RATNER'S THEOREM ON $\mathrm{SL}(2,\mathbb{R})$-INVARIANT MEASURES

MANFRED EINSIEDLER

Abstract. We give a relatively short and self contained proof of Ratner's theorem in the special case of $\mathrm{SL}(2,\mathbb{R})$-invariant measures on $\Gamma\backslash G$.

1. Introduction

M. Ratner proved in a series of papers [34, 33, 28, 29, 35] very strong results on invariant measures and orbit closures for certain subgroups H of a Lie group G — where H acts on the right of $X = \Gamma\backslash G$ and $\Gamma < G$ is a lattice. More concretely, H needs to be generated by one parameter unipotent subgroups, and the statements all are of the form that invariant measures and orbit closures are always algebraic as conjectured earlier by Raghunathan. Today these theorems are applied in many different areas of mathematics. We will motivate these questions and study some more concrete cases in Section 2.

While there are some very special cases for which the proof simplifies [36, 26], the general proof requires a deep understanding of the structure of Lie groups and ergodic theory. The aim of this paper is to give a self-contained more accessible proof of the classification of invariant and ergodic measures for subgroups H isomorphic to $\mathrm{SL}(2,\mathbb{R})$. While this is as well a special case of Ratner's theorem, it is a rich class since G can be much larger than H. Moreover, the proof for this class is more accessible in terms of its requirements. The methods used are not new, in particular most appear also in earlier work [32, 33] of Ratner, but it does not seem to be known that the following theorem allows also a relatively simple and short proof.

Date: May 19, 2006.

2000 *Mathematics Subject Classification.* 37D40, 37A17, 22E40.

Key words and phrases. Ratner's theorem, invariant measures, measure rigidity.

The author acknowledges support of NSF Grant 0509350. This research was partially conducted while the author was employed by the Clay Mathematics Institute as a Research Scholar.

Theorem 1.1. *Let G be a Lie group, $\Gamma < G$ a discrete subgroup, and $H < G$ a subgroup isomorphic to* $\mathrm{SL}(2,\mathbb{R})$*. Then any H-invariant and ergodic probability measure μ on $X = \Gamma\backslash G$ is homogeneous, i.e. there exists a closed connected subgroup $L < G$ containing H such that μ is L-invariant and some $x_0 \in X$ such that the L-orbit x_0L is closed and supports μ. In other words μ is an L-invariant volume measure on x_0L.*

As we will see a graduate student, who started to learn or is willing to learn the very basics of Lie groups and ergodic theory, should be able to follow the argument (no knowledge of radicals or structure theory of Lie groups and no knowledge of entropy is necessary). We will discuss the requirements in Section 3. The initiated reader will notice that this approach generalizes without too much work to other semisimple groups $H < G$ without compact factors. However, to keep the idea simple and to avoid unnecessary technicalities we only treat the above case.

In the next section we give some motivation for the above and related questions. In particular, we will discuss an application of Ratner's theorems where the above special case is sufficient. This paper is best described as an introduction to Ratner's theorem on invariant measures, and is not a comprehensive survey of this area of research. The reader seeking such a survey is referred to [17] and for some more recent developments to [8].

The author would like to thank M. Ratner for comments on an earlier draft of this paper.

2. Motivation[1]

2.1. **The geodesic flow and the horocycle flow on hyperbolic surfaces.** The hyperbolic plane $\mathbb{H}$ is defined by the upper half plane $\{z = x + iy \in \mathbb{C} : \mathrm{Im}(z) = y > 0\}$ together with the hyperbolic metric. The latter is easiest described locally: An element $v \in T_z\mathbb{H}$ of the tangent plane of the hyperbolic plane at z is simply an element of $\mathbb{R}^2$. Two elements $v, w \in T_z\mathbb{H}$ with the same base point $z = x + iy$ are given an inner product $(v, w)_z = \frac{1}{y^2}(v \cdot w)$ that depends on the base point z (actually just on the imaginary part y of z). Using this inner product we can define the length of a tangent vector $\|v\|_z = \sqrt{(v, v)_z}$ as usual, and define the length of a (differentiably parameterized) curve by integration (of the length of the derivative over the parameter). Finally

[1]The motivated reader who is also familiar with the setup is welcome to skip this section.

the hyperbolic distance between two points in the hyperbolic plane is then the infimum of the lengths of paths linking the two points.

To see the connection to Lie groups, we introduce an action of $\mathrm{SL}(2,\mathbb{R})$ on $\mathbb{H}$. An element $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ acts on z by a linear fractional transformation $g.z = \frac{az+b}{cz+d}$, also called a Möbius transformation. Note that this expression is always defined since $cz + d \neq 0$ for $\mathrm{Im}(z) > 0$. Moreover, an easy calculation shows that g maps $\mathbb{H}$ to itself and that it really gives an action of $\mathrm{SL}(2,\mathbb{R})$, i.e. that when we compose two such Möbius transformations we get the Möbius transformation for the product of the matrices. To understand the action better we ask what happens to the distance of two points in $\mathbb{H}$ when we apply g to both of them. However, distances of points we defined using the inner product on the tangent plane, so we should ask instead what happens to the inner product $(v, w)_z$ when we apply g. More precisely, we apply the derivative Dg to the tangent vectors v, w to get tangent vectors $Dg(v), Dg(w)$ that are based at the new point $g.z$. We claim that we actually have $(Dg(v), Dg(w))_{g.z} = (v, w)_z$, i.e. the inner product does not change (when we use the correct inner product at the points z and $g.z$) and g acts as an isometry on $\mathbb{H}$. This can be shown by a direct computation which can be simplified by studying only $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$ for all $b \in \mathbb{R}$, that together generate $\mathrm{SL}(2,\mathbb{R})$. (For the second type this is quite easy since the inner product only depends on y which is not changed by the corresponding transformation.)

The unit tangent bundle $T^1\mathbb{H}$ of the hyperbolic plane is defined as the collection of all vectors $v \in T_z\mathbb{H}$ of length one $\|v\|_z = 1$ for all possible base points $z \in \mathbb{H}$. By the above the derivative of the action of $g \in \mathrm{SL}(2,\mathbb{R})$ maps $T^1\mathbb{H}$ to itself, i.e. we have a new action of $\mathrm{SL}(2,\mathbb{R})$ which extends the previous one. Even more is true, the action is transitive: This means that for any two vectors $v, v' \in T^1\mathbb{H}$ there exists a $g \in \mathrm{SL}(2,\mathbb{R})$ for which Dg maps v to v'. Note that v and v' might have different base points z and z' and the above g has to map z to $g.z = z'$. To see transitivity, one should first find for any z an element g with $g.i = z$ — where we have started to use somewhat arbitrary the point i as a reference point. (Let us note that it suffices to consider one fixed reference point and one arbitrary point instead of two arbitrary points: If similarly $z' = g'.i$ then $z' = g'g^{-1}.z$.) We already saw above a one parameter family of matrices that just translate points to the left or right, and similarly one sees that diagonal matrices act by multiplication with a real number on $\mathbb{H}$. Together one can in fact reach any $z = g.i$ starting from i by using some $g \in \mathrm{SL}(2,\mathbb{R})$. To prove the original claim one has to be able to rotate a given vector based at i without moving the base point. And in fact one can verify that $\mathrm{SO}(2)$ maps i to i but rotates tangent vectors. To summarize, one can almost identify $T^1\mathbb{H}$ with $\mathrm{SL}(2,\mathbb{R})$ — only "almost" for two reasons: First one needs to pick somewhat arbitrary a reference vector in $T^1\mathbb{H}$ that will correspond to the identity in $\mathrm{SL}(2,\mathbb{R})$, we chose the unit vector pointing up at the base point i. Second the action of $\mathrm{SL}(2,\mathbb{R})$ has a kernel, the matrix $-\mathrm{I} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ acts as the identity on $T^1\mathbb{H}$ but this is the only such nontrivial element and we get an identification of $\mathrm{SL}(2,\mathbb{R})/\{\pm \mathrm{I}\}$ and $T^1\mathbb{H}$.

We can start to discuss the geodesic flow on $\mathbb{H}$. Let v be a unit vector based at a point z. Then the geodesic flow is defined by following the hyperbolic line — the geodesic — going through z in the direction of v for the given time t. Let us first consider our reference vector: the unit vector pointing upwards at i. For this vector the hyperbolic line is just the y-axis and applying the matrix $g_t = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$ moves the vector along the line at unit speed (again using the hyperbolic metric) — in other words we have the canonical parametrization of the geodesic line through our reference point, see Figure 1. When we apply now

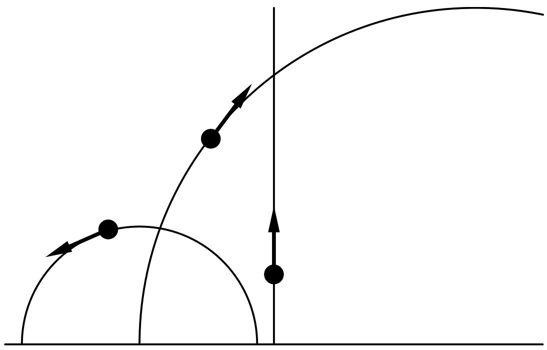

FIGURE 1. The geodesic lines determined by various vectors including the reference vector at i.

elements $g \in \mathrm{SL}(2,\mathbb{R})$ to this curve we will get the parametrization of any other geodesic line, most of these look like half circles that hit the real line in a normal angle, see Figure 1. Since we apply the isometry corresponding to g after applying the parameterization g_t, the geodesic flow corresponds to right multiplication by g_t on $\mathrm{SL}(2,\mathbb{R})/\{\pm \mathrm{I}\}$.

The horocycle flow h_t, vaguely speaking, is defined by the property that applied to any given vector it gives new vectors that have almost the same future for the geodesic flow as the original point — in other words the horocycle flow describes the stable manifold for the geodesic flow. More concretely, let us consider our reference vector at i and another vector pointing upwards at $x+i$. Both of the geodesics going through these vectors are straight parallel lines. From the Euclidean perspective the base points seem to stay away from each other as we move along the geodesic — after all the lines are parallel, see Figure 2. However, in the hyperbolic metric the two base points approach each

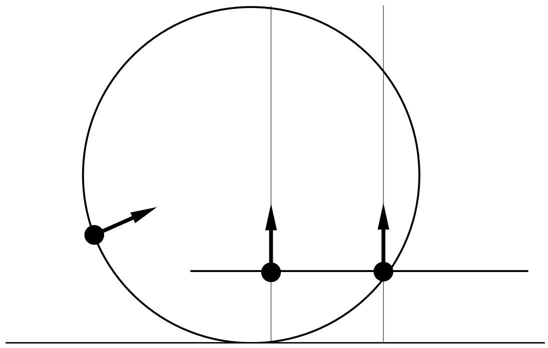

FIGURE 2. The two geodesics for two vectors pointing straight up approach each other in the hyperbolic metric. The horocycles are either lines parallel to the real axis or circles tangent to it.

other since we divide the Euclidean inner product by the square of the y-coordinate of the base points to get the hyperbolic inner product. In other words, the one-parameter subgroup consisting of the elements $h_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$ produces the points of the stable manifold through the reference vector. By the same reasoning as above, right multiplication by h_s on $\mathrm{SL}(2,\mathbb{R})/\{\pm \mathrm{I}\}$ defines the horocycle flow.

Dynamically the two flows defined above on $T^1\mathbb{H}$ are rather dull: Every orbit is a closed manifold escaping to infinity. To get a more interesting picture we need to take quotients by a discrete group of isometries. This should be thought of as an analogue to how one gets the circle $\mathbb{R}/\mathbb{Z}$ from $\mathbb{R}$, or the two-dimensional torus $\mathbb{R}^2/\mathbb{Z}^2$ from $\mathbb{R}^2$ using the discrete group of integer translations. The most natural discrete subgroup of $\mathrm{SL}(2,\mathbb{R})$ is $\mathrm{SL}(2,\mathbb{Z})$ which gives us the quotient $\mathrm{SL}(2,\mathbb{Z})\backslash\mathbb{H}$ of the hyperbolic plane and as well the quotient $\mathrm{SL}(2,\mathbb{Z})\backslash \mathrm{SL}(2,\mathbb{R})$. The former is a hyperbolic surface called the modular surface, and the latter we can (almost correctly) think of as the unit tangent bundle of the modular surface. In Figure 3 we see the fundamental domain for this

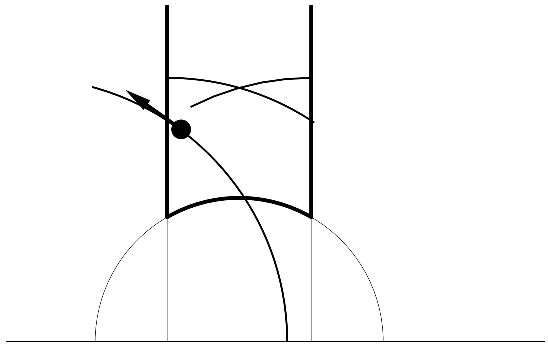

FIGURE 3. The fundamental domain of $\mathrm{SL}(2,\mathbb{R})$ acting on $\mathbb{H}$ (thick) is obtained from three bounding geodesics: two vertical lines at real part equal to $\pm 1/2$ and the unit circle. A typical geodesic produces a complicated picture if whenever it moves out of the fundamental domain we bring it back by one of the isometries from $\mathrm{SL}(2,\mathbb{Z})$.

discrete subgroup acting on $\mathbb{H}$ — the two left sides are being identified by the isometry defined by $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and the bottom part is being folded up to itself by the isometry defined by $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. The two flows we discussed before are still defined on the quotient $\mathrm{SL}(2,\mathbb{Z})\backslash \mathrm{SL}(2,\mathbb{R})$, in fact right multiplication by elements of $\mathrm{SL}(2,\mathbb{R})$ still makes sense on $\mathrm{SL}(2,\mathbb{Z})\backslash \mathrm{SL}(2,\mathbb{R})$. What we gained from this changed setup is that now we have interesting dynamics: The geodesics and horocycles of a given unit vector based at a point in the fundamental domain leave the fundamental domain. When this happens we need to apply the corresponding isometry to get back to the fundamental domain — doing this over and over produces the geodesics and horocycles inside the surface. This time the dynamics is interesting: There are dense geodesics and non-dense ones, there are dense horocycles and non-dense ones. We will be discussing the possibilities of such orbits and related question below after introducing a more general framework.

2.2. **A more general setup.** Let G be a closed linear group, i.e. a closed subgroup of $\mathrm{GL}(n,\mathbb{R})$, e.g. $G = \mathrm{SL}(n,\mathbb{R})$. Let $\Gamma < G$ be a discrete subgroup, e.g. $\Gamma = \mathrm{SL}(n,\mathbb{Z})$, so that $X = \Gamma\backslash G$ is a locally compact space with a natural G-action:

$$\text{for } g \in G, x \in X \text{ let } g.x = xg^{-1}.$$

Now let $H < G$ be a closed subgroup and restrict the above action to the subgroup H. The question about the properties of the resulting H-action has many connections to various a priori non-dynamical mathematical problems, and is from that point of view but also in its own light highly interesting. As we have discussed above in the case of $G = \mathrm{SL}(2,\mathbb{R})$ two different subgroups H give rise to the geodesic flow and the horocycle flow.

The most basic question, vaguely formulated, is how H-orbits $H.x_0$ for various points $x_0 \in X$ look like. Here one can make restrictions on x_0 or not, and ask, more precisely, either about the distribution properties of the orbit or about the nature of the closure of the orbit.

If X carries a G-invariant probability measure m_X, which in many situations is the case, then m_X is called the *Haar measure* of X and Γ is by definition a *lattice*. The best example for such a lattice is $\mathrm{SL}(n,\mathbb{Z}) < \mathrm{SL}(n,\mathbb{R})$. For $n = 2$ one can verify that the fundamental domain of Figure 3 has finite hyperbolic volume which is the reason for $\mathrm{SL}(2,\mathbb{Z}) < \mathrm{SL}(2,\mathbb{R})$ being a lattice.

If Γ is a lattice one can restrict our questions on the H-orbits to m_X-typical points and by doing so one has entered the realm of ergodic theory. Rephrased the basic question is now whether m_X is H-ergodic. Recall that by definition m_X is H-ergodic if every measurable f that is H-invariant is constant a.e. with respect to m_X, or equivalently that every measurable H-invariant set must be trivial in the sense that the set has either measure zero or measure one. Note that while H-invariance of m_X is inherited from G-invariance, the same is not true for H-ergodicity. The characterization of H-ergodicity in this context has been given in varying degrees of generality by many authors mostly before 1980, see [17, Chpt. 2] for a detailed account. The power of this characterization is that often — unless there are obvious reasons for failure of ergodicity — the Haar measure turns out to be H-ergodic. For example any non-compact closed subgroup $H < \mathrm{SL}(n,\mathbb{R})$ acts ergodically on $\mathrm{SL}(n,\mathbb{Z})\backslash \mathrm{SL}(n,\mathbb{R})$, and in particular the geodesic flow and the horocycle flow on $\mathrm{SL}(2,\mathbb{Z})\backslash \mathrm{SL}(2,\mathbb{R})$ are ergodic. Moreover, assuming ergodicity one of the most basic theorems in abstract ergodic theory, the ergodic theorem, states that a.e. point equidistributes in X. (For the notion of equidistribution we need that H is an amenable subgroup. The fact that $\mathrm{SL}(2,\mathbb{R})$ is not amenable actually makes it harder to apply Theorem 1.1 — and is the reason why the proof simplifies.) In particular, a.e. H-orbit is dense. This can be seen as the first answer to our original question.

Let us assume from now on that Γ is a lattice (some of what follows holds more generally for any discrete Γ but not all). The above discussion around H-ergodicity is only the first step in understanding the structure of H-orbits. In general, there is no reason to believe that a similarly simple answer is possible for all points of X. This is especially true for general dynamical systems but as we will discuss also in the algebraic setting we consider here. More surprisingly, there are cases where we can understand **all** H-orbits respectively believe that it is possible to understand **all** H-orbits. To be able to describe this we need to recall a few notions: $u \in G$ is *unipotent* if 1 is the only eigenvalue of the matrix u, $a \in G$ is $\mathbb{R}$-*diagonalizable* if it is diagonalizable as a matrix over $\mathbb{R}$. A subgroup $U < G$ is a *one-parameter unipotent subgroup* if U is the image of a homomorphism $t \in \mathbb{R} \mapsto u_t \in U$ with u_t unipotent for all $t \in \mathbb{R}$ — the prime example for such a group is the subgroup we encountered in the discussion of the horocycle flow. Another example in $G = \mathrm{SL}(n, \mathbb{R})$ is the subgroup

$$U = \begin{pmatrix} 1 & * & \cdots & * \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

which contains only unipotent elements and is generated by one-parameter unipotent subgroups. A subgroup $A < G$ is $\mathbb{R}$-*diagonalizable* if for some $g \in \mathrm{GL}(n, \mathbb{R})$ the conjugate gAg^{-1} is a subgroup of the diagonal subgroup. The prime example is the subgroup corresponding to the geodesic flow.

2.3. **Unipotently generated subgroups and Oppenheim's conjecture.** One case of subgroups we today understand quite well is when H is generated by one-parameter unipotent subgroups (as it is the case for $H = \mathrm{SL}(2, \mathbb{R})$). In the case of $G = \mathrm{SL}(2, \mathbb{R})$ we can think of the subgroup corresponding to the horocycle flow. As mentioned before we understand this case well thanks to the theorem of M. Ratner [29] which says that all H-orbits are well-behaved as conjectured by Raghunathan earlier. She proved that every H-orbit $H.x_0$ is dense in the closed orbit $L.x_0$ of a closed connected group $L > H$, and the latter orbit $L.x_0$ supports a finite L-invariant volume measure $m_{L.x_0}$. If H is itself a unipotent one-parameter subgroup the orbit $H.x_0$ is equidistributed with respect to this measure $m_{L.x_0}$. In the case of the horocycle flow on a compact quotient this result is due to Furstenberg [12] and says that all orbits are equidistributed with respect to the Haar measure of SL$(2,\mathbb{R})$ and in particular that all horocycles are dense. More generally these theorems were later extended by Ratner [30] herself, as well as by Margulis and Tomanov [22] to the setting of products of linear algebraic groups over various local fields (S-algebraic groups).

A bit more technical is the following question: What are the H-invariant probability measures? Here it suffices to restrict to H-invariant and ergodic measures — the general theorem of the ergodic decomposition states that any H-invariant probability measure can be obtained by averaging H-invariant and ergodic measures. Therefore, if we understand the latter measures we understand all of them. Ratner showed that all H-invariant and ergodic probability measures are of the form $m_{L.x_0}$ as discussed above — Theorem 1.1 is a special case of this. The special case of the horocycle flow on e.g. $\mathrm{SL}(2,\mathbb{Z})\backslash\mathrm{SL}(2,\mathbb{R})$ is due to Dani [2]. As it turns out this question is crucial for the proof of the topological theorem regarding orbit closures mentioned above. Namely using her theorem on measure classification Ratner then proves first that the orbit of a unipotent group always equidistributes with respect to an ergodic measure, and finally uses this to prove her topological theorem — this is similar to the earlier discussion of m_X-typical points. However, these steps are quite involved: First of all it is not clear that a limit measure coming from the orbit of a one-parameter unipotent subgroup is a probability measure since the space might not be compact. However, earlier work of Dani [1, 3, 4] (which extend work by Margulis [23]) shows precisely this. Then it is not clear why such a limit is ergodic and why it is independent of the times used in the converging subsequence — without going into details let us just say that the proof relies heavily on the structure of the ergodic measures and the properties of unipotent subgroup.

Before Ratner classified in her work all orbit closures Margulis used a special case of this to prove Oppenheim's conjecture, which by that time was a long standing open conjecture. This conjecture concerns the values of an indeterminate irrational quadratic form Q in n variables at the integer lattice $\mathbb{Z}^n$, and Margulis theorem [20] states that under these assumptions $Q(\mathbb{Z}^n)$ is dense in $\mathbb{R}$ if $n \geq 3$. (It is not hard to see that all of these assumptions including $n \geq 3$ are necessary for the density conclusion.) The proof consists of analyzing all possible orbits of $\mathrm{SO}(2,1)$ on $X_3 = \mathrm{SL}(3,\mathbb{Z})\backslash\mathrm{SL}(3,\mathbb{R})$. Even though $\mathrm{SO}(2,1)$ is essentially a quotient of $\mathrm{SL}(2,\mathbb{R})$ Theorem 1.1 does not imply immediately Oppenheim's conjecture since for the non-amenable group $\mathrm{SL}(2,\mathbb{R})$ it is not obvious how to find an invariant measure on the closure of an orbit.

2.4. **Diagonalizable subgroups.** The opposite extreme to unipotent elements are $\mathbb{R}$-diagonalizable elements, so it is natural to ask next about the case of a $\mathbb{R}$-diagonalizable subgroups $A < G$: What do the closures of A-orbits look like? What are the A-invariant and ergodic probability measures? As we will see this case is more difficult in various ways, in particular we do not have complete answers to these questions.

Rank one: Let $G = \mathrm{SL}(2, \mathbb{R})$, $\Gamma = \mathrm{SL}(2, \mathbb{Z})$, and set $X_2 = \Gamma\backslash G$. Let $A < \mathrm{SL}(2, \mathbb{R})$ be the diagonal subgroup. As discussed above the action of A on X_2 can also be described as the geodesic flow on the unit tangent bundle of the modular surface. Up to the fact that the underlying space X_2 is not compact this is a very good example of an Anosov flow. The corresponding theory can be used to show that there is a huge variety of orbit closures and A-invariant ergodic probability measures. So the answer is in that case that there is no simple answer to our questions — but at least we know that.

Higher rank: We replace '2' by '3' and encounter very different behaviour. Let $G = \mathrm{SL}(3, \mathbb{R})$, $\Gamma = \mathrm{SL}(3, \mathbb{Z})$, and set $X_3 = \Gamma\backslash G$. Let $A < \mathrm{SL}(3, \mathbb{R})$ be the diagonal subgroup, which this time up to finite index is isomorphic to $\mathbb{R}^2$. Margulis, Furstenberg, Katok, and Spatzier conjectured that for the action of A on X_3 there are very few A-invariant and ergodic probability measures, in particular that they again are all of the type $m_{L.x_0}$ for some $L > A$ and $x_0 \in X_3$ with closed orbit $L.x_0$. One motivation for that conjecture is Furstenberg's theorem [11] on $\times 2$, $\times 3$-invariant closed subsets of $\mathbb{R}/\mathbb{Z}$ which states that all such sets are either finite unions of rational points or the whole space. This can be seen as an abelian and topological analogue to the above problem.

For orbit closures the situation is a bit more complicated: G contains an isomorphic copy L of the subgroup $\mathrm{GL}(2, \mathbb{R})$ embedded into the upper left 2-by-2 block (where the lower right entry is used to fix the determinant). Now ΓL is a closed orbit of L and the A-action inside this orbit consists of the rank one action discussed above and one extra direction which moves everything towards infinity. (This behaviour of the L-orbit and the A-orbits needs of course justification, which in this case can be done algebraically.) Therefore, in terms of orbit closures the situation for A-orbits inside this L-orbit is as bad as for the corresponding action on X_2. However, it is possible to avoid this issue and to formulate a meaningful topological conjecture: Margulis conjectured that all bounded A-orbits on X_3 are compact, i.e. the only bounded orbits are periodic orbits for the A-action (which in fact all arise from a number theoretic construction).

Margulis [21] also noted that the question regarding orbits in this setting is related to a long standing conjecture by Littlewood. Littlewood conjectured around 1930 that for any two real numbers $\alpha, \beta \in \mathbb{R}$ the vector (α, β) is well approximable by rational vectors in the following multiplicative manner:

$$\liminf_{n\to\infty} n\|n\alpha\|\|n\beta\| = 0,$$

where $\|u\|$ denotes the distance of a real number $u \in \mathbb{R}$ to $\mathbb{Z}$. Here n is the common denominator of the components of the rational vector that approximates (α, β), and instead of taking the maximum of the differences along the x-axis and the y-axis we instead measure the quality of approximation by taking the product of the differences. The corresponding dynamical conjecture states that certain points (defined in terms of the vectors (α, β)) all have unbounded orbit (where actually only a quarter of the acting group A is used).

Building on earlier work of E. Lindenstrauss [18] and a joint work of the author with A. Katok [5] we have obtained together [7] a partial answer to the conjecture on A-invariant and ergodic probability measures: If the measure has in addition positive entropy for some element of the action, then it must be the Haar measure m_{X_3} — this generalizes earlier work of Katok, Spatzier, and Kalinin [15, 16, 14], and related work by Lyons [19], Rudolph [37], and Johnson [13] in the abelian setting of $\times 2, \times 3$. For Littlewood's conjecture we show also in [7] that the exceptions form at most a set of Hausdorff dimension zero. Roughly speaking the classification of all A-invariant probability measures with positive entropy can be used to show that very few A-orbits can stay within a compact subset of X_3, which by the mentioned dynamical formulation of Littlewood's conjecture is what is needed.

Most of the proof of this theorem consists in showing that positive entropy implies invariance of the measure μ under some subgroup $H < \mathrm{SL}(3, \mathbb{R})$ that is generated by one-parameter unipotent subgroups. Then one can apply Ratner's classification of invariant measures to the H-ergodic components of μ. However, in this case (unless we are in the easy case of $H = \mathrm{SL}(3, \mathbb{R})$) the subgroup H is actually isomorphic to $\mathrm{SL}(2, \mathbb{R})$. Therefore, Theorem 1.1 is sufficient to analyze the H-ergodic components.

The more general case of the maximal diagonal subgroup A acting on X_n for $n \geq 3$ is also treated in [7] (always assuming positive entropy). Even more generally one can ask about any $\mathbb{R}$-diagonalizable subgroup of an (algebraic) linear group. However, here there are unsolved technical difficulties that prevent so far a complete generalization. Joint ongoing work [9] of the author with E. Lindenstrauss solves these problems for maximally $\mathbb{R}$-diagonalizable subgroups (more technically speaking, for maximal $\mathbb{R}$-split tori A in algebraic groups G over $\mathbb{R}$ and similarly also for S-algebraic groups). This approach uses results from [6] and [10]. For a more complete overview of these results and related applications see the survey [8].

3. Ingredients of the proof

We list the facts and notions needed for the proof of Theorem 1.1, all of which, except for the last one, can be found in any introduction to Lie groups resp. ergodic theory.

3.1. The Lie group and its Lie algebra.

The Lie algebra $\mathfrak{g}$ of G is the tangent space to G at the identity element $e \in G$. The exponential map $\exp : \mathfrak{g} \to G$ and the locally defined inverse, the logarithm map, give local isomorphisms between $\mathfrak{g}$ and G. For any $g \in G$ the derivative of the conjugation map is the adjoint transformation $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$ and satisfies $\exp \mathrm{Ad}_g(v) = g \exp(v) g^{-1}$ for $g \in G$ and $v \in \mathfrak{g}$. For linear groups this could not be easier, the Lie algebra is a linear subspace of the space of matrices, $\exp(\cdot)$ and $\log(\cdot)$ are defined as usual by power series, and the adjoint transformation Ad_g is still conjugation by g.

Closed subgroups $L < G$ are almost completely described by their respective Lie algebras $\mathfrak{l}$ inside $\mathfrak{g}$ as follows. Let L° be the connected component of L (that contains the identity e). Then the Lie algebra $\mathfrak{l}$ of L (and L°) uniquely determines L° — L° is the subgroup generated by $\exp(\mathfrak{l})$. (Moreover, any element $\ell \in L$ sufficiently close to e is actually in L° and equals $\ell = \exp(v)$ for some small $v \in \mathfrak{l}$.)

Using an inner product on $\mathfrak{g}$ we can define a left invariant Riemannian metric $d(\cdot,\cdot)$ on G. We will be using the restriction of $d(\cdot,\cdot)$ to subgroups $L < G$ and denote by B_r^L the r-ball in L around $e \in L$.

If $\Gamma < G$ is a discrete subgroup, then $X = \Gamma\backslash G$ has a natural topology and in fact a metric defined by $d(\Gamma g, \Gamma h) = \min_{\gamma \in \Gamma} d(g, \gamma h)$ for any $g, h \in G$ (which uses left invariance of $d(\cdot,\cdot)$). With this metric and topology X can locally be described by G as follows. For any $x \in X$ there is an $r > 0$ such that the map $\imath : g \mapsto xg$ is an homeomorphism between B_r^G and a neighborhood of y. Moreover, if r is small enough $\imath : B_r^G \to X$ is in fact an isometric embedding. For a given x a number $r > 0$ with these properties we call an *injectivity radius at x*.

3.2. **Complete reducibility and the irreducible representations of** $\mathrm{SL}(2,\mathbb{R})$**.** The first property of $\mathrm{SL}(2,\mathbb{R})$ we will need is the following standard fact. Let V be a finite dimensional real vector space and suppose $\mathrm{SL}(2,\mathbb{R})$ acts on V. Then any $\mathrm{SL}(2,\mathbb{R})$-invariant subspace $W < V$ has an $\mathrm{SL}(2,\mathbb{R})$-invariant complement $W' < V$ with $V = W \oplus W'$.

The above implies that all finite dimensional representations of $\mathrm{SL}(2,\mathbb{R})$ can be written as a direct sum of irreducible representations. The second fact we need is the description of these irreducible representations. Let $A = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ denotes the standard basis of $\mathbb{R}^2$ so that $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} A = A$ and $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} B = B + tA$. Any irreducible representation is obtained as a symmetric tensor product $\mathrm{Sym}^n(\mathbb{R}^2)$ of the standard representation on $\mathbb{R}^2$ for some n. $\mathrm{Sym}^n(\mathbb{R}^2)$ has $A^n, A^{n-1}B, \ldots, B^n$ as a basis, and every element we can view as a homogeneous polynomial $p(A,B)$ of degree n. The action of $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ can now be described by substitution, $p(A,B)$ is mapped to $p(A, B+tA)$. More concretely, $p(A,B) = c_0A^n + c_1A^{n-1}B + \cdots + c_nB^n$ is mapped to

$$\begin{aligned} p(A, B+tA) = &(c_0 + c_1 t + \cdots + c_n t^n)A^n + \\ &(c_1 + \cdots + c_n n t^{n-1})A^{n-1}B + \\ &\cdots + c_n B^n, \end{aligned}$$

where the coefficients in front of the various powers of t are the original components of the vector $p(A,B)$ multiplied by binomial coefficients. Notice that all components of $p(A,B)$ appear in the image vector in the component corresponding to A^n. Moreover, for any component of $p(A,B)$ the highest power of t it gets multiplied by appears in the resulting component corresponding to A^n. For that reason, when t grows (and say $p(A,B)$ is not just a multiple of A^n) the image of $p(A,B)$ under $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ will always grow fastest in the direction of A^n when $t \to \infty$.

3.3. **Recurrence and the ergodic theorem.** Let (X,μ) be a probability space, and let $T : X \to X$ be measure preserving. Then for any set $B \subset X$ of positive measure and a.e. $x \in B$ there are infinitely many n with $T^n x \in B$ by Poincaré recurrence.

Now suppose $u_t : X \to X$ for $t \in \mathbb{R}$ is a one parameter flow acting on X such that μ is $u_{\mathbb{R}}$-invariant and ergodic. Then for any $f \in L^1(X,\mu)$

and μ-a.e. $x \in X$ we have

$$\frac{1}{T}\int_0^T f(u_t(x))\,\mathrm{d}t \to \int_X f\,\mathrm{d}\,\mu \text{ for } T \to \infty$$

This is Birkhoff's pointwise ergodic theorem for flows.

3.4. **Mautner's phenomenon for** $\mathrm{SL}(2,\mathbb{R})$**.** To be able to apply the ergodic theorem as stated in the last section in the proof of Theorem 1.1 we will need to know that the $\mathrm{SL}(2,\mathbb{R})$-invariant and ergodic probability measure is also ergodic under a one-parameter flow. The corresponding fact is best formulated in terms of unitary representations and is due to Moore [25] and is known as the *Mautner phenomenon*. For completeness we prove the special case needed.

Proposition 3.1. *Let $\mathfrak{H}$ be a Hilbert space, and suppose $\phi : \mathrm{SL}(2,\mathbb{R}) \to U(\mathfrak{H})$ is a continuous unitary representation on $\mathfrak{H}$. In other words, ϕ is a homomorphism into the group of unitary automorphisms $\mathbb{U}(\mathfrak{H})$ of $\mathfrak{H}$ such that for every $v \in \mathfrak{H}$ the vector $\phi(g)(v) \in \mathfrak{H}$ depends continuously on $g \in \mathrm{SL}(2,\mathbb{R})$. Then any vector $v \in \mathfrak{H}$ that is invariant under the upper unipotent matrix group $U = \left\{\begin{pmatrix} 1 & * \\ 0 & 1\end{pmatrix}\right\}$ is in fact invariant under $\mathrm{SL}(2,\mathbb{R})$.*

Since any measure preserving action on (X,μ) gives rise to a continuous unitary representation on $\mathfrak{H} = L^2(X,\mu)$ the above gives immediately what we need (see also [33, Prop. 5.2] for another elementary treatment):

Corollary 3.2. *Let μ be an H-invariant and ergodic probability measure on $X = \Gamma\backslash G$ with $\Gamma < G$ discrete, and $H < G$ isomorphic to $\mathrm{SL}(2,\mathbb{R})$. Then μ is also ergodic with respect to the one-parameter unipotent subgroup U of H corresponding to the upper unipotent subgroup in $\mathrm{SL}(2,\mathbb{R})$.*

In fact, an invariant function $f \in L^2(X,\mu)$ that is invariant under U must be invariant under $\mathrm{SL}(2,\mathbb{R})$ by Proposition 3.1. Since the latter group is assumed to be ergodic, the function must be constant as required. (We leave it to the reader to check the continuity requirement.)

Proof of Proposition 3.1. Following Margulis [24] we define the auxiliary function $p(g) = (\phi(g)v, v)$. Notice first that the function $p(\cdot)$ characterizes invariance in the sense that $p(g) = (v,v)$ implies $\phi(g)v = v$. By continuity of the representation $p(\cdot)$ is also continuous. Moreover, by our assumption on v the map $p(\cdot)$ is bi-U-invariant since

$$p(ugu') = (\phi(u)\phi(g)\phi(u')v, v) = (\phi(g)v, \phi(u^{-1})v) = p(g).$$

Let $\epsilon, r, s \in \mathbb{R}$ and calculate

$$\begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \epsilon & 1 \end{pmatrix} \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 + r\epsilon & r + s + rs\epsilon \\ \epsilon & 1 + s\epsilon \end{pmatrix}.$$

Now fix some $t \in \mathbb{R}$, let ϵ be close to zero but nonzero, choose $r = \frac{e^t - 1}{\epsilon}$ and $s = \frac{-r}{1+r\epsilon}$. Then the above matrix simplifies to

$$\begin{pmatrix} e^t & 0 \\ \epsilon & e^{-t} \end{pmatrix}$$

In particular, this shows that

$$p\left(\begin{pmatrix} 1 & 0 \\ \epsilon & 1 \end{pmatrix}\right) = p\left(\begin{pmatrix} e^t & 0 \\ \epsilon & e^{-t} \end{pmatrix}\right)$$

is both close to $p(e)$ and to

$$p\left(\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}\right).$$

Therefore, the latter equals (v, v) which implies that v is invariant under $\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$ as mentioned before.

The above implies now that $p(\cdot)$ is bi-invariant under the diagonal subgroup. Using this and the above argument once more, it follows that v is also invariant under $\begin{pmatrix} 1 & 0 \\ s & 1 \end{pmatrix}$ for all $s \in \mathbb{R}$. □

4. The proof of Theorem 1.1

In this section we prove Theorem 1.1 using the prerequisites discussed in the last section. Let us mention again that the general outline of the proof is very similar to the strategy M. Ratner [33] used to prove her theorems.

From now on let μ be an H-invariant and ergodic probability measure on $X = \Gamma\backslash G$.

4.1. **The goal and the first steps.**

It is easy to check that

$$\mathrm{Stab}(\mu) = \{g \in G : \text{ right multiplication with } g \text{ on } X \text{ preserves } \mu\}$$

is a closed subgroup of G. Let $L = \mathrm{Stab}(\mu)^\circ$ be the connected component. Then as discussed any element of $\mathrm{Stab}(\mu)$ sufficiently close to e belongs to L. Also since $\mathrm{SL}(2, \mathbb{R})$ is connected we have $H < L$.

We will show that μ is concentrated on **a single orbit** of L, i.e. that there is some L-orbit $L.x_0$ of measure one $\mu(L.x_0) = 1$. Then by L-invariance of μ and uniqueness of Haar measure, μ would have to be the L-invariant volume form on this orbit $L.x_0$. However, since μ

is assumed to be a probability measure this also implies that the orbit $L.x_0$ is closed as seen in the next lemma.

Lemma 4.1. *If μ is concentrated on a single L-orbit $L.x_0$ and is L-invariant, then $L.x_0$ is closed and μ is supported on $L.x_0$.*

In the course of the proof we will recall a few facts about Haar measures and also prove that a Lie group which admits a lattice is unimodular, i.e. satisfies that the Haar measure is left and right invariant. For a more comprehensive treatment of the relationship between lattices and Haar measures see [27].

Proof. Suppose $x_i \in \ell_i.x_0 \in L.x_0$ converges to y. We have to show that $y \in L.x_0$. Now either $x_i \in \overline{B_1^L}.y$ for some i — i.e. the convergence is along L and the lemma follows — or $x_i \notin \overline{B_1^L}.y$ for all i. In the latter case we may choose a subsequence so that $x_i \notin \overline{B_1^L}.x_j$ for $i \neq j$. Now let $r < 1$ be an injectivity radius of X at y. Then $x_i \in B_{r/2}.y$ for large enough i, say for $i \geq i_0$. It follows that the sets $B_{r/2}^L.x_i \subset X$ are disjoint for $i \geq i_0$. Since $x_i = \ell_i.x_0$ it follows that these sets are all of the form $B_{r/2}^L(\ell_i).x_0$. We claim that the existence of the finite volume orbit implies that L is unimodular. If this is so, then we see that the sets $B_{r/2}^L.x_i$ all have the same measure, which contradicts the finite volume assumption.

It remains to show that L is unimodular if the L-orbit of $x_0 = \Gamma g_0$ has finite L-invariant volume, or equivalently if $\Gamma_L = g_0^{-1}\Gamma g_0 \cap L < L$ is a lattice. So suppose μ is an L-invariant probability measure on $\Gamma_L \backslash L$, where L is acting on the right. Then the following gives the relationship between μ and a right Haar measure m_L on L. Let f be a compactly supported continuous function on L, then

$$\int f \,\mathrm{d}m_L = \int \sum_{\gamma \in \Gamma_L} f(\gamma\ell)\,\mathrm{d}\mu.$$

Here note first that the sum $F(\ell)$ inside the integral on the right is finite for every ℓ, satisfies $F(\gamma\ell) = F(\ell)$, and so defines a function on $\Gamma_L \backslash L$ which is also compactly supported and continuous. Therefore, the right integral is well-defined. Using invariance of μ and the uniqueness of the right Haar measure on L, the equation follows (after possibly rescaling m_L).

The above formula immediately implies that m_L is left-invariant under Γ_L. We use Poincaré recurrence to extend this to all of L. Let $K \subset L$ be a compact subset of positive measure. Let $\ell \in L$ be arbitrary and consider the map $T : \Gamma_L \backslash L \to \Gamma_L \backslash L$ defined by right multiplication by ℓ^{-1}. By assumption this preserves the probability measure μ, therefore there exists some fixed $k \in K$, infinitely many n_i, and $\gamma_i \in \Gamma_L$ such that $\gamma_i k \ell^{-n_i} \in K$. In other words there exist infinitely many $k_i \in K$ with $\ell^{n_i} = k_i^{-1}\gamma_i k$. The measure obtained by left multiplication by ℓ from a right Haar measure m_L is again a right Haar measure and so must be a multiple $c(\ell)m_L$. The constant $c(\ell) \in \mathbb{R}^\times$ defines a character, i.e. a continuous homomorphism $c : L \to \mathbb{R}^\times$. It follows that $c(\ell)^{n_i} = c(k_i)c(\gamma_i)c(k) = c(k_i)c(k)$ remains bounded as $n_i \to \infty$. Therefore, $c(\ell) = 1$. Since ℓ was arbitrary this proves the claim and the lemma. □

The main argument will be to show that if μ is not concentrated on a single orbit of L, then there are other elements of $\mathrm{Stab}(\mu)$ close to e. This shows that we should have started with a bigger subgroup L'. If we repeat the argument with this bigger L', we will either achieve our goal or make L' even bigger. We start by giving a local condition for a measure μ to be concentrated on a single orbit.

Lemma 4.2. *Suppose $x_0 \in X$ has the property that $\mu(B_\delta^L.x_0) > 0$ for some $\delta > 0$, then μ is concentrated on $L.x_0$. So either the conclusion of Theorem 1.1 holds for L and x_0, or for every x_0 we have $\mu(B_\delta^L.x_0) = 0$.*

Proof. This follows immediately from the definition of ergodicity of μ and the fact that $L.x_0$ is an H-invariant measurable set. □

We will be achieving the assumption to the last lemma by studying large sets $X' \subset X$ of points with good properties. Let $x_0 \in X'$ be such that all balls around x_0 have positive measure. Suppose X' has the property that points x' close to x_0 that also belong to X' give 'rise to additional invariance' of μ unless x and x' are locally on the same L-orbit (i.e. $x' = \ell.x$ for some $\ell \in L$ close to e). Then either L can be made bigger or $B_\delta^L.x \cap X' \subset B_\delta^L.x_0$ for some $\delta > 0$ and therefore the latter has positive measure. However, to carry that argument through requires a lot more work. We start by a less ambitious statement where two close by points in a special position from each other give 'rise to invariance' of μ. Recall that $U = \left\{\begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}\right\}$.

Proposition 4.3. *There is a set $X' \subset X$ of μ-measure one such that if $x, x' \in X'$ and $x' = c.x$ with*

$$c \in C_G(U) = \{g \in G : gu = ug \textit{ for all } u \in U\},$$

then c preserves μ.

The set X' in the above proposition we define to be the set of μ-generic points (for the one paramenter subgroup defined by U). A point $x \in X$ is *μ-generic* if

$$\frac{1}{T}\int_0^T f(u_t.x)\,\mathrm{d}t \to \int f\,\mathrm{d}\mu \text{ for } T \to \infty$$

for all compactly supported, continuous functions $f : X \to \mathbb{R}$. Recall that by the Mautner phenomenon μ is U-ergodic. Now the ergodic theorem implies that the set X' of all μ-generic points has measure one. (Here one first applies the ergodic theorem for a countable dense set of compactly supported, continuous functions and then extends the statement to all such functions by approximation.)

Proof. For $c \in C_G(U)$ and a compactly supported, continuous function $f : X \to \mathbb{R}$ define the function $f_c(x) = f(c.x)$ of the same type. Now assume $x, x' = c.x \in X'$ are μ-generic. Since $u_t c = c u_t$ we have $f(u_t.x') = f(cu_t.x) = f_c(u_t.x)$ and so the limits

$$\frac{1}{T}\int_0^T f(u_t.x')\,\mathrm{d}t \to \int f\,\mathrm{d}\mu \text{ and}$$
$$\frac{1}{T}\int_0^T f_c(u_t.x)\,\mathrm{d}t \to \int f_c\,\mathrm{d}\mu$$

are equal. However, the last integral equals $\int f_c\,\mathrm{d}\mu = \int f\,\mathrm{d}c_*\mu$ were $c_*\mu$ is the push forward of μ under c. Since f was any compactly supported, continuous function, $\mu = c_*\mu$ as claimed. □

4.2. **Outline of the H-principle.** In Proposition 4.3 we derived invariance of μ but only if we have two points $x, x' \in X'$ that are in a very special relationship to each other. On the other hand if μ is not supported on the single L-orbit, then we know that we can find many $y, y' \in X'$ that are close together but are not on the same L-leaf locally by Lemma 4.2. Without too much work we will see that we can assume

$$y' = \exp(v).y \text{ with } v \in \mathfrak{l}'$$

where $\mathfrak{l}'$ is an $\mathrm{SL}(2,\mathbb{R})$-invariant complement in $\mathfrak{g}$ of the Lie algebra $\mathfrak{l}$ of L, see Lemma 4.5. What we are going to describe is a version of the so-called H-principle as introduced by Ratner [31, 32] and generalized by Witte [38], see also [26].

By applying the same unipotent matrix $u \in U$ to y and y' we get

$$u.y' = (u\exp(v)u^{-1}).(u.y) = \exp(\mathrm{Ad}_u(v)).(u.y).$$

In other words, the divergence of the orbits through y and y' can be described by conjugation in G — or even by the adjoint representation on $\mathfrak{g}$. Since H is assumed to be isomorphic to $\mathrm{SL}(2,\mathbb{R})$ we will be able to use the theory on representations as in Section 3.2. In particular, recall that the fastest divergence is happening along a direction which is stabilized by U. Since all points on the orbit of a μ-generic point are also μ-generic, one could hope to flow along U until the two points $x = u.y, x' = u.x'$ differ significantly but not yet to much. Then $y' = u.x' = h.(u.x) = h.y$ with h almost in $C_G(U)$. To fix the 'almost' in this statement we will consider points that are even closer to each other, flow along U for a longer time, and get a sequence of pairs of μ-generic points that differ more and more by some element of $C_G(U)$. In the limit we hope to get two points that differ precisely by some element of $C_G(U)$ which is not in L.

The main problem is that limits of μ-generic points need not be μ-generic (even for actions of unipotent groups). Therefore, we need to introduce quite early in the argument a compact subset $K \subset X'$ of almost full measure that consists entirely of μ-generic points. When constructing $u.x', u.x$ we will make sure that they belong to K — this way we will be able to go to the limit and get μ-generic points that differ by some element of $C_G(U)$.

We are now ready to proceed more rigorously.

4.3. **Formal preparations, the sets K, X_1, and X_2.** Let X' be the sets of μ-generic points as above, and let $K \subset X'$ be compact with $\mu(K) > 0.9$. By the ergodic theorem

$$\frac{1}{T}\int_0^T 1_K(u_t.y)\,\mathrm{d}t \to \mu(K)$$

for μ-a.e. $y \in X$. In particular, we must have for a.e. $y \in X$

$$\frac{1}{T}\int_0^T 1_K(u_t.y)\,\mathrm{d}t > 0.8 \text{ for large enough } T.$$

Here T may depend on y but by choosing T_0 large enough we may assume that the set

$$X_1 = \left\{ y \in X : \frac{1}{T}\int_0^T 1_K(u_t.y)\,\mathrm{d}t > 0.8 \text{ for all } T \geq T_0 \right\}$$

has measure $\mu(X_1) > 0.99$. By definition points in X_1 visit K often enough so that we will be able to find for any $y, y' \in X_1$ many common values of t with $u_t.y, u_t.y' \in K$.

The last preparation we need will allow us to find $y, y' \in X_1$ that differ by some $\exp(v)$ with $v \in \mathfrak{l}'$. For this we define

$$X_2 = \left\{ z \in X : \frac{1}{m_L(B_1^L)}\int_{B_1^L} 1_{X_1}(\ell.z)\,\mathrm{d}m_L(\ell) > 0.9 \right\}$$

where m_L is a Haar measure on L. (Any other smooth measure would do here as well.)

Lemma 4.4. $\mu(X_2) > 0.9$.

Proof. Define $Y = X \times B_1^L$ and consider the product measure $\nu = \frac{1}{m_L(B_1^L)}\mu \times m_L$ on Y. The function $f(z,\ell) = 1_{X_1}(\ell.z)$ integrated over z gives independently of ℓ allways $\mu(X_1)$ since μ is L-invariant. By integrating over ℓ first we get therefore the function

$$g(z) = \frac{1}{m_L(B_1^L)} \int_{B_1^L} 1_{X_1}(\ell.z)\,\mathrm{d}m_L(\ell)$$

whose integral satisfies $\int g\,\mathrm{d}\mu = \mu(X_1)$. Since $g(z) \in [0,1]$ for all z

$$0.99 < \int_{X_2} g\,\mathrm{d}\mu + \int_{X\setminus X_2} g\,\mathrm{d}\mu \le \mu(X_2) + 0.9\mu(X\setminus X_2) = 0.1\mu(X_2) + 0.9$$

which implies the lemma. $\square$

Let as before $\mathfrak{l} \subset \mathfrak{g}$ be the Lie algebra of $L < G$ and let $\mathfrak{l}' \subset \mathfrak{g}$ be an $\mathrm{SL}(2,\mathbb{R})$-invariant complement of $\mathfrak{l}$ in $\mathfrak{g}$. Then the map $\phi : \mathfrak{l}' \times \mathfrak{l} \to G$ defined by $\phi(v,w) = \exp(v)\exp(w)$ is C^∞ and its derivative at $(0,0)$ is the embedding of $\mathfrak{l}' \times \mathfrak{l}$ into $\mathfrak{g}$. Therefore, ϕ is locally invertible so that every $g \in G$ close to e is a unique product $g = \exp(v)\ell$ for some $\ell \in L$ close to e and some small $v \in \mathfrak{l}'$. We define $\pi_L(g) = \ell$. For simplicity of notation we assume that this map is defined on an open set containing $\overline{B_1^L}$ (if necessary we rescale the metric).

Lemma 4.5. *For any $\epsilon > 0$ there exists $\delta > 0$ such that for $g \in B_\delta^G$, and $z, z' = g.z \in X_2$ there are $\ell_2 \in B_1^L$ and $\ell_1 \in B_\epsilon^L(\ell_2)$ with $\ell_1.z, \ell_2.z' \in X_1$ and $\ell_2 g \ell_1^{-1} = \exp(v)$ for some $v \in B_\epsilon^{\mathfrak{l}'}(0)$.*

Proof. Let $g \in B_\delta^G$ and consider the C^∞-function $\psi(\ell) = \pi_L(\ell g)$. If $g = e$ then ψ is the identity, therefore if δ is small enough the derivative of ψ is close to the identity and its Jacobian is close to one. In particular, we can ensure that $m_L(\psi(E')) > 0.9 m_L(E')$ for any measurable subset $E' \subset B_1^L$. Moreover, for δ small enough we have $\psi(\ell) \in B_\epsilon^L(\ell)$ for any $\ell \in B_1^L$ and so $\psi(B_1^L) \subset B_{1+\epsilon}^L$.

Now define the sets $E = \{\ell \in B_1^L : \ell.z \in X_1\}$ and $E' = \{\ell \in B_1^L : \ell.z' \in X_1\}$ which satisfy $m_L(E) > 0.9 m_L(B_1^L)$ and $m_L(E') > 0.9 m_L(B_1^L)$ by definition of X_2. We may assume that ϵ is small enough so that $m_L(B_{1+\epsilon}^L) < 1.1 m_L(B_1^L)$. Together with the above estimate we now have $m_L(\psi(E')) > 0.5 m_L(B_{1+\epsilon}^L)$ and $m_L(E) > 0.5 m_L(B_{1+\epsilon}^L)$. Therefore, there exists some $\ell_2 \in E'$ with $\ell_1 = \psi(\ell_2) \in E$. By definition of E, E' we have $\ell_1.y, \ell_2.y' \in X_1$. Finally, by definition of π_L

$$\ell_2 g \ell_1^{-1} = \ell_2 g \pi_L(\ell_2 g)^{-1} = \exp(v)$$

for some $v \in \mathfrak{l}'$. Again for sufficiently small δ we will have $v \in B_\epsilon^{\mathfrak{l}'}(0)$. □

4.4. **H-principle for** $\mathrm{SL}(2,\mathbb{R})$**.** Let $x_0 \in X_2 \cap \operatorname{supp} \mu|_{X_2}$ so that

$$\mu\big((B_\delta^G.x_0) \cap X_2\big) > 0 \text{ for all } \delta > 0.$$

Now one of the following two statements must hold:

(1) there exists some $\delta > 0$ such that $B_\delta^G.x_0 \cap X_2 \subset B_\delta^L.x_0$, or
(2) for all $\delta > 0$ we have $B_\delta^G.x_0 \cap X_2 \not\subset B_\delta^L.x_0$.

We claim that actually only (1) above is possible if L is really the connected component of $\mathrm{Stab}(\mu)$. Assuming this has been shown, then we have $\mu(B_\delta^L.x_0) > 0$ which was the assumption to Lemma 4.2. Therefore, $\mu(L.x_0) = 1$ and by Lemma 4.1 $L.x_0 \subset X$ is closed — Theorem 1.1 follows. So what we really have to show is that (2) implies that μ is invariant under a one parameter subgroup that does not belong to L.

Lemma 4.6. *Assuming (2) there are for every $\epsilon > 0$ two points $y, y' \in X_1$ with $d(y, y') < \epsilon$ and $y' = \exp(v).y$ for some nonzero $v \in B_\epsilon^{\mathfrak{l}'}(0)$.*

Proof. Let $z = x_0$. By (2) there exists a point $z' \in X_2$ with $d(z, z') < \delta$ and $z' \notin B_\delta^L.z$. Let $g \in B_\delta^G$ be such that $z' = g.z$ and $g \notin L$. Applying Lemma 4.5 the statement follows since $g \notin L$ and so $v \neq 0$ by our choice of z'. □

Using $y, y' \in X_1$ and $v \in \mathfrak{l}'$ for all $\epsilon > 0$ as in the above lemma we will show that μ is invariant under a one-parameter subgroup that does not belong to L. For this it is enough to show the following:

Claim: For any $\eta > 0$ there exists a nonzero $w \in B_\eta^{\mathfrak{l}'}(0)$ such that μ is invariant under $\exp(w)$.

To see that this is the remaining assertion, notice that we then also have invariance of μ under the subgroup $\exp(\mathbb{Z}w)$. While this subgroup could still be discrete, when $\eta \to 0$ we find by compactness of the unit ball in $\mathfrak{l}'$ a limiting one parameter subgroup $\exp(\mathbb{R}w)$ that leaves μ invariant.

We start proving the claim. Let $\eta > 0$ be fixed, and let $\epsilon > 0$, $y, y' \in X_1$, and $v \in B_\epsilon^{\mathfrak{l}'}(0)$ as above. We will think of ϵ as much smaller than η since we will below let ϵ shrink to zero while not changing η. Let $\mathrm{Sym}^n(\mathbb{R}^2)$ be an irreducible representation as in Section 3.2, and let $p = p(A, B) \in \mathrm{Sym}^n(\mathbb{R}^2)$. Recall that $u_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ applied to $p(A, B)$

gives $p(A, B+tA)$. We define

$$T_p = \frac{\eta}{\max(|c_1|, \dots, |c_n|^{1/n})}$$

and set $T_p = \infty$ if the expression on the right is not defined. The significance of T_p is that for $t = T_p$ at least one term in the sum $(c_0 + c_1 t + \dots + c_n t^n)$ is of absolute value one while all others are less than that — recall that this sum is the coefficient of A^n in $p(A, B+tA)$. To extend this definition to $\mathfrak{l}'$ which is not necessarily irreducible we split $\mathfrak{l}'$ into irreducible representations $\mathfrak{l}' = \bigoplus_{j=1}^k V_j$ and define for $v = (p_j)_{j=1,\dots,k}$

$$T_v = \min_j T_{p_j}.$$

Lemma 4.7. *There exists constants $n > 0$ and $C > 0$ that only depend on $\mathfrak{l}'$ such that for $v \in B_\epsilon^{\mathfrak{l}'}(0)$ and $t \in [0, T_v]$ we have*

$$\mathrm{Ad}_{u_t}(v) = w + O(\epsilon^{1/n})$$

where $w = w(t) \in B_{C\eta}^{\mathfrak{l}'}(0)$ is fixed under the subgroup $U = u_{\mathbb{R}}$. Here we write $O(\epsilon^{1/n})$ to indicate a vector in $\mathfrak{l}'$ of norm less than $C\epsilon^{1/n}$.

Proof. We first show the statement for irreducible representations $\mathrm{Sym}^n(\mathbb{R}^2)$ inside $\mathfrak{l}'$. There any multiple of A^n is fixed under U. Similar to the earlier discussion the coefficient $(c_0 + c_1 t + \dots + c_n t^n)$ is for $t \in [0, T_p]$ bounded by $n\eta$. For the other coefficients note first that these are sums of terms of the form $c_i t^j$ for $j < i$. Since $t < \eta |c_i|^{-1/i}$ we have that each such term is bounded by $|c_i t^j| \le |c_i||c_i|^{-j/i} \le |c_i|^{1/i} \ll \epsilon^{1/n}$ where the implied constant only depends on the norm on $\mathfrak{l}'$ and the way we split $\mathfrak{l}'$ into irreducibles (and we assumed $\eta < 1$). This proves the lemma for irreducible representations.

The general case is now straightforward. If $t < T_v = \min_j T_{p_j}$ then since $\mathrm{Ad}_{u_t}(p_j)$ is of the required form for $j = 1, \dots, k$ the lemma follows by taking sums. □

If v is already fixed by U then $T_v = \infty$ (and other way around) and the above statement is rather trivial since $w = v$. Moreover, by definition of X_1 we have $\frac{1}{T}\int_0^T 1_K(u_t x_i)\,\mathrm{d}t > 0.8$ for $i = 1, 2$. From this it follows that there is some $t \in [0, T_0]$ with $u_t.x_1, u_t.x_2 \in K$. Since $K \subset X'$ Lemma 4.3 proves (assuming $\epsilon < \eta$) the claim in that case and we may from now on assume that v is not fixed under the action of U and so $T_v < \infty$.

Lemma 4.8. *There exists a constant $c > 0$ that only depends on $\mathfrak{l}'$ such that the decomposition $\mathrm{Ad}_{u_t}(v) = w + O(\epsilon^{1/n})$ as above satisfies*

$\|w\| > c\eta$ for $t \in E_v$ where $E_v \subset [0, T_v]$ has Lebesgue measure at least $0.9T_v$.

Proof. We only have to look at the irreducible representations $V_j = \mathrm{Sym}^n(\mathbb{R}^2)$ in $\mathfrak{l}'$ with $T_v = T_{p_j}$. The size of the corresponding component of w is determined by the value of the polynomial $c_0 + c_1 t + \cdots + c_n t^n$. We change our variable by setting $s = \frac{t}{T_v}$ and get the polynomial $q(s) = c_0 + \frac{c_1}{T_v}s + \cdots + \frac{c_n}{T_v^n}s^n$. By definition of T_v the polynomial $q(s)$ has at least one coefficient of absolute value one while the others are in absolute value less or equal than one. Therefore, we have reduced the problem to finding a constant c such that for every such polynomial we have that $E_1 = \{s \in [0,1] : |q(s)| \geq c\}$ has measure bigger than 0.9.

This can be done in various ways — we will use the following property of polynomials for the proof. Every polynomial of degree n is determined by $n+1$ values (by the standard interpolation procedure). Moreover, we can give an upper bound on the coefficients in terms of these values unless the values of s used for the data points are very close together. (The determinant of the Vandermonde determinant is the product of the differences of the values of s used in the interpolation.) Suppose $s_0 \in [0,1] \setminus E_1$, then all points close to s_0 give unsuprisingly also small values. So we now look for $s_1 \in [0,1] \setminus (E_1 \cup [s_n - \frac{1}{20n}, s_n + \frac{1}{20n}])$ — if there is no such point then E_1 is as big as required. We repeat this search until we have found $n+1$ points $s_0, \ldots, s_n \in [0,1] \setminus E_1$ that are all separated from each other by at least $\frac{1}{20n}$. Again if we are not able to find these points, then E_1 is sufficiently big. However, as discussed the coefficients of q are due to $|q(s_0)|, \ldots, |q(s_n)| < c$ bounded by a multiple of c (that also involves some power of the degree n). If c is small enough compared to that coefficient we get a contradiction to the assumption that q has at least one coefficient of absolute value one. It follows that for that choice of c we can find at most n points in the above search and so E_1 has Lebesgue measure at least $1 - n\frac{2}{20n} = 0.9$. $\square$

Recall that case (1) from the beginning of this section implies Theorem 1.1 and that we are assuming case (2). Moreover, recall that this implies for all $\epsilon > 0$ the existence of $y, y' \in X_1$ with $y' = \exp(v).y$ for some nonzero $v \in B_\epsilon^{\mathfrak{l}'}(0)$ by Lemma 4.6. By definition of X_1 the sets

$$E_T = \{t \in [0,T] : u_t.y \in K\} \text{and}$$
$$E'_T = \{t \in [0,T] : u_t.y' \in K\}$$

have Lebesgue measure bigger than $0.8T$ whenever $T \geq T_0$. From the definition it is easy to see that $T_v \geq T_0$ once ϵ and therefore v are sufficiently small, so we can set to $T = T_v$. Moreover, let E_v be as in Lemma 4.8. Then the union of the complements of these three sets in $[0, T_v]$ has Lebesgue measure less than $0.5T_v$. Therefore, there exists some $t \in E_{T_v} \cap E'_{T_v} \cap E_v$. We set $x = u_t.y$, $x' = u_t.y'$ which both belong to K by definition of E_{T_v} and E'_{T_v}. Moreover, $x' = \exp(w + O(\epsilon^{1/n})).X$ where $w \in \mathfrak{l}'$ is stabilized by U and satisfies $c\eta \le \|w\| \le C\eta$ by Lemma 4.7– 4.8. We let $\epsilon \to 0$ and choose converging subsequences for x, x', and w. This shows the existence of $x, x' = \exp(w).x \in K$ and $w \in \mathfrak{l}'$ with $c\eta \le \|w\| \le C\eta$ which is stabilized by U. This is in effect our earlier claim which as we have shown implies that μ is invariant under a oneparameter subgroup not belong into L. This concludes the proof of Theorem 1.1.

References

[1] S. G. Dani. On invariant measures, minimal sets and a lemma of Margulis. *Invent. Math.*, 51(3):239–260, 1979.

[2] S. G. Dani. On uniformly distributed orbits of certain horocycle flows. *Ergodic Theory Dynamical Systems*, 2(2): 139–158, 1982 (1983).

[3] S. G. Dani. On orbits of unipotent flows on homogeneous spaces. *Ergodic Theory Dynam. Systems*, 4(1):25–34, 1984.

[4] S. G. Dani. On orbits of unipotent flows on homogeneous spaces. II. *Ergodic Theory Dynam. Systems*, 6(2):167–182, 1986.

[5] M. Einsiedler and A. Katok. Invariant measures on G/Γ for split simple Lie-groups *G. Comm. Pure Appl. Math.*, 56(8):1184–1221, 2003.

[6] M. Einsiedler and A. Katok. Rigidity of measures—the high entropy case and non-commuting foliations. *Israel J. Math.*, 148:169–238, 2005. Probability in mathematics.

[7] M. Einsiedler, A. Katok, and E. Lindenstrauss. Invariant measures and the set of exceptions to Littlewood's conjecture. to appear in Ann. of Math.

[8] M. Einsiedler and E. Lindenstrauss. Diagonalizable flows on locally homogeneous spaces and number theory. to appear in Proc. ICM 2006.

[9] M. Einsiedler and E. Lindenstrauss. On measures invariant under maximal split tori for semisimple S-algebraic groups. in preparation 2006.

[10] M. Einsiedler and E. Lindenstrauss. Rigidity of measures invariant under a diagonalizable group — the general low entropy method. in preparation 2006.

[11] H. Furstenberg. Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math. Systems Theory*, 1:1–49, 1967.

[12] H. Furstenberg. The unique ergodicity of the horocycle flow. *Recent advances in topological dynamics (Proc. Conf., Yale Univ., New Haven, Conn., 1972; in honor of Gustav Arnold Hedlund)*, 95–115. Lecture Notes in Math., Vol. 318, Springer, Berlin, 1973.

[13] A. Johnson. Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of the integers. *Israel J. Math.*, 77(1-2):211–240, 1992.

[14] Boris Kalinin and Ralf Spatzier. Rigidity of the measurable structure for algebraic actions of higher-rank Abelian groups. *Ergodic Theory Dynam. Systems*, 25(1):175–200, 2005.

[15] A. Katok and R. Spatzier. Invariant measures for higher-rank hyperbolic abelian actions. *Ergodic Theory Dynam. Systems*, 16(4):751–778, 1996.

[16] A. Katok and R. Spatzier. Corrections to: "Invariant measures for higher-rank hyperbolic abelian actions" [Ergodic Theory Dynam. Systems **16** (1996), no. 4, 751–778]. *Ergodic Theory Dynam. Systems*, 18(2):503–507, 1998.

[17] Dmitry Kleinbock, Nimish Shah, and Alexander Starkov. Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory. In *Handbook of dynamical systems, Vol. 1A*, pages 813–930. North-Holland, Amsterdam, 2002.

[18] E. Lindenstrauss. Invariant measures and arithmetic quantum unique ergodicity. to appear in Ann. of Math.

[19] R. Lyons. On measures simultaneously 2- and 3-invariant. *Israel J. Math.*, 61(2):219–224, 1988.

[20] G. Margulis. Discrete subgroups and ergodic theory. In *Number theory, trace formulas and discrete groups (Oslo, 1987)*, pages 377–398. Academic Press, Boston, MA, 1989.

[21] G. Margulis. Oppenheim conjecture. In *Fields Medallists' lectures*, volume 5 of *World Sci. Ser. 20th Century Math.*, pages 272–327. World Sci. Publishing, River Edge, NJ, 1997.

[22] G. Margulis and G. Tomanov. Invariant measures for actions of unipotent groups over local fields on homogeneous spaces. *Invent. Math.*, 116(1-3):347–392, 1994.

[23] G. A. Margulis. On the action of unipotent groups in the space of lattices. In *Lie groups and their representations (Proc. Summer School, Bolyai, János Math. Soc., Budapest, 1971)*, pages 365–370. Halsted, New York, 1975.

[24] Grigorii A. Margulis. Dynamical and ergodic properties of subgroup actions on homogeneous spaces with applications to number theory. In *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*, pages 193–215, Tokyo, 1991. Math. Soc. Japan.

[25] Calvin C. Moore. The Mautner phenomenon for general unitary representations. *Pacific J. Math.*, 86(1):155–169, 1980.

[26] Dave Witte Morris. *Ratner's theorems on unipotent flows*. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 2005.

[27] M. S. Raghunathan. *Discrete subgroups of Lie groups*. Springer-Verlag, New York, 1972. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 68.

[28] M. Ratner. On Raghunathan's measure conjecture. *Ann. of Math. (2)*, 134(3):545–607, 1991.

[29] M. Ratner. Raghunathan's topological conjecture and distributions of unipotent flows. *Duke Math. J.*, 63(1):235–280, 1991.

[30] M. Ratner. Raghunathan's conjectures for Cartesian products of real and p-adic Lie groups. *Duke Math. J.*, 77(2):275–382, 1995.

[31] Marina Ratner. Factors of horocycle flows. *Ergodic Theory Dynam. Systems*, 2(3-4):465–489 (1983), 1982.

[32] Marina Ratner. Horocycle flows, joinings and rigidity of products. *Ann. of Math. (2)*, 118(2):277–313, 1983.

[33] Marina Ratner. On measure rigidity of unipotent subgroups of semisimple groups. *Acta Math.*, 165(3-4):229–309, 1990.

[34] Marina Ratner. Strict measure rigidity for unipotent subgroups of solvable groups. *Invent. Math.*, 101(2):449–482, 1990.

[35] Marina Ratner. Distribution rigidity for unipotent actions on homogeneous spaces. *Bull. Amer. Math. Soc. (N.S.)*, 24(2):321–325, 1991.

[36] Marina Ratner. Raghunathan's conjectures for $\mathrm{SL}(2,\mathbf{R})$. *Israel J. Math.*, 80(1-2):1–31, 1992.

[37] D. Rudolph. $\times 2$ and $\times 3$ invariant measures and entropy. *Ergodic Theory Dynam. Systems*, 10(2):395–406, 1990.

[38] Dave Witte. Rigidity of some translations on homogeneous spaces. *Bull. Amer. Math. Soc. (N.S.)*, 12(1):117–119, 1985.

Mathematics Department, The Ohio State University, 231 W. 18th Avenue, Columbus, Ohio 43210