi, \mu\times\nu)$ is ergodic.*

It might be worth for the reader to compare conditions 2–4 of Proposition 1.26 with the following equivalent definition of ergodicity.

Exercise 1.27. *Let (M,φ,μ) be a ppf. Show that it is ergodic if and only if for any $f,g \in L^2(M)$ we have*
$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \left(\int_M f\circ\varphi_t \cdot \bar{g}\,\mathrm{d}\mu \right) \mathrm{d}t = \mu(f)\,\mu(\bar{g}).$$
Deduce in particular that a weak mixing ppf is ergodic.

1.4.2 Mixing

Mixing, sometimes called strong mixing, is an even stronger property that, roughly speaking, says that any two events become asymptotically independent.

Definition 1.28. *We say that the ppf (M,φ,μ) is* (strong) mixing *if for any two observables $f,g \in L^2(M)$, the* correlations decay*, namely if*
$$\langle f\circ\varphi_t, g\rangle = \int_M f\circ\varphi_t \cdot \bar{g}\,\mathrm{d}\mu \to \mu(f)\,\mu(\bar{g}),$$
as $t\to\infty$.

It is clear from Proposition 1.26-(4) that any mixing ppf is also weak mixing. The converse, however, is not true: there are weak mixing ppf's which are not strong mixing. The first examples of weak mixing but not mixing transformations were constructed by cutting-and-stacking methods. In the context of flows, typical translation flows on translation surfaces and typical minimal area-preserving flows on higher genus surfaces are also natural classes of examples of weak mixing flows that are not mixing. It is also interesting to notice that, by the Halmos-Rokhlin Theorem, weak mixing is a *generic* property, whereas mixing is *meager*. In this course, however, the flows we will encounter are either not weak mixing (the nilflows in Chapter 3) or mixing (the geodesic and horocycle flows in Chapters 4–7).

Returning to our case study, we already know from Lemma 1.25 that irrational linear flows are not weak mixing, hence they cannot possibly be mixing. We can actually prove a stronger result, namely they have the so-called *rigidity property*.

Exercise 1.29. *Let* $\mathbf{v} = (v, 1) \in \mathbb{R}^2$*, with* $v \notin \mathbb{Q}$*.*

(a) *Prove that the linear flow* $\varphi^{\mathbf{v}}$ *is* rigid*, namely there exists an increasing sequence* $t_n \to \infty$ *such that for any measurable set* $A \subset \mathbb{T}^2$ *we have*

$$\mathrm{Leb}(A \triangle \varphi^{\mathbf{v}}_{t_n}(A)) \to 0, \quad \text{as } n \to \infty.$$

(b*) *Even more, find an explicit increasing sequence* $t_n \to \infty$ *such that for any set* Q *of the form* $Q = I_1 \times I_2 + \mathbb{Z}^2$*, where* $I_1, I_2 \subset [0,1)$ *are intervals, we have*

$$\mathrm{Leb}(Q \cap \varphi^{\mathbf{v}}_{t_n}(Q)) \geq \mathrm{Leb}(Q) - t_n^{-2},$$

for all $n \in \mathbb{N}$ *sufficiently large (Hint: it might be useful to consider the continued fraction expansion of* v*).*

(c) *Conclude in particular that* $\varphi^{\mathbf{v}}$ *is not mixing.*

One can also ask about the correlations of several events or observables, leading to the following definition.

Definition 1.30. *We say that the ppf* (M, φ, μ) *is* mixing of order k *or* k-mixing *if for any* k *(real-valued) observables* $f_1, \dots, f_k \in L^2(M)$ *we have*

$$\int_M f_1 \cdot f_2 \circ \varphi_{t_2} \cdots f_k \circ \varphi_{t_k} \, \mathrm{d}\mu \to \mu(f_1) \cdots \mu(f_k),$$

as $t_2, t_3 - t_2, \dots, t_k - t_{k-1} \to \infty$*.*

We say that the ppf (M, φ, μ) *is* mixing of all orders *if it is mixing of order* k *for all* $k \geq 2$*.*

It is currently unknown whether mixing implies mixing of all orders. This open question is known as the "Rokhlin Problem".

1.5 Outline of the course

In Chapter 2, we present all the relevant background material on matrix Lie groups. We will introduce their associated Lie algebras, which can be described as the space of all left-invariant vector fields. We will then study the induced flows using the exponential map. In Section 2.3, we introduce the Haar measure, which is the invariant measure we will be interested in, the Killing form and the Casimir operator. These last two objects will play a role in the final chapter of these notes. Finally, we define homogeneous spaces as the smooth manifolds obtained as quotients of Lie groups by lattices.

In Chapter 3, we focus on Heisenberg nilflows. We describe them using the so-called exponential coordinates and we classify all possible Heisenberg nilmanifolds. We then show that Heisenberg nilflows are never mixing, but typically relatively mixing and uniquely ergodic. In Section 3.3, we point out an interesting connection between Heisenberg nilflows and theta sums (or quadratic Weyl sums), which are classical objects in analytic number theory.

In Chapter 4, we study in detail the action of $\mathrm{PSL}(2, \mathbb{R})$ on the hyperbolic plane (namely, on its upper-half plane model), which is first introduced in §2.3. We define the geodesic and horocycle flow as particular cases of homogeneous flows on quotients of $\mathrm{PSL}(2, \mathbb{R})$. As an important example, we introduce the Modular Surface.

Chapter 5 is devoted to the study of the ergodic properties of geodesic and horocycle flow. We prove that they are ergodic and mixing.

In Chapter 6, we study the connection between the geodesic flow on the Modular Surface and continued fractions. This fascinating topic dates back to Artin [1], but we will follow an elegant presentation by Series [17].

The final part, Chapter 7, is devoted to the treatment of more advanced material. We prove unique ergodicity of the horocycle flow on compact manifolds, a result originally due to Furstenberg [8]. The proof we present in these notes is due to Coudène [3]. We then discuss the generalizations of this result to finite volume, noncompact spaces and we state Ratner's Theorem [15] on measure classification in the case of unipotent flows. We then study some quantitative properties. We present a special case of Ratner's quantitative mixing result [14] for geodesic and horocycle flow and a special case of Flaminio and Forni's result [6] on asymptotics of horocycle averages, but following the proof in [16].

Chapter 2

Lie Groups

In this section, we introduce and study Lie groups, in particular matrix Lie groups, and their Lie algebras. We introduce the objects and some fundamental tools we are going to study in this course: homogeneous flows, Haar measures, the Adjoint representation and the Casimir operator. Excellent references for these topics are the books [13] and [12]

The reader will benefit from some familiarity with basic notions in differential topology, such as tangent spaces, vector fields, and differential forms.

2.1 Matrix Lie groups

2.1.1 Definitions

Definition 2.1. *A* Lie group *G is a group $(G,\cdot)$ endowed with a differential structure such that both the multiplication map and the inverse map*

$$\begin{aligned} G\times G &\to G, \\ (g,h) &\mapsto gh \end{aligned} \qquad \textit{and} \qquad \begin{aligned} G &\to G \\ g &\mapsto g^{-1} \end{aligned}$$

are smooth.

If G is a Lie group, it follows immediately from the definition that, for any $g\in G$, the *left multiplication map* $L_g\colon G\to G$ given by $L_g(h)=gh$ and the *right multiplication map* $R_g\colon G\to G$ given by $R_g(h)=hg$ are smooth maps.

The simplest example of a Lie group is $(\mathbb{R}^n,+)$ equipped with the trivial atlas. It is clear that the maps

$$\begin{aligned} \mathbb{R}^n\times\mathbb{R}^n &\to \mathbb{R}^n, \\ (\mathbf{x},\mathbf{y}) &\mapsto \mathbf{x}+\mathbf{y} \end{aligned} \qquad \text{and} \qquad \begin{aligned} \mathbb{R}^n &\to \mathbb{R}^n \\ \mathbf{x} &\mapsto -\mathbf{x} \end{aligned}$$

are smooth. The space $\mathrm{Mat}(n,\mathbb{R})$ of square matrices of size n with real coefficients is also a Lie group for the addition operation, since it is isomorphic to $\mathbb{R}^{n^2}$ (not only as Abelian groups, but also as vector spaces). Similarly, we can see that the torus $\mathbb{T}^n=\mathbb{R}^n/\mathbb{Z}^n$ is a Lie group. Notice that the projection map $\pi\colon \mathbb{R}^n\to\mathbb{R}^n/\mathbb{Z}^n$, $\pi(\mathbf{x})=[\![\mathbf{x}]\!]=\mathbf{x}+\mathbb{Z}^n$, is a local homeomorphism; that is, for every $p\in\mathbb{T}^n$ and any $r\in(0,1)$ there exists $\mathbf{x}\in\mathbb{R}^n$ such that the restriction $\pi|_{B(\mathbf{x},r)}\colon B(\mathbf{x},r)\to B(p,r)$ of π to the ball centered at $\mathbf{x}$ of radius r is a homeomorphism on its image. Then, one can construct an atlas on $\mathbb{T}^n$ by means of $(\pi|_{B(\mathbf{x},r)})^{-1}\colon B(p,r)\to\mathbb{R}^n$. The transition maps between charts are translations by elements of $\mathbb{Z}^n$. In this atlas, the group operations on $\mathbb{T}^n$, as a quotient group of $\mathbb{R}^n$ are smooth.

Exercise 2.2. (a) *Show that the circle* $\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}$ *is a Lie group.*

(b) *Characterize for which* $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$ *the subgroup* $\{[\![t\mathbf{v}]\!] : t \in \mathbb{R}\}$ *of* $\mathbb{T}^2$ *is a Lie group (with respect to the induced topology of* $\mathbb{T}^2$*).*

In all the simple examples we have seen above, the group is Abelian. This is not an interesting situation; our focus will be on non-Abelian groups.

Lemma 2.3. *The* general linear group

$$\mathrm{GL}(n,\mathbb{R}) = \{g \in \mathrm{Mat}(n,\mathbb{R}) : \det g \neq 0\}$$

with matrix multiplication is a (non-Abelian) Lie group.

Proof. It is clear that $\mathrm{GL}(n,\mathbb{R})$ is a smooth manifold, since it is an open subset of $\mathrm{Mat}(n,\mathbb{R}) \simeq \mathbb{R}^{n^2}$, and the restriction of the coordinate maps to this open subset defines an atlas of smooth charts. We only need to verify that multiplication and inversion are smooth with respect to this atlas. Matrix multiplication is smooth since, in these charts, it is a polynomial map; similarly, taking the inverse is also a polynomial map in coordinates by the Cramer's rule. □

As a consequence of the following lemma, whose proof can be found, for example, in [13, Proposition 7.11], we get many more examples of Lie groups.

Proposition 2.4. *Let H be a* closed *subgroup of a Lie group G. Then H is an embedded submanifold of G and hence a Lie group.*

Proposition 2.4 motivates the following definition.

Definition 2.5. *A* (real) matrix Lie group *is a closed subgroup G of* $\mathrm{GL}(n,\mathbb{R})$.

A matrix Lie group is thus a Lie group according to Definition 2.1. We remark that the converse is not true, namely there exists Lie groups that are not matrix Lie groups[1]. However, we will not deal with them in this course; actually, the examples of matrix Lie groups that we will mostly be interested in are the *special linear group of degree 2*

$$\mathrm{SL}(2,\mathbb{R}) = \{g \in \mathrm{GL}(2,\mathbb{R}) : \det g = 1\},$$

and the *Heisenberg group*

$$\mathrm{Heis} = \left\{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : x,y,z \in \mathbb{R} \right\}.$$

2.1.2 Tangent spaces and geometric tangent vectors

Proposition 2.4 states that a matrix Lie group G is an embedded submanifold in $\mathrm{Mat}(n,\mathbb{R}) \simeq \mathbb{R}^{n^2}$. In particular, we can look at the set of *geometric tangent vectors* at any point $g \in G$; that is, at the set of vectors in $\mathbb{R}^{n^2}$ which are parallel to the tangent space at g. We will denote by $\mathfrak{g}$ the set of geometric tangent vectors at the identity $e \in G$, more precisely we define

$$\mathfrak{g} := \left\{ \dot\gamma(0) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \gamma(t) \ : \ \gamma\colon \mathbb{R} \to G \text{ is a smooth curve with } \gamma(0) = e \right\}.$$

Lemma 2.6. *If G is a matrix Lie group, the set* $\mathfrak{g}$ *is a subspace of* $\mathrm{Mat}(n,\mathbb{R})$.

[1] For example, the universal cover of $\mathrm{SL}(2,\mathbb{R})$.

Proof. The zero matrix $\mathbf{0} \in \mathrm{Mat}(n,\mathbb{R})$ belongs to $\mathfrak{g}$, since it is the derivative of the constant curve $t \mapsto e \in G$. Let $\dot{\gamma}(0)$ and $\dot{\eta}(0)$ be two geometric tangent vectors at e, and let us check that their sum is a geometric tangent vector as well. Define the curve $(\gamma \cdot \eta)(t) := \gamma(t)\eta(t)$. Then, $(\gamma \cdot \eta)(0) = e$, and, using the product rule, its geometric tangent vector is

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (\gamma \cdot \eta)(t) = \dot{\gamma}(0)\eta(0) + \gamma(0)\dot{\eta}(0) = \dot{\gamma}(0) + \dot{\eta}(0),$$

hence $\dot{\gamma}(0) + \dot{\eta}(0) \in \mathfrak{g}$. Finally, let us check that $\mathfrak{g}$ is closed under scalar multiplication. Let $\dot{\gamma}(0) \in \mathfrak{g}$, and let $a \in \mathbb{R}$. Then, the curve $(a\gamma)(t) := \gamma(at)$ is a smooth curve such that $(a\gamma)(0) = e$ and

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (a\gamma)(t) = a\dot{\gamma}(0) \in \mathfrak{g},$$

which completes the proof. □

Let us see a concrete example: let us find the space $\mathfrak{sl}(2,\mathbb{R})$ of geometric tangent vectors of $\mathrm{SL}(2,\mathbb{R})$ at the identity $e = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. If we define $u(t) = \left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$, then we obtain a smooth function $u\colon \mathbb{R} \to \mathrm{SL}(2,\mathbb{R})$ and

$$\mathbf{u} := \dot{u}(0) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{R}).$$

In a similar way, by looking at the smooth curves $a(t) := \left(\begin{smallmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{smallmatrix}\right)$ and $v(t) := \left(\begin{smallmatrix} 1 & 0 \\ t & 1 \end{smallmatrix}\right)$, we find that $\mathbf{u}, \mathbf{a}, \mathbf{v} \in \mathfrak{sl}(2,\mathbb{R})$, where

$$\mathbf{a} := \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}, \quad \text{and} \quad \mathbf{v} := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

In particular, if we denote by $\mathrm{tr}(\mathbf{x})$ the trace of the matrix $\mathbf{x}$, by Lemma 2.6 we get

$$\mathrm{span}\{\mathbf{u}, \mathbf{a}, \mathbf{v}\} = \{\mathbf{x} \in \mathrm{Mat}(2,\mathbb{R}) : \mathrm{tr}(\mathbf{x}) = 0\} \subseteq \mathfrak{sl}(2,\mathbb{R}).$$

Let us show that equality holds. If $\gamma(t) = \left(\begin{smallmatrix} a(t) & b(t) \\ c(t) & d(t) \end{smallmatrix}\right)$ is a smooth curve in $\mathrm{SL}(2,\mathbb{R})$ with $\gamma(0) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, then we have $a(t)d(t) - b(t)c(t) = 1$ for all $t \in \mathbb{R}$. Differentiating at $t = 0$, we get

$$0 = \dot{a}(0)d(0) + a(0)\dot{d}(0) - \dot{b}(0)c(0) - b(0)\dot{c}(0) = \dot{a}(0) + \dot{d}(0).$$

This shows that $\dot{\gamma}(0) \in \{\mathbf{x} \in \mathrm{Mat}(2,\mathbb{R}) : \mathrm{tr}(\mathbf{x}) = 0\}$, and hence proves the equality

$$\mathfrak{sl}(2,\mathbb{R}) = \{\mathbf{x} \in \mathrm{Mat}(2,\mathbb{R}) : \mathrm{tr}(\mathbf{x}) = 0\}.$$

Exercise 2.7. *Find the space* $\mathfrak{h}$ *of geometric tangent vectors of* Heis *at the identity.*

If $\mathbf{x}, \mathbf{y} \in \mathfrak{g}$ are two geometric tangent vectors of a matrix Lie group G, by Lemma 2.6, their sum $\mathbf{x} + \mathbf{y}$ is still in $\mathfrak{g}$, as well as any of their scalar multiples. We define now a bilinear, antisymmetric operation on geometric tangent vectors, which we call the *bracket*, and will turn the space $\mathfrak{g}$ into an *algebra*. The geometric interpretation of this operation will become clear in a little while.

If $G \subset \mathrm{GL}(n,\mathbb{R})$ is a matrix Lie group, define the *bracket operation*

$$[\cdot,\cdot]_{\mathfrak{g}}\colon \mathfrak{g} \times \mathfrak{g} \to \mathrm{Mat}(n,\mathbb{R}) \quad \text{by} \quad [\mathbf{x},\mathbf{y}]_{\mathfrak{g}} := \mathbf{x} \cdot \mathbf{y} - \mathbf{y} \cdot \mathbf{x},$$

where $\cdot$ denotes the matrix multiplication in $\mathrm{Mat}(n,\mathbb{R})$. From now on, we will often suppress the symbol $\cdot$, which should be clear from the context. The definition immediately implies that the bracket of a vector with itself is zero.

Exercise 2.8. *Show that* $[\cdot,\cdot]_{\mathfrak{g}}$ *is bilinear, antisymmetric and satisfies the* Jacobi identity: *for all* $\mathbf{x},\mathbf{y},\mathbf{z}\in\mathfrak{g}$, *we have*

$$[[\mathbf{x},\mathbf{y}]_{\mathfrak{g}},\mathbf{z}]_{\mathfrak{g}}+[[\mathbf{y},\mathbf{z}]_{\mathfrak{g}},\mathbf{x}]_{\mathfrak{g}}+[[\mathbf{z},\mathbf{x}]_{\mathfrak{g}},\mathbf{y}]_{\mathfrak{g}}=0.$$

In $\mathfrak{sl}(2,\mathbb{R})$, one can check that the nontrivial possible brackets of the basis elements $\mathbf{u},\mathbf{a},\mathbf{v}$ are

$$[\mathbf{a},\mathbf{u}]_{\mathfrak{sl}(2,\mathbb{R})}=-[\mathbf{u},\mathbf{a}]_{\mathfrak{sl}(2,\mathbb{R})}=\mathbf{u},\qquad [\mathbf{a},\mathbf{v}]_{\mathfrak{sl}(2,\mathbb{R})}=-[\mathbf{v},\mathbf{a}]_{\mathfrak{sl}(2,\mathbb{R})}=-\mathbf{v}, \tag{2.1}$$

$$[\mathbf{u},\mathbf{v}]_{\mathfrak{sl}(2,\mathbb{R})}=-[\mathbf{v},\mathbf{u}]_{\mathfrak{sl}(2,\mathbb{R})}=2\mathbf{a}. \tag{2.2}$$

Notice in particular that the bracket of any two vectors in $\mathfrak{sl}(2,\mathbb{R})$ is again an element of $\mathfrak{sl}(2,\mathbb{R})$. Indeed, this is no coincidence, as the next lemma shows.

Lemma 2.9. *(a) For all* $g\in G$ *and all* $\mathbf{x}\in\mathfrak{g}$, *we have*

$$g^{-1}\mathbf{x}g\in\mathfrak{g}.$$

(b) The space $\mathfrak{g}$ *is closed under bracket* $[\cdot,\cdot]_{\mathfrak{g}}$.

Before proving the lemma, let us note that, by part (a), the map

$$\mathrm{Ad}(g)\colon \mathbf{x}\mapsto g^{-1}\mathbf{x}g$$

is an invertible linear transformation of $\mathfrak{g}$ for any $g\in G$, the inverse being $(\mathrm{Ad}(g))^{-1}=\mathrm{Ad}(g^{-1})$. We call $\mathrm{Ad}(g)\in\mathrm{GL}(\mathfrak{g})$ the *Adjoint of* g. As we mentioned, a geometric interpretation of these facts will come later on.

Proof of Lemma 2.9. Let γ be a smooth curve such that $\dot\gamma(0)=\mathbf{x}$. Then, for all $g\in G$, the map $\gamma_g(t):=g^{-1}\gamma(t)g$ is a smooth curve in G with $\gamma_g(0)=g^{-1}g=e$. Differentiating at $t=0$, we get

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\gamma_g(t)=g^{-1}\left(\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\gamma(t)\right)g=g^{-1}\mathbf{x}g\in\mathfrak{g},$$

which proves (a).

In order to prove (b), let $\mathbf{x},\mathbf{y}\in\mathfrak{g}$; we need to show that $[\mathbf{x},\mathbf{y}]_{\mathfrak{g}}\in\mathfrak{g}$. Let $\gamma(t)$ be such that $\dot\gamma(0)=\mathbf{x}$. By (a), we know that

$$\eta(t):=\gamma(t)\mathbf{y}\gamma(t)^{-1}\in\mathfrak{g}\qquad\text{for all } t\in\mathbb{R}.$$

Moreover, since γ is smooth and multiplying matrices is a smooth map as well, the function η defines a smooth curve in $\mathfrak{g}$. In particular, the derivative $\dot\eta(0)$ of η at $t=0$ exists. By Lemma 2.6, $\mathfrak{g}$ is a closed subset, hence the limit

$$\dot\eta(0)=\lim_{t\to0}\frac{\eta(t)-\eta(0)}{t}$$

belongs to $\mathfrak{g}$. Let us compute it. First of all, since $\gamma(t)^{-1}\gamma(t)=e$ for all $t\in\mathbb{R}$ and $\gamma(0)=e$, by differentiating, we get

$$0=\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}(\gamma(t)^{-1}\gamma(t))=\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}(\gamma(t)^{-1})+\dot\gamma(0).$$

From this, we conclude

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\eta(t)=\left(\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\gamma(t)\right)\mathbf{y}\gamma(0)^{-1}+\gamma(0)\mathbf{y}\left(\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\gamma(t)^{-1}\right)=\mathbf{x}\mathbf{y}-\mathbf{y}\mathbf{x},$$

which shows that $[\mathbf{x},\mathbf{y}]_{\mathfrak{g}}\in\mathfrak{g}$. □

Finally, let us recall that the definition of tangent space in the general context of differentiable manifolds is given in terms of *derivations* as follows. For any $g \in G$, the *tangent space* T_gG at g is the vector space of all possible derivations at g; that is, the space of all linear maps $X_g \colon \mathscr{C}^\infty(G) \to \mathbb{R}$ on smooth functions on G which satisfy the Leibniz rule

$$X_g(f_1 f_2) = X_g(f_1)\, f_2(g) + f_1(g) X_g(f_2).$$

It is not hard to see that the tangent space at any point can be identified with the space of its geometric tangent vectors; in particular we have the following identification.

Lemma 2.10. *If G is a matrix Lie group, then $T_eG \simeq \mathfrak{g}$.*

Proof. The isomorphism between $\mathfrak{g}$ and T_eG is defined as follows: if $\dot\gamma(0) \in \mathfrak{g}$, then we associate the derivation

$$X_e \colon f \mapsto \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f \circ \gamma(t).$$

In coordinates, if $\dot\gamma(0) = \mathbf{x}$, where $\mathbf{x} = (x_{i,j})_{i,j=1}^n \in \mathrm{Mat}(n,\mathbb{R})$, then the associated derivation is $X_e(f) = \sum_{i,j=1}^n x_{i,j}\partial_{i,j} f(e)$, where $\partial_{i,j}$ is the partial derivative with respect to the (i,j)-coordinate. It is an easy exercise to verify that this map is indeed a linear isomoprhism, for details see, e.g., [13, Proposition 3.2]. □

2.2 The Lie algebra of a Lie group

2.2.1 Left-invariant vector fields

Let G be a Lie group, and let $F \colon G \to G$ be a smooth map. The *differential* DF of F is a smooth map on the tangent bundle TG of G defined as follows: the differential $DF(g)$ at $g \in G$ is a linear map from the tangent space T_gG at g to the tangent space $T_{F(g)}G$ at $F(g)$ which sends a derivation $X_g \in T_gG$ to the derivation $DF(g)X_g \in T_{F(g)}G$ given by

$$[DF(g)X_g](f) = X_g(f \circ F).$$

The reader can check that $DF(g)X_g$ is indeed a derivation (namely, it is a linear map on $\mathscr{C}^\infty(G)$ which satisfies the Leibniz rule).

If G is a matrix Lie group, then we can define the differential DF of F in terms of geometric tangent vectors as well, following Lemma 2.10. Let $\mathbf{x} = \dot\gamma(0) \in \mathfrak{g}$ be a geometric tangent vector at the identity. We define

$$DF(e)\mathbf{x} := \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} F \circ \gamma(t),$$

which is a geometric tangent vector at $F(e) \in G$.

Using smooth maps, we can therefore "move" tangent vectors around G. Recall that, for any $g \in G$, the left-multiplication map $L_g \colon h \mapsto gh$ is smooth. Then, for any element $\mathbf{x} \in \mathfrak{g}$ we can associate a tangent vector at any other point g by the aid of the differential of L_g at e. In other words, we can define a *vector field* X on G by setting

$$X_g := DL_g(e)X_e, \qquad \text{for all } g \in G, \tag{2.3}$$

where, again, $X_e \in T_eG$ is the derivation associated to $\mathbf{x}$ according to Lemma 2.10. Let us express this in coordinates. If $\mathbf{x} = \dot\gamma(0) \in \mathfrak{g}$, by Taylor's Theorem we can write

$$\gamma(t) = e + t\mathbf{x} + tR(t),$$

where $t \mapsto R(t) \in \mathrm{Mat}(n,\mathbb{R})$ is a smooth map and $R(t) \to \mathbf{0}$ as $t \to 0$. Then

$$DL_g(e)\mathbf{x} = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} L_g \circ \gamma(t) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} (g + tg\mathbf{x} + tgR(t)) = g\mathbf{x} \in \mathrm{Mat}(n,\mathbb{R}). \tag{2.4}$$

Proposition 2.11. *For any $\mathbf{x} = X_e \in T_eG$, the associated vector field X defined by* (2.3) *is smooth and* left-invariant*, namely for all $g \in G$ we have $DL(g)X = X$.*

Proof. In coordinates, the fact that X is smooth comes from expression (2.4). More formally, in order to check that X is smooth it is enough to show that $Xf\colon G \to \mathbb{R}$ is a smooth function for any $f \in \mathscr{C}^\infty(G)$. Let $\gamma(t)$ be a smooth curve such that $\dot\gamma(0) = X_e$, and fix any such $f \in \mathscr{C}^\infty(G)$. Define $F(g,t) := f(g\gamma(t))$; then it is clear that $F\colon G \times \mathbb{R} \to \mathbb{R}$ is a smooth function, and so is its derivative $\frac{\partial}{\partial t}F(g,0)$ at $t = 0$. Thus,

$$X_g f = (DL_g(e)X_e)(f) = X_e(f \circ L_g) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f \circ L_g(\gamma(t)) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f(g\gamma(t)) = \frac{\partial}{\partial t}F(g,0)$$

is a smooth function in g, which proves that X is smooth.

The fact that X is left-invariant is an easy exercise that we leave to the reader. □

So far, we have seen that in a matrix Lie group G we can identify the tangent space at the identity with the space of geometric tangent vectors $\mathfrak{g} \subset \mathrm{Mat}(n,\mathbb{R})$. Moreover, given any element $X_e \in T_eG$, we can define a smooth vector field X on G which is left-invariant.

Definition 2.12. *Let G be a Lie group. The vector space of all smooth, left-invariant vector fields*

$$\mathrm{Lie}(G) := \{X\colon G \to TG : X \text{ is smooth and } DL_gX = X \text{ for all } g \in G\}$$

is called the Lie algebra *of G.*

The use of the word "algebra" will become clear later on. The following result should come as no surprise.

Lemma 2.13. *The evaluation at the identity map* $\mathrm{ev}\colon \mathrm{Lie}(G) \to T_eG$ *defined by* $\mathrm{ev}(X) = X_e$ *is a linear isomorphism.*

Proof. It immediately follows from the definition that ev is linear. If $\mathrm{ev}(X) = X_e = 0$, then for every $g \in G$ we have $X_g = DL_g(e)0 = 0$, which implies $X = 0$. This shows that ev is injective.

Let now $X_e \in T_eG$, and define a vector field X as in (2.3). By Proposition 2.11, X is smooth and left-invariant, hence $X \in \mathrm{Lie}(G)$. By construction, $\mathrm{ev}(X) = X_e$, which proves surjectivity and completes the proof. □

In the following, we will often identify $\mathbf{x} \in \mathfrak{g}$, $X_e \in T_eG$ and $X \in \mathrm{Lie}(G)$. We now know that (geometric) tangent vectors at the identity are in 1-to-1 correspondence with smooth left-invariant vector fields. Each of the latters generate a smooth flow on G; more precisely, if $X \in \mathrm{Lie}(G)$, there exists a unique $\varphi^X = \{\varphi_t^X\}_{t\in I}$ (which is defined at least on small intervals I containing 0) whose infinitesimal generator is X, namely such that

$$X_g f = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f \circ \varphi_t^X(g), \qquad \text{for all } g \in G,$$

where f is any smooth function. The goal now is to find an expression for φ^X and show that it is defined for all $t \in \mathbb{R}$.

2.2.2 The exponential map

On the space of matrices $\mathrm{Mat}(n,\mathbb{R})$, we introduce the following map

$$\begin{aligned} \exp\colon\ \mathrm{Mat}(n,\mathbb{R}) &\to \mathrm{Mat}(n,\mathbb{R}) \\ \mathbf{x} &\mapsto \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{x}^k, \end{aligned} \tag{2.5}$$

called the *matrix exponential*.

Proposition 2.14. *The matrix exponential is well-defined and satisfies the following properties:*

1. $\exp(\mathbf{x})\exp(\mathbf{y}) = \exp(\mathbf{x}+\mathbf{y})$ *if* $\mathbf{x}$ *and* $\mathbf{y}$ *commute,*
2. $\exp(\mathbf{x}) \in \mathrm{GL}(n,\mathbb{R})$ *and* $\exp(\mathbf{x})^{-1} = \exp(-\mathbf{x})$*,*
3. $\gamma_{\mathbf{x}}(t) := \exp(t\mathbf{x})$ *is a smooth curve in* $\mathrm{GL}(n,\mathbb{R})$ *whose geometric tangent vector at e is* $\mathbf{x}$*,*
4. $\exp(\mathrm{Ad}(g)\mathbf{x}) = \exp(g^{-1}\mathbf{x}g) = g^{-1}\exp(\mathbf{x})g$ *for all* $g \in \mathrm{GL}(n,\mathbb{R})$*,*
5. $\det(\exp(\mathbf{x})) = e^{\mathrm{tr}(\mathbf{x})}$*,*
6. $\exp\colon\ \mathrm{Mat}(n,\mathbb{R}) \to \mathrm{GL}(n,\mathbb{R})$ *is a smooth map.*

Proof. In oder to show that exp is well-defined, we prove that for any $\mathbf{x} \in \mathrm{Mat}(n,\mathbb{R})$, the series $\sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{x}^k$ converges. For any submultiplicative norm $\|\cdot\|$, we have

$$\sum_{k=0}^{\infty} \left\| \frac{1}{k!}\mathbf{x}^k \right\| \leq \sum_{k=0}^{\infty} \frac{1}{k!} \|\mathbf{x}\|^k < \infty,$$

that is, $\sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{x}^k$ is absolutely convergent. Since $\mathrm{Mat}(n,\mathbb{R})$ is complete, this shows that the series is convergent. We now verify the other claims.

1. if $\mathbf{x}$ and $\mathbf{y}$ commute, we have

$$\begin{aligned} \exp(\mathbf{x})\exp(\mathbf{y}) &= \left(\sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{x}^k\right)\left(\sum_{l=0}^{\infty} \frac{1}{l!}\mathbf{y}^l\right) = \sum_{k,l=0}^{\infty} \frac{1}{k!l!}\mathbf{x}^k\mathbf{y}^l = \sum_{k=0}^{\infty}\sum_{l=0}^{k} \frac{1}{l!(k-l)!}\mathbf{x}^l\mathbf{y}^{k-l} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\sum_{l=0}^{k} \binom{k}{l}\mathbf{x}^l\mathbf{y}^{k-l} = \sum_{k=0}^{\infty} \frac{1}{k!}(\mathbf{x}+\mathbf{y})^k = \exp(\mathbf{x}+\mathbf{y}). \end{aligned}$$

2. Take $\mathbf{y} = -\mathbf{x}$ in part 1, and notice that $\exp(\mathbf{0}) = e$.
3. We proved that $\sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{x}^k$ is absolutely convergent, so the following equalities hold

$$\dot{\gamma}_{\mathbf{x}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\exp(t\mathbf{x}) = \sum_{k=0}^{\infty} \frac{1}{k!}\frac{\mathrm{d}}{\mathrm{d}t}(t\mathbf{x})^k = \sum_{k=0}^{\infty} \frac{1}{(k-1)!}\mathbf{x}^k t^{k-1} = \mathbf{x}\sum_{k=0}^{\infty} \frac{1}{k!}(t\mathbf{x})^k = \mathbf{x}\gamma_{\mathbf{x}}(t).$$

   Hence $\dot{\gamma}_{\mathbf{x}}(0) = \mathbf{x}$.

4. For any $g \in \mathrm{GL}(n,\mathbb{R})$,

$$\exp(g^{-1}\mathbf{x}g) = \sum_{k=0}^{\infty} \frac{1}{k!}(g^{-1}\mathbf{x}g)^k = \sum_{k=0}^{\infty} \frac{1}{k!}(g^{-1}\mathbf{x}^k g) = g^{-1}\left(\sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{x}^k\right)g = g^{-1}\exp(\mathbf{x})g.$$

5. If $\mathbf{x} \in \mathrm{Mat}(n,\mathbb{R})$ is an upper triangular matrix, it is easy to see that the diagonal entries of $\exp(\mathbf{x})$ are the exponentials of the diagonal entries of $\mathbf{x}$; in particular the equality $\det(\exp(\mathbf{x})) = e^{\mathrm{tr}(\mathbf{x})}$ holds for these matrices. If $\mathbf{x}$ is arbitrary, write $\mathbf{x} = g^{-1}\mathbf{y}g$ in Jordan normal form and apply part 4.

6. The fact that exp is a smooth map follows from the rules of differentiations for series of functions.

□

For any $\mathbf{x} \in \mathfrak{g}$ and $t \in \mathbb{R}$ let us define

$$\varphi_t^{\mathbf{x}}(g) = g\exp(t\mathbf{x}). \tag{2.6}$$

Notice that, a priori, $\varphi_t^{\mathbf{x}} \colon G \to \mathrm{GL}(n,\mathbb{R})$. We shall now see that $\varphi_t^{\mathbf{x}}$ has values in G and defined the integral curves of the vector field $X \in \mathrm{Lie}(G)$.

Proposition 2.15. *The map $\varphi^{\mathbf{x}} \colon \mathbb{R} \times G \to \mathrm{GL}(n,\mathbb{R})$ defined by $\varphi^{\mathbf{x}}(t,g) = \varphi_t^{\mathbf{x}}(g)$ as in* (2.6) *is a smooth flow with infinitesimal generator $X \in \mathrm{Lie}(G)$, where $X_e = \mathbf{x}$.*

Proof. The fact that $\varphi^{\mathbf{x}}$ is a smooth map follows from Proposition 2.14-(6). For $t = 0$, we have $\varphi_0^{\mathbf{x}}(g) = g\exp(\mathbf{0}) = g$ and for any $t,s \in \mathbb{R}$ we have

$$\varphi_{t+s}^{\mathbf{x}}(g) = g\exp((t+s)\mathbf{x}) = g\exp(t\mathbf{x}+s\mathbf{x}) = g\exp(t\mathbf{x})\exp(s\mathbf{x}) = \varphi_s^{\mathbf{x}} \circ \varphi_t^{\mathbf{x}}(g),$$

where we have used Proposition 2.14-(1). Hence, $\varphi^{\mathbf{x}}$ is a smooth flow.

We now check that its infinitesimal generator is $X \in \mathrm{Lie}(G)$. By Proposition 2.14-(3), we know that the tangent vector at the identity associated to $\varphi^{\mathbf{x}}$ is $\mathbf{x} = X_e$. Then, for any smooth function f and any $g \in G$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f \circ \varphi_t^{\mathbf{x}}(g) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f(g\exp(t\mathbf{x})) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} (f \circ L_g)(\exp(t\mathbf{x})) = [DL_g(e)X_e](f) = X_g f,$$

which proves that the infinitesimal generator is X. □

Corollary 2.16. *We have* $\exp \colon \mathfrak{g} \to G$. *Moreover,* exp *is a smooth diffeomorphism between a neighbourhood of $\mathbf{0} \in \mathfrak{g}$ and a neighbourhood of $e \in G$.*

Proof. Let $X \in \mathfrak{g} = \mathrm{Lie}(G)$. By the standard theory of ODEs, there exists a unique smooth solution $\gamma \colon I \to G$ to the equation $\dot{\gamma}(t) = X(\gamma(t))$ with the initial condition $\gamma(0) = e$. The solution γ is a smooth curve in G defined on an interval $I = (-\varepsilon,\varepsilon)$ for some $\varepsilon > 0$. By Proposition 2.15, the smooth curve $t \mapsto \varphi_t^{\mathbf{x}}(e) = \exp(t\mathbf{x})$ satisfies the ODE $\dot{\gamma} = X(\gamma)$, hence $\exp(t\mathbf{x}) \in G$ for all $t \in (-\varepsilon,\varepsilon)$. Let $N \in \mathbb{Z}$ be such that $|N|^{-1} < \varepsilon$. By Proposition 2.14-(1), we conclude

$$\exp(\mathbf{x}) = \exp(N^{-1}\mathbf{x})^N \in G,$$

since G is a group. This shows that exp maps $\mathfrak{g}$ into G.

Let us show it is a local diffeomorphism. By the Inverse Function Theorem, it is enough to show that the differential $D\exp(\mathbf{0})$ from $T_{\mathbf{0}}\mathfrak{g} \simeq \mathfrak{g}$ to $T_eG = \mathfrak{g}$ is invertible. Indeed, for any $\mathbf{x} \in \mathfrak{g}$ we have $D\exp(\mathbf{0})\mathbf{x} = \frac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\exp(t\mathbf{x}) = \mathbf{x}$; that is, $D\exp(\mathbf{0})$ is the identity. This completes the proof. □

We have then shown that for any $\mathbf{x} \in \mathfrak{g}$ there exists a smooth flow $\varphi^{\mathbf{x}}$ on G defined for all times $t \in \mathbb{R}$ given by the action by multiplication on the right by the 1-parameter subgroup $\{\exp(t\mathbf{x}) : t \in \mathbb{R}\}$ generated by the left-invariant vector field X associated to $\mathbf{x}$.

Definition 2.17. *For any* $\mathbf{x} \in \mathfrak{g} \setminus \{\mathbf{0}\}$*, the flow* $\{\varphi_t^{\mathbf{x}}\}_{t\in\mathbb{R}}$ *given by* $\varphi_t^{\mathbf{x}}(g) = g\exp(t\mathbf{x})$ *is called the* homogeneous flow generated by $\mathbf{x}$. *We will write interchangeably* $\varphi_t^{\mathbf{x}}$ *and* φ_t^X.

For example, we can check that

$$\exp(t\mathbf{u}) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \qquad \exp(t\mathbf{a}) = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}, \qquad \exp(t\mathbf{v}) = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}, \qquad \text{for all } t \in \mathbb{R}.$$

As we saw, exp is a diffeomorphism between a neighbourhood of $\mathbf{0} \in \mathfrak{g}$ and $e \in G$, but in general exp is neither injective nor surjective.

Exercise 2.18. *(a) Show that* $\exp\colon \mathfrak{h} \to \mathrm{Heis}$ *is a global diffeomorphism.*

(b) Show that* $\exp\colon \mathfrak{sl}(2,\mathbb{R}) \to \mathrm{SL}(2,\mathbb{R})$ *is neither injective nor surjective: find countably many* $\mathbf{x}_n \in \mathfrak{sl}(2,\mathbb{R})$ *such that* $\exp(\mathbf{x}_n) = e$ *and find a matrix* $g \in \mathrm{SL}(2,\mathbb{R})$ *which cannot be written as* $\exp(\mathbf{x})$ *for* $\mathbf{x} \in \mathfrak{sl}(2,\mathbb{R})$.

2.2.3 Adjoints, Lie derivatives and Lie brackets

Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra; call $k = \dim \mathfrak{g}$. Let us recall, from Lemma 2.9, that, for all $g \in G$, we can define a linear map $\mathrm{Ad}(g)\colon \mathfrak{g} \to \mathfrak{g}$ called the Adjoint of g by $\mathrm{Ad}(g)\mathbf{x} = g^{-1}\mathbf{x}g$. The map

$$\begin{aligned} \mathrm{Ad}\colon G &\to \mathrm{GL}(\mathfrak{g}) \simeq \mathrm{GL}(k,\mathbb{R}) \\ g &\mapsto \mathrm{Ad}(g) \end{aligned} \tag{2.7}$$

is a group anti-homomorphism, since

$$\mathrm{Ad}(gh)\mathbf{x} = (gh)^{-1}\mathbf{x}(gh) = h^{-1}(g^{-1}\mathbf{x}g)h = (\mathrm{Ad}(h)\circ\mathrm{Ad}(g))\mathbf{x},$$

for all $\mathbf{x} \in \mathfrak{g}$. We call Ad the *Adjoint representation of* G.

Let us comment again on its dynamical significance. Let $\mathbf{x} \in \mathfrak{g} \setminus \{\mathbf{0}\}$, and let $\varphi^{\mathbf{x}}$ be the associated flow. We want to study the divergence of nearby points under $\varphi^{\mathbf{x}}$. By Proposition 2.14-(6), the exponential map is a smooth diffeomorphism when restricted to a sufficiently small neighbourhood $\mathcal{U}$ of $\mathbf{0} \in \mathfrak{g}$. Let $g \in \exp(\mathcal{U}) \subset G$ be a point sufficiently close to the identity e, so that we can write $g = \exp(\mathbf{z})$ for some $\mathbf{z} \in \mathcal{U}$. If we want to move between the points $\varphi_t^{\mathbf{x}}(e) = \exp(t\mathbf{x})$ and $\varphi_t^{\mathbf{x}}(g) = g\exp(t\mathbf{x})$, we need to multiply by

$$\exp(-t\mathbf{x})g\exp(tx) = \exp(-t\mathbf{x})\exp(\mathbf{z})\exp(tx) = \exp(\exp(-t\mathbf{x})\mathbf{z}\exp(tx)) = \exp(\mathrm{Ad}(\exp(t\mathbf{x}))\mathbf{z}),$$

where we used Proposition 2.14-(4). In other words, the exponential of the Adjoint tells us how nearby points diverge.

Let us be more precise. Let us fix $t \in \mathbb{R}$ and consider the time-t map $\varphi_t^{\mathbf{x}}\colon G \to G$. We compute its differential $D\varphi_t^{\mathbf{x}}$ acting on tangent vectors. Fix $g \in G$ and $Z \in \mathrm{Lie}(G)$, which we identify with $\mathbf{z} \in \mathfrak{g}$ as usual. In order to compute the image $[D\varphi_t^{\mathbf{x}}(Z)]_g$ of the vector field Z at the point g, we fix an arbitrary smooth function f on G so that

$$\begin{aligned} [D\varphi_t^{\mathbf{x}}(Z)]_g(f) = Z_{\varphi_{-t}^{\mathbf{x}}(g)}(f\circ\varphi_t^{\mathbf{x}}) &= \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0}(f\circ\varphi_t^{\mathbf{x}})(\varphi_{-t}^{\mathbf{x}}(g)\exp(s\mathbf{z})) \\ &= \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} f(g\exp(-t\mathbf{x})\exp(s\mathbf{z})\exp(t\mathbf{x})) = \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} f\big(g\exp[s\,\mathrm{Ad}(\exp(t\mathbf{x}))\mathbf{z}]\big) \\ &= [\mathrm{Ad}(\exp(t\mathbf{x}))\mathbf{z}]_g(f), \end{aligned}$$

where, in the last equality, we have used Proposition 2.14-(4). Therefore, we conclude

$$D\varphi_t^{\mathbf{x}}(Z) = \mathrm{Ad}(\exp(t\mathbf{x}))\mathbf{z}, \tag{2.8}$$

that is, the action of $\mathrm{Ad}(\exp(t\mathbf{x}))$ on $\mathfrak{g}$ describes how tangent vectors evolve under $\varphi^{\mathbf{x}}$.

Exercise 2.19. *Let $\mathscr{B} := \{\mathbf{u}, \mathbf{a}, \mathbf{v}\}$ the basis of $\mathfrak{sl}(2,\mathbb{R})$ we introduced in §2.1.2. For all $\mathbf{x} \in \mathscr{B}$ and any given $t \in \mathbb{R}$, compute explicitly the matrix associated to $D\varphi_t^{\mathbf{x}}$ with respect to $\mathscr{B}$. What is the difference between* **a** *and the other two elements of $\mathscr{B}$?*

For all $\mathbf{x} \in \mathfrak{g}$, let us call $\mathfrak{ad}_{\mathbf{x}} := [\mathbf{x}, \cdot]_{\mathfrak{g}}$ the linear map $\mathfrak{ad}_{\mathbf{x}} \colon \mathfrak{g} \to \mathfrak{g}$. It is called the *adjoint endomorphism of* **x**, and it can be expressed by a matrix $\mathfrak{ad} \in \mathrm{Mat}(k,\mathbb{R})$. Clearly, this matrix is not invertible. From the proof of Lemma 2.9, it follows that

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathrm{Ad}(\exp(-t\mathbf{x})) = \mathfrak{ad}_{\mathbf{x}}, \tag{2.9}$$

from which one deduces

$$\mathrm{Ad}(\exp(t\mathbf{x})) = \exp(t\,\mathfrak{ad}_{\mathbf{x}}) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathfrak{ad}_{\mathbf{x}}^k.$$

While $\mathrm{Ad}(\exp(t\mathbf{x}))$ describes the divergence of left-invariant vector fields (and nearby points) under the flow $\varphi^{\mathbf{x}}$, the map $\mathfrak{ad}_{\mathbf{x}} = \mathscr{L}_X$ describes its infinitesimal version, that is how vector fields diverge "infinitesimally".

Again, let us be more precise. Let us consider two flows φ_t^X and φ_t^Y generated by the vector fields $X, Y \in \mathrm{Lie}(G)$. If φ_t^X and φ_t^Y commute, that is if $\varphi_t^X \circ \varphi_s^Y = \varphi_s^Y \circ \varphi_t^X$ for all $t, s \in \mathbb{R}$, then for any fixed $t \in \mathbb{R}$, the differential $D\varphi_t^X$ of the smooth map φ_t^X maps the vector field Y into itself. If the two flows do not commute, then $D\varphi_t^X$ maps Y smoothly into another smooth vector field $Z = Z(t)$. The Lie derivative describes the "infinitesimal change" of Y when moved by $D\varphi_t^X$. More precisely, the *Lie derivative $\mathscr{L}_X(Y)$ of Y with respect to X at $g \in G$* is defined by

$$\mathscr{L}_X(Y) := \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} D\varphi_{-t}^X(Y) = \lim_{t\to 0} \frac{D\varphi_{-t}^X(Y) - Y}{t}.$$

The reader might be familiar with the formula

$$\mathscr{L}_X(Y) = XY - YX,$$

we now show that the Lie derivative coincides with the bracket operation we defined in §2.1.2.

Proposition 2.20. *Let $\mathbf{x}, \mathbf{y} \in \mathfrak{g}$, identified with $X, Y \in \mathrm{Lie}(G)$. Then,*

$$\mathscr{L}_X(Y) = [\mathbf{x}, \mathbf{y}]_{\mathfrak{g}},$$

which is called the Lie bracket *of* **x** *and* **y**.

Proof. The claim follows immediately from (2.8) and (2.9), since

$$\mathscr{L}_X(Y) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} D\varphi_{-t}^X(Y) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{y} = \mathfrak{ad}_{\mathbf{x}}(\mathbf{y}) = [\mathbf{x}, \mathbf{y}]_{\mathfrak{g}}.$$

□

Therefore, the geometric interpretation of the Lie bracket $[\mathbf{x}, \mathbf{y}]_{\mathfrak{g}}$ is to describe the infinitesimal distortion of the left-invariant vector field $Y = \mathbf{y}$ under the action of the flow generated by $X = \mathbf{x}$. The Lie derivative (the Lie bracket) makes the space $\mathrm{Lie}(G) = \mathfrak{g}$ in an algebra.

Recall that, by definition, an *ideal* $\mathfrak{k}$ of $\mathfrak{g}$ is a vector subspace with the property that $[\mathfrak{k}, \mathfrak{g}]_{\mathfrak{g}} \subset \mathfrak{k}$.

Definition 2.21. *A Lie algebra $\mathfrak{g}$ is* simple *if has no non-trivial ideals, namely if $\mathfrak{k}$ is an ideal of $\mathfrak{g}$, then $\mathfrak{k} = \mathfrak{g}$ or $\mathfrak{k} = \{\mathbf{0}\}$. A matrix Lie group G is* simple *if its lie algebra $\mathfrak{g}$ is simple.*

Exercise 2.22. *Show that* $\mathrm{SL}(2,\mathbb{R})$ *is simple.*

2.3 Haar, Killing, Casimir

2.3.1 The Haar measure

Using the differential of the left-multiplication maps we can not only define vector fields starting from a single vector at the identity, but we can also construct a measure on G starting from "a determinant" on $\mathfrak{g}$. This measure will be one of the fundamental objects of this course. Let us see how to do this.

Let V be a k-dimensional real vector space. It is a standard fact from linear algebra that there exists a k-multilinear alternating form on V which is unique up to scalar multiplication; that is, the space $\wedge^k V^*$ is one dimensional. In order to explicilty write one of such multilinear alternating forms ω, one can do the following: choose a basis $\{v_1,\dots,v_k\}$ of V and identify V with $\mathbb{R}^k$ by means of this basis (i.e., identify $w = a_1v_1+\cdots+a_kv_k$ with $(a_1,\dots,a_k)\in\mathbb{R}^k$). Then, for any k vectors $w^{(1)},\dots,w^{(k)}$, with $w^{(j)}=\sum_i a_i^{(j)}v_i$, consider the matrix W whose j-th row is $(a_1^{(j)},\dots,a_k^{(j)})$. Associated to this choice of basis, we can define ω by

$$\omega(w^{(1)},\dots,w^{(k)}) = \det W.$$

A different choice of basis would have the effect of multiplying ω by the determinant of the matrix expressing the change of basis; in particular all possible multilinear alternating forms are multiples of each other.

Let us now turn to matrix Lie groups. Let k be the dimension of $\mathfrak{g}$, and fix a basis $\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$ of $\mathfrak{g}$. Let ω_e be the associated multilinear alternating form on $\mathfrak{g}\simeq T_eG$. Using the left-multiplication maps L_g, we can define a multilinear alternating form ω_g on the tangent space of any other point $g\in G$ by pull-back, namely

$$\omega_g(X_g^{(1)},\dots,X_g^{(k)}) := \omega_e(DL_{g^{-1}}(g)X_g^{(1)},\dots,DL_{g^{-1}}(g)X_g^{(k)}) \quad \text{for any } X_g^{(1)},\dots,X_g^{(k)}\in T_gG.$$

Notice that, indeed, since $L_{g^{-1}}(g)=e$, its differential $DL_{g^{-1}}(g)$ maps T_gG to T_eG.

In other words, from any choice of basis on $\mathfrak{g}$, we have defined a *k-differential form*; in formal terms, a section of the vector bundle $\wedge^k T^*G\to G$. From a k-differential form ω, we obtain a positive measure μ by taking its absolute value, $\mathrm{d}\mu = |\omega|$.

Concretely, we fix a basis $X_e^{(1)},\dots,X_e^{(k)}$ of T_eG. Using the differentials of L_g, we obtain vector fields $X^{(1)},\dots,X^{(k)}\in\mathrm{Lie}(G)$ by $X_g^{(j)}=DL_g(e)X_e^{(j)}$. Then, we take their dual $\mathrm{d}X^{(j)}$: these are differential 1-forms defined by saying that for all $g\in G$, we have $\mathrm{d}X_g^{(j)}(X_g^{(i)})=\delta_{i,j}$ (i.e., 1 if $i=j$ and 0 otherwise). We define the measure μ by saying that for all continuous functions $f\colon G\to\mathbb{R}$,

$$\int_G f(g)\,\mathrm{d}\mu(g) = \int_G f(g)\,\mathrm{d}X_g^{(1)}\cdots\mathrm{d}X_g^{(k)}.$$

With a little extra effort, we can complete the proof of the following important result.

Theorem 2.23. *Let G be a matrix Lie group. There exists a smooth measure μ on G which is invariant by all left-multiplication maps, namely for all continuous functions $f\colon G\to\mathbb{R}$ and for all $h\in G$ we have*

$$\int_G f(hg)\,\mathrm{d}\mu(g) = \int_G f(g)\,\mathrm{d}\mu(g).$$

This measure μ is unique up to scalar and is called the (left) Haar measure on G.

Proof. Let us consider a measure μ constructed as above. The fact that μ is a smooth measure is a consequence of Proposition 2.11, since the cotangent vector fields $\mathrm{d}X^{(j)}$ are smooth. Let us verify that $\mu = L_g^*\mu$. First of all, we claim that the pullback $(L_g)^*(\mathrm{d}X^{(i)})$ of $\mathrm{d}X^{(i)}$ is again $\mathrm{d}X^{(i)}$

for all $i = 1, \dots, k$. In order to show this, it is enough to show that for all fixed $h \in G$ we have $[(L_g)^*(\mathrm{d}X^{(i)})]_h(X_h^{(j)}) = \delta_{i,j}$, by the definition of $\mathrm{d}X^{(i)}$. Indeed, we have that

$$[(L_g)^*(\mathrm{d}X^{(i)})]_h(X_h^{(j)}) = \mathrm{d}X^{(i)}_{L_g(h)}(DL_g(X_h^{(j)})) = \mathrm{d}X^{(i)}_{gh}(X_{gh}^{(j)}) = \delta_{i,j},$$

hence our claim is proved.

Now, for any continuous function $f\colon G \to \mathbb{R}$, by the change of variable formula, we have

$$\begin{aligned}\int_G f \circ L_h \,\mathrm{d}\mu &= \int_{L_h(G)} f\,(L_h)^*(\mathrm{d}X^{(1)} \cdots \mathrm{d}X^{(k)}) = \int_{L_h(G)} f\,[(L_h)^*(\mathrm{d}X^{(1)})] \cdots [(L_h)^*(\mathrm{d}X^{(k)})] \\ &= \int_{hG} f\,\mathrm{d}X^{(1)} \cdots \mathrm{d}X^{(k)} = \int_G f\,\mathrm{d}\mu.\end{aligned}$$

This completes the proof of the existence of a measure as in the statement of the theorem.

Let us verify the uniqueness claim. The idea is that a left-invariant differential k-form is uniquely determined by its restriction to T_eG, and, by the previous discussion, all multilinear alternating forms on T_eG are multiples of each other. Formally, let ν be another smooth measure as in the statement of the theorem. Then, ν is defined by integrating a smooth k-differential form. In particular, there exists a smooth function $f\colon G \to \mathbb{R}_{\geq 0}$ such that for all $g \in G$ we have $\mathrm{d}\nu(g) = f(g)\,\mathrm{d}X^{(1)} \cdots \mathrm{d}X^{(k)}$. By invariance under L_h for all $h \in G$, we deduce that f must be constant. This proves the uniqueness claim and hence completes the proof. □

The important point to remember is that on any matrix Lie group, up to a normalization factor, there is a unique smooth measure that is invariant by all left translations $g \mapsto hg$. We will come back to the case of $\mathrm{SL}(2,\mathbb{R})$ in Chapter 4. In the case of the Heisenberg group, the Haar measure is actually the Lebesgue measure on $\mathbb{R}^3$, as the next exercise shows.

Exercise 2.24. *Let μ denote the Haar measure on* Heis*, normalized so that*

$$\mu\left(\left\{\begin{pmatrix}1 & x & z\\ 0 & 1 & y\\ 0 & 0 & 1\end{pmatrix} : x,y,z \in [0,1]\right\}\right) = 1.$$

For any function $f\colon$ Heis $\to \mathbb{R}$ and any $g = \left(\begin{smallmatrix}1 & x & z\\ 0 & 1 & y\\ 0 & 0 & 1\end{smallmatrix}\right) \in$ Heis, write $f(g) = f(x,y,z)$. Show that for any continuous function $f\colon$ Heis $\to \mathbb{R}$ we have

$$\int_{\mathrm{Heis}} f(g)\,\mathrm{d}\mu(g) = \int_{\mathbb{R}^3} f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z.$$

In other words, μ coincides with the Lebesgue measure on $\mathbb{R}^3$.

2.3.2 The Killing form

We can define a symmetric bilinear form on $\mathfrak{g}$ as follows. Recall that, for all $\mathbf{x} \in \mathfrak{g}$, its adjoint is given by $\mathfrak{ad}_{\mathbf{x}} = [\mathbf{x}, \cdot]_{\mathfrak{g}} \in \mathrm{Mat}(k,\mathbb{R})$, where $k = \dim \mathfrak{g}$.

Definition 2.25. *The Killing form* B *is a bilinear symmetric form on $\mathfrak{g}$ defined by*

$$\mathrm{B}(\mathbf{x},\mathbf{y}) := \mathrm{tr}(\mathfrak{ad}_{\mathbf{x}} \circ \mathfrak{ad}_{\mathbf{y}}), \qquad \textit{for all } \mathbf{x},\mathbf{y} \in \mathfrak{g}.$$

The fact that B is bilinear follows from the linearity of the Lie bracket $\mathfrak{ad}_{a\mathbf{x}+b\mathbf{y}} = [a\mathbf{x}+b\mathbf{y}, \cdot]_{\mathfrak{g}} = a[\mathbf{x},\cdot]_{\mathfrak{g}} + b[\mathbf{y},\cdot]_{\mathfrak{g}} = a\mathfrak{ad}_{\mathbf{x}} + b\mathfrak{ad}_{\mathbf{y}}$, as the reader can easily check. The symmetry of B follows from the properties of the trace: for any matrices A, B we have $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

The Killing form is Ad-invariant, as the next lemma shows.

Lemma 2.26. *For all $g \in G$, we have*

$$\mathrm{B}(\mathrm{Ad}(g)\mathbf{x}, \mathrm{Ad}(g)\mathbf{y}) = \mathrm{B}(\mathbf{x}, \mathbf{y}), \qquad \textit{for all } \mathbf{x}, \mathbf{y} \in \mathfrak{g}.$$

Proof. We first claim that, for all $g \in G$ and $\mathbf{x} \in \mathfrak{g}$, we have

$$\mathfrak{ad}_{\mathrm{Ad}(g)\mathbf{x}} = \mathrm{Ad}(g) \circ \mathfrak{ad}_{\mathbf{x}} \circ \mathrm{Ad}(g)^{-1}.$$

Indeed, let $\mathbf{y} \in \mathfrak{g}$. Straightforward computations give us

$$\begin{aligned}\mathfrak{ad}_{\mathrm{Ad}(g)\mathbf{x}}(\mathbf{y}) &= [\mathrm{Ad}(g)\mathbf{x}, \mathbf{y}]_{\mathfrak{g}} = g^{-1}\mathbf{x}g\mathbf{y} - \mathbf{y}g^{-1}\mathbf{x}g = g^{-1}(\mathbf{x}g\mathbf{y}g^{-1} - g\mathbf{y}g^{-1}\mathbf{x})g \\ &= \mathrm{Ad}(g)(\mathbf{x}\,\mathrm{Ad}(g^{-1})\mathbf{y} - \mathrm{Ad}(g^{-1})\mathbf{y}\mathbf{x}) = (\mathrm{Ad}(g) \circ \mathfrak{ad}_{\mathbf{x}} \circ \mathrm{Ad}(g)^{-1})(\mathbf{y}),\end{aligned}$$

which proves our claim. Then,

$$\begin{aligned}\mathrm{B}(\mathrm{Ad}(g)\mathbf{x}, \mathrm{Ad}(g)\mathbf{y}) &= \mathrm{tr}(\mathfrak{ad}_{\mathrm{Ad}(g)\mathbf{x}} \circ \mathfrak{ad}_{\mathrm{Ad}(g)\mathbf{y}}) \\ &= \mathrm{tr}\left((\mathrm{Ad}(g) \circ \mathfrak{ad}_{\mathbf{x}} \circ \mathrm{Ad}(g)^{-1}) \circ (\mathrm{Ad}(g) \circ \mathfrak{ad}_{\mathbf{y}} \circ \mathrm{Ad}(g)^{-1})\right) \\ &= \mathrm{tr}(\mathrm{Ad}(g) \circ \mathfrak{ad}_{\mathbf{x}} \circ \mathfrak{ad}_{\mathbf{y}} \circ \mathrm{Ad}(g)^{-1}) \\ &= \mathrm{tr}(\mathfrak{ad}_{\mathbf{x}} \circ \mathfrak{ad}_{\mathbf{y}}) = \mathrm{B}(\mathbf{x}, \mathbf{y}),\end{aligned}$$

where we used the fact that the trace is invariant under conjugation. □

Let us compute the Killing form in the case of $\mathfrak{sl}(2,\mathbb{R})$. Let us fix the usual basis $\{\mathbf{u}, \mathbf{a}, \mathbf{v}\}$ as in §2.1.2. Then, using the computations (2.1), we can write

$$\mathfrak{ad}_{\mathbf{u}} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathfrak{ad}_{\mathbf{a}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad \mathfrak{ad}_{\mathbf{v}} = \begin{pmatrix} 0 & 0 & 0 \\ -2 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

In order to compute the Killing form, it is enough to compute six matrices, for example $\mathfrak{ad}_{\mathbf{u}}^2, \mathfrak{ad}_{\mathbf{a}}^2, \mathfrak{ad}_{\mathbf{v}}^2$, and $\mathfrak{ad}_{\mathbf{u}} \circ \mathfrak{ad}_{\mathbf{a}}, \mathfrak{ad}_{\mathbf{a}} \circ \mathfrak{ad}_{\mathbf{v}}, \mathfrak{ad}_{\mathbf{u}} \circ \mathfrak{ad}_{\mathbf{v}}$, and look at their traces. In matrix form, we get

$$\mathrm{B}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \begin{pmatrix} 0 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 0 \end{pmatrix} \mathbf{y} \tag{2.10}$$

where ^{T} denotes the transpose. We conclude that the Killing form B on $\mathfrak{sl}(2,\mathbb{R})$ is non-degenerate and has signature $(2,1)$. From this, we can prove an important geometrical fact that links the algebraic properties of $\mathrm{SL}(2,\mathbb{R})$ to hyperbolic geometry.

Proposition 2.27. *Let $\mathscr{H}$ be the hyperboloid model of the hyperbolic plane, that is the set*

$$\mathscr{H} := \{\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_1 > 0 \textit{ and } x_1^2 - x_2^2 - x_3^2 = 1\},$$

equipped with the hyperbolic distance

$$d_{\mathscr{H}}(\mathbf{x}, \mathbf{y}) := \mathrm{arcosh}(x_1y_1 - x_2y_2 - x_3y_3).$$

The group

$$\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm e\}$$

acts on $\mathscr{H}$ by hyperbolic isometries.

Proof. We start by diagonalizing the Killing form, namely we can find positive constants a_1, a_2, a_3 such that the Killing form with respect to the basis $\{a_1(\mathbf{u}-\mathbf{v}), a_2\mathbf{a}, a_3(\mathbf{u}+\mathbf{v})\}$ can be expressed as

$$\mathrm{B}(\mathbf{x},\mathbf{y}) = -x_1y_1 + x_2y_2 + x_3y_3.$$

By Lemma 2.26, since $\mathrm{Ad}(g)$ preserves B, we have that $\mathrm{Ad}(g)$ maps the set of vectors $\mathbf{x} \in \mathbb{R}^3$ which satisfy $\mathrm{B}(\mathbf{x},\mathbf{x}) = -1$ into itself. Moreover, one can verify by hand that if $x_1 > 0$, then the first coordinate of $\mathrm{Ad}(g)\mathbf{x}$ also is positive. Therefore, $\mathrm{Ad}(g)$ maps $\mathscr{H} = \{\mathbf{x} \in \mathbb{R}^3 : \mathrm{B}(\mathbf{x},\mathbf{x}) = -1 \text{ and } x_1 > 0\}$ into itself. Again by Lemma 2.26, $\mathrm{Ad}(g)$ is an isometry with respect to $d_{\mathscr{H}}$.

We have shown that

$$\mathrm{Ad}\colon \mathrm{SL}(2,\mathbb{R}) \to O(2,1)$$

is a smooth homomorphism into the orthogonal group of signature $(2,1)$. It remains to show that its kernel is $\{\pm e\}$. Clearly, $-e \in \ker(\mathrm{Ad})$, so we need to verify the other inclusion. Let $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \ker(\mathrm{Ad})$, then $\mathrm{Ad}(g)\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ implies $d = \pm 1$ and $c = 0$, so that $a = \pm 1$. Repeating the same argument with $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ shows that $b = 0$ and hence $g = \pm e$, which concludes the proof. □

As we just saw, in the case of $\mathrm{SL}(2,\mathbb{R})$, the Killing form is non-degenerate. This is not always the case.

Exercise 2.28. *Show that the Killing form on* Heis *is identically zero, that is* $\mathrm{B}(\mathbf{x},\mathbf{y}) = 0$ *for all* $\mathbf{x},\mathbf{y} \in \mathfrak{h}$.

Lie groups for which the Killing form is non-degenerate have a special name. Proposition 2.30 below, which we will not prove, gives some equivalent conditions.

Definition 2.29. *A matrix Lie group G for which the Killing form on its Lie algebra $\mathfrak{g}$ is non-degenerate is called* semisimple.

Proposition 2.30 (Cartan's Criterion for semisimplicity)**.** *Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. The following are equivalent:*

1. *G is semisimple,*
2. *$\mathfrak{g}$ is a direct sum of simple algebras,*
3. *$\mathfrak{g}$ has no non-zero abelian ideals.*

2.3.3 The Casimir operator

We conclude this section by introducing the Casimir operator, which will play a key role when we discuss the quantitative ergodic properties of homogeneous flows on $\mathrm{SL}(2,\mathbb{R})$ in Chapter 7.

Recall that any element $\mathbf{x} \in \mathfrak{g}$ can be seen as a (left-invariant) vector field $X = \mathbf{x} \in \mathrm{Lie}(G)$ on G, and hence as a first order differential operator. The Casimir operator is a second order differential operator on G, which, roughly speaking, plays the same role in the harmonic analysis on G that the operator $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ on $\mathbb{R}$ does in the Fourier analysis in one variable.

Let G be a semisimple matrix Lie group, and let B be the Killing form. Let $\mathscr{B} = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ be a basis of $\mathfrak{g}$. Since B is non-degenerate, we can construct the *dual basis* $\widehat{\mathscr{B}} = \{\widehat{\mathbf{x}}_1, \dots, \widehat{\mathbf{x}}_k\}$ given by the condition $\mathrm{B}(\mathbf{x}_i, \widehat{\mathbf{x}}_j) = \delta_{i,j}$.

Definition 2.31. *The* Casimir operator $\Box = \Box_{\mathscr{B}}$ *associated to the basis $\mathscr{B}$ is the second order differential operator on G given by*

$$\Box = \sum_{i=1}^{k} \mathbf{x}_i \widehat{\mathbf{x}}_i.$$

The definition is actually independent of the choice of basis: the matrix expressing the change of basis of the dual basis is the inverse transpose of the matrix expressing the change of the original basis, and these cancel out when computing the Casimir in the new basis.

Let us compute the Casimir operator for $\mathrm{SL}(2,\mathbb{R})$. Since the choice of basis is irrelevant, we continue working with $\{\mathbf{u},\mathbf{a},\mathbf{v}\}$. From (2.10), it immediately follows that $\widehat{\mathbf{u}} = \frac{1}{4}\mathbf{v}, \widehat{\mathbf{a}} = \frac{1}{2}\mathbf{a}$, and $\widehat{\mathbf{v}} = \frac{1}{4}\mathbf{u}$, so that

$$\square = \frac{1}{4}\mathbf{u}\,\mathbf{v} + \frac{1}{2}\mathbf{a}^2 + \frac{1}{4}\mathbf{v}\,\mathbf{u}.$$

The most important property of the Casimir operator, and the one we will use in this course, is that it commutes with all the elements of $\mathrm{Lie}(G) = \mathfrak{g}$.

Proposition 2.32. *For every $\mathbf{y} \in \mathfrak{g}$ we have $\square\mathbf{y} = \mathbf{y}\square$.*

Proof. We first show that for all $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathfrak{g}$ we have

$$\mathrm{B}(\mathfrak{ad}_{\mathbf{x}}(\mathbf{y}),\mathbf{z}) + \mathrm{B}(\mathbf{y},\mathfrak{ad}_{\mathbf{x}}(\mathbf{z})) = 0. \tag{2.11}$$

Indeed, from Lemma 2.26, for all $t \in \mathbb{R}$ we have

$$\mathrm{B}\left(\mathbf{y},\mathbf{z}\right) = \mathrm{B}\left(\mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{y}, \mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{z}\right).$$

Differentiating at $t = 0$, we get

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \mathrm{B}\left(\mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{y}, \mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{z}\right) \\
&= \mathrm{B}\left(\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{y}, \mathbf{z}\right) + \mathrm{B}\left(\mathbf{y}, \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \mathrm{Ad}(\exp(-t\mathbf{x}))\mathbf{z}\right) \\
&= \mathrm{B}(\mathfrak{ad}_{\mathbf{x}}(\mathbf{y}),\mathbf{z}) + \mathrm{B}(\mathbf{y},\mathfrak{ad}_{\mathbf{x}}(\mathbf{z})),
\end{aligned}$$

where we used (2.9).

Let us now fix a basis $\{\mathbf{x}_1,\dots,\mathbf{x}_k\}$ and its dual $\{\widehat{\mathbf{x}}_1,\dots,\widehat{\mathbf{x}}_k\}$, and let $\mathbf{y} \in \mathfrak{g}$. Let $c_{i,j}, d_{i,j} \in \mathbb{R}$ be such that

$$[\mathbf{y},\mathbf{x}_i]_{\mathfrak{g}} = \sum_{j=1}^{k} c_{i,j}\mathbf{x}_j, \qquad \text{and} \qquad [\mathbf{y},\widehat{\mathbf{x}}_i]_{\mathfrak{g}} = \sum_{j=1}^{k} d_{i,j}\widehat{\mathbf{x}}_j.$$

Thus,

$$\mathrm{B}(\mathfrak{ad}_{\mathbf{y}}(\mathbf{x}_i),\widehat{\mathbf{x}}_j) = c_{i,j}, \qquad \text{and} \qquad \mathrm{B}(\mathbf{x}_i,\mathfrak{ad}_{\mathbf{y}}(\widehat{\mathbf{x}}_j)) = d_{j,i},$$

which, by (2.11), yields $c_{i,j} + d_{j,i} = 0$.

We conclude

$$\begin{aligned}
[\square,\mathbf{y}] &= \sum_{i=1}^{k} \mathbf{x}_i\widehat{\mathbf{x}}_i\mathbf{y} - \mathbf{y}\,\mathbf{x}_i\widehat{\mathbf{x}}_i = \sum_{i=1}^{k} \mathbf{x}_i\left[\widehat{\mathbf{x}}_i,\mathbf{y}\right] - \left[\mathbf{y},\mathbf{x}_i\right]\widehat{\mathbf{x}}_i \\
&= \sum_{i,j=1}^{k} \mathbf{x}_i\left(-d_{i,j}\widehat{\mathbf{x}}_j\right) - \sum_{i,j=1}^{k} c_{i,j}\mathbf{x}_j\widehat{\mathbf{x}}_i = 0,
\end{aligned}$$

hence the proof is complete. □

Corollary 2.33. *For every $f \in \mathscr{C}^\infty(G)$ and for every $\mathbf{x} \in \mathfrak{g}\setminus\{\mathbf{0}\}$, we have*

$$(\square f)\circ\varphi_t^{\mathbf{x}} = \square(f\circ\varphi_t^{\mathbf{x}}), \quad \textit{for all } t \in \mathbb{R}.$$

Proof. By Proposition 2.32, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0}(\Box f)\circ\varphi_t^{\mathbf{x}} = \mathbf{x}(\Box f) = \Box(\mathbf{x}f) = \Box\left(\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} f\circ\varphi_t^{\mathbf{x}}\right).$$

Integrating both sides with respect to t gives the result (notice that we can interchange the Casimir with the integration since f is smooth and hence locally absolutely integrable). $\Box$

One could prove a stronger result, namely that, up to normalization, the Casimir is the only operator with such property, but this goes beyond the scope of this course.

2.4 Homogeneous spaces

We have defined homogeneous flows and we have a smooth measure on the group. At the moment, we still do not know whether homogeneous flows preserve the Haar measure, but we also have another issue to worry about. The problem we face now is that, in the examples we are interested in, the group G has *infinite* measure. This is the same situations as for linear flows in Chapter 1: $\mathbb{R}^n$ has infinite volume and the dynamics of the linear flows is not interesting, since every points escapes to infinity. To have some recurrence, as we did in Chapter 1 by looking at linear flows on tori, we want to quotient the group G by a discrete subgroup and study the homogeneous flows on these finite-volume quotients of G.

2.4.1 Discrete subgroups

Let us start with an important observation. In the same way as we did for both tangent vectors and for multilinear alternating forms, we can "move" inner products from the tangent space at the identity to the tangent space at any other point using left-multiplication maps. More precisely, let us fix a basis $\mathscr{B} = \{\mathbf{x}_1,\dots,\mathbf{x}_k\}$ of $\mathfrak{g}$, and let us define an inner product $\langle\cdot,\cdot\rangle_{\mathfrak{g}}$ on $\mathfrak{g}$ by declaring that $\mathscr{B}$ is orthonormal and extending it by linearity. For any $g\in G$, we can define an inner product on the tangent space T_gG at g by

$$\langle X,Y\rangle_{T_gG} = \langle DL_{g^{-1}}(g)X, DL_{g^{-1}}(g)Y\rangle_{\mathfrak{g}}.$$

This definition gives us a *metric* on G, which allows us to compute the length $\ell(\gamma)$ of smooth curves $\gamma\colon [a,b]\to G$ by

$$\ell(\gamma) = \int_a^b \sqrt{\langle\dot\gamma(t),\dot\gamma(t)\rangle_{T_{\gamma(t)}G}}\,\mathrm{d}t.$$

Since the metric is left-invariant by construction (as in the case of vector fields and of multilinear alternating forms), it satisfies $\ell(g\gamma)=\ell(\gamma)$ for all $g\in G$. The resulting *distance* d_G on G

$$d_G(g,h) = \inf\{\ell(\gamma):\ \gamma \text{ is a smooth curve from } g \text{ to } h\}$$

is left-invariant as well, in other words

$$d_G(g_1,g_2) = d_G(hg_1,hg_2), \qquad \text{for all } h\in G.$$

At this point, the metric d_G depends on the choice of basis, but this will not be important for what follows.

Remark 2.34. *1. If the Lie group G is not connected, then the distance between two points in different connected components is infinite. This will not be a problem for us since we will only consider connected Lie groups.*

2. *It is not hard to see that the topology induced by any left-invariant metric is equivalent to the Euclidean topology on* $\mathrm{GL}(n,\mathbb{R}) \subset \mathbb{R}^{n^2}$. *For a proof of this fact, the reader can refer to [4, Lemma 9.12].*

Let $\Gamma < G$ be a *discrete subgroup* of G, that is, a subgroup consisting of isolated points. Then, Γ acts on G by left-multiplications, and all these multiplication maps are isometries, since d_G is left-invariant. With this in mind, we can prove the following lemma.

Lemma 2.35. *Let Γ be a discrete subgroup of a matrix Lie group G. Then, Γ acts properly discontinuously on G. In other words, for every compact set $K \subset G$, the set $\{h \in \Gamma : hK \cap K \neq \emptyset\}$ is finite.*

Proof. Assume that this is not the case, namely there exists a compact set $K \subset G$ and an infinite set of elements $h_i \in \Gamma$ such that $h_i K \cap K \neq \emptyset$. Therefore, for every $p \in K$, we have $d_G(p, h_i p) \leq R$, where R is twice the diameter of K. This implies that the isometries L_{h_i} are equibounded on K. Since they are clearly equicontinous (they are isometries!), by the Ascoli-Arzelà Theorem, there is a subsequence $h_{i_j} \in \Gamma$ such that the maps $L_{h_{i_j}} : g \mapsto h_{i_j} g$ converge uniformly on K. Let us consider the sequence of elements $\ell_j = h_{i_j}^{-1} h_{i_{j+1}} \in \Gamma$. The elements h_i are assumed to be all distinct, so $\ell_j \neq e$. Moreover, the sequence of isometries L_{ℓ_j} converges uniformly on K to the identity. By Remark 2.34, the right-multiplication map $R_{g^{-1}}$ is continuous with respect to d_G; hence, the sequence $\ell_j \to e$. This violates the assumptions that Γ is discrete, hence proves the lemma. □

2.4.2 Quotients

Let $\Gamma < G$ be a discrete subgroup of G, and let $M = \Gamma \backslash G$ the quotient space. From Lemma 2.35, we can deduce that M is a smooth manifold.

Lemma 2.36. *If $\Gamma < G$ is discrete, then $M = \Gamma \backslash G$ is a smooth manifold. Moreover, if Γ is normal in G, then $M = G/\Gamma$ is a Lie group.*

Proof. By Lemma 2.35, the quotient map $\pi \colon G \to M$ is a covering map; in other words, for every $p = \Gamma g \in M$ there exists an open neighbourhood U_p of p such that $\pi^{-1}(U_p)$ is a disjoint union of open neighbourhoods U_i of points in $g_i \in \pi^{-1}\{p\}$. We can define charts on M as follows. For every $p \in M$, choose $g \in G$ and open neighbourhoods U_p of p and U_g of g such that $\pi(g) = p$ and $\pi \colon U_g \to U_p$ is a homeomorphism. The atlas on M is defined by composing the local sections $\pi^{-1} \colon U_p \to U_g$ with the charts of G. The transition maps are given by the left multiplication maps L_h for $h \in \Gamma$, which are smooth since G is a Lie group. This proves the first part.

If Γ is a normal subgroup, then the quotient inherits a group structure which is clearly smooth in the charts we defined. □

Corollary 2.37. *The group*

$$\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm e\}$$

is a Lie group. Its Lie algebra coincides with the Lie algebra of $\mathrm{SL}(2,\mathbb{R})$.

Proof. We apply the previous lemma with $\Gamma = \{\pm e\} = Z(\mathrm{SL}(2,\mathbb{R}))$, which is the centre of $\mathrm{SL}(2,\mathbb{R})$. Since $-e$ acts properly discontinuously on $\mathrm{SL}(2,\mathbb{R})$, the tangent space at the identity e is the same, hence $\mathrm{Lie}(\mathrm{PSL}(2,\mathbb{R})) = \mathfrak{sl}(2,\mathbb{R})$. □

Once we fixed a left-invariant distance and a Haar measure on G, we obtain a well-defined distance and measure on M as well by composing them with local sections $\pi_i^{-1} \colon M \to G$. The definitions are well-posed, since they do not depend on the choice of local section, by the left-invariance properties of the distance and the Haar measure. Therefore, M is a metric and measured space which is *locally isometric* to G.

Definition 2.38. 1. *The quotients $M = \Gamma\backslash G$ of matrix Lie groups by discrete subgroups Γ are called* homogeneous manifolds.

2. *A discrete subgroup Γ such that the quotient M has finite volume with respect to a (equivalently, any) Haar measure is called a* lattice. *A discrete subgroup Γ such that the quotient M is compact is called a* uniform *or* co-compact lattice.

We will see examples of lattices in the next chapters. Notice that any uniform lattice is also a lattice. The converse is not true, as we will see later in this course.

Exercise 2.39. $(^*)$ *Let $M = \Gamma\backslash G$ be a compact homogeneous manifold, and let μ be the probability measure on M induced by the Haar measure on G. Let $X \in \mathrm{Lie}(G)$, seen as a first order differential operator (i.e., a derivation) on M. Show that for all $f \in \mathscr{C}^\infty(M)$ we have*

$$\int_M Xf \, \mathrm{d}\mu = 0.$$

We conclude this chapter with the following simple but important observation. Let $\varphi^{\mathbf{v}}$ be the homoegeneous flows on G defined by $\mathbf{v} \in \mathfrak{g} \setminus \{0\}$, and let $M = \Gamma\backslash G$ be a homogeneous manifold. Since $\varphi^{\mathbf{v}}$ is obtained by multiplying *on the right* $\varphi^{\mathbf{v}}_t(g) = g\exp(t\mathbf{v})$ and M consists of left cosets, the flow (which, by a little abuse of notation, we denote with the same symbol)

$$\varphi^{\mathbf{v}} \colon \mathbb{R} \times M \to M, \qquad \varphi^{\mathbf{v}}_t(\Gamma g) = \Gamma g \exp(t\mathbf{v})$$

is well-defined and is called the *homogeneous flow on M generated by* $\mathbf{v}$.

Chapter 3

Heisenberg nilflows

In this chapter we will study homogeneous flows on quotients of the Heisenberg group and we will look into a nice application to a problem in number theory. More advanced material on general nilpotent Lie groups can be found in [2].

3.1 Heisenberg nilmanifolds

3.1.1 Preliminaries on the Heisenberg group

Let us recall that the Heisenberg group is the 3 dimensional matrix Lie group

$$\mathrm{Heis} = \left\{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : x, y, z \in \mathbb{R} \right\}.$$

Although it is not an Abelian group, it is not very far from it. Let us explain what we mean. Recall that, for any group G, the *commutator* (or *derived*) *subgroup* $G' = [G,G]$ is the subgroup of G generated by all elements of the form $g^{-1}h^{-1}gh$ for $g,h \in G$. It is not hard to see that G' is a normal subgroup of G and the quotient $\mathrm{ab}(G) := G/G'$ is the largest Abelian quotient of G, which is called the *Abelianization of* G. In particular, G is Abelian if and only if $G' = \{e\}$, and hence $\mathrm{ab}(G) = G$.

In our case, if $g,h \in \mathrm{Heis}$ with

$$g = \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}, \; h = \begin{pmatrix} 1 & u & w \\ 0 & 1 & v \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{then } g^{-1}h^{-1}gh = \begin{pmatrix} 1 & 0 & vx-yu \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It follows that the commutator subgroup is

$$\mathrm{Heis}' = \left\{ \begin{pmatrix} 1 & 0 & z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} : z \in \mathbb{R} \right\} = Z(\mathrm{Heis}).$$

which coincides with the centre $Z(\mathrm{Heis})$ of Heis. We conclude that $\mathrm{ab}(\mathrm{Heis}) \simeq \mathbb{R}^2$, where the isomorphism is induced by

$$g = \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \mapsto (x,y) \in \mathbb{R}^2,$$

whose kernel is exactly $Z(\mathrm{Heis})$. From a geometrical point of view, this means that the Heisenberg group is a connected and simply connected manifold, which can be expressed as a (non-trivial) line bundle over $\mathbb{R}^2 \simeq \mathrm{ab}(\mathrm{Heis})$.

Let us look at its Lie algebra $\mathfrak{h}$. From Exercise 2.7, we get that $\mathfrak{h}$ is the 3 dimensional vector space

$$\mathfrak{h} = \left\{ \begin{pmatrix} 0 & x & z \\ 0 & 0 & y \\ 0 & 0 & 0 \end{pmatrix} : x, y, z \in \mathbb{R} \right\} = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\},$$

where

$$\mathbf{e}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

With the aid of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, we will from now on identify $\mathfrak{h}$ with $\mathbb{R}^3$.

Notice that the centre $\mathfrak{z}(\mathfrak{h})$ of $\mathfrak{h}$, namely the set of vectors whose Lie bracket with any other vector is zero, is $\operatorname{span}\{\mathbf{e}_3\}$, which also coincides with the non-trivial ideal $\mathfrak{h}' = [\mathfrak{h}, \mathfrak{h}]_{\mathfrak{h}}$. Since $[\mathfrak{h}, \mathfrak{h}']_{\mathfrak{h}} = \{\mathbf{0}\}$, we say that $\mathfrak{h}$ is a *nilpotent Lie algebra of step 2*.

3.1.2 The exponential coordinates

Recall that the exponential map exp maps $\mathfrak{h}$ into Heis. Let $\mathbf{v} = (v_1, v_2, v_3) \in \mathfrak{h}$, and let us compute its exponential. We have

$$\exp(\mathbf{v}) = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{v}^k = e + \mathbf{v} + \frac{1}{2}\mathbf{v}^2 = \begin{pmatrix} 1 & v_1 & v_3 + \frac{1}{2}v_1 v_2 \\ 0 & 1 & v_2 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.1}$$

since $\mathbf{v}^k = \mathbf{0}$ for all $k \geq 3$. We now prove the following simple but important result, which the reader might have already proved by themselves in Exercise 2.18-(a).

Lemma 3.1. *The exponential map is a global diffeomorphism between* $\mathfrak{h}$ *and* Heis.

Proof. From (3.1), it is easy to see that the map

$$\log\colon \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \mapsto (x, y, z - xy/2)$$

is the inverse of exp. Both exp and log are smooth since they are polynomial maps in coordinates. □

From Lemma 3.1, it follows that we can define a new atlas on Heis consisting of a single chart given by the exponential map. This new set of coordinates are called *exponential coordinates*. Let us now compute how the matrix multiplication on Heis translates in exponential coordinates.

Lemma 3.2 (Baker-Campbell-Hausdorff Formula for the Heisenberg group)**.** *For any* $\mathbf{v}, \mathbf{w} \in \mathfrak{h}$*, we have* $\exp(\mathbf{v})\exp(\mathbf{w}) = \exp(\mathbf{v} * \mathbf{w})$*, where*

$$\mathbf{v} * \mathbf{w} = \mathbf{v} + \mathbf{w} + \frac{1}{2}[\mathbf{v}, \mathbf{w}]_{\mathfrak{h}}.$$

Proof. Let $\mathbf{v} = (v_1, v_2, v_3)$ and $\mathbf{w} = (w_1, w_2, w_3)$. From (3.1) we get

$$\exp(\mathbf{v})\exp(\mathbf{w}) = \begin{pmatrix} 1 & v_1 + w_1 & v_3 + \frac{1}{2}v_1 v_2 + v_1 w_2 + w_3 + \frac{1}{2}w_1 w_2 \\ 0 & 1 & v_2 + w_2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Using the formula for log in Lemma 3.1, we conclude

$$\mathbf{v} * \mathbf{w} = \log(\exp(\mathbf{v})\exp(\mathbf{w})) = (v_1 + w_1, v_2 + w_2, v_3 + w_3 + (v_1 w_2 - v_2 w_1)/2),$$

which proves the formula. □

It remains to find an expression for any given Haar measure on Heis in exponential coordinates. In a similar way as the reader did in Exercise 2.24, we prove the following lemma.

Lemma 3.3. *Let μ be a Haar measure on* Heis. *There exists $\lambda > 0$ such that $\mu = \lambda\,\mathrm{Leb}$ on $\mathfrak{h}$. In particular, we can identify* $(\mathrm{Heis}, \cdot, \mu)$ *with* $(\mathfrak{h}, *, \lambda\,\mathrm{Leb})$.

Proof. By Theorem 2.23, it is enough to show that Leb on $\mathfrak{h} = \mathbb{R}^3$ is invariant under all left-multiplication maps. Fix $\mathbf{v} \in \mathfrak{h}$, and let $A \subset \mathfrak{h}$ be a measurable set. If we denote $L_{\mathbf{v}}(\mathbf{x}) = \mathbf{v} * \mathbf{x}$, we need to show that $\mathrm{Leb}(L_{\mathbf{v}}A) = \mathrm{Leb}(A)$. Applying the change of variable formula, we get

$$\mathrm{Leb}(\mathbf{v} * A) = \int_{L_{\mathbf{v}}(A)} \mathrm{d}\mathbf{x} = \int_{\mathfrak{h}} \mathbb{1}_A \circ L_{\mathbf{v}}^{-1}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathfrak{h}} \mathbb{1}_A(\mathbf{x})\,\mathrm{d}L_{\mathbf{v}}(\mathbf{x}) = \int_A |\det DL_{\mathbf{v}}(\mathbf{x})|\,\mathrm{d}\mathbf{x}.$$

From Lemma 3.2, the Jacobian matrix of $L_{\mathbf{v}}$ at $\mathbf{x}$ is

$$DL_{\mathbf{v}}(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -v_2/2 & v_1/2 & 1 \end{pmatrix},$$

which implies that $|\det DL_{\mathbf{v}}(\mathbf{x})| = 1$ and hence proves the invariance of Leb under $L_{\mathbf{v}}$. □

3.1.3 Lattices

By definition, a *Heisenberg nilmanifold* is a quotient $M = \Gamma\backslash\,\mathrm{Heis}$ of Heis by a discrete subgroup Γ. We are interested in the case where Γ is a lattice, namely when the quotient manifold M has finite volume. By the definition of the push-forward measure on M, this is equivalent to asking that a fundamental domain for M in Heis has finite volume. Given a lattice Γ, we choose the normalization of the Haar (Lebesgue) measure μ on Heis so that (any fundamental domain for) M has volume 1. By a little abuse of notation, we will use the letter μ to denote both the Haar measure on Heis, which we identify with $\mathfrak{h}$, and the induced measure on M.

Let us classify all possible Heisenberg nilmanifolds of finite volume. The following exercise will become useful in the proof.

Exercise 3.4. *Show that the following linear transformations of $\mathfrak{h}$ are* group automorphisms:

- $F_{a,b}\colon \mathfrak{h} \to \mathfrak{h}$ *given by*

$$F_{a,b}(x_1, x_2, x_3) := (x_1, x_2, x_3 - ax_1 - bx_2),$$

  where $a, b \in \mathbb{R}$;

- $F_A\colon \mathfrak{h} \to \mathfrak{h}$ *given by*

$$F_A(x_1, x_2, x_3) := (ax_1 + bx_2, cx_1 + dx_2, (\det A)x_3),$$

  where $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}(2, \mathbb{R})$ is an invertible matrix.

Proposition 3.5. *Let $\Gamma <$* Heis *be a lattice. Up to a group automorphism, we have*

$$\Gamma = \Gamma_E := \mathbb{Z}\mathbf{e}_1 * \mathbb{Z}\mathbf{e}_2 * \frac{1}{E}\mathbb{Z}\mathbf{e}_3 = \left\{ \begin{pmatrix} 1 & p & \frac{r}{E} \\ 0 & 1 & q \\ 0 & 0 & 1 \end{pmatrix} : p, q, r \in \mathbb{Z} \right\}.$$

where $E \geq 1$ is a positive integer.

Proof. Let $\Gamma < \mathrm{Heis} = \mathfrak{h}$ be a lattice. We notice that $\Gamma \neq \{(0,0,0)\}$, since $\mathfrak{h}$ has infinite Haar measure. Let us call $\pi\colon \mathfrak{h} \to \Gamma\backslash\mathfrak{h}$ the canonical projection.

Assume that $\Gamma \subset \mathbb{R}\mathbf{e}_3$. Since Γ is a discrete subgroup of $\mathbb{R}\mathbf{e}_3$, it follows that $\Gamma = \{(0,0,mz) : m \in \mathbb{Z}\}$ for some $z \in \mathbb{R}\setminus\{0\}$. Then, the projection π is injective on the set $\mathbb{R}^2 \times (0,z)$, which has infinite Haar measure. This is in contradiction with the assumption that Γ is a lattice.

Since Γ is not contained in $\mathbb{R}\mathbf{e}_3$, there exists $\mathbf{v} = (v_1, v_2, v_3) \in \Gamma$ with $v_1 v_2 \neq 0$. Without loss of generality, let us assume $v_2 \neq 0$. We can find a matrix $A_1 \in \mathrm{SL}(2,\mathbb{R})$ that maps (v_1, v_2) to $(0,1)$. Thus, by Exercise 3.4, up to a group automorphism we can assume $\mathbf{v} = (0,1,v_3)$ and, by the first part of the same exercise, $\mathbf{v} = \mathbf{e}_2 \in \Gamma$.

Let E be the subgroup of $\mathfrak{h}$ given by $E = \mathbb{R}\mathbf{e}_2 * \mathbb{R}\mathbf{e}_3 = \{(0,y,z) : y,z \in \mathbb{R}\}$, which is isomorphic to $\mathbb{R}^2$, and let us assume that $\Gamma \subset E$. Since Γ is a discrete subgroup, there exists an $\varepsilon > 0$ such that there are no elements of Γ in the set $\{(0,y,z) : y,z \in (0,\varepsilon)\}$. This implies that the projection π is injective when restricted to the set $\{(x,y,z) : y,z \in (0,\varepsilon)\}$, which has infinite measure. The contradiction with the lattice assumption implies that Γ is not fully contained in E.

Let $\mathbf{w} = (w_1, w_2, w_3) \in \Gamma \setminus E$, so that $w_1 \neq 0$. Then, we can find $A_2 \in \mathrm{GL}(2,\mathbb{R})$ that maps (w_1, w_2) to $(1,0)$ and fixes $(0,1)$, so that, using both parts of Exercise 3.4, up to a group automorphism we can assume that $\mathbf{e}_1, \mathbf{e}_2 \in \Gamma$. Moreover, we also have that

$$\mathbf{e}_3 = \mathbf{e}_1 * \mathbf{e}_2 * (-\mathbf{e}_1) * (-\mathbf{e}_2) \in \Gamma.$$

Notice that $\Gamma \cap \mathbb{R}\mathbf{e}_3$ is a discrete subgroup of $\mathbb{R} = \mathbb{R}\mathbf{e}_3$ containing $\mathbb{Z}\mathbf{e}_3$, hence $\Gamma \cap \mathbb{R}\mathbf{e}_3 = \frac{1}{E}\mathbb{Z}\mathbf{e}_3$ for some integer $E \geq 1$. In particular, this proves

$$\Gamma_E := \mathbb{Z}\mathbf{e}_1 * \mathbb{Z}\mathbf{e}_2 * \frac{1}{E}\mathbb{Z}\mathbf{e}_3 \subseteq \Gamma.$$

Let us show the other inclusion. The Abelianization $\mathrm{ab}\colon \mathfrak{h} \to \mathfrak{h}/Z(\mathfrak{h}) = \mathbb{R}^2$, in exponential coordinates, is the projection onto the first two coordinates. The projection $\Gamma_{\mathrm{ab}} = \mathrm{ab}(\Gamma)$ of Γ is a subgroup of $\mathbb{R}^2$ which contains $\mathbb{Z}^2$. We claim that the index $[\mathbb{Z}^2 : \Gamma_{\mathrm{ab}}]$ (equivalently, the cardinality of the set $\Gamma_{\mathrm{ab}} \cap [0,1)^2$) is finite. Otherwise, there would exist a sequence $\mathbf{v}_k = (x_k, y_k, z_k) \in \Gamma$ such that the projections $\mathrm{ab}(\mathbf{v}_k) = (x_k, y_k)$ are all distinct elements in $[0,1)^2$. Thus, we could construct a sequence of distinct elements

$$\mathbf{v}_k * (-\lfloor z_k \rfloor \mathbf{e}_3) = (x_k, y_k, z_k - \lfloor z_k \rfloor) \in \Gamma \cap [0,1)^3,$$

from which we could extract a converging subsequence. This contradicts the discreteness of Γ. Therefore $[\mathbb{Z}^2 : \Gamma_{\mathrm{ab}}] < \infty$, which implies in particular that Γ_{ab} has rank 2. Up to a change of basis of $\mathbb{R}^2$, and hence using once more Exercise 3.4, up to a group automorphism we can assume that $\Gamma_{\mathrm{ab}} = \mathbb{Z}^2$.

Let $\mathbf{x} = (x,y,z) \in \Gamma$. Since $\mathrm{ab}(\mathbf{x}) = (x,y) \in \Gamma_{\mathrm{ab}} = \mathbb{Z}^2$, it follows that $x, y \in \mathbb{Z}$, which yields

$$\mathbf{x} * (-y\mathbf{e}_2) * (-x\mathbf{e}_1) = (0,0,z - xy/2) \in \Gamma \cap \mathbb{R}\mathbf{e}_3 = \frac{1}{E}\mathbb{Z}\mathbf{e}_3.$$

This implies that $z - xy/2 = m/E$ for some $m \in \mathbb{Z}$. Therefore

$$\mathbf{x} = (x,y,z) = (x,y,m/E + xy/2) = (x\mathbf{e}_1) * (y\mathbf{e}_2) * \left(\frac{m}{E}\mathbf{e}_3\right) \in \mathbb{Z}\mathbf{e}_1 * \mathbb{Z}\mathbf{e}_2 * \frac{1}{E}\mathbb{Z}\mathbf{e}_3,$$

which proves the other inclusion and completes the proof.

□

A consequence of Proposition 3.5 is that all Heisenberg nilmanifolds of finite volume are also compact, that is to say that all lattices in Heis are uniform. This property holds for all nilpotent matrix Lie groups, although we will see it does not hold in $\mathrm{SL}(2,\mathbb{R})$. For the rest of these notes, all Heisenberg nilmanifolds are assumed to be compact. Then, up to a group automorphism, they are of the form $\Gamma_E\backslash\mathrm{Heis} = \Gamma_E\backslash\mathfrak{h}$ for some integer $E \geq 1$.

Corollary 3.6. *Let $M = \Gamma\backslash\mathfrak{h}$ be a Heisenberg nilmanifold. Then, M is a circle bundle over $\mathbb{T}^2$ with fibers parallel to $\mathbf{e}_3$.*

Proof. By Proposition 3.5, up to an automorphism of Heis (that is, up to a change of coordinates), we have that $\Gamma = \mathbb{Z}\mathbf{e}_1 * \mathbb{Z}\mathbf{e}_2 * \frac{1}{E}\mathbb{Z}\mathbf{e}_3 = \Gamma_E$. Thus, $\mathrm{ab}(\Gamma) = \mathbb{Z}^2$. It is immediate to check that ab induces a well-defined map

$$\begin{aligned}\overline{\mathrm{ab}}\colon M = \Gamma\backslash\mathfrak{h} &\to \mathbb{T}^2\\ \Gamma(x,y,z) &\mapsto [\![x,y]\!]\end{aligned}$$

which expresses M as a bundle over $\mathbb{R}/(E^{-1}\mathbb{Z}) \to M \xrightarrow{\overline{\mathrm{ab}}} \mathbb{T}^2$. Indeed, the preimage of any point $\overline{\mathrm{ab}}(p) \in \mathbb{T}^2$ is a circle $p * (s\mathbf{e}_3)$, with $s \in \mathbb{R}/(E^{-1}\mathbb{Z})$, which is homeomorphic to $\mathbb{R}/(E^{-1}\mathbb{Z})$. □

3.2 Ergodic properties of Heisenberg nilflows

A *Heisenberg nilflow* is a homogeneous flow on a Heisenberg nilmanifold. Let $M = \Gamma\backslash\mathrm{Heis}$ and let $\mathbf{v} \in \mathrm{Lie}(\mathrm{Heis}) = \mathfrak{h}$, then the nilflow generated by $\mathbf{v}$ is the flow $\varphi_t^{\mathbf{v}}(p) = p\exp(t\mathbf{v})$ where $p \in M$. In exponential coordinates, if $\mathbf{v} = (v_1, v_2, v_3)$, we can write

$$\varphi_t^{\mathbf{v}}(p) = p * t\mathbf{v} = \Gamma(x+tv_1, y+tv_2, z+tv_3+t(xv_2-yv_1)/2), \qquad \text{where } p = \Gamma(x,y,z). \tag{3.2}$$

By Proposition 3.5, up to an automorphism of Heis, we have $\Gamma = \Gamma_E$. In the following, just for simplicity of notation, we will assume that $E = 1$; in other words we will consider the Heisenberg nilmanifold

$$M = \Gamma_1\backslash\mathfrak{h} = \{\Gamma(x,y,z) : x,y,z \in [0,1)\}, \qquad \Gamma_1 = \mathbb{Z} * \mathbb{Z} * \mathbb{Z}.$$

By Lemma 3.3, we can fix a normalization of the Haar measure so that M has volume 1, so that the smooth measure on M coincides with the Lebesgue volume of $[0,1)^3$, which we denote by μ.

Let us verify that $\varphi^{\mathbf{v}}$ preserves μ. Indeed, as we did in Lemma 3.3, we check that, for all $t \in \mathbb{R}$, the Jacobian matrix

$$D\varphi_t^{\mathbf{v}}(p) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ tv_2/2 & -tv_1/2 & 1 \end{pmatrix}$$

of $\varphi_t^{\mathbf{v}}$ at p has determinant 1. Thus, by the change of variable formula, it follows that $\varphi_t^{\mathbf{v}}$ preserves the Lebesgue measure μ. Indeed, the same reasoning yields the following result.

Lemma 3.7. *The Haar measure μ on M is* right-invariant*, namely for every continuous function $f\colon M \to \mathbb{C}$ and for every $\mathbf{x} \in \mathfrak{h}$ we have*

$$\int_M f(p * \mathbf{x})\,\mathrm{d}\mu = \int_M f(p)\,\mathrm{d}\mu.$$

We now investigate the ergodic properties of $\varphi^{\mathbf{v}}$ with respect to the invariant measure μ.

3.2.1 Mixing properties

We start from the strong chaotic properties, since it is quite easy to see that nilflows are never weakly mixing. Actually, there is a clear obstruction to weak-mixing, coming from the Abelian quotient of Heis.

Proposition 3.8. *Let $\phi_t^{\overline{\mathbf{v}}}$ denote the linear flow on $\mathbb{T}^2$ in direction $\overline{\mathbf{v}} = (v_1, v_2) \in \mathbb{R}^2$. Then, $\phi_t^{\overline{\mathbf{v}}}$ is a factor of $\varphi_t^{\mathbf{v}}$; that is, the following diagram*

$$\begin{tikzcd} (M,\mu) \arrow[r, "\varphi_t^{\mathbf{v}}"] \arrow[d, "\overline{\mathrm{ab}}"'] & (M,\mu) \arrow[d, "\overline{\mathrm{ab}}"] \\ (\mathbb{T}^2, \mathrm{Leb}) \arrow[r, "\phi_t^{\overline{\mathbf{v}}}"] & (\mathbb{T}^2, \mathrm{Leb}) \end{tikzcd}$$

commutes for every $t \in \mathbb{R}$. In particular, $\varphi_t^{\mathbf{v}}$ is not weak-mixing.

Proof. The commutativity of the diagram is clear from Corollary 3.6 and (3.2). Let us show $\varphi_t^{\mathbf{v}}$ is not weak-mixing.

For all $(m,n) \in \mathbb{Z}^2 \setminus \{(0,0)\}$, we define $\overline{f}_{m,n}$ on $\mathbb{T}^2$ by $\overline{f}_{m,n} \colon [\![x,y]\!] \mapsto e^{2\pi i(mx+ny)}$. We showed in Lemma 1.25 that $\overline{f}_{m,n}$ is a non-constant eigenfunction of $\phi_t^{\overline{\mathbf{v}}}$ with eigenvalue $e^{2\pi i t\alpha}$, where $\alpha = mv_1 + nv_2$. Then, the function $f_{m,n} := \overline{f}_{m,n} \circ \overline{\mathrm{ab}}$ is a non constant eigenfunction for $\varphi^{\mathbf{v}}$; indeed

$$f_{m,n} \circ \varphi_t^{\mathbf{v}} = \overline{f}_{m,n} \circ \phi_t^{\overline{\mathbf{v}}} \circ \overline{\mathrm{ab}} = e^{2\pi i t\alpha}\, \overline{f}_{m,n} \circ \overline{\mathrm{ab}} = e^{2\pi i t\alpha} f_{m,n}.$$

This shows that $\varphi_t^{\mathbf{v}}$ is not weak-mixing. □

Let $\overline{\mathrm{ab}} \colon (M,\mu) \to (\mathbb{T}^2, \mathrm{Leb})$ be the factor map as in the previous proposition. If we denote by $\overline{\mathrm{ab}}^* \colon L^2(\mathbb{T}^2) \to L^2(M)$ the pull-back $\overline{\mathrm{ab}}^*(f) = f \circ \overline{\mathrm{ab}}$, we have an orthogonal decomposition

$$L^2(M) = \overline{\mathrm{ab}}^* L^2(\mathbb{T}^2) \oplus L^2_0(M), \tag{3.3}$$

where $L^2_0(M)$ consists of all L^2 functions on M whose integral along all circles parallel to $\mathbf{e}_3$ is zero, namely

$$L^2_0(M) = \left\{ f \in L^2(M) : \int_0^1 f(p * s\mathbf{e}_3)\, \mathrm{d}s = 0 \text{ for all } p \in M \right\}.$$

We leave as en exercise to the reader, Exercise 3.10 below, to establish the decomposition (3.3).

In the proof of Proposition 3.8 we showed that there exists a dense set of functions in $\overline{\mathrm{ab}}^* L^2(\mathbb{T}^2)$ which are eigenfunctions for the nilflow. We now prove that, in the orthogonal complement $L^2_0(M)$, typical nilflows are mixing. In some sense, this suggests that the toral factor of Proposition 3.8 is *the only* obstruction to mixing.

Theorem 3.9. *Let $\varphi_t^{\mathbf{v}}$ be a nilflow such either v_1 or v_2 is not zero. Then, for every $f, g \in L^2_0(M)$ we have*

$$\lim_{t\to\infty} \langle f \circ \varphi_t^{\mathbf{v}}, g \rangle = 0.$$

Let us notice that all functions in $L^2_0(M)$ have zero integral, so that the result above indeed shows the asymptotic decorrelations of the observables in $L^2_0(M)$ (i.e., mixing). Moreover, it is clear that if both v_1 and v_2 are zero, the nilflow consists simply of parallel translations along the fibers of the projection on $\mathbb{T}^2$, which is a periodic flow. Hence the assumption on $\mathbf{v}$ is necessary.

The proof of Theorem 3.9 is based on a "wrapping mechanism": short segments in direction $\mathbf{e}_2$ gets sheared along the circle fibers $\mathbb{T}\mathbf{e}_3$ at linear speed (if $v_1 \neq 0$, otherwise one can consider $\mathbf{e}_1$ instead), so that they get wrapped along these fibers. Since the integral of functions in $L^2_0(M)$ on any circle $\mathbb{T}\mathbf{e}_3$ is zero, this implies that, for any continuous function $f \in L^2_0(M)$, the integral of $f \circ \varphi_t^{\mathbf{v}}$ along any short segment in direction $\mathbf{e}_2$ is small. Let us now formalize this idea.

Proof of Theorem 3.9. Since the space of continuous functions $\mathscr{C}_0(M) = \mathscr{C}(M) \cap L^2_0(M)$ in $L^2_0(M)$ is dense in $L^2_0(M)$, it is enough to prove the claim for every $f, g \in \mathscr{C}_0(M)$. Let us assume that $v_1 \neq 0$, the proof in the case $v_2 \neq 0$ is identical.

Let $a = \min\{\|f\|_\infty^{-1}, 1\} \cdot \min\{\|g\|_\infty^{-1}, 1\}$ and let us fix $\varepsilon > 0$. Since f and g are uniformly continuous on M, let $\delta > 0$ be such that if $d(p,q) < \delta$ then both $|f(p) - f(q)| < a\varepsilon$ and $|g(p) - g(q)| < a\varepsilon$. Call $\sigma = a\delta$ and fix $T > (a\varepsilon\sigma v_1)^{-1}$. Let $t \geq T$.

By Lemma 3.7, for any $s \in \mathbb{R}$ we have

$$\langle f \circ \varphi_t^{\mathbf{v}}, g \rangle = \int_M f \circ \varphi_t^{\mathbf{v}}(p) \overline{g}(p) \, \mathrm{d}\mu = \int_M f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2) \overline{g}(p * s\mathbf{e}_2) \, \mathrm{d}\mu.$$

Moreover, for any $s \in [0, \sigma]$, by uniform continuity,

$$|f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2)\overline{g}(p * s\mathbf{e}_2) - f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2)\overline{g}(p)| \leq \|f\|_\infty |g(p * s\mathbf{e}_2) - g(p)| < \varepsilon.$$

In particular, averaging from 0 to σ,

$$\begin{aligned} |\langle f \circ \varphi_t^{\mathbf{v}}, g \rangle| &= \left| \frac{1}{\sigma} \int_0^\sigma \int_M f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2) \overline{g}(p * s\mathbf{e}_2) \, \mathrm{d}\mu \right| \\ &\leq \left| \frac{1}{\sigma} \int_0^\sigma \int_M f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2) \overline{g}(p) \, \mathrm{d}\mu \, \mathrm{d}s \right| + \|f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2)\overline{g}(p * s\mathbf{e}_2) - f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2)\overline{g}(p)\|_\infty \\ &< \int_M \left| \frac{1}{\sigma} \int_0^\sigma f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2) \, \mathrm{d}s \right| \cdot |\overline{g}(p)| \, \mathrm{d}\mu + \varepsilon. \end{aligned} \tag{3.4}$$

We now focus on the integral in absolute value above. Geometrically, it corresponds to integrating the function f along the push-forward of a segment of length σ in direction $\mathbf{e}_2$ under the nilflow $\varphi_t^{\mathbf{v}}$. Let us see that it is "almost vertical". Since

$$s\mathbf{e}_2 * t\mathbf{v} = (tv_1, s + tv_2, tv_3 - (tv_1)/2) \quad \text{and} \quad t\mathbf{v} * s\mathbf{e}_2 = (tv_1, s + tv_2, tv_3 + (tv_1)/2),$$

we can rewrite $s\mathbf{e}_2 * t\mathbf{v} = t\mathbf{v} * (-stv_1)\mathbf{e}_3 * s\mathbf{e}_2$. By uniform continuity of f, we get

$$|f \circ \varphi_t^{\mathbf{v}}(p * s\mathbf{e}_2) - f(p * t\mathbf{v} * (-stv_1)\mathbf{e}_3)| \leq \varepsilon.$$

Thus, if we write $p_t := \varphi_t^{\mathbf{v}}(p) = p * t\mathbf{v}$, from (3.4) we get

$$|\langle f \circ \varphi_t^{\mathbf{v}}, g \rangle| < \int_M \left| \frac{1}{\sigma} \int_0^\sigma f(p_t * (-stv_1)\mathbf{e}_3) \, \mathrm{d}s \right| \cdot |\overline{g}(p)| \, \mathrm{d}\mu + 2\varepsilon. \tag{3.5}$$

For any $q \in M$ and $r \in \mathbb{R}$, from the assumption $f \in \mathscr{C}_0(M)$ it follows that

$$\left| \int_0^r f(q * s\mathbf{e}_3) \, \mathrm{d}s \right| = \left| \int_{\lfloor r \rfloor}^r f(q * s\mathbf{e}_3) \, \mathrm{d}s \right| \leq \|f\|_\infty,$$

since the integral of along a fiber $\mathbb{R}\mathbf{e}_3/\mathbb{Z}\mathbf{e}_3$ is zero. By a change of variable, we get

$$\left| \frac{1}{\sigma} \int_0^\sigma f(p_t * (-stv_1)\mathbf{e}_3) \, \mathrm{d}s \right| = \left| \frac{1}{\sigma t v_1} \int_0^{\sigma t v_1} f(p_t * (-s)\mathbf{e}_3) \, \mathrm{d}s \right| \leq \frac{\|f\|_\infty}{\sigma T v_1} < a\varepsilon \, \|f\|_\infty.$$

Using this observation in (3.5), we conclude

$$|\langle f \circ \varphi_t^{\mathbf{v}}, g \rangle| < a\varepsilon \, \|f\|_\infty \|g\|_\infty + 2\varepsilon < 3\varepsilon,$$

which finishes the proof. □

Exercise 3.10. *Prove the decomposition* (3.3).

3.2.2 Ergodicity

We now look at ergodicity. The result we will prove is the following.

Theorem 3.11. *Let $\varphi^{\mathbf{v}}$ be a Heisenberg nilflow on the Heisenberg nilmanifold $M = \Gamma_1 \backslash \mathfrak{h}$, where $\mathbf{v} = (v_1, v_2, v_3) \in \mathfrak{h}$. The following are equivalent.*

(a) The nilflow $\varphi^{\mathbf{v}}$ is uniquely ergodic.

(b) The Haar measure μ is ergodic for $\varphi^{\mathbf{v}}$.

(c) The projected linear flow $\phi_t^{\overline{\mathbf{v}}} \colon \mathbb{T}^2 \to \mathbb{T}^2$, with $\overline{\mathbf{v}} = (v_1, v_2) \in \mathbb{R}^2$, as in Proposition 3.8, is minimal.

(d) The coordinates v_1 and v_2 are rationally independent.

Clearly, (a) implies (b). By Theorem 1.5, (c) and (d) are equivalent. We leave as an exercise to the reader, Exercise 3.12, to check that (b) implies (d).

Exercise 3.12. *Show that if v_1 and v_2 are rationally dependent, then $\varphi^{\mathbf{v}}$ is not ergodic; in particular (b) implies (d) in Theorem 3.11.*

Our goal is to verify that (d) implies (a). We will actually first prove that (d) implies (b), and then we will show (a) using (b). In order to do this, in Lemma 3.13 below, we provide a *special flow representation* of our nilflow. Let us recall the general definition of special flows first. Given an invertible probability preserving system $T \colon (X, \mu) \to (X, \mu)$ and a positive measurable function $f \colon X \to \mathbb{R}_{>0}$, we define the space

$$X^f := \{(x, r) : x \in X,\ 0 \le r \le f(x)\} \ / \sim,$$

where the relation $\sim$ identifies the points $(x, f(x))$ with $(Tx, 0)$. The special flow over T with roof function f is the measurable flow $T_t = T_t^f$ on X^f defined by $T_t(x, r) = (x, r + t)$, equipped with the invariant measure $\mu \times \mathrm{d}r$, where $\mathrm{d}r$ is the Lebesgue measure on the second coordinate.

Lemma 3.13. *Let $\varphi^{\mathbf{v}}$ be a Heisenberg nilflow on M as above. Assume that $v_2 \neq 0$. Then, $\varphi^{\mathbf{v}}$ is smoothly isomorphic to the special flow over the skew-translation*

$$\begin{aligned} T \colon \mathbb{T}^2 &\to \mathbb{T}^2 \\ [\![x_1, x_2]\!] &\mapsto [\![x_1 + \alpha, x_2 + x_1 + \beta]\!], \end{aligned}$$

where $\alpha = v_1/v_2$ and $\beta = (v_1 + 2v_3)/(2v_2)$, with constant roof function $f = 1/v_2$.

Proof. We will find an embedded submanifold isomorphic to $\mathbb{T}^2$ which intersects all orbits of the nilflow (i.e., a global cross-section for the flow) and whose return time is constant and equal to $1/v_2$. This will suffice to prove our claim.

The set

$$\Sigma = \{\Gamma_1(x, 0, z) : x, z \in [0, 1)\} \subset M$$

is an embedded submanifold in M, and the map $\Gamma_1(x, 0, z) \mapsto [\![x, z]\!]$ realizes an isomorphism between Σ and $\mathbb{T}^2$. Under the assumption $v_2 \neq 0$, all orbits of the nilflow intersect Σ, more precisely if $p = \Gamma_1(x, y, z) \in M$ with $x, y, z \in [0, 1)$, then it is easy to see that $\varphi^{\mathbf{v}}_{-y/v_2}(p) \in \Sigma$.

If $p = \Gamma_1(x, 0, z) \in \Sigma$, then $\varphi^{\mathbf{v}}_t(p) \in \Sigma$ if and only if $tv_2 \in \mathbb{Z}$. It follows that the first return time to Σ is constant and equal to $1/v_2$. The first return map $T = \varphi^{\mathbf{v}}_{1/v_2}$ on Σ is

$$\begin{aligned} Tp = \varphi^{\mathbf{v}}_{1/v_2}(p) &= \Gamma_1(x, 0, z) * (v_1/v_2, 1, v_3/v_2) = \Gamma_1(x + v_1/v_2, 1, z + v_3/v_2 + x/2) \\ &= \Gamma_1 * (-\mathbf{e}_2) * (x + v_1/v_2, 1, z + v_3/v_2 + x/2) = \Gamma_1(x + v_1/v_2, 0, z + v_3/v_2 + x/2 + v_1/(2v_2)), \end{aligned}$$

that is, T is a skew-translation on $\Sigma \simeq \mathbb{T}^2$ of the desired form. By Fubini, the Haar measure μ on M is identified with the product of the Lebesgue measure on Σ and the rescaled Lebesgue measure $v_2\,\mathrm{d}r$ on the second component. □

Assumption (d) is equivalent to say that α in Lemma 3.13 is irrational. Our goal is now to show that if $\alpha \notin \mathbb{Q}$, then the special flow T_t is uniquely ergodic. We first verify that (d) implies (b) in Theorem 3.11 by showing that the Lebesgue measure on $\mathbb{T}^2$ is ergodic for T (the fact that ergodicity of the base map implies ergodicity of the special flow is a general standard fact, but we still show it for the sake of completeness).

Lemma 3.14. *Assume that $\alpha \notin \mathbb{Q}$. Then the skew-translation $T\colon (\mathbb{T}^2, \mathrm{Leb}) \to (\mathbb{T}^2, \mathrm{Leb})$ is ergodic. In particular, if v_1 and v_2 are rationally independent, then the Haar measure μ is ergodic for $\varphi^{\mathbf{v}}$.*

Proof. Similarly to Theorem 1.15, we use Fourier analysis. For any $f \in L^2(\mathbb{T}^2)$, let us consider the Fourier expansion

$$f([\![\mathbf{x}]\!]) = \sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i \mathbf{n}\cdot\mathbf{x}}, \qquad \text{with} \sum_{\mathbf{n}\in\mathbb{Z}^2} |f_{\mathbf{n}}|^2 = \|f\|_2^2.$$

In particular, notice that $|f_{\mathbf{n}}| \to 0$ when $\|\mathbf{n}\|_\infty \to \infty$.

Assume that f is an invariant function, that is assume that $f \circ T = f$ in $L^2(\mathbb{T}^2)$. For all $\mathbf{x} \in \mathbb{R}^2$, by definition

$$f \circ T([\![\mathbf{x}]\!]) = \sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i \mathbf{n}\cdot(x_1+\alpha, x_2+x_1+\beta)} = \sum_{\mathbf{n}\in\mathbb{Z}^2} f_{\mathbf{n}}\, e^{2\pi i n_1 \alpha}\, e^{2\pi i n_2 \beta}\, e^{2\pi i((n_1+n_2)x_1 + n_2 x_2)}.$$

By uniqueness of the coefficients, we must have

$$f_{(n_1+n_2,n_2)} = f_{(n_1,n_2)}\, e^{2\pi i n_1 \alpha}\, e^{2\pi i n_2 \beta} \qquad \text{for all } n_1, n_2 \in \mathbb{Z}. \tag{3.6}$$

In particular, (3.6) implies that $|f_{(n_1+n_2,n_2)}| = |f_{(n_1,n_2)}|$ and, by induction, for all $k \in \mathbb{Z}$, we also have $|f_{(n_1+kn_2,n_2)}| = |f_{(n_1,n_2)}|$. Since we know that $|f_{\mathbf{n}}| \to 0$ when $\|\mathbf{n}\|_\infty \to \infty$, we deduce that $f_{(n_1,n_2)} = 0$ for all $n_2 \neq 0$. On the other hand, when $n_2 = 0$, again from (3.6), we get $f_{(n_1,0)} = f_{(n_1,0)}\, e^{2\pi i n_1 \alpha}$. Since $\alpha \notin \mathbb{Q}$, this forces $f_{(n_1,0)} = 0$ for all $n_1 \neq 0$. We conclude that $f = f_{(0,0)}$ is constant.

We have shown that $\alpha = v_1/v_2 \notin \mathbb{Q}$ implies the ergodicity of the base skew-product. We want to show that this implies ergodicity of the special flow T_t. Let A be an invariant set for T_t. In particular, A is foliated by orbits of T_t. Therefore, the intersection $\Sigma \cap A$ is an invariant set for T. This implies that $\mathrm{Leb}(\Sigma \cap A)$ is 0 or 1. Since the measure μ for the special flow is a product of the Lebesgue measure on the base and of $v_2\,\mathrm{d}r$ on the fibers, by Fubini, it follows that $\mu(A)$ is either 0 or 1. This proves ergodicity of the special flow, and hence of $\varphi^{\mathbf{v}}$. □

We are ready to finish the proof of Theorem 3.11.

Lemma 3.15. *If $\alpha \notin \mathbb{Q}$, then the skew-translation T is uniquely ergodic.*

Proof. Let ν be a probability ergodic invariant measure for S; we need to show that $\nu = \mathrm{Leb}_2$, the two-dimensional Lebesgue measure on $\mathbb{T}^2$. Fix $f \in \mathscr{C}(\mathbb{T}^2)$; we will show that $\nu(f) = \mathrm{Leb}_2(f)$, which implies our result.

Since the Lebesgue measure Leb_2 is ergodic, there exists a set $B \subseteq \mathbb{T}^2$ with $\mathrm{Leb}_2(B) = 1$ such that for all $[\![x_1, x_2]\!] \in B$ we have

$$A_N f([\![x_1,x_2]\!]) := \frac{1}{N}\sum_{n=0}^{N-1} f \circ S^n([\![x_1,x_2]\!]) \to \mathrm{Leb}_2(f) \qquad \text{as } N \to \infty.$$

By Fubini, for Lebesgue almost all $[\![x_1]\!] \in \mathbb{T}^1$ we must have that $\mathrm{Leb}_1(B \cap \{[\![x_1]\!]\} \times \mathbb{T}^1) = 1$. We claim that $\{[\![x_1]\!]\} \times \mathbb{T}^1 \subset B$ for almost all $x \in \mathbb{T}^1$.

For any fixed $[\![x_1]\!] \in \mathbb{T}^1$, the family of functions $\mathscr{A}_{[\![x_1]\!]} := \{A_N f([\![x_1,\cdot]\!])\colon \mathbb{T}^1 \to \mathbb{C}\}_{N\in\mathbb{N}}$ is equibounded and equicontinuous, since f is continuous (hence bounded) and the restriction of T to any fixed fiber is an isometry. By the Ascoli-Arzelà Theorem, the closure of $\mathscr{A}_{[\![x_1]\!]}$ is compact, in particular the sequence $A_N f([\![x_1,\cdot]\!])$ has limit points in $\mathscr{C}(\mathbb{T}^1)$. Since we showed that for almost all $[\![x_1]\!] \in \mathbb{T}^1$, the sequence of functions $A_N f([\![x_1,\cdot]\!])$ converges to $\mathrm{Leb}_2(f)$ almost everywhere in $\mathbb{T}^1$, we deduce that $A_N f([\![x_1,\cdot]\!])$ converges to $\mathrm{Leb}_2(f)$ uniformly on $\mathbb{T}^1$ for almost all $[\![x_1]\!] \in \mathbb{T}^1$. In other words, we can write $B = B_1 \times \mathbb{T}^1$ for a set $B_1 \subseteq \mathbb{T}^1$ with $\mathrm{Leb}_1(B_1) = 1$.

Let $\pi\colon \mathbb{T}^2 \to \mathbb{T}^1$ denote the projection on the first coordinate. Then, the push-forward measure $\pi_*\nu$ is invariant under $x \mapsto x+\alpha$. Since $\alpha \notin \mathbb{Q}$, by Theorem 1.21 we get $\pi_*\nu = \mathrm{Leb}_1$. This implies that

$$\nu(B) = \nu(B_1 \times \mathbb{T}^1) = \pi_*\nu(B_1) = \mathrm{Leb}_1(B_1) = 1.$$

Since, by the Ergodic Theorem, the set E of points $[\![x_1,x_2]\!] \in \mathbb{T}^2$ for which $A_N f([\![x_1,x_2]\!]) \to \nu(f)$ has ν-measure 1, it follows that $\nu(E\cap B) \neq 0$, in particular there exists a point $[\![x_1,x_2]\!] \in E\cap B$. For such a point, we have $\mathrm{Leb}_2(f) = \lim_{N\to\infty} A_N f([\![x_1,x_2]\!]) = \nu(f)$, which implies $\nu(f) = \mathrm{Leb}_2(f)$ and completes the proof of the lemma. □

Corollary 3.16. *If v_1 and v_2 are rationally independent, then $\varphi^{\mathbf{v}}$ is uniquely ergodic.*

Proof. As before, let $M = \{p = \Gamma_1(x,y,z) : x,y,z \in [0,1)\}$, $\Sigma = \{q = \Gamma_1(x,0,z) : x,z \in [0,1)\}$, and let $\varphi^{\mathbf{v}}$ be a Heisenberg niflow, with $\mathbf{v} = (v_1,v_2,v_3) \in \mathfrak{h}$ and v_1,v_2 rationally independent. It is clear that unique ergodicity is preserved by all positive constant rescalings; that is to say, $\{\varphi^{\mathbf{v}}_t\}_{t\in\mathbb{R}}$ is uniquely ergodic if and only if $\{\varphi^{\mathbf{v}}_{at}\}_{t\in\mathbb{R}}$ is uniquely ergodic for all $a > 0$. Therefore, without loss of generality, we can assume that $v_2 = 1$.

Let $f \in \mathscr{C}(M)$. By Proposition 1.20, we need to show that, for all $p \in M$, $\frac{1}{T}\int_0^T f\circ\varphi^{\mathbf{v}}_t(p)\,\mathrm{d}t$ converges to a constant as $T \to \infty$. Let us define $g\colon \Sigma \to \mathbb{C}$ by

$$g(q) = \int_0^1 f\circ\varphi^{\mathbf{v}}_t(q)\,\mathrm{d}t.$$

Clearly, g is continuous on Σ. By unique ergodicity of the skew-translation $T = \varphi^{\mathbf{v}}_1$ (see Lemma 3.15), which by definition is the time-one map of the flow $\varphi^{\mathbf{v}}$, we have

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} g\circ T^n(q) = \int_\Sigma g\,\mathrm{d}\,\mathrm{Leb}. \tag{3.7}$$

We can rewrite the average in left-hand side of (3.7) as

$$\frac{1}{N}\sum_{n=0}^{N-1} g\circ T^n(q) = \frac{1}{N}\sum_{n=0}^{N-1}\int_0^1 f\circ\varphi^{\mathbf{v}}_t(\varphi^{\mathbf{v}}_n(q))\,\mathrm{d}t = \frac{1}{N}\int_0^N f\circ\varphi^{\mathbf{v}}_t(q)\,\mathrm{d}t,$$

so that, for all $q \in \Sigma$, we deduce

$$\frac{1}{N}\int_0^N f\circ\varphi^{\mathbf{v}}_t(q)\,\mathrm{d}t \to \int_\Sigma g\,\mathrm{d}\,\mathrm{Leb} = \int_\Sigma\int_0^1 f\circ\varphi^{\mathbf{v}}_t\,\mathrm{d}\,\mathrm{Leb} = \int_M f\,\mathrm{d}\mu, \qquad \text{as } N\to\infty,\ N\in\mathbb{N}. \tag{3.8}$$

Let now $p = \Gamma_1(x,y,z) \in M$, with $x,y,z \in [0,1)$, be arbitrary. Notice that

$$q = \varphi^{\mathbf{v}}_{-y}(p) = \Gamma_1(x-y,0,z-xy/2) \in \Sigma,$$

thus, from (3.8), after a change of variable, we get

$$\lim_{N\to\infty,\ N\in\mathbb{N}} \frac{1}{N}\int_{-y}^{N-y} f\circ\varphi^{\mathbf{v}}_t(p)\,\mathrm{d}t = \mu(f).$$

If $T > 0$, define $N = \lfloor T \rfloor + 1$. Then,

$$\begin{aligned}
&\left| \frac{1}{T} \int_0^T f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t - \frac{1}{N} \int_{-y}^{N-y} f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t \right| \\
&\quad \leq \frac{1}{T} \left| \int_0^T f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t - \int_{-y}^{N-y} f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t \right| + \left| \frac{1}{T} - \frac{1}{N} \right| \cdot \left| \int_{-y}^{N-y} f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t \right| \\
&\quad \leq \frac{2\,\|f\|_\infty}{T} + \frac{N+1}{NT}\,\|f\|_\infty,
\end{aligned}$$

which tends to zero (uniformly in p) as $T \to \infty$. We conclude that

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t = \lim_{N\to\infty,\ N\in\mathbb{N}} \frac{1}{N} \int_{-y}^{N-y} f \circ \varphi_t^{\mathbf{v}}(p)\,\mathrm{d}t = \mu(f),$$

which proves unique ergodicity. □

Exercise 3.17. *Generalize the proof Lemma* 3.15 *to compact Abelian extensions of uniquely ergodic systems; that is, prove the following version of a theorem of Furstenberg.*

Let $T\colon X \to X$ be a uniquely ergodic homeomorphism of a compact metric space, and let μ be the unique invariant measure. Let $n \geq 1$ and let $f\colon X \to \mathbb{T}^n$ be continuous. Define the skew-product $S\colon X \times \mathbb{T}^n \to X \times \mathbb{T}^n$ by $S(x,p) = (Tx, p + f(x))$. If S is ergodic with respect to $\mu \times \mathrm{Leb}$, then it is uniquely ergodic.

3.3 A connection to number theory

In this chapter, we studied the ergodic properties of Heisenberg nilflows $\varphi^{\mathbf{v}}$, with $\mathbf{v} = (v_1, v_2, v_3) \in \mathfrak{h} \setminus \{\mathbf{0}\}$ on the nilmanifold $M = \Gamma_1 \backslash \mathrm{Heis} = \Gamma_1 \backslash \mathfrak{h}$, where $\Gamma_1 = \mathbb{Z} * \mathbb{Z} * \mathbb{Z}$. In Theorem 3.11, we showed that if v_1 and v_2 are rationally independent, then the nilflow is uniquely ergodic.

Let us take $v_2 = 1$. In the proof of Theorem 3.11, we studied the time-one map $T = \varphi_1^{\mathbf{v}}$ restricted to the submanifold Σ. The latter is isomorphic to a skew-translation $T[\![x,y]\!] = [\![x+\alpha, y+x+\beta]\!]$ on the torus $\mathbb{T}^2$ equipped with the Lebesgue measure, where $\beta = v_1/2 + v_3$. We showed that for all $\alpha, \beta \in \mathbb{R}$, if $\alpha \notin \mathbb{Q}$, then $(\mathbb{T}^2, \mathrm{Leb}, T)$ is a uniquely ergodic transformation. As a consequence, we get the following result.

Proposition 3.18. *Let $P(X) = aX^2 + bX + c \in \mathbb{R}[X]$ be a quadratic polynomial with real coefficients. If $a \notin \mathbb{Q}$, then for all positive integer $k \geq 1$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e^{2\pi i k P(n)} = 0. \tag{3.9}$$

In particular, the sequence $(P(n))_{n\in\mathbb{N}}$ is equidistributed mod 1.

Proof. The equidistribution of the fractional parts of $P(n)$ follows from (3.9) by the Weyl's Criterium. Thus, we only need to prove (3.9).

For any fixed $\alpha, \beta \in \mathbb{R}$, it is easy to see, for example by induction, that the associated skew-product $T[\![x,y]\!] = [\![x+\alpha, x+y+\beta]\!]$ on $\mathbb{T}^2$ satisfies

$$T^n[\![x,y]\!] = [\![x+n\alpha, y+nx+n\beta+n(n-1)/2\alpha]\!] \qquad \text{for all } n \in \mathbb{Z}.$$

We consider the continuous function $f_k([\![x,y]\!]) = e^{2\pi i k y}$, which has zero integral $\mathrm{Leb}(f_k) = 0$. Let $P(X)$ be a polynomial as in the statement of the proposition. Let $\alpha = 2a \notin \mathbb{Q}$ and $\beta = b + a$, and consider the point $p = [\![0,c]\!] \in \mathbb{T}^2$. By Lemma 3.15, we deduce

$$0 = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_k \circ T^n(p) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f_k([\![n\alpha, c+n(b+a)+n(n-1)a]\!]) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e^{2\pi i k P(n)},$$

which proves the result. □

Sums of the form

$$\sum_{n=0}^{N-1} e^{2\pi i P(n)},$$

where P is a polynomial with real coefficients, are called *Weyl sums* and they are classical objects in number theory. In the case of a quadratic polynomial, they are called *quadratic Weyl sums* or *theta sums*, and they have been studied since the work of Hardy and Littlewood [9, 10]. Recently, the connection of Weyl sums to the ergodic theory of nilflows has become the subject of a lot of research. For example, we highlight the work of Flaminio and Forni [7], who recovered the optimal bounds

$$\left|\sum_{n=0}^{N-1} e^{2\pi i P(n)}\right| = o(N^{1/2+\varepsilon}) \qquad \text{for all } \varepsilon > 0,$$

(originally proved by Fiedler, Jurkat and Körner [5] using analytic number theory) by studing the quantitative ergodic properties of Heisenberg nilflows. Although this goes beyond the scope of this course, it is interesting to notice how the number theoretic properties of polynomial sequences and the ergodic theory of nilflows are intimately related.

Chapter 4

Elements of hyperbolic geometry

We have seen in Proposition 2.27 that the Lie group $\mathrm{PSL}(2,\mathbb{R})$ acts on the hyperbolic plane by isometries; in this chapter, we will study this action in detail. We will work, however, in another model of the hyperbolic plane, namely the upper-half space, which we now introduce.

4.1 The hyperbolic plane

4.1.1 The upper-half space model

Let

$$\mathbb{H} := \{z = x + iy \in \mathbb{C} : y = \Im(z) > 0\}.$$

To fix notation, we will always denote elements of $\mathbb{H}$ by $z = x + iy$, where $x = \Re(z)$ is the real part of z and $y = \Im(z) > 0$ is its imaginary part.

For any $z \in \mathbb{H}$, we identify its tangent space $T_z\mathbb{H}$ with $\mathbb{R}^2 \simeq \mathbb{C}$. The hyperbolic metric on $\mathbb{H}$ is defined by

$$\langle v, w\rangle_z := \frac{1}{y^2} v \cdot w, \qquad \text{for } v, w \in T_z\mathbb{H},\ z \in \mathbb{H}.$$

We remark that, for any $z \in \mathbb{H}$, the scalar product $\langle \cdot, \cdot \rangle_z$ is a rescaled version of the Euclidean scalar product on $\mathbb{R}^2$; in particular, the topology induced by the hyperbolic metric on $\mathbb{H}$ is the same as the Euclidean topology. As we will see shortly, the *geometry*, however, is very different.

As usual, the hyperbolic metric allows to define the lengths of curves $\gamma\colon [a,b] \to \mathbb{H}$ by

$$\ell(\gamma) = \int_a^b \sqrt{\langle \dot\gamma(t), \dot\gamma(t)\rangle_{\gamma(t)}}\, \mathrm{d}t,$$

and, in turn, we define the *hyperbolic distance* by

$$d_{\mathbb{H}}(z_0, z_1) = \inf\{\ell(\gamma) :\ \gamma \text{ is a smooth curve from } z_0 \text{ to } z_1\}. \tag{4.1}$$

We will see shortly that the hyperbolic distance between two points z_0 and z_1 is always realized by a smooth curve, which is called the *hyperbolic geodesic* between z_0 and z_1.

4.1.2 The action of $\mathrm{PSL}(2,\mathbb{R})$ on $\mathbb{H}$

Let $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{R})$, and let

$$g.z := \frac{az+b}{cz+d}, \qquad \text{for } z \in \mathbb{H} \tag{4.2}$$

be the associated *Möbius transformation*.

Lemma 4.1. *Equation* (4.2) *gives a well-defined action of* $\mathrm{PSL}(2,\mathbb{R})$ *on* $\mathbb{H}$.

Proof. We first check that the definition is well-posed: the denominator $cz+d$ in (4.2) is zero if and only if $z=-d/c\in\mathbb{R}$ or $d=c=0$; both possibilities cannot occur by definition.

Let us verify that all Möbius transformations map $\mathbb{H}$ to $\mathbb{H}$. Indeed, for any $g=\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)\in\mathrm{SL}(2,\mathbb{R})$ and $z\in\mathbb{H}$, we compute the imaginary part of $g.z$ by

$$\begin{aligned}\Im(g.z)&=\frac{1}{2}(g.z-\overline{g.z})=\frac{1}{2}\left(\frac{az+b}{cz+d}-\frac{a\bar{z}+b}{c\bar{z}+d}\right)=\frac{(az+b)((c\bar{z}+d))-(a\bar{z}+b)(cz+d)}{2(cz+d)(c\bar{z}+d)}\\&=\frac{ad-bc}{|cz+d|^2}\,\frac{z-\bar{z}}{2}=\frac{\Im(z)}{|cz+d|^2}.\end{aligned}\tag{4.3}$$

Thus, since $\Im(z)>0$, we get $\Im(g.z)>0$, which shows that $g.z\in\mathbb{H}$.

Finally, we verify it is an action. Let $g=\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)$, $\widetilde{g}=\left(\begin{smallmatrix} \widetilde{a} & \widetilde{b} \\ \widetilde{c} & \widetilde{d}\end{smallmatrix}\right)\in\mathrm{SL}(2,\mathbb{R})$ and $z\in\mathbb{H}$. Then, we have

$$\begin{aligned}\widetilde{g}.(g.z)&=\widetilde{g}.\left(\frac{az+b}{cz+d}\right)=\frac{\widetilde{a}(\frac{az+b}{cz+d})+\widetilde{b}}{\widetilde{c}(\frac{az+b}{cz+d})+\widetilde{d}}=\frac{(a\widetilde{a}+\widetilde{b}c)z+(\widetilde{a}b+\widetilde{b}d)}{(\widetilde{c}a+\widetilde{d}c)z+(\widetilde{c}b+\widetilde{d}d)}=\begin{pmatrix} a\widetilde{a}+\widetilde{b}c & \widetilde{a}b+\widetilde{b}d\\ \widetilde{c}a+\widetilde{d}c & \widetilde{c}b+\widetilde{d}d\end{pmatrix}.z\\&=(\widetilde{g}g).z,\end{aligned}$$

which proves the claim. The matrix $-e=\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1\end{smallmatrix}\right)$, as Möbius transformation, is the identity, hence (4.2) descends to an action of the Lie group $\mathrm{PSL}(2,\mathbb{R})$ on $\mathbb{H}$. □

Möbius transformations on $\mathbb{H}$ are smooth maps and their derivative can be expresses as

$$g'(z)=\frac{a(cz+d)-(az+d)c}{(cz+d)^2}=\frac{1}{(cz+d)^2}.\tag{4.4}$$

In particular, we can define an action D of $\mathrm{PSL}(2,\mathbb{R})$ on the tangent bundle $T\mathbb{H}=\mathbb{H}\times\mathbb{C}$ by

$$Dg.(z,w)=(g.z,g'(z)w)=\left(\frac{az+b}{cz+d},\frac{w}{(cz+d)^2}\right).$$

Proposition 4.2.
1. *For every* $g\in\mathrm{PSL}(2,\mathbb{R})$*, the map* Dg *preserves the hyperbolic metric. In particular,* g *acts on* $\mathbb{H}$ *by isometries.*
2. *The stabilizer of* $i\in\mathbb{H}$ *is the compact group* $\mathrm{PSO}(2,\mathbb{R})=\mathrm{SO}(2,\mathbb{R})/\{\pm e\}$.
3. *The action on* $\mathbb{H}$ *is transitive.*

Proof. Let us fix $z\in\mathbb{H}$ and $v,w\in T_z\mathbb{H}$. Notice that under the identification $\mathbb{R}^2=\mathbb{C}$ we can write the scalar product $v\cdot w$ as $\Re(v\overline{w})$. Using the definition of the hyperbolic metric, (4.3), and (4.4), we compute

$$\langle g'(z)v,g'(z)w\rangle_{g.z}=\frac{1}{(\Im(g.z))^2}\Re(g'(z)v\overline{g'(z)w})=\frac{|cz+d|^4}{(\Im z)^2}|g'(z)|^2\Re(v\overline{w})=\frac{1}{(\Im z)^2}v\cdot w=\langle v,w\rangle_z.$$

This shows that Dg preserves the hyperbolic metric, hence g is an isometry.

Let us compute the stabilizer of the point $i\in\mathbb{H}$. By definition, $g.i=i$ if and only if $i=\frac{ai+b}{ci+d}$, which can be rewritten as $(b+c)+i(a-d)=0$. Thus, $g=\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right)\in\mathrm{Stab}(i)$ if and only if $b=-c$ and $a=d$. Since $ad-bc=1$, we get $a^2+b^2=1$. In particular, we can choose $\theta\in[0,2\pi)$ such that $a=\cos\theta$ and $b=\sin\theta$. From this we conclude that

$$g=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\in\mathrm{SO}(2,\mathbb{R}),$$

which proves the claim.

Finally, let us verify that the action is transitive. Let $z = x + iy \in \mathbb{H}$; it is enough to show that there exists $g \in \mathrm{SL}(2,\mathbb{R})$ such that $g.i = z$. By choosing the element $g = \left(\begin{smallmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{smallmatrix}\right)$, it is easy to check that we indeed get $g.i = x + iy = z$ as desired. □

Corollary 4.3. *We can identify* $\mathbb{H} = \mathrm{PSL}(2,\mathbb{R})/\mathrm{PSO}(2,\mathbb{R})$.

Proof. The identification is given by associating to the point $z \in \mathbb{H}$ the element $g\,\mathrm{PSO}(2,\mathbb{R}) \in \mathrm{PSL}(2,\mathbb{R})/\mathrm{PSO}(2,\mathbb{R})$, where $g \in \mathrm{PSL}(2,\mathbb{R})$ is such that $g.i = z$. □

Let us denote by $T^1\mathbb{H}$ the *unit tangent bundle of* $\mathbb{H}$, that is the subset of the tangent bundle consisting of vectors of norm 1,

$$T^1\mathbb{H} = \{(z,v) \in T\mathbb{H} : \|v\| = 1\}.$$

By Proposition 4.2, the action of $\mathrm{PSL}(2,\mathbb{R})$ on $T\mathbb{H}$ restricts to $T^1\mathbb{H}$. We now show that the this latter action is simply transitive.

Proposition 4.4. *The action of* $\mathrm{PSL}(2,\mathbb{R})$ *on* $T^1\mathbb{H}$ *is simply transitive.*

Proof. Let us fix $(z,v) \in T^1\mathbb{H}$. Since the tangent vector $v \in T_z\mathbb{H} = \mathbb{C}$ has unit (hyperbolic) norm, we can identify it with the angle 2θ that it forms with the half line $i\mathbb{R}_{>0}$; In other words, we can write $v = yi(\cos(2\theta) + i\sin(2\theta))$, where that the factor $y = \Im z$ is to ensure that v has norm 1.

Let

$$g = \begin{pmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \in \mathrm{PSL}(2,\mathbb{R}).$$

Using Proposition 4.2, we get

$$Dg.(i,i) = D\begin{pmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{pmatrix} . \left(i, \frac{i}{(\cos\theta - i\sin\theta)^2}\right) = \left(z, yi\frac{(\cos\theta + i\sin\theta)^2}{(\cos^2\theta - i^2\sin\theta)^2}\right) = (z,v),$$

which shows that the action is transitive.

Let us show that the stabilizer of (i,i) is $\{\pm e\}$. By Proposition 4.2, if $Dg.(i,i) = (i,i)$, then $g \in \mathrm{SO}(2,\mathbb{R})$. Let us write $g = \left(\begin{smallmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{smallmatrix}\right)$. Similarly as above, $g'(i)i = i$ implies

$$i = \frac{i}{(\cos\theta - i\sin\theta)^2} = i(\cos\theta + i\sin\theta)^2.$$

Therefore, we deduce $2\theta = 0 \bmod 2\pi$, which means $g = \pm e$. This completes the proof. □

By Proposition 4.4, for any $(z,v) \in T^1\mathbb{H}$ there exists a unique $g \in \mathrm{PSL}(2,\mathbb{R})$ such that $Dg.(i,i) = (z,v)$. In this way, we have an identification

$$T^1\mathbb{H} \simeq \mathrm{PSL}(2,\mathbb{R}). \tag{4.5}$$

4.1.3 Hyperbolic geodesics

Recall from (4.1) that the hyperbolic distance between two points $z_0, z_1 \in \mathbb{H}$ is defined as the infimum of the hyperbolic lengths of all smooth curves connecting z_0 to z_1. We will now see that the infimum is always realized by a unique (up to reparametrization) curve, which will be said to be a *geodesic*.

Let us start by looking at a simple case.

Lemma 4.5. *The distance between $z_0 = iy_0$ and $z_1 = iy_1$ is $d_{\mathbb{H}}(z_0, z_1) = |\log(y_1/y_0)|$ and is realized by the curve*

$$\gamma(t) = y_0 \left(\frac{y_1}{y_0}\right)^t i, \qquad \text{for } t \in [0,1].$$

Any other smooth curve $\widetilde{\gamma}\colon [a,b] \to \mathbb{H}$ with the same length is a smooth reparametrization of γ; in other words, there exists an increasing, piecewise differentiable map $f\colon [a,b] \to [0,1]$ such that $\widetilde{\gamma} = \gamma \circ f$.

Proof. First of all, we notice that $\gamma(t) = i\Im\gamma(t)$ and $\dot{\gamma}(t) = \gamma(t)\log(y_1/y_0)$. On one hand,

$$\begin{aligned} d_{\mathbb{H}}(z_0, z_1) \leq \ell(\gamma) &= \int_0^1 \|\dot{\gamma}(t)\|_{T_{\gamma(t)}\mathbb{H}}\,\mathrm{d}t = \int_0^1 \frac{|\gamma(t)|}{\Im\gamma(t)} \left|\log\left(\frac{y_1}{y_0}\right)\right| \mathrm{d}t = \int_0^1 \left|\log\left(\frac{y_1}{y_0}\right)\right| \mathrm{d}t \\ &= \left|\log\left(\frac{y_1}{y_0}\right)\right|. \end{aligned}$$

On the other hand, let now $\eta\colon [a,b] \to \mathbb{H}$ be any other smooth curve connecting z_0 to z_1. Up to a reparametrization, we can assume that $[a,b] = [0,1]$ and, up to reversing the orientation, that $\Im\eta(0) \leq \Im\eta(1)$. Let us write $\eta_x(t) = \Re\eta(t)$ and $\eta_y(y) = \Im\eta(t)$. Then,

$$\begin{aligned} \ell(\eta) &= \int_0^1 \|\dot{\eta}(t)\|_{T_{\eta(t)}\mathbb{H}}\,\mathrm{d}t = \int_0^1 \frac{\sqrt{\dot{\eta}_x^2(t) + \dot{\eta}_y^2(t)}}{\Im\eta(t)}\,\mathrm{d}t \geq \int_0^1 \frac{|\dot{\eta}_y(t)|}{\eta_y(t)}\,\mathrm{d}t \geq \int_0^1 \frac{\dot{\eta}_y(t)}{\eta_y(t)}\,\mathrm{d}t \\ &= \log\left(\frac{\eta_y(1)}{\eta_y(0)}\right) = \left|\log\left(\frac{y_1}{y_0}\right)\right|. \end{aligned} \tag{4.6}$$

This shows that $d_{\mathbb{H}}(z_0, z_1) = \ell(\gamma) = |\log(y_1/y_0)|$. Finally, notice that equality holds in (4.6) if and only if $\dot{\eta}_x(t) = 0$ and $\dot{\eta}_y(t) \geq 0$ for all $t \in [0,1]$, which proves the second claim. □

The infinite path $\gamma(t) = y_0 e^t i$ starting at $y_0 i \in \mathbb{H}$ is a *geodesic ray*, that is, a curve realising the minimum distance between its points and parametrized with unit speed.

Proposition 4.6. *Any two distinct points $z_0, z_1 \in \mathbb{H}$ are connected by a geodesic γ. This curve is unique: there exists a unique $g \in \mathrm{PSL}(2,\mathbb{R})$ such that $\gamma(t) = g.(e^t i)$.*

Proof. By Proposition 4.2, we can choose $g_0 \in \mathrm{PSL}(2,\mathbb{R})$ such that $g_0^{-1}.z_0 = i$; let us denote $z = g_0^{-1}.z_1$. For any $\widetilde{g} = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right) \in \mathrm{PSO}(2,\mathbb{R})$, we have $\widetilde{g}g_0^{-1}.z_0 = i$; we claim that we can choose $\widetilde{g}$ so that $\widetilde{g}.z$ is purely imaginary. In order to do this, we have to solve

$$0 = \Re(\widetilde{g}.z) = \frac{1}{2}(\widetilde{g}.z + \overline{\widetilde{g}.z}) = \frac{1}{2}\left(\frac{\cos\theta z - \sin\theta}{\sin\theta z + \cos\theta} + \frac{\cos\theta\bar{z} - \sin\theta}{\sin\theta\bar{z} + \cos\theta}\right).$$

After some calculations, we get that $\Re(\widetilde{g}.z) = 0$ if and only if

$$(|z|^2 - 1)\sin(2\theta) + 2\Re(z)\cos(2\theta) = 0.$$

If $|z| = 1$, we can choose $\theta = \pi/4$, otherwise we can choose θ such that $\tan(2\theta) = -\frac{2\Re z}{|z|^2 - 1}$. We have then showed that there exists $g = g_0\widetilde{g}^{-1} \in \mathrm{PSL}(2,\mathbb{R})$ such that $g^{-1}.z_0 = i$ and $g^{-1}.z_1 = iy$ for some $y > 0$. Moreover, we can assume that $y > 1$: since z_0 and z_1 are distinct, we have $y \neq 1$, and if $y < 1$, then $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right).(iy) = iy^{-1}$.

By Lemma 4.5, there exists a unique geodesic curve between i and iy with unit speed given by $e^t i$ for $t \in [0, \log y]$, where $\log y = d_{\mathbb{H}}(i, iy)$. Hence, since g is an isometry, $\gamma(t) = g.(e^t i)$ is a geodesic curve between z_0 and z_1.

Finally, one can check that the element g as above is unique: if h is another element such that $h.(e^t i)$ is a geodesic between z_0 and z_1, then $h^{-1}g.(e^t i) = e^t i$. Letting $h^{-1}g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, we get $e^t i = \frac{ae^t i+b}{ce^t i+d}$ and, taking the derivative at $t = 0$, we obtain

$$i = -i\frac{ad-bc}{(c-id)^2} = -i\frac{1}{(c-id)^2}.$$

This gives us $c = 0$ and $d = \pm 1$, from which we get $a = \pm 1$ and $b = 0$. Hence $h = g$ in $\mathrm{PSL}(2,\mathbb{R})$. This completes the proof. □

By Proposition 4.6, geodesics can be described as the images of the vertical line $e^t i$ under Möbius transformations. It is a standard fact from elementary geometry that Möbius transformations map vertical lines into either vertical lines or into semicircles with centre on the real axis. These are all the hyperbolic geodesics.

4.2 Geodesic and horocycle flows

We are going to define the two flows which will be the subject of the rest of this course: the geodesic and horocycle flow.

4.2.1 Algebraic and geometric definitions

We have seen that for any $p = (z,v) \in T^1\mathbb{H}$ there exists a unique geodesic ray starting at p; that is, a unique smooth curve $\gamma\colon \mathbb{R} \to \mathbb{H}$ such that $\gamma(0) = z$, $\dot\gamma(0) = v$ and which realizes the smallest distance between any two of its points. The image $\gamma(\mathbb{R})$ is either a vertical line or a semicircle with centre on the real axis. The *geodesic flow* $g_t : T^1\mathbb{H} \to T^1\mathbb{H}$ consists in following this unique geodesic for time t. We are now going to show it can be seen as a homogeneous flow as discussed in Chapter 2.

For $(i,i) \in T^1\mathbb{H}$, we know that the geodesic is given by $\gamma(t) = e^t i$. Once again, notice that the tangent vector has norm 1, since

$$\|\dot\gamma(t)\|_{T_{\gamma(t)}\mathbb{H}} = \frac{1}{\Im\gamma(t)}|\dot\gamma(t)| = 1.$$

Therefore, $g_t(i,i) = (e^t i, e^t i)$. It is immediate to check that

$$(e^t i, e^t i) = Da_t.(i,i), \qquad \text{where } a_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R}).$$

Let now $(z,v) \in T^1\mathbb{H}$ be arbitrary. We have shown that there exists a unique $g \in \mathrm{PSL}(2,\mathbb{R})$ such that $(z,v) = Dg.(i,i)$ and the geodesic starting at that point is the image under g of the vertical geodesic at (i,i). In particular

$$g_t(g.(i,i)) = g_t(z,v) = Dg.(e^t i, e^t i) = Dg.(Da_t.(i,i)) = D(ga_t).(i,i).$$

Thus, under the identification (4.5), the geodesic flow on $T^1\mathbb{H}$ translates into the flow on $\mathrm{PSL}(2,\mathbb{R})$ that maps the element g into ga_t. Using the notation of Chapter 2, the *geodesic flow* is the homogeneous flow

$$\varphi_t^{\mathbf{a}} : g \mapsto g\begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} = g\exp(t\mathbf{a}), \qquad \text{where } \mathbf{a} = \begin{pmatrix} 1/2 & 0 \\ 0 & -1/2 \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{R}).$$

The other homogeneous flows on $\mathrm{SL}(2,\mathbb{R})$ we encountered in Chapter 2 were the ones generated by $\mathbf{u} = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ and $\mathbf{v} = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, namely the flows $\varphi_t^{\mathbf{u}}(g) = g\left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$ and $\varphi_t^{\mathbf{v}}(g) = g\left(\begin{smallmatrix} 1 & 0 \\ t & 1 \end{smallmatrix}\right)$. We now see what are the geometric interpretations of these two flows.

Proposition 4.7. *1. For all $g \in \mathrm{SL}(2,\mathbb{R})$ and all $t,s \in \mathbb{R}$ we have*

$$\varphi_s^{\mathbf{a}} \circ \varphi_t^{\mathbf{u}}(g) = \varphi_{e^{-s}t}^{\mathbf{u}} \circ \varphi_s^{\mathbf{a}}(g). \tag{4.7}$$

2. The flow $\varphi_t^{\mathbf{u}}$ consists of unit-speed translations along the stable manifolds *of the geodesic flow: points in the same orbit of $\varphi_t^{\mathbf{u}}$ get exponentially closed under the action of the geodesic flow.*

Proof. A straightforward computation gives us

$$\begin{aligned}\varphi_s^{\mathbf{a}} \circ \varphi_t^{\mathbf{u}}(g) &= g\exp(t\mathbf{u})\exp(s\mathbf{a}) = g\begin{pmatrix}1 & t\\ 0 & 1\end{pmatrix}\begin{pmatrix}e^{s/2} & 0\\ 0 & e^{-s/2}\end{pmatrix} = g\begin{pmatrix}e^{s/2} & te^{-s/2}\\ 0 & e^{-s/2}\end{pmatrix}\\ &= g\begin{pmatrix}e^{s/2} & 0\\ 0 & e^{-s/2}\end{pmatrix}\begin{pmatrix}1 & te^{-s}\\ 0 & 1\end{pmatrix} = g\exp(s\mathbf{a})\exp(e^{-s}t\mathbf{u}) = \varphi_{e^{-s}t}^{\mathbf{u}} \circ \varphi_s^{\mathbf{a}}(g),\end{aligned}$$

which proves (4.7).

Let us now verify the second claim. We fix a left-invariant metric on $\mathrm{PSL}(2,\mathbb{R})$ as in §2.4.1 given by the usual basis $\{\mathbf{a},\mathbf{u},\mathbf{v}\}$. Then, the homogeneous flow $\varphi_t^{\mathbf{u}}$ has unit speed. Let us show that points on the same orbit get exponentially close under the geodesic flow: fix $g \in \mathrm{PSL}(2,\mathbb{R})$ and let $g_1 = \varphi_\ell^{\mathbf{u}}(g) = g\exp(\ell\mathbf{u})$ another point in the orbit of $\varphi^{\mathbf{u}}$. Using (4.7), we get

$$\begin{aligned}d(\varphi_t^{\mathbf{a}}(g), \varphi_t^{\mathbf{a}}(g_1)) &= d(\varphi_t^{\mathbf{a}}(g), \varphi_{e^{-t}\ell}^{\mathbf{u}} \circ \varphi_t^{\mathbf{a}}(g)) = d(g\exp(t\mathbf{a}), g\exp(t\mathbf{a})\exp(e^{-t}\ell\mathbf{u}))\\ &= d(e, \exp(e^{-t}\ell\mathbf{u})) = e^{-t}\ell,\end{aligned}$$

where we used the left-invariance of the metric on $\mathrm{PSL}(2,\mathbb{R})$. Since the latter term goes to zero as $t \to \infty$ exponentially, the proof is complete. □

The flow $\varphi_t^{\mathbf{u}}$ is called the *(stable) horocycle flow*. On the hyperbolic plane $\mathbb{H}$, the orbits of the horocycle flow (called horocycles) are either horizontal lines or circles tangent to the real axis (the former case can be seen as a circle tangent to infinity). The point of tangency is the limit point of the geodesic starting at any point of the horcycle orbit.

A similar characterization holds for $\mathbf{v}$, we leave it as an exercise to the reader.

Exercise 4.8. *Show that $\varphi_s^{\mathbf{a}} \circ \varphi_t^{\mathbf{v}}(g) = \varphi_{e^s t}^{\mathbf{v}} \circ \varphi_s^{\mathbf{a}}(g)$ for all $g \in \mathrm{SL}(2,\mathbb{R})$ and all $t,s \in \mathbb{R}$. Deduce that $\varphi_t^{\mathbf{v}}$ parametrizes the unstable manifolds of the geodesic flow. It is called the* (unstable) horocycle flow.

Geodesic and horocycle flows on $T^1\mathbb{H}$ are rather boring, as every orbits escape to infinity (we are in the same situation as a linear flow on $\mathbb{R}^n$). In order to have some recurrence, we need to look at the projections of these flows on finite volume quotients of $\mathrm{PSL}(2,\mathbb{R})$, which we are going to focus on in the next section.

4.2.2 Finite volume homogeneous manifolds

Let us call $G = \mathrm{PSL}(2,\mathbb{R})$, and let $\Gamma \leq G$ be a discrete subgroup. We have seen in Lemma 2.36 that the quotient space $M = \Gamma\backslash G$ is a smooth manifold. Moreover, for every $x = \Gamma g \in M$, there exists $r > 0$ such that the restriction of the canonical projection $\pi\colon G \to M$ to the ball centered at g with radius r is a homeomorphism onto its image; an atlas of charts on M can be obtained by considering these local inverses of π. The push-forward under π of any fixed left-invariant metric on G gives us a distance on M by

$$d_M(x_1,x_2) = \inf_{\gamma\in\Gamma} d_G(g_1,\gamma g_2), \qquad \text{where } x_i = \Gamma g_i,\ i = 1,2,$$

and the infimum above is actually a minimum by Lemma 2.35.

In a similar way, the push-forward of any Haar measure μ_G on G induces a well-defined measure on M, which we will denote by μ. Recall that a discrete subgroup Γ is said to be a lattice if M has finite measure, and is said to be co-compact if M is compact.

Before looking at examples of lattices, we are now going to show that the Haar measure μ_G is invariant under *right-multiplication*, and therefore is an invariant measure for the geodesic and horocycle flows. Given a discrete subgroup Γ of G, a subset $F \subset G$ is called a *fundamental domain* for Γ if

$$G = \bigcup_{\gamma\in\Gamma} \gamma F, \qquad \text{and} \qquad \mu(F \cap \gamma F) = 0 \ \text{ for all } \gamma \in \Gamma \setminus \{e\}.$$

In other words, a fundamental domain is a set which, up to a zero measure subset, contains one representative per orbit of Γ.

Exercise 4.9. *Show that if $F \subset G$ is a fundamental domain for Γ, then so is Fg for all $g \in G$.*

Proposition 4.10. *Let Γ be a discrete subgroup of G, and assume that there exists a fundamental domain F with finite Haar measure. Then, Γ is a lattice, all fundamental domains have the same measure and for all measurable subsets $B \subset M$ we have*

$$\mu(B) = \mu_G(\pi^{-1}(B) \cap F).$$

Moreover, μ is invariant under multiplication on the right and hence is an invariant measure for all homogeneous flows on M.

Proof. Let $A, A' \subset G$ be two measurable sets such that the restrictions of the canonical projection $\pi\colon G \to M$ to A and A' are both injective up to zero measure sets and $\pi(A) = \pi(A')$. We now show that $\mu_G(A) = \mu_G(A')$; in particular this will imply that any two fundamental domains F, F' for Γ have the same Haar measure.

Notice that, for almost all $g \in A$, there exists a unique $\gamma \in \Gamma$ such that $\gamma^{-1}g \in A'$. Thus, up to a zero measure set, we can write

$$A = \bigsqcup_{\gamma\in\Gamma} A \cap \gamma A',$$

and, similarly,

$$A' = \bigsqcup_{\gamma\in\Gamma} A' \cap \gamma A.$$

Using the left-invariance property of the Haar measure, we get

$$\mu_G(A) = \sum_{\gamma\in\Gamma} \mu_G(A \cap \gamma A') = \sum_{\gamma\in\Gamma} \mu_G(\gamma^{-1}A \cap A') = \mu_G(A').$$

In particular, all fundamental domains have the same measure. By definition, the push-forward measure μ on M can be expressed as in the statement of the proposition, and this also proves that $\mu(M) = \mu_G(F)$ is finite, hence Γ is a lattice.

We are left to show that μ is right-invariant. We will prove that the Haar measure μ_G is right-invariant, from which the claim follows. Let us fix $g \in G$, and consider the measure ν_g on G defined by $\nu_g(A) = \mu_G(Ag)$. Clearly, ν_g is invariant by left-multiplication of any element of G, since the Haar measure μ_G is. By uniqueness (see Theorem 2.23), there exists a positive constant $m(g) > 0$ such that $\nu_g = m(g)\mu_G$. Let now F be a fundamental domain for Γ. By Exercise 4.9, the set Fg is also a fundamental domain; therefore, using what we proved so far,

$$\mu_G(F) = \mu_G(Fg) = \nu_g(F) = m(g)\mu_G(F).$$

This implies that $m(g) = 1$ for all $g \in G$, that is, μ_G is right-invariant. The proof is then complete. □

Given a finite volume homogeneous manifold $M = \Gamma\backslash G$, we will always consider the normalization of the Haar measure μ_G that makes μ a probability measure on M; that is, such that $\mu(M) = \mu_G(F) = 1$, where F is any fundamental domain for M.

In order to determine whether a discrete subgroup is a lattice, it is enough to find a fundamental domain $F \subset G = T^1\mathbb{H}$ and compute its Haar measure. Let us try to express the Haar measure on $\mathrm{PSL}(2,\mathbb{R})$ in terms of the coordinates on $\mathbb{T}^1\mathbb{H}$.

We have seen in Proposition 4.2 that the action of G on $\mathbb{H}$ by Möbius transformations is isometric with respect to the hyperbolic metric $y^{-2}(\mathrm{d}x^2 + \mathrm{d}y^2)$. Thus, the action of G preserves the Riemannian volume

$$\mathrm{d}m_{\mathbb{H}} = \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y.$$

Once the base point $z \in \mathbb{H}$ is fixed, the restriction of the action of G on tangent vectors in $T_z^1\mathbb{H}$ is given by $v \mapsto g'(z)v$, which is a *rotation* (a multiplication by a constant of modulus one with respect to the hyperbolic metric). Therefore, if we choose the angle θ that v makes with the vertical as a coordinate on $T_z^1\mathbb{H}$, all maps Dg for $g \in G$ preserve the modulus of the 1-form $\mathrm{d}\theta$. Therefore, we have shown that the maps Dg on $T^1\mathbb{H}$ preserve the measure

$$\mathrm{d}\mu_{\mathbb{H}} := \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}\theta.$$

By the identification (4.5), the action of Dg on $\mathbb{H}$ corresponds to the left-multiplication action on $G = \mathrm{PSL}(2,\mathbb{R})$. By uniqueness of the Haar measure in Theorem 2.23, we conclude that $\mu_G = \mu_{\mathbb{H}}$ (as usual, up to a constant factor).

The expression we just found for the Haar measure on $\mathrm{PSL}(2,\mathbb{R})$ allows us to estimate explicilty the measure of fundamental domains for discrete subgroups. We remark that, since the stabilizer of $i \in \mathbb{H}$ acts on the right (see Corollary 4.3), while the discrete subgroup Γ on the left, we have that

$$\Gamma\backslash\,\mathrm{PSL}(2,\mathbb{R}) = \Gamma\backslash T^1\mathbb{H} = T^1(\Gamma\backslash\mathbb{H}),$$

that is, the quotient M can be identified with the unit tangent bundle of the hyperbolic surface $S := \Gamma\backslash\mathbb{H}$. In order to show that the measure of $\Gamma\backslash\,\mathrm{PSL}(2,\mathbb{R})$ is finite, it is enough to show that the hyperbolic area of S is finite:

$$m_{\mathbb{H}}(\Gamma\backslash\mathbb{H}) < \infty.$$

4.2.3 An important example: the Modular Surface

We can now see an important example of a lattice in $\mathrm{PSL}(2,\mathbb{R})$. Consider

$$\Gamma = \mathrm{PSL}(2,\mathbb{Z}) = \left\{\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}(2,\mathbb{R}) : a,b,c,d \in \mathbb{Z}\right\}.$$

Clearly, Γ is a discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$. It will be useful to consider the elements

$$\tau = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},\ \sigma = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \in \Gamma.$$

Notice that $\tau.z = z+1$ and $\sigma.z = -1/z$.

We now describe a fundamental domain for Γ and show it has finite measure, thus showing that Γ is a lattice.

Proposition 4.11. *Let*

$$E = \{z \in \mathbb{H} : |z| \geq 1,\ |\Re z| \leq 1/2\}.$$

Then, $F = T^1(E) \subset T^1\mathbb{H}$ is a fundamental domain for $\Gamma = \mathrm{PSL}(2,\mathbb{Z})$. Moreover, Γ is a lattice in $\mathrm{PSL}(2,\mathbb{R})$.

Proof. Let us first assume that F is a fundamental domain for Γ, and let us estimate its measure. Clearly, E is a subset of $\widetilde{E} = \{z \in \mathbb{H} : |\Re z| \leq 1/2,\ |\Im z| \geq \sqrt{3}/2\}$, so that its hyperbolic area can be bounded by

$$m_{\mathbb{H}}(E) \leq \int_{\widetilde{E}} \mathrm{d}m_{\mathbb{H}} = \left(\int_{-1/2}^{1/2} \mathrm{d}x\right)\left(\int_{\sqrt{3}/2}^{\infty} \frac{\mathrm{d}y}{y^2}\right) \leq \frac{2}{\sqrt{3}} < \infty.$$

where we used Fubini's Theorem. Thus, $\mu_{T^1\mathbb{H}}(F) \leq 2\pi m_{\mathbb{H}}(E) < \infty$, which proves that Γ is a lattice.

Let us now verify that F is a fundamental domain for Γ. In order to do this, it is enough to show that all Γ-orbits intersect E and, up to a zero measure set, they intersect E exactly once.

Fix $z \in \mathbb{H}$, and let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$. By (4.3), $\Im(\gamma.z) \to 0$ when either c or d go to infinity. Thus, there exists $\gamma \in \Gamma$ that maximizes $\Im(\gamma.z)$. Choose $k \in \mathbb{Z}$ such that $|\Re w| \leq 1/2$, where $w = \tau^k \gamma.z$. If $|w| < 1$, then the element $\sigma.w$ is such that $\Im(\sigma.w) = \Im(w)/|w| > \Im(w)$, contradicting the maximality assumption. Hence $|w| \geq 1$. This shows that $w \in E$ and hence $\Gamma z \cap E \neq \emptyset$.

It remains to show that if two points in E are in the same Γ-orbit, then they belong to a zero measure set; in particular we will show that they belong to the boundary of E. Let us consider $z, w \in E$ such that $w = \gamma.z$ for some $\gamma \in \Gamma$. Up to exchanging z and w, and considering γ^{-1}, we can assume that $\Im(w) \geq \Im(z)$. We write again $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, with the sign assumption that $c \geq 0$. By (4.3), from $\Im(\gamma.z) \geq \Im(z)$ we get

$$1 \geq |cz+d| \geq \Im(cz+d) = c\Im z \geq c\sqrt{3}/2 > c/2,$$

from which it follows $c = 0$ or $c = 1$.

If $c = 0$, then $1 = \det\gamma = ad$ implies that $a = d = \pm 1$ so that $\gamma.z = z \pm b$. Since $w = \gamma.z, z \in E$ both have real part in $[-1/2, 1/2]$, the only possibilities are $b = 0$, which implies $w = z$, or $b = \pm 1$ and $\Re z = \mp 1/2$, which means that w and z belongs to the boundary of E, as claimed.

If $c = 1$, then from the inequality $|z+d| \leq 1$ we deduce that either $d = 0$, or $d = \pm 1$. In the first case, we get $|z| \leq 1$, which, since $z \in E$, forces $|z| = 1$; moreover, the condition on the determinant implies $b = -1$. Thus, $w = \gamma.z = a - 1/z \in E$ implies that $a = 0$ and $|w| = 1$ as well, or $a = 1$ and z is the primitive sixth root of unity in E. Either way, both z and w belong to the boundary of E. In the second case, similarly, from $|z \pm 1| \leq 1$ and $|z| \geq 1$, we deduce that z is one of the sixth roots of unity in E. A straightforward computation allows to conclude that also w is a sixth root of unity in E. In both cases the claim is proved, hence the proof is complete. □

The hyperbolic surface $S = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathbb{H}$ is called the *Modular Surface*, and, as we mentioned above, its unit tangent bundle can be identified with the quotient $T^1(S) = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathrm{PSL}(2,\mathbb{R})$. It is an example of a homogeneous space with finite measure but which is not compact.

Chapter 5

Ergodic properties of geodesic and horocycle flows

In this chaper, we are going to study the ergodic theory of geodesic and horocycle flows on finite-volume quotients M of $G = \mathrm{PSL}(2,\mathbb{R})$. We showed in Proposition 4.10 that the normalized Haar measure is an invariant probability measure for all homogeneous flows on M. We are going to show that it is ergodic and, in fact, mixing for both the geodesic and the horocycle flow.

5.1 Ergodicity

5.1.1 Hopf's argument

Let us fix a lattice Γ in G, and let $M = \Gamma\backslash G$ be a homogeneous space. We equip M with the measure μ (which, locally, coincide with the Haar measure μ_G on G), normalized so that M is a probability space. The goal is to show that the geodesic flow $\phi_t^{\mathbf{x}} \colon \Gamma g \mapsto \Gamma g \exp(t\mathbf{x})$ is ergodic with respect to μ.

One ingredient that we will need to use is that the subgroups $\exp(\mathbb{R}\mathbf{u})$ and $\exp(\mathbb{R}\mathbf{v})$ generate $\mathrm{SL}(2,\mathbb{R})$. One could show this fact using the general theory of Lie groups, or in a direct, computation-based, way. We leave this as an exercise to the reader.

Exercise 5.1. *(a) Show that all matrices of the form* $\begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$ *or* $\begin{pmatrix} 0 & a \\ -1/a & 0 \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R})$ *can be written as a product* $\exp(x_1\mathbf{u})\exp(x_2\mathbf{v})\exp(x_3\mathbf{u})\exp(x_4\mathbf{v})$ *for some* $x_i \in \mathbb{R}$, $i = 1,2,3,4$.

(b) Show that all elements $g \in \mathrm{SL}(2,\mathbb{R})$ *can be written as a product* $g = g_0\exp(y_1\mathbf{v})\exp(y_2\mathbf{u})$ *for some* $y_1, y_2 \in \mathbb{R}$, *where* g_0 *is either of the form* $\begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$ *or* $\begin{pmatrix} 0 & a \\ -1/a & 0 \end{pmatrix}$.

(c) Deduce that all for all $g \in \mathrm{SL}(2,\mathbb{R})$ *there exist* $x_i \in \mathbb{R}$, $i = 1,\dots,x_5$ *such that* $g = \exp(x_1\mathbf{u})\exp(x_2\mathbf{v})\exp(x_3\mathbf{u})\exp(x_4\mathbf{v})\exp(x_5\mathbf{u})$.

Another result that we will use is the following, whose proof relies on Fubini's Theorem and the right-invariance of the Haar measure (see Proposition 4.10).

Lemma 5.2. *Let* $A_1, A_2 \subset M$ *be two measurable sets of positive measure. Then, the set*

$$\{g \in G : \mu(A_1 g \cap A_2) > 0\}$$

has positive measure; in particular it is not empty.

Proof. Let $F \subset G$ be a fundamental domain for Γ. In particular, for any subset $E \subset F$ we have $\mu_G(E) = \mu(\pi(E))$. We can find two subsets $B_1, B_2 \subset F$ such that, up to zero measure sets, $\pi(B_1) = A_1$ and $\pi(B_2) = A_2$. Since $\pi(B_1 g \cap B_2) \subset A_1 g \cap A_2$ for all $g \in G$, we obtain that

$$\mu(A_1 g \cap A_2) \geq \mu(\pi(B_1 g \cap B_2)) = \mu_G(B_1 g \cap B_2),$$

thus, in order to conclude, it is enough to show that the set $\{g \in G : \mu_G(B_1 g \cap B_2) > 0\}$ has positive measure.

Before doing that, let us make a few observations:

1. for any $g, h \in G$, $h \in B_1 g$ if and only if $g \in B_1^{-1} h$, where $B_1^{-1} := \{b^{-1} : b \in B_1\}$.
2. $\mu_G(B_1) > 0$ if and only if $\mu_G(B_1^{-1}) > 0$; indeed, it is easy to see that the measure $\nu(A) := \mu_G(A^{-1})$ is a (right-invariant) Haar measure on G, and hence must be proportional to μ_G.
3. if $\int_G \mu_G(B_1 g \cap B_2)\, \mathrm{d}\mu(g) > 0$, then the set $\{g \in G : \mu_G(B_1 g \cap B_2) > 0\}$ has positive measure.

The third observation tells us that we should estimate $\int_G \mu_G(B_1 g \cap B_2)\, \mathrm{d}\mu(g)$. Using Fubini's Theorem and the first observation,

$$\begin{aligned}
\int_G \mu_G(B_1 g \cap B_2)\, \mathrm{d}\mu(g) &= \int_G \left(\int_G \mathbb{1}_{B_1 g}(h)\, \mathbb{1}_{B_2}(h)\, \mathrm{d}\mu(h) \right) \mathrm{d}\mu(g) \\
&= \int_G \mathbb{1}_{B_2}(h) \left(\int_G \mathbb{1}_{B_1 g}(h)\, \mathrm{d}\mu(g) \right) \mathrm{d}\mu(h) = \int_G \mathbb{1}_{B_2}(h) \left(\int_G \mathbb{1}_{B_1^{-1} h}(g)\, \mathrm{d}\mu(g) \right) \mathrm{d}\mu(h) \\
&= \int_G \mathbb{1}_{B_2}(h) \mu_G(B_1^{-1} h)\, \mathrm{d}\mu(h).
\end{aligned}$$

From the right invariance of the Haar measure and from the second observation, we conclude

$$\int_G \mu_G(B_1 g \cap B_2)\, \mathrm{d}\mu(g) = \int_G \mathbb{1}_{B_2}(h) \mu_G(B_1^{-1})\, \mathrm{d}\mu(h) = \mu_G(B_1^{-1}) \mu_G(B_2) > 0,$$

which proves our claim. □

In order to prove ergodicity, we need to show that any measurable function which is invariant by the geodesic flow is almost everywhere constant. The first step in this direction is given by the following lemma

Lemma 5.3. *Let $f \colon M \to \mathbb{C}$ be a measurable function such that $f \circ \phi_t^{\mathbf{a}} = f$ almost everywhere for all $t \in \mathbb{R}$. For every $s \in \mathbb{R}$, there exists a set $M_s \subset M$ of measure $\mu(M_s) = 1$ such that for all $p \in M$ for which both $p \in M_s$ and $\phi_s^{\mathbf{u}}(p) \in M_s$, we have $f(p) = f(\phi_s^{\mathbf{u}}(p))$.*

The same conclusion holds with $\phi_s^{\mathbf{u}}$ replaced by $\phi_s^{\mathbf{v}}$.

Proof. Let $f \colon M \to \mathbb{C}$ be as in the assumptions, and let $s \in \mathbb{R}$. Fix $\varepsilon > 0$, and choose a compact set $K \subset M$ of measure $\mu(K) \geq 1 - \varepsilon$ such that the restriction of f to K is continuous.

By the Ergodic Theorem, there exists a measurable function $\ell \colon M \to \mathbb{R}$ such that

$$\ell(p) = \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{1}_K \circ \phi_r^{\mathbf{a}}(p)\, \mathrm{d}r$$

almost everywhere, and $\int_M \ell\, \mathrm{d}\mu = \int_M \mathbb{1}_K\, \mathrm{d}\mu = \mu(K) \geq 1 - \varepsilon$. Let

$$B = \left\{ p \in M : \ell(p) > \frac{1}{2} \right\}.$$

Then, since $0 \leq \ell(p) \leq 1$ almost everywhere,

$$\begin{aligned} 1-\varepsilon \leq \int_M \ell \, \mathrm{d}\mu = \int_B \ell \, \mathrm{d}\mu + \int_{M\setminus B} \ell \, \mathrm{d}\mu \leq \mu(B) + \mu(M\setminus B) \cdot \operatorname{ess\,sup}_{p\in M\setminus B} \ell(p) \\ \leq \mu(B) + \frac{1}{2}\mu(M\setminus B) \leq \frac{1}{2} + \frac{1}{2}\mu(B), \end{aligned}$$

from which we get $\mu(B) \geq 1-2\varepsilon$.

Let now $p \in B$ and $q = \phi_s^{\mathbf{u}}(p) \in B$. Since, by definition of B, the geodesic orbits of p and q spend more than half the times in K, we can find an increasing subsequence of times $t_n \to \infty$ such that $\phi_{t_n}^{\mathbf{a}}(p) \in K$ and $\phi_{t_n}^{\mathbf{a}}(q) \in K$. Moreover, we have

$$\phi_{t_n}^{\mathbf{a}}(q) = \phi_{t_n}^{\mathbf{a}} \circ \phi_s^{\mathbf{u}}(p) = \phi_{e^{-t_n}s}^{\mathbf{u}} \circ \phi_{t_n}^{\mathbf{a}},$$

hence

$$d_G(\phi_{t_n}^{\mathbf{a}}(p), \phi_{t_n}^{\mathbf{a}}(q)) \to 0, \qquad \text{as } n \to \infty,$$

exponentially fast. By the invariance assumption on f and the continuity on K, we deduce

$$|f(p) - f(q)| = |f(\phi_{t_n}^{\mathbf{a}}(p)) - f(\phi_{t_n}^{\mathbf{a}}(q))| \to 0,$$

which means that $f(p) = f(q)$ whenever p and $q = \phi_s^{\mathbf{u}}(p)$ are both in B.

If now we consider a smaller $\widetilde{\varepsilon} < \varepsilon$, we can repeat the same argument and find a larger $\widetilde{B}$ on which the same conclusion holds. Therefore, taking the union over all $\varepsilon > 0$, we deduce that there exists a set M_s of full measure on which the conclusion holds.

The same argument can be repeated for $\phi_s^{\mathbf{v}}$ up to considering negative times $t_n \to -\infty$. □

We now have all the tools to prove ergodicity of the geodesic flow.

Theorem 5.4. *Let Γ be a lattice in $G = \mathrm{PSL}(2,\mathbb{R})$. Then, the geodesic flow $\phi_t^{\mathbf{a}}$ on $M = \Gamma\backslash G$ is ergodic with respect to μ.*

Proof. Let $f\colon M \to \mathbb{C}$ be a measurable function which is invariant by the geodesic flow $\phi_t^{\mathbf{a}}$. By Lemma 5.3, for every fixed $s \in \mathbb{R}$, there exists a set M_s' of full measure such that

$$f(p) = f(p\exp(s\mathbf{u})) = f(p\exp(s\mathbf{v})),$$

for all $p \in M_s'$.

We need to show that f is constant almost everywhere. Assume that this is not the case, namely there exists two measurable subsets $A_1, A_2 \subset M$, both of positive measure, and two disjoint balls $I_1, I_2 \subset \mathbb{C}$ such that $f(A_1) \subset I_1$ and $f(A_2) \subset I_2$. By Lemma 5.2, there exists $\widetilde{g} \in G$ such that $\mu(A_1\widetilde{g} \cap A_2) > 0$. According to Exercise 5.1, we can write $\widetilde{g} = \exp(s_1\mathbf{u})\exp(s_2\mathbf{v})\exp(s_3\mathbf{u})\exp(s_4\mathbf{v})\exp(s_5\mathbf{u})$. By Lemma 5.3 applied (at most) 5 times, there exists a set

$$\begin{aligned} \widetilde{M} = & M_{s_1} \cap M_{s_2}\exp(-s_1\mathbf{u}) \cap M_{s_3}\exp(-s_2\mathbf{v})\exp(-s_1\mathbf{u}) \cap M_{s_4}\exp(-s_3\mathbf{u})\exp(-s_2\mathbf{v})\exp(-s_1\mathbf{u}) \\ & \cap M_{s_5}\exp(-s_4\mathbf{v})\exp(-s_3\mathbf{u})\exp(-s_2\mathbf{v})\exp(-s_1\mathbf{u}), \end{aligned}$$

with $\mu(\widetilde{M}) = 1$, such that $f(p) = f(p\widetilde{g})$ for all $p \in \widetilde{M}$.

The set

$$A(\widetilde{g}) := A_1\widetilde{g}^{-1} \cap A_2 \cap \widetilde{M}$$

has positive measure, in particular it is not empty. Let $p \in A(\widetilde{g})$. From $p \in A_2$ we get $f(p) \in I_2$; on the other hand, since $p \in A_1\widetilde{g}^{-1}$, we get also $f(p\widetilde{g}) \in I_1$. However, $p \in \widetilde{M}$ implies $f(p) = f(p\widetilde{g})$, which contradicts the fact that I_1 and I_2 are disjoint. The proof is then complete. □

5.1.2 Mautner's Phenomenon

In this section we will see an instance of the so-called *Mautner's Phenomenon*. Roughly speaking, this refers to the situation when the invariance of an observable with respect to a flow implies an additional invariance with respect to a transverse flow. In our case, we will show that a measurable function which is invariant for the horocycle flow is also invariant for the geodesic flow. From this we will deduce that the horocycle flow is ergodic.

As before, $M = \Gamma\backslash G$ is a finite-volume homogeneous manifold. We will need a couple of preliminary results.

Lemma 5.5. *Let (X,d) be a metric space and let μ be a Borel probability measure on X. Assume that $\{T_n\}_{n\in\mathbb{N}}$ is a sequence of continuous measure preserving self-maps on X which converges uniformly to a continuous measure preserving map $T\colon X \to X$. Then, for every $f \in L^2(X)$, we have $f \circ T_n \to f \circ T$ in $L^2(X)$.*

Proof. Let $f \in L^2(X)$ and let us fix $\varepsilon > 0$. By Lusin's Theorem, there exists a compact set $K \subset X$ of measure $\mu(K) \geq 1-\varepsilon$ such that the restriction of f to K is uniformly continuous. Let $\delta > 0$ be such that $|f(x) - f(y)| < \varepsilon$ whenever $x, y \in K$ and $d(x,y) < \delta$.

By assumption, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ and for all $x \in X$ we have $d(T_n(x), T(x)) \leq \delta$. Let $n \geq N$, and consider $K(n) := T^{-1}K \cap T_n^{-1}K$. By construction, and since T_n and T are measure preserving, $\mu(K(n)) \geq 1-2\varepsilon$ and $|f \circ T_n(x) - f \circ T(x)| < \varepsilon$ for all $x \in K(n)$. Therefore, we get

$$\begin{aligned}
\|f \circ T_n - f \circ T\|_2^2 &= \int_X |f \circ T_n(x) - f \circ T(x)|^2 \,\mathrm{d}\mu \\
&= \int_{X\setminus K(n)} |f \circ T_n(x) - f \circ T(x)|^2 \,\mathrm{d}\mu + \int_{K(n)} |f \circ T_n(x) - f \circ T(x)|^2 \,\mathrm{d}\mu \\
&\leq 2\|f\|_2^2 \mu(X \setminus K(n)) + \varepsilon^2 \mu(K(n)) \leq (4\|f\|_2^2 + \varepsilon)\varepsilon,
\end{aligned}$$

which proves our claim. □

In the following proposition, we prove that if a function is invariant for the horocycle flow, it is also invariant for any other homogeneous flow that satisfies a certain condition.

Proposition 5.6. *Assume that $\mathbf{w} \in \mathfrak{sl}(2,\mathbb{R})$ satisfies the following condition:*

> *for all $t \in \mathbb{R}$, there exist sequences $(r_n)_{n\in\mathbb{N}}, (s_n)_{n\in\mathbb{N}}$ of real numbers, and $\{g_n\}_{n\in\mathbb{N}} \subset G$ with $g_n \to e$, such that*
> $$\exp(s_n\mathbf{u})\, g_n \exp(r_n\mathbf{u}) \to \exp(t\mathbf{w}).$$

Then, if a function $f \in L^2(M)$ is invariant under $\{\phi_t^{\mathbf{u}}\}_{t\in\mathbb{R}}$, it is also invariant under $\{\phi_t^{\mathbf{w}}\}_{t\in\mathbb{R}}$.

Proof. Let $f \in L^2(M)$ be any real-valued measurable function. We are going to define an auxiliary function $p\colon G \to \mathbb{R}$ as follows

$$p(g) = \langle f \circ R_g, f\rangle,$$

where, we recall, $R_g\colon M \to M$ is the right-multiplication map $p \mapsto pg$. The function p satisfies the following properties:

1. p is continuous. Indeed, for any convergent sequence $g_n \to g$ in G, by the Cauchy-Schwartz inequality

   $$|p(g_n) - p(g)| = |\langle f \circ R_{g_n} - f \circ R_g, f\rangle| \leq \|f\|_2 \cdot \|f \circ R_{g_n} - f \circ R_g\| \to 0,$$

   which follows from Lemma 5.5.

2. if $p(g) = \|f\|_2^2$, then $f \circ R_g = \pm f$. This follows again from the Cauchy-Schwartz inequality:

$$\|f\|_2^2 = p(g) \leq |\langle f \circ R_g, f\rangle| \leq \|f\|_2 \cdot \|f \circ R_g\|_2 = \|f_2\|_2^2.$$

Since equality holds, the terms $f \circ R_g$ and f must be linearly dependent, $f \circ R_g = \lambda f$. Since they have the same norm, $\lambda = \pm 1$.

Assume now that $f \in L^2(M)$ is invariant by the horocycle flow, and let $\mathbf{w} \in \mathfrak{sl}(2,\mathbb{R})$ be as in the assumption of the lemma. We need to show that $f \circ \phi_t^{\mathbf{w}} = f$ in $L^2(M)$ for all $t \in \mathbb{R}$.

Fix $t \in \mathbb{R}$, and let $(r_n)_{n\in\mathbb{N}}, (s_n)_{n\in\mathbb{N}}$, and $\{g_n\}_{n\in\mathbb{N}}$ be such that $\exp(s_n\mathbf{u})\, g_n \exp(r_n\mathbf{u}) \to \exp((t/2)\mathbf{w})$. On one hand, by the continuity of p, we have

$$p\left(\exp(s_n\mathbf{u})\, g_n \exp(r_n\mathbf{u})\right) \to p\left(\exp((t/2)\mathbf{w})\right).$$

On the other hand, by assumption on f and again by the continuity of p, we have

$$\begin{aligned} p\left(\exp(s_n\mathbf{u})\, g_n \exp(r_n\mathbf{u})\right) &= \langle f \circ R_{\exp(s_n\mathbf{u})\, g_n \exp(r_n\mathbf{u})}, f\rangle = \langle f \circ \phi_{r_n}^{\mathbf{u}} \circ R_{g_n} \circ \phi_{s_n}^{\mathbf{u}}, f\rangle = \langle f \circ R_{g_n} \circ \phi_{s_n}^{\mathbf{u}}, f\rangle \\ &= \langle f \circ R_{g_n}, f \circ \phi_{-s_n}^{\mathbf{u}}\rangle = \langle f \circ R_{g_n}, f\rangle = p(g_n) \to p(e) = \|f\|_2^2. \end{aligned}$$

Combining these two observations, we deduce $p\left(\exp((t/2)\mathbf{w})\right) = \|f\|_2^2$. By the second property of p, we get $f \circ \phi_{t/2}^{\mathbf{w}} = \pm f$, from which it follows

$$f \circ \phi_t^{\mathbf{w}} = f \circ \phi_{t/2}^{\mathbf{w}} \circ \phi_{t/2}^{\mathbf{w}} = \pm(f \circ \phi_{t/2}^{\mathbf{w}}) = f,$$

and the proof is complete. □

Theorem 5.7. *Let Γ be a lattice in $G = \mathrm{PSL}(2,\mathbb{R})$. Then, the horocycle flow $\phi_t^{\mathbf{u}}$ on $M = \Gamma\backslash G$ is ergodic with respect to μ.*

Proof. Let $f \in L^2(M)$ be an invariant function for the horocycle flow. We want to apply Proposition 5.6 to show that f is also invariant for the geodesic flow. Once we have done this, it follows from Theorem 5.4 that f is constant almost everywhere, which proves the result.

Let $\mathbf{a} = \begin{pmatrix} 1/2 & 0 \\ 0 & -1/2 \end{pmatrix}$ be the generator of the geodesic flow. We show that we can find sequences $(r_n)_{n\in\mathbb{N}}, (s_n)_{n\in\mathbb{N}}$ such that

$$\begin{pmatrix} 1 & s_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{1}{n} & 1 \end{pmatrix} \begin{pmatrix} 1 & r_n \\ 0 & 1 \end{pmatrix} \to \begin{pmatrix} e^{\frac{t}{2}} & 0 \\ 0 & e^{-\frac{t}{2}} \end{pmatrix} = \exp(t\mathbf{a}).$$

It is immediate to verify that $s_n = n(e^{\frac{t}{2}} - 1)$ and $r_n = n(e^{-\frac{t}{2}} - 1)$ satisfy the required property. □

5.2 Mixing

In this section we prove that the geodesic and the horocycle flow are mixing. The proof relies on the ergodicity of the horocycle flow that we showed in Theorem 5.7.

Let us first prove mixing for the geodesic flow.

Theorem 5.8. *Let Γ be a lattice in $G = \mathrm{PSL}(2,\mathbb{R})$. Then, the geodesic flow $\phi_t^{\mathbf{a}}$ on $M = \Gamma\backslash G$ is mixing with respect to μ.*

Proof. Let $f \in L^2(M)$. Mixing for the geodesic flow is equivalent to the claim that $f \circ \phi_t^{\mathbf{a}}$ converges to $\mu(f) = \int_M f \, \mathrm{d}\mu$ in the weak-$*$ topology.

Since $\|f \circ \phi_t^{\mathbf{a}}\|_2 = \|f\|_2$, the family $\{f \circ \phi_t^{\mathbf{a}} : t \in \mathbb{R}\}$ is contained in a closed ball in the dual $L^2(M)^* = L^2(M)$. By Banach-Alaoglu's Theorem, such set is relatively compact in the weak-$*$ topology; in particular it has limit points. We claim that the only limit is the constant $\mu(f)$.

Let $t_n \to \infty$ be an increasing sequence of times such that $f \circ \phi_{t_n}^{\mathbf{a}} \to f_0$, for some $f_0 \in L^2(M)^*$. Notice that

$$\int_M f \, \mathrm{d}\mu = \langle f, \mathbb{1} \rangle = \langle f \circ \phi_{t_n}^{\mathbf{a}}, \mathbb{1} \rangle \to \langle f_0, \mathbb{1} \rangle = \int_M f_0 \, \mathrm{d}\mu,$$

therefore it is enough to show that f_0 is constant.

We show that f_0 is invariant under the horocycle flow. Theorem 5.7 implies that it is constant, hence finishing the proof. Let us fix $T \in \mathbb{R}$, and let us prove that $f_0 \circ \phi_T^{\mathbf{u}} = f_0$. Let $\ell \in L^2(M)$; then, using measure invariance,

$$\langle f \circ \phi_{t_n}^{\mathbf{a}}, \ell \rangle = \langle f, \ell \circ \phi_{-t_n}^{\mathbf{a}} \rangle = \langle f \circ \phi_{-T/e^{t_n}}^{\mathbf{u}}, \ell \circ \phi_{-t_n}^{\mathbf{a}} \circ \phi_{-T/e^{t_n}}^{\mathbf{u}} \rangle = \langle f, \ell \circ \phi_{-t_n}^{\mathbf{a}} \circ \phi_{-T/e^{t_n}}^{\mathbf{u}} \rangle + E(t_n),$$

where the term

$$E(t_n) = \langle (f \circ \phi_{-T/e^{t_n}}^{\mathbf{u}} - f), \ell \circ \phi_{-t_n}^{\mathbf{a}} \circ \phi_{-T/e^{t_n}}^{\mathbf{u}} \rangle$$

satisfies

$$|E(t_n)| \leq \|f \circ \phi_{-T/e^{t_n}}^{\mathbf{u}} - f\|_2 \cdot \|\ell\|_2 \to 0,$$

where we used Cauchy-Schwartz inequality and Lemma 5.5.

We now apply the commutation relation (4.7) and we get

$$\langle f \circ \phi_{t_n}^{\mathbf{a}}, \ell \rangle = \langle f, \ell \circ \phi_{-T}^{\mathbf{u}} \circ \phi_{-t_n}^{\mathbf{a}} \rangle + E(t_n) = \langle f \circ \phi_{t_n}^{\mathbf{a}}, \ell \circ \phi_{-T}^{\mathbf{u}} \rangle + E(t_n).$$

Taking the limits for $n \to \infty$, since we showed $E(t_n) \to 0$, we obtain

$$\langle f_0, \ell \rangle = \lim_{n \to \infty} \langle f \circ \phi_{t_n}^{\mathbf{a}}, \ell \rangle = \lim_{n \to \infty} \langle f \circ \phi_{t_n}^{\mathbf{a}}, \ell \circ \phi_{-T}^{\mathbf{u}} \rangle + E(t_n) = \langle f_0, \ell \circ \phi_{-T}^{\mathbf{u}} \rangle = \langle f_0 \circ \phi_T^{\mathbf{u}}, \ell \rangle.$$

Since the equality holds for all $\ell \in L^2(M)$, we conclude that $f_0 = f_0 \circ \phi_T^{\mathbf{u}}$ in $L^2(M)$; in other words we proved that f_0 is invariant under the horocycle flow, which proves the theorem. □

With the same strategy, we can prove that the horocycle flow is mixing.

Theorem 5.9. *Let Γ be a lattice in $G = \mathrm{PSL}(2, \mathbb{R})$. Then, the horocycle flow $\phi_t^{\mathbf{u}}$ on $M = \Gamma \backslash G$ is mixing with respect to μ.*

Proof. We proceed as in Theorem 5.8: let $f \in L^2(M)$, and let $f \circ \phi_{t_n}^{\mathbf{u}} \to f_0$ be a converging subsequence in the weak-$*$ topology of $L^2(M)$. We prove that f_0 is invariant under the horocycle flow. This forces f_0 to be constant and equal to $\mu(f)$, which proves mixing.

Let us fix $T \in \mathbb{R}$. We will use again (4.7), more precisely we will apply the relation

$$\phi_{-t_n}^{\mathbf{u}} \circ \phi_{2\log(1+T/(2t_n))}^{\mathbf{a}}(p) = \phi_{2\log(1+T/(2t_n))}^{\mathbf{a}} \circ \phi_{-T^2/(2t_n)}^{\mathbf{u}} \circ \phi_{-T}^{\mathbf{u}} \circ \phi_{-t_n}^{\mathbf{u}}(p), \tag{5.1}$$

for all $p \in M$. As in the proof of Theorem 5.8, using (5.1), we have

$$\begin{aligned}
\langle f \circ \phi_{t_n}^{\mathbf{u}}, \ell \rangle &= \langle f \circ \phi_{2\log(1+T/(2t_n))}^{\mathbf{a}}, \ell \circ \phi_{-t_n}^{\mathbf{u}} \circ \phi_{2\log(1+T/(2t_n))}^{\mathbf{a}} \rangle \\
&= \langle f \circ \phi_{2\log(1+T/(2t_n))}^{\mathbf{a}}, \ell \circ \phi_{2\log(1+T/(2t_n))}^{\mathbf{a}} \circ \phi_{-T^2/(2t_n)}^{\mathbf{u}} \circ \phi_{-T}^{\mathbf{u}} \circ \phi_{-t_n}^{\mathbf{u}} \rangle \\
&= \langle f, \ell \circ \phi_{-T}^{\mathbf{u}} \circ \phi_{-t_n}^{\mathbf{u}} \rangle + E_1(t_n) + E_2(t_n),
\end{aligned}$$

where we have

$$\begin{aligned}|E_1(t_n)| &= |\langle (f\circ\phi^{\mathbf{a}}_{2\log(1+T/(2t_n))}-f),\ell\circ\phi^{\mathbf{a}}_{2\log(1+T/(2t_n))}\circ\phi^{\mathbf{u}}_{-T^2/(2t_n)}\circ\phi^{\mathbf{u}}_{-T}\circ\phi^{\mathbf{u}}_{-t_n}\rangle| \\ &\leq \|f\circ\phi^{\mathbf{a}}_{2\log(1+T/(2t_n))}-f\|_2\cdot\|\ell\|_2\to 0,\end{aligned}$$

by Lemma 5.5, and similarly

$$\begin{aligned}|E_2(t_n)| &= |\langle f,(\ell\circ\phi^{\mathbf{a}}_{2\log(1+T/(2t_n))}\circ\phi^{\mathbf{u}}_{-T^2/(2t_n)}-\ell)\circ\phi^{\mathbf{u}}_{-T}\circ\phi^{\mathbf{u}}_{-t_n}\rangle| \\ &\leq \|f\|_2\cdot\|\ell\circ\phi^{\mathbf{a}}_{2\log(1+T/(2t_n))}\circ\phi^{\mathbf{u}}_{-T^2/(2t_n)}-\ell\|_2\to 0.\end{aligned}$$

Therefore, we get

$$\langle f_0,\ell\rangle = \lim_{n\to\infty}\langle f\circ\phi^{\mathbf{u}}_{t_n},\ell\rangle = \lim_{n\to\infty}\langle f\circ\phi^{\mathbf{u}}_{t_n},\ell\circ\phi^{\mathbf{u}}_{-T}\rangle + E_1(t_n)+E_2(t_n) = \langle f_0,\ell\circ\phi^{\mathbf{u}}_{-T}\rangle = \langle f_0\circ\phi^{\mathbf{u}}_{T},\ell\rangle.$$

Since $\ell\in L^2(M)$ was arbitrary, we conclude $f_0 = f_0\circ\phi^{\mathbf{u}}_T$, which completes the proof. □

Chapter 6

The geodesic flow on the Modular Surface and continued fractions

In this chapter, we are going to explore an interesting and rather pleasing connection between the geodesic flow on the Modular Surface and the classical continued fraction expansion of real numbers. This connection was first noted by Artin [1], who used it to construct dense orbits of the geodesic flow. We will follow the treatement by Caroline Series in [17].

6.1 Background on continued fractions

Let us recall some background on the theory of standard continued fractions. Let us consider a positive rational number $p/q \in \mathbb{Q}$, with $p,q \in \mathbb{N}$ coprime. By applying the standard Euclidean algorithm, we obtain finite sequences $a_0,\dots,a_n,r_1,\dots,r_{n-1} \in \mathbb{N}$ with $q > r_1 > \cdots > r_n > 1$ such that

$$p = a_0 q + r_1, \quad q = a_1 r_1 + r_2, \quad r_1 = a_2 r_2 + r_3, \quad \dots, \quad r_{n-2} = a_{n-1} r_{n-1} + 1, \quad r_{n-1} = a_n \cdot 1.$$

The second to last remainder is 1 because of the coprimality assumption. If we combine these equalities together, we can write

$$\frac{p}{q} = a_0 + \frac{r_1}{q} = a_0 + \frac{1}{\frac{q}{r_1}} = a_0 + \cfrac{1}{a_1 + \cfrac{r_2}{r_1}} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{r_3}{r_2}}} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}}.$$

For short, we write $p/q = [a_0; a_1, \dots, a_n]$. Note that this expression is almost unique, as we could also write $p/q = [a_0; a_1, \dots, a_n - 1, 1]$.

The same algorithmic procedure can be carried out for irrational numbers. Let $\lfloor \cdot \rfloor$ and $\{\cdot\}$ denote respectively the integer and fractional part. For any positive real $x > 0$, we inductively define

$$a_0 = \lfloor x \rfloor,\ G^1(x) = \{1/x\}, \qquad a_n = \lfloor G^n(x) \rfloor,\ G^{n+1} = \{1/G^n(x)\}.$$

In this way, for any $n \in \mathbb{N}$, we can express x as

$$x = a_0 + \frac{1}{G^1(x)} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{G^2(x)}} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_n + \cfrac{1}{G^{n+1}(x)}}}}.$$

This procedure stops if and only if $x \in \mathbb{Q}$. If $x \in \mathbb{R} \setminus \mathbb{Q}$, we obtain an infinite sequence of positive integers $a_i \in \mathbb{N}$, and we write $x = [a_0; a_1, a_2, \dots]$. This is called the *continued fraction expansion of x*.

The previous discussion can easily be extended to negative reals if we allow a_0 to possibly be a negative integer (while still $a_i \in \mathbb{N}$ for all $i \geq 1$).

If $x \in \mathbb{R} \setminus \mathbb{Q}$, the rational numbers p_n/q_n we get by truncating the continued fraction expansion at step n, namely

$$\frac{p_n}{q_n} = [a_0; a_1, \dots, a_n]$$

are called the *convergents of x*. Convergents satisfy the following properties, see, for example, [11].

Proposition 6.1. *Let $x = [a_0; a_1, a_2, \dots]$, and let p_n/q_n denote the sequence of its convergents (written in reduced terms). Then,*

1. *the matrix $\begin{pmatrix} p_n & p_{n+1} \\ q_n & q_{n+1} \end{pmatrix}$ has determinant ± 1 for all $n \geq 0$,*
2. *for $n \geq 1$,*
$$p_{n+1} = a_{n+1} p_n + p_{n-1} \qquad \text{and} \qquad q_{n+1} = a_{n+1} q_n + q_{n-1}$$
where $p_{-1} = 1$ and $q_{-1} = 0$,
3. *for all $n \geq 0$,*
$$\frac{p_{2n}}{q_{2n}} \leq \frac{p_{2n+2}}{q_{2n+2}} \leq x \leq \frac{p_{2n+1}}{q_{2n+1}} \leq \frac{p_{2n-1}}{q_{2n-1}},$$
with equality on one side or the other if and only if $x \in \mathbb{Q}$ and the sequence terminates,
4. *for all $n \geq 1$, we have $|x - p_n/q_n| \leq (q_n q_{n+1})^{-1}$ and $q_n \geq n$.*

The continued fraction expansion of a number can be recovered by looking at its itinerary under the *Gauss map* $G\colon [0,1) \to [0,1)$, defined by $G(0) = 0$ and $G(x) = \{1/x\}$ for $x \neq 0$, as follows. Define

$$I_n := \left[\frac{1}{n+1}, \frac{1}{n}\right), \qquad \text{for } n \geq 1;$$

note that the intervals I_n form a partition of $(0,1)$. For $x \in \mathbb{R}$, let $a_0 = \lfloor x \rfloor$, and $x_0 = \{x\} \in [0,1)$. Then, $a_n = k$ if and only if the $(n-1)$-th iterate $G^{n-1}(x_0)$ of x_0 belongs to I_k. The orbit of x_0 under G is infinite if and only if x_0 (and hence x) is irrational.

6.2 The Farey tessellation

Given two rational numbers $p/q, r/s \in \mathbb{Q}$, we define their *Farey sum* $p/q \oplus r/s \in \mathbb{Q}$ to be

$$\frac{p}{q} \oplus \frac{r}{s} := \frac{p+r}{q+s}.$$

Notice that $p/q < p/q \oplus r/s < r/s$. We say that p/q and r/s are *neighbors* if $|ps - rq| = 1$. It is not hard to see that, if p/q and r/s are neighbors, then so are p/q and $p/q \oplus r/s$, and $p/q \oplus r/s$ and r/s.

For any $n \in \mathbb{N}$, we define the *n-th Farey Sequence* $\mathcal{F}_n$ to be the set of all rationals p/q with $|p|, |q| \leq n$ arranged in increasing order (with the convention $\pm\infty = \pm 1/0$). The first few Farey

Sequences are

$$\mathcal{F}_1 \ : \ -\infty < -1 < 0 < 1 < \infty,$$
$$\mathcal{F}_2 \ : \ -\infty < -2 < -1 < -\frac{1}{2} < 0 < \frac{1}{2} < 1 < 2 < \infty,$$
$$\mathcal{F}_3 \ : \ -\infty < -3 < -2 < -\frac{3}{2} < -1 < -\frac{2}{3} < -\frac{1}{2} < -\frac{1}{3} < 0 < \frac{1}{3} < \frac{1}{2} < \frac{2}{3} < 1 < \frac{3}{2} < 2 < 3 < \infty.$$

One can see that the n-th Farey Sequence is obtained by adding to the $(n-1)$-th Farey Sequence all the Farey sums between consecutive terms.

Using the Farey Sequences, we now describe an algorithmic procedure that will result in a *tessellation* of the hyperbolic plane $\mathbb{H}$. The procedure goes as follows:

1. draw vertical lines from each $n = n/1 \in \mathbb{Z}$,
2. join each pair $(\frac{n}{1}, \frac{n+1}{1})$ by a semicircle with center on $\mathbb{R}$,
3. mark the points $\frac{n}{1} \oplus \frac{n+1}{1}$ and join them with their neighbors $\frac{n}{1}$ and $\frac{n+1}{1}$ by semicircles centered on $\mathbb{R}$,
4. inductively, if p/q and r/s are neighbors joined by a semicircle, join p/q with $p/q \oplus r/s$ and $p/q \oplus r/s$ with r/s by semicircles centered on $\mathbb{R}$,
5. continue in this way.

Notice that the lines drawn in this procedure are all geodesics with respect to the hyperbolic metric. In this way, we subdivide $\mathbb{H}$ into infinitely many *ideal triangles* Δ; that is, hyperbolic triangles whose vertices lie in the boundary $\partial\mathbb{H} = \mathbb{R} \cup \{\infty\}$ of $\mathbb{H}$. We call $\mathcal{T}$ the resulting subdivision. Alternatively, one can obtain $\mathcal{T}$ by drawing the vertical line through 0 and the joining adjacent points in each Farey Sequence $\mathcal{F}_n$ by a hyperbolic geodesic (with $-\infty$ identified with ∞).

We now show that $\mathcal{T}$ is a tessellation of $\mathbb{H}$, namely we show that $\mathcal{T}$ is obtained as the images of a single ideal triangle Δ_e under a group of symmetries.

Proposition 6.2. *Let Δ_e be the ideal triangle with vertices $0, 1, \infty$. For all $g \in \mathrm{PSL}(2,\mathbb{Z})$, the triangle $g(\Delta_e)$ is an element of $\mathcal{T}$. Moreover, the triangles in $\mathcal{T}$ cover $\mathbb{H}$ without overlaps (except at their boundaries).*

Proof. Let us show that all triangles in $\mathcal{T}$ are the image of Δ_e under some element of $\mathrm{PSL}(2,\mathbb{Z})$. If $p/q > r/s$ are joined by an arc of $\mathcal{T}$, then, by construction, they are neighbors in some Farey Sequence $\mathcal{F}_n$. Therefore, $\det\left(\begin{smallmatrix} p & r \\ q & s \end{smallmatrix}\right) = 1$, so that $g = \left(\begin{smallmatrix} p & r \\ q & s \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$. Since g is a hyperbolic isometry, it maps the triangle Δ_e into the hyperbolic triangle Δ_g with vertices $g.0$, $g.1$ and $g.\infty$. These are r/s, $p/q \oplus r/s$ and p/q respectively, hence our claim follows.

Let us now verify the second claim. Clearly, all points in $\mathbb{H}$ belong to some triangle in $\mathcal{T}$. Since any triangle in $\mathcal{T}$ is $\Delta_g = g(\Delta_e)$ for some $g \in \mathrm{PSL}(2,\mathbb{Z})$, in order to check that elements of $\mathcal{T}$ do not overlap it is enough to show that the interior of Δ_e and of $g(\Delta_e)$ do not intersect when $g \neq e$. Let us first note that the matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right)$ maps Δ to itself, with the effect of rotating its vertices.

Let $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$, with $a/c > b/d$. By translating and rotating Δ_e if needed, that is, up to applying matrices of the form $\left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right)$, we may assume that the side of $g(\Delta_e)$ from a/c to b/d cuts the imaginary axis. This implies that $a/c > 0 > b/d$, in particular ad and bc have opposite signs. Since they are all not 0, we would obtain $1 = |ad - bc| = |ad| + |bc| \geq 2$, which is a contradiction. $\square$

We have shown that the Farey Tessellation $\mathcal{T}$ is invariant under the action of $\mathrm{PSL}(2,\mathbb{Z})$. The following exercise provides some additional properties of $\mathcal{T}$.

Exercise 6.3. (a) *If $p/q, r/s \in \mathbb{Q}$ are neighbors, then they are vertices of some triangle in $\mathcal{T}$.*

(b) *Every $p/q \in \mathbb{Q}$ is a vertex of a triangle in $\mathcal{T}$.*

In §4.2.3, we have introduced the Modular Surface $S = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathbb{H}$ and its unit tangent bundle $M = \mathrm{PSL}(2,\mathbb{Z})\backslash\mathrm{PSL}(2,\mathbb{R})$. We showed that the set $T^1(E)$, where

$$E = \{z \in \mathbb{H} : |z| \geq 1,\ |\Re z| \leq 1/2\},$$

is a fundamental domain for the action of $\mathrm{PSL}(2,\mathbb{Z})$. Let us cut E is half at the imaginary axis, and let us move the left half over using the transformation $z \mapsto z+1$ and glue the two halves together. This cut-and-paste operation give us a new fundamental domain F, which is a quadrilateral with vertices i, $(1+i\sqrt{3})/2$, $1+i$, and ∞. Once can write

$$\Delta_e = F \cup g_0(F) \cup g_0^2(F), \quad \text{where} \quad g_0 = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}$$

is a hyperbolic isometry of order 3 (i.e., $g_0^3 = e$) that stabilizes Δ.

6.3 Cutting sequences and continued fractions

Let γ be an oriented geodesic on $T^1\mathbb{H} = \mathrm{PSL}(2,\mathbb{R})$ passing through some element $g \in \mathrm{PSL}(2,\mathbb{R})$, namely let $\gamma = \gamma_g(t)$ be of the form $\gamma_g(t) = g\exp(t\mathbf{x})$. For any $\Delta \in \mathcal{T}$, let $s_\Delta = \gamma \cap \Delta$ be the oriented segment of geodesic contained in Δ. Then, s_Δ cuts two sides of Δ that meet in a vertex. We say that s_Δ has type L if this latter vertex lies on the left of s_Δ, and we say s_Δ has type R if the vertex lies on its right. There is a little ambiguity if the segment s_Δ ends in a vertex of Δ; in which case the type could be either L or R; we will come back to this in a few moments.

Any geodesic $\gamma = \gamma_g(t)$ as above cuts a sequence of triangles $\Delta_1, \Delta_2, \dots$ when moving along future times $t \geq 0$, and similarly its "past" cuts a sequence of triangles $\dots, \Delta_{-1}, \Delta_0$. We define the *cutting sequence of* γ to be the sequence of symbols

$$\cdots L^{n_{-2}} R^{n_{-1}}\, g\, L^{n_0} R^{n_1} \cdots,$$

where the i-th symbol is L or R if the type of the segment $s_{\Delta_i} = \gamma \cap \Delta_i$ is L or R respectively. The symbol g in the sequence is there only to denote the starting point of the geodesic. In other words, the cutting sequence above means that the future geodesic from g intersects n_0 triangles of $\mathcal{T}$ in segments of type L, then n_1 triangles in segments of type R, and so on. A similar description holds for the past of the geodesic, ending in g.

For any geodesic γ as above, let

$$\gamma_{\pm\infty} := \lim_{t\to\pm\infty} \gamma_g(t) \in \mathbb{R} \cup \{\infty\}$$

denote the endpoints of γ. The cutting sequence of γ is finite on the right if and only if $\gamma_{+\infty}$ a vertex of a triangle of $\mathcal{T}$, in which case the cutting sequence could end with either L^{n_j+1} or with $L^{n_j}R$, or similarly with either R^{n_j+1} or with $R^{n_j}L$. Analogously, it is finite on the left if $\gamma_{-\infty}$ is a vertex of a triangle in $\mathcal{T}$.

By Proposition 6.2, any element $\ell \in \Gamma = \mathrm{PSL}(2,\mathbb{Z})$ preserves $\mathcal{T}$ and moreover ℓ preserve orientation. Therefore, the cutting sequence of the geodesic $\gamma_g(t)$ must coincide with the cutting sequence of the geodesic $\gamma_{\ell(g)}(t)$. This implies that, if $\overline{\gamma}$ is a geodesic on the Modular Surface S, then its cutting sequence is well-defined: the cutting sequence of any two lifts of $\overline{\gamma}$ to $T^1\mathbb{H}$ must coincide, up to a choice of the starting point.

Exercise 6.4. *Let $\overline{\gamma}$ be a geodesic in S. Show that there exists a lift $\gamma = \gamma_g(t)$ on $tT^1\mathbb{H}$ that satisfies the following properties*

(A) *the starting point g is on the imaginary axis (i.e., $g.i \in i\mathbb{R}_{>0}$), $0 < |\gamma_{-\infty}| \leq 1$, $|\gamma_{+\infty}| \geq 1$, and $\gamma_{-\infty}$ and $\gamma_{+\infty}$ have opposite signs.*

The following result makes precise the connection between geodesics on M and continued fractions.

Theorem 6.5. *Let $\overline{\gamma}$ be a geodesic on $M = T^1(S)$, and let $\gamma_g(t)$ be a lift of $\overline{\gamma}$ which satisfies* (A).

1. *If $\cdots L^{n_{-2}} R^{n_{-1}} g L^{n_0} R^{n_1} \cdots$ is the cutting sequence of $\gamma_g(t)$, then*

$$\gamma_{+\infty} = [n_0; n_1, \ldots], \qquad \textit{and} \qquad \frac{-1}{\gamma_{-\infty}} = [n_{-1}; n_{-2}, \ldots].$$

2. *If $\cdots R^{n_{-2}} L^{n_{-1}} g R^{n_0} L^{n_1} \cdots$ is the cutting sequence of $\gamma_g(t)$, then*

$$\gamma_{+\infty} = -[n_0; n_1, \ldots], \qquad \textit{and} \qquad \frac{1}{\gamma_{-\infty}} = [n_{-1}; n_{-2}, \ldots].$$

Proof. Let us assume that we are in the first case, namely the cutting sequence of $\gamma_g(t)$ is $\cdots L^{n_{-2}} R^{n_{-1}} g L^{n_0} R^{n_1} \cdots$; the other case is analogous. We write $\gamma_{+\infty} = [a_0; a_1, a_2, \ldots]$, and we want to show that $a_i = n_i$.

Since the cutting sequence after $g.i \in i\mathbb{R}_{>0}$ starts with $L^{n_0} R$, this implies that $n_0 < \gamma_{+\infty} \leq n_0 + 1$, and the equality $\gamma_{+\infty} = n_0 + 1$ holds if and only if the cutting sequence stops. In this latter case, the claim is proved, hence let us assume it does not terminate at this point. Therefore, we get $n_0 = \lfloor \gamma_{+\infty} \rfloor = a_0$.

Call x_1 the point on γ that lies on the boundary of the triangle $\Delta_1 := \tau^{n_0}\Delta_e$ and where the change from L to R in the cutting sequence occurs. We now make use of the distinguished elements $\tau = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\sigma = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ of $\mathrm{SL}(2,\mathbb{Z})$, as we denoted them in §4.2.3. By definition, the element $\tau^{-n_0} x_1$ belongs to the imaginary axis. The geodesic $\tau^{-n_0}(\gamma)$ still projects onto $\overline{\gamma}$ and, up to a time reparametrization, can be identified with the geodesic $\gamma_{\tau^{-n_0} x_1}(t)$ starting at $\tau^{-n_0} x_1$ and ending at $\gamma_{+\infty} - n_0 \in (0,1)$. The cutting sequence of this new lift of $\overline{\gamma}$ starts with $\tau^{-n_0} x_1 R^{n_1} \cdots$.

We now apply σ. Call $g_1 := \sigma\tau^{-n_0} x_1$ and $\gamma_1 = \sigma\tau^{-n_0}(\gamma)$ the new geodesic (which, again, is a lift of $\overline{\gamma}$). Since the imaginary axis is fixed by σ, the starting point g_1 still belongs to $i\mathbb{R}_{>0}$; on the other hand, the endpoint of γ_1 is now

$$\sigma\tau^{-n_0}(\gamma_{+\infty}) = \sigma(\gamma_{+\infty} - n_0) = -1/(\gamma_{+\infty} - n_0) < -1.$$

Similarly, the starting point $\sigma\tau^{-n_0}(\gamma_{-\infty}) \in (0,1)$. This implies that γ_1 is a lift of $\overline{\gamma}$ which satisfies (A). Since the cutting sequence of γ_1 starts with $R^{n_1} L$, as before, we deduce that $-n_1 - 1 \leq \sigma\tau^{-n_0}(\gamma_{+\infty}) < -n_1$, with equality if and only if the cutting sequence terminates (in which case the claim is proved). If it does not terminate, we deduce

$$n_1 = \lfloor -\sigma\tau^{-n_0}(\gamma_{+\infty}) = 1/(\gamma_{+\infty} - a_0) = a_1.$$

At this point, we apply $\sigma\tau^{n_1}$ to γ_1, and we obtain a new geodesic $\gamma_2 = \sigma\tau^{n_1}\gamma_1 = \sigma\tau^{n_1}\sigma\tau^{-n_0}\gamma$, which projects onto $\overline{\gamma}$, and satisfies (A) with endpoint $\sigma\tau^{n_1}\sigma\tau^{-n_0}\gamma_{+\infty} \geq 1$ and starting point $\sigma\tau^{n_1}\sigma\tau^{-n_0}\gamma_{-\infty} \in [-1,0)$. The argument repeats, and hence the claim is proved for $\gamma_{+\infty}$.

To study the starting point $\gamma_{-\infty}$, one applies the map σ and considers the geodesic $\sigma(\gamma)$ parametrized with negative times. In this way, the starting point is $\sigma(\gamma_{+\infty}$, the endpoint is $\sigma(\gamma_{-\infty}) = -1/\gamma_{-\infty}$, and the cutting sequence is $\cdots R^{n_1} L^{n_0} \sigma(g) R^{n_{-1}} L^{n_{-2}} \cdots$. The first part of the argument applies and yields the result $-1/\gamma_{-\infty} = [n_{-1}; n_{-2}, \ldots]$. □

Remark 6.6. *1. The endpoint $\gamma_{+\infty}$ is independent of $\gamma_{-\infty}$ and on the part of the cutting sequence that precedes g, and vice-versa.*

2. The ambiguity in the cutting sequence that arises if the geodesic ends in a vertex of a triangle of $\mathcal{T}$ corresponds to the ambiguity in the continued fraction expansion of rational numbers that we mentioned in §6.1.

The remarkable connection described in Theorem 6.5 betweend geodesics on M and continued fractions helps in both ways: on one hand, one could use properties of continued fractions to deduce properties of geodesics on M; on the other hand, it is possible to recover simple and streamlined proofs of facts on continued fractions from geometric considerations or from facts on the geodesic flow. We give here a couple of simple examples of this phenomenon, but the story goes on well beyond what we mention here. The interested reader can find more material for example in [17] and in further, more recent works.

We can use the properties of the golden mean $(1+\sqrt{5})/2 = [1;1,1,\dots]$ to construct a periodic geodesic on M. More in general, as the next lemma shows, any real number with a periodic continued fraction expansion gives rise to a periodic geodesic on M.

Lemma 6.7. *Let $\alpha > 1$ be a real number with a periodic continued fraction expansion. Then, the geodesic γ in $\mathbb{T}^1\mathbb{H}$ with endpoints $\gamma_{+\infty} = \alpha$ and $\gamma_{-\infty} = -1/\alpha$ projects onto a periodic geodesic on M.*

Proof. Up to repeating the period, we can write $\alpha = [\overline{a_0; a_1, \dots, a_{2r+1}}]$ be a periodic continued fraction expansion of α with an even period. Consider the geodesic γ on M with endpoints $\gamma_{+\infty} = \alpha$ and $\gamma_{-\infty} = \beta$, where $-1/\beta = [\overline{a_{2r+1}; a_{2r}, \dots, a_{-1}}]$. Then, γ has a periodic cutting sequence

$$\cdots L^{a_{2r}} R^{a_{2r+1}} g L^{a_0} R^{a_1} \cdots L^{a_{2r}} R^{a_{2r+1}} L^{a_0} \cdots$$

which is fixed by the element

$$\ell := \sigma\tau^{a_{2r+1}}\sigma\tau^{-a_{2r}} \cdots \sigma\tau^{a_1}\sigma\tau^{-a_0} \in \mathrm{SL}(2,\mathbb{Z}).$$

This implies that the geodesic γ is fixed by ℓ, i.e., $\ell(\gamma) = \gamma$. It follows that the projection on M is periodic. □

On the other hand, we can give a simple proof of the following fact on continued fractions using what we have seen so far. We say that two numbers $\alpha, \beta \in \mathbb{R}$, with $\alpha = [a_0; a_1, \dots]$ and $\beta = [b_0; b_1, \dots]$, *have the same tail* if there exists $k, l \in \mathbb{N}$ such that $a_{k+n} = b_{l+n}$ for all $n \in \mathbb{N}$; we say that *they have the same tails* mod 2 if $k+l$ is even.

Lemma 6.8. *Two reals α and β have the same tails* mod 2 *if and only if there exists $\ell \in \mathrm{SL}(2,\mathbb{Z})$ such that $\ell.\alpha = \beta$.*

Proof. We leave as an exercise to the reader to show the left-to-right implication.

We start by verifying the following claim: if two geodesic have the same endpoint, then their cutting sequences eventually coincide. Indeed, if two geodesic have the same endpoint, we can find a side of a triangle in $\mathcal{T}$ which is cut by both of them. By applying an element of $\mathrm{SL}(2,\mathbb{Z})$, we can map that side to the imaginary axis, hence getting two geodesics starting on the imaginary axis with the same endpoint. Applying this element of $\mathrm{SL}(2,\mathbb{Z})$ had the effect of shifting the cutting sequences of the original geodesics. On the other hand, these two new geodesics, by Theorem 6.5, have the same future cutting sequence, hence the claim is proved.

Let us assume that there exists $\ell \in \mathrm{SL}(2,\mathbb{Z})$ such that $\ell.\alpha = \beta$. Up to applying σ and some appropriate power of τ, we can assume that both $\alpha, \beta > 1$. Choose $\omega \in (-1,1)$, and consider

the geodesics γ_α, γ_β with starting from ω and with endpoints α and β respectively. Now, γ_α and $\ell(\gamma_\alpha)$ have the same cutting sequence, up to a shift in the symbols (that is, up to choosing the starting point). Moreover, by the previous claim, the cutting sequences of $\ell(\gamma_\alpha)$ and γ_β eventually coincide, since they have the same endpoint. Moreover, they must have the same tails mod 2 since the alternating symbols L and R have to match. □

Exercise 6.9. *Prove that, if $\alpha, \beta \in \mathbb{R}$ have the same tails* mod 2*, then there exists $\ell \in \mathrm{SL}(2,\mathbb{Z})$ such that $\ell.\alpha = \beta$.*

Let us finish this chapter by giving a geometric interpretations of the sequence of convergents of a real number α.

Lemma 6.10. *Let $s_0, s_1, \dots$ be the sides of the triangles in $\mathcal{T}$ which mark the changes in the cutting sequence of $\gamma_g(t)$ for $t > 0$ from L to R and vice-versa. Then, for $n \geq 0$, the endpoints of s_n are p_n/q_n and p_{n-1}/q_{n-1}.*

Proof. Let $\cdots L^{n_{-2}} R^{n_{-1}} g L^{n_0} R^{n_1} \cdots$ be the cutting sequence of $\gamma_g(t)$, and let us look at positive times $t > 0$. As in the proof of Theorem 6.5, the endpoints of the side s_0 are a_0 and ∞. Define, as in Proposition 6.1, $p_{-1} = 1$ and $q_{-1} = 0$, so that $\infty = p_{-1}/q_{-1}$ and $a_0 = p_0/q_0$. After the side s_0, there are a_1 segments of type R. Thus, the left endpoint of s_1 is still $a_0 = p_0/q_0$; however, to find the right endpoint, we need to perform the Farey addition n_1 times, and hence we find

$$\infty \oplus p_0/q_0 \oplus \cdots \oplus p_0/q_0 = \frac{p_{-1} + n_1 p_0}{q_{-1} + n_1 q_0} = \frac{p_1}{q_1},$$

where we have used Proposition 6.1-(2). We now continue in the same way: the right endpoint of s_2 is the same as s_1, namely $\frac{p_1}{q_1}$, whereas to find the left endpoint we need to perform n_2 times the Farey addition, since there are n_2 segments labeled L. We then deduce that the left endpoint of s_2 is

$$\frac{p_0}{q_0} \oplus \frac{p_1}{q_1} \oplus \cdots \oplus \frac{p_1}{q_1} = \frac{p_0 + n_2 p_1}{q_0 + n_2 q_1} = \frac{p_2}{q_2}.$$

The proof carries on for all $n \geq 1$. □

Chapter 7

Further topics

7.1 Invariant measures and Ratner's Theorems

7.1.1 The horocycle flow on compact quotients

7.1.2 Ratner's Theorem on measure classification

7.2 Quantitative properties

7.2.1 Quantitative mixing

7.2.2 Quantitative unique ergodicity

Bibliography

[1] E. Artin. *Ein Mechanisches System mit quasi-ergodischen Bahnen* Collected Papers, Addison Wesley, Reading, Mass., 1965.

[2] L. Corwin, F.P. Greenleaf. *Representations of nilpotent Lie groups and their applications. Part I: Basic theory and examples.* Cambridge Studies in Advanced Mathematics **18**, Cambridge University Press, Cambridge, UK, 1990.

[3] Y. Coudène. *A short proof of the unique ergodicity of horocyclic flows.* Contemporary Mathematics, Amer. Math. Soc., **485**:85–89, 2009.

[4] M. Einsiedler, T. Ward. *Ergodic Theory with a view towards Number Theory.* Graduate Texts in Mathematics 259, Springer, London, 2011.

[5] H. Fiedler, J. Jurkat, O.K.. Körner *Asymptotic expansions of finite theta series.* Acta Arith., **32**(2):129–146, 1977.

[6] L. Flaminio, G. Forni, *Invariant distributions and time averages for horocycle flows.* Duke Math. J., **119**(3):465–536, 2003.

[7] L. Flaminio, G. Forni. *Equidistribution of nilflows and applications to theta sums.* Ergod. Th. Dyn. Syst., **26**:409–433, 2006.

[8] H. Furstenberg. *The unique ergodicity of the horocycle flow.* In: "Recent Advances in Topological Dynamics" (New Haven, Conn., 1972), Lecture Notes in Math. 318, Springer, Berlin, 95–115, 1973.

[9] G.H. Hardy, J.E. Littlewood. *The trigonometrical series associated with elliptic theta-functions.* Acta Math., **37**:193–238, 1914.

[10] G.H. Hardy, J.E. Littlewood. *A further note on the trigonometrical series associated with the elliptic theta-functions.* Proc. Cambridge Philos. Soc., **21**:1–5, 1923.

[11] A.Ya. Khinchin *Continued Fractions.* University of Chicago Press, Chicago, 1935.

[12] A.W. Knapp *Lie Groups Beyond an Introduction.* Birkhäuser, Boston, 1996.

[13] J.M. Lee *Introduction to Smooth Manifolds.* Graduate Texts in Mathematics 218, Springer, New York, NY, 2012.

[14] M. Ratner, *The rate of mixing for geodesic and horocycle flows.* Ergodic Theory Dynam. Systems **7**:267–288, 1987.

[15] M. Ratner, *On Raghunathan's measure conjecture.* Ann. Math **134**(3):545–607, 1991.

[16] D. Ravotti. *Asymptotics and limit theorems for horocycle ergodic integrals à la Ratner.* Preprint arXiv:2107.02090, 2021.

[17] C. Series, *The Modular Surface and continued fractions.* J. London Math. Soc. **31**:69–80, 1985.