EFFECTIVE EQUIDISTRIBUTION AND SPECTRAL GAP

MANFRED EINSIEDLER

ABSTRACT. In these notes we discuss some equidistribution problems with the aim to give reasonable error rates, i.e. we are interested in effective statements. We motivate some arguments by studying a concrete problem on a two-torus, and then describe recent results on the equidistribution of semisimple orbits obtained in joint work with G. Margulis and A. Venkatesh. We end by studying the relationship between equidistribution of closed orbits and mixing properties. This leads to a way of transporting spectral gap from one group — via an effective equidistribution result on a quotient by an irreducible lattice – to another group. The latter topic is ongoing joint work with G. Margulis and A. Venkatesh.

1. PURPOSE

These notes are the combination of a few lectures given on an effective equidistribution theorem and related material. The main theorem that we discuss e.g. describes how dense closed orbits xH of $H = \mathrm{SO}(2,1)(\mathbb{R})^\circ$ on $\mathrm{SL}(3,\mathbb{Z})\backslash\mathrm{SL}(3,\mathbb{R})$ with big volume have to be. A more general version of this was obtained in joint work [7] with G. Margulis and A. Venkatesh and will be described in §6. A crucial input to the method that we used in [7] was spectral gap — in §4 we state what is used in general and prove the statement in the special case of $\mathrm{SL}(3,\mathbb{R})$ being the acting group. We motivate these questions and give a brief historical discussion in §2–§3. In §7 we outline the idea of ongoing joint work with G. Margulis and A.Venkatesh. Most of the material herein is well known to experts, but we think that assembling the material in these notes is worthwhile as it may help someone reading [7].

The author would like to thank his coauthors and students for many discussions on these topics. This research has been supported by the NSF-FRG collaborative grant 0554373.

2. MODEL CASES OF EQUIDISTRIBUTION PROBLEMS

2.1. **(Too) General setup.** Let us start with the following kind of equidistribution problems (which we specialize later to a more concrete setup). Suppose $T : X \to X$ is a continuous map on a compact metrizable space and $x \in X$. Then one can ask about the distribution properties of the finite sequence of points

$$x, T(x), T^2(x), \ldots, T^{n-1}(x) \in X.$$

We can specify the question more concretely by defining the measure

$$\int f d\delta_{x,n} = \frac{1}{n}\sum_{\ell=0}^{n-1} f(T^\ell(x)) \text{ for all } f \in C(X),$$

Date: December 2008.

and asking about the behavior of $\delta_{x,n}$ for large n and a given x. If $\delta_{x,n}$ converges for $n \to \infty$ in the weak* topology to some measure μ, then we say that the orbit of x *equidistributes* (w.r.t. μ). If we have a reasonable error for the expression $\left|\int f d\delta_{x,n} - \int f d\mu\right|$ for smooth functions, then we speak of *effective equidistribution.*

If T is ergodic with respect to an invariant probability measure μ, one knows that $\delta_{x,n}$ converges to μ in the weak* topology as $n \to \infty$ for μ-a.e. $x \in X$ by Birkhoff's pointwise ergodic theorem, i.e. a.e. orbit equidistributes w.r.t. μ. This is an interesting statement and can be quite useful in applications, but it does by no means provide a complete answer to the problem. For instance, it does not say anything about orbits of points $x \in X$ that are not typical for μ. Moreover, if we want to work with large but fixed n then again this provides no information about $\delta_{x,n}$ as the general ergodic theorem does not provide an effective error rate.

2.2. **Rotation on the circle $\mathbb{T}$.**

In the generality discussed above, one cannot hope to say anything about a given point $x \in X$ and also not anything — even if x is typical for an ergodic measure μ — about the speed of approximation. However, there are cases where both can be achieved.

Still in the same generality as above, if T has only one invariant probability measure on X, say μ, then more is true: For every point x one has that $\delta_{x,n}$ converges in the weak* topology to μ — and does so uniformly. It is easy to give an example for this, as e.g. the circle rotation defined by $T(x) = x + \alpha$ for $x \in \mathbb{T} = \mathbb{R}/\mathbb{Z}$ and some fixed irrational $\alpha \in \mathbb{R}$ (with addition being understood modulo $\mathbb{Z}$) preserves only the Lebesuge measure $m_{\mathbb{T}}$ and $\delta_{x,n}$ converges to $m_{\mathbb{T}}$ for any $x \in \mathbb{T}$. Moreover, one can also answer the refined question for an error rate but this requires[1] additional assumptions on α: If α is not a Liouville number and $f \in C^\infty(X)$, then

$$\left|\frac{1}{n}\sum_{\ell=0}^{n-1} f(x+\ell\alpha) - \int_{\mathbb{T}} f d\lambda\right| < \frac{1}{n} S(f),$$

where $S(f)$ depends on the function f (respectively on the sizes of the first few derivatives $f, f', f'', ..., f^{(L)}$ with L dependent on α). This is quite well known and can be proven directly for characters $e_k(x) = \exp(2\pi i k x)$ (using the geometric series and the assumption on α), and then can be boot-strapped to any smooth function f by an application of the Cauchy-Schwarz inequality. Instead of proving this, we give a proof of a different effective equidistribution statement on $\mathbb{T}^2$ below.

2.3. **Polynomial curves on $\mathbb{T}^2$.**

Let us study now two[2] polynomials $p_1(t), p_2(t)$ for $t \in [0,1]$ of degree $\leq D$ and how the corresponding curve[3] $\{(p_1(t), p_2(t)) : t \in [0,1]\}$ behaves modulo $\mathbb{Z}^2$ as a subset of $\mathbb{T}^2$. I.e. we wish to estimate

$$\left|\int_0^1 f(p_1(t), p_2(t))dt - \int_{\mathbb{T}^2} f(\mathbf{x})d\mathbf{x}\right| \tag{1}$$

for a given smooth f defined on $\mathbb{T}^2$. Clearly, if e.g. $p_2 = 0$ then there will not be a reasonable estimate for (1) as the curve in questions stays in a subtorus, more

[1] If one asks for a weaker form of an effective error rate, then one can do any irrational α. We refer to the work of Green and Tao [11] for what one can say without the Liouville-assumption.

[2] The only reason for restricting the dimension to 2 is just to restrict the number of parameters in this discussion.

[3] The continuous setting is in some aspects easier than the discrete one considered before, but is also more relevant to the following discussion.

generally the same holds if $(p_1(t), p_2(t))$ is close to a rational line for all $t \in [0,1]$. To avoid this problem, let us assume that there exists some L and T such that for any integer $\mathbf{n} \in \mathbb{Z}^2$ we have that the polynomial

$$p_{\mathbf{n}}(t) = (n_1 p_1 + n_2 p_2)(t) = c_0 + c_1 t + \cdots + c_D t^D$$

is nonconstant and moreover that[4]

$$\max_{j=1,\ldots,D} |c_j| \geq T \text{ for all } \mathbf{n} \in \mathbb{Z}^2 \text{ with } \|\mathbf{n}\|^L \leq T. \tag{2}$$

We fix L and think of T being a very large number (but without actually taking the limit $T \to \infty$ as one often would do in ergodic theory). We wish to estimate (1) by a negative power of T (which will depend on D) times a constant that depends on the sizes of the first few partial derivatives of f (where the number of derivatives used will depend on L). As this section's main purpose is to introduce some ideas in a very concrete setup, we make no claims regarding the optimality[5] of the estimates. We refer to [2] for the case of expanded images of a fixed curve where the Van der Corput lemma (see e.g. [10, pg. 146]) is used to prove a sharper estimate. Instead of using Van der Corput we will combine harmonic with more geometric arguments as this generalize more easily to the context considered later.

2.3.1. *Characters first.* We start the calculation for the desired estimate by the case where $f(\mathbf{x}) = e_{\mathbf{n}}(\mathbf{x}) = \exp(2\pi i(n_1 x_1 + n_2 x_2))$ is a character. The following argument is relatively simple but requires some games with exponents and hence a few constants that we will optimize at the end.

By definition of $p_{\mathbf{n}}$ we have $e_{\mathbf{n}}((p_1(t), p_2(t)) = \exp(2\pi i p_{\mathbf{n}}(t))$. We assume first that $\|\mathbf{n}\|^L \leq T$. By our assumption (2) and the equivalence of norms on the space of polynomials of degree $\leq D-1$ we have[6]

$$S = \max_{t \in [0,1]} |p'_{\mathbf{n}}(t)| \gg T$$

and similarly

$$\max_{t \in [0,1]} |p''_{\mathbf{n}}(t)| \ll S.$$

As $p'_{\mathbf{n}}(t)$ is a polynomial of degree $D-1$ we can easily convince ourselves that the Lebesgue measure of the points $t \in [0,1]$ where $|p'_{\mathbf{n}}(t)|$ is much smaller than S must in fact be small. In fact, for the polynomial $\frac{1}{S} p'_{\mathbf{n}}$ on $[0,1]$ of supremum norm about equal to one, the value of this polynomial can only be small, say smaller than $S^{\alpha-1}$, roughly speaking, close to the roots. Here the worst case happens if all the $D-1$ roots are equal, in which case $\frac{1}{S} p'_{\mathbf{n}}(t)$ is smaller than $S^{\alpha-1}$ on an interval of size $S^{\frac{\alpha-1}{D-1}}$. More formally, this estimate follows from the interpolation formula for polynomials, see for instance [12, Prop. 3.2], and gives that

$$m_{\mathbb{R}}\big(\{t \in [0,1] : |p'_{\mathbf{n}}(t)| < S^\alpha\}\big) \ll S^{\frac{\alpha-1}{D}}, \tag{3}$$

[4] Clearly, making a restriction on the $\mathbf{n}$ for which we require (2), will lead to a stronger result. In particular, with this restriction one can apply the result (5) to the case of a flow $(p_1(t), p_2(t)) = T(t, \alpha t)$ whenever α is not a Liouville number.

[5] This is partly but not only because we will be wasteful at places in the estimates if this helps to keep the expressions tidy.

[6] Implicit constants in the $\ll$-notation we allow to depend on D.

where $\alpha \in (0,1)$ is to be determined later. Vaguely speaking, for any α we will be able to ignore those t with $|p'_{\mathbf{n}}(t)| < S^\alpha$ as we are aiming to obtain an estimate involving a negative power of T.

Next we fix some $\beta \in (0,1)$, again to be determined later, and divide $[0,1]$ into subintervals of size $S^{-\beta}$ and one interval that may be shorter than that. Let I be one such interval of length $\leq S^{-\beta}$ and assume that for some $t_0 \in I$ we have $|p'_{\mathbf{n}}(t_0)| \geq S^\alpha$. Then as the second derivative is bounded by $\ll S$ on $[0,1]$ we have for any $t \in [0,1]$ that

$$p_{\mathbf{n}}(t) = p_{\mathbf{n}}(t_0) + (t-t_0)p'_{\mathbf{n}}(t_0) + O((t-t_0)^2 S).$$

By choosing $t \in I$ and $\beta > \frac{1}{2}$ we can make the error term here of the form

$$|(t-t_0)^2|S \leq S^{-2\beta+1}.$$

As the derivative of $\exp(2\pi i\cdot)$ is of absolute value 2π we have with these choice that

$$\Big|e_{\mathbf{n}}\big((p_1(t),p_2(t)\big) - \exp\big(2\pi i\|\mathbf{n}\|(p_{\mathbf{n}}(t_0) + (t-t_0)p'_{\mathbf{n}}(t_0))\big)\Big| \ll S^{-2\beta+1}.$$

We make that approximation because it is trivial to integrate an exponential function, which leads to

$$\Big|\int_I \exp\big(2\pi i(p_{\mathbf{n}}(t_0) + (t-t_0)p'_{\mathbf{n}}(t_0))\big)dt\Big| \ll \big|p'_{\mathbf{n}}(t_0)\big|^{-1} \leq S^{-\alpha}$$

Together we get

$$\Big|\int_I e_{\mathbf{n}}\big((p_1(t),p_2(t)\big)dt\Big| \ll S^{-\alpha} + S^{-2\beta+1}m_{\mathbb{R}}(I).$$

We are summing this estimate over all intervals I that contain some t_0 with $|p'_{\mathbf{n}}(t_0)| \geq S^\alpha$, of which there are at most $S^\beta + 1 \ll S^\beta$, and add the integral of the trivial estimate $\|e_{\mathbf{n}}\|_\infty = 1$ over the remaining intervals. As the union of the latter intervals is contained in the set in (3), we obtain

$$\Big|\int_0^1 e_{\mathbf{n}}\big((p_1(t),p_2(t)\big)dt\Big| \ll S^{-\alpha}S^\beta + S^{-2\beta+1} + S^{\frac{\alpha-1}{D}}.$$

It is clear that if we choose e.g. $\beta = \frac{3}{5}$ and $\alpha = \frac{4}{5}$, then all of the exponents are negative. A slightly better negative exponent is achieved by setting all the exponents equal and solving for α and β, which then turns the right hand side into $\ll S^{-\frac{1}{2D+3}}$. Using in addition $S \gg T$ gives for all $\mathbf{n} \in \mathbb{Z}^2 \setminus \{0\}$ that

$$\Big|\int_0^1 e_{\mathbf{n}}\big((p_1(t),p_2(t)\big)dt\Big| \ll T^{-\frac{1}{2D+3}}\|\mathbf{n}\|^L. \tag{4}$$

2.3.2. *Bootstrapping to any smooth function.* Using Cauchy-Schwarz and the relation between smoothness and Fourier-coefficients we can now generalize (4) to an estimate for (1). In fact, we claim that

$$\Big|\int_0^1 f(p_1(t),p_2(t))dt - \int_{\mathbb{T}^2} f(\mathbf{x})d\mathbf{x}\Big| \ll T^{-\frac{1}{2D+3}}S_{L+2}(f) \tag{5}$$

whenever p_1, p_2 are polynomials of degree $\leq D$ satisfying (2) and $f \in C^\infty(\mathbb{T}^2)$. We note that the error is independent of the starting point of the curve $(p_1(0), p_2(0))$

and of the particular polynomial as long as it satisfies our assumptions. Here S_{L+2} is the L^2-Sobolev norm of f of degree $L+2$ defined by

$$\begin{aligned} S_{L+2}(f)^2 &= \int |f(\mathbf{x})|^2 d\mathbf{x} + \int \big|\big(\frac{\partial}{\partial x_1}\big)^{L+2} f(\mathbf{x})\big|^2 d\mathbf{x} + \int \big|\big(\frac{\partial}{\partial x_2}\big)^{L+2} f(\mathbf{x})\big|^2 d\mathbf{x} \\ &= \sum_{\mathbf{n}\in\mathbb{Z}^2} \big(1 + (2\pi)^{L+2} \|\mathbf{n}\|^{2(L+2)}\big) |\hat{f}(\mathbf{n})|^2 \end{aligned}$$

To obtain (5) recall also that the Fourier series $f = \sum_{\mathbf{n}\in\mathbb{Z}^2} \hat{f}(\mathbf{n}) e_{\mathbf{n}}$ converges uniformly for $f \in C^\infty(\mathbb{T}^2)$. Hence we may sum (4) multiplied by $\hat{f}(\mathbf{n})$ over all $\mathbf{n} \in \mathbb{Z}^2 \setminus \{0\}$ to obtain

$$\Big|\int_0^1 f(p_1(t), p_2(t))dt - \int_{\mathbb{T}^2} f(\mathbf{x})d\mathbf{x}\Big| \ll T^{-\frac{1}{2D+3}} \sum_{\mathbf{n}\in\mathbb{Z}^2\setminus\{0\}} |\hat{f}(\mathbf{n})| \|\mathbf{n}\|^L$$

Here the sum on the right hand side can be estimated via Cauchy-Schwarz

$$\sum_{\mathbf{n}\in\mathbb{Z}^2\setminus\{0\}} |\hat{f}(\mathbf{n})| \|\mathbf{n}\|^L = \sum_{\mathbf{n}\in\mathbb{Z}^2\setminus\{0\}} |\hat{f}(\mathbf{n})| \|\mathbf{n}\|^{L+2} \frac{1}{\|\mathbf{n}\|^2} \ll \Big(\sum_{\mathbf{n}\in\mathbb{Z}^2\setminus\{0\}} |\hat{f}(\mathbf{n})|^2 \|\mathbf{n}\|^{2(L+2)}\Big)^{\frac{1}{2}}$$

where we used that $(\frac{1}{\|\mathbf{n}\|^2})_{\mathbf{n}\in\mathbb{Z}^2\setminus\{0\}}$ belons to ℓ^2. As the last expression is $\le S_{L+2}(f)$ this finishes the proof of (5).

3. Equidistribution of unipotent and closed orbits on homogeneous spaces

3.1. **Unipotent orbits.**

We replace the setup of a single transformation on a compact space discussed in §2.1 by a one-parameter flow, i.e. an action of $\mathbb{R}$, on a homogeneous space. Let $X = \Gamma\backslash G$ be a quotient of a linear group G by a lattice Γ, and let $U = \big\{u_t = \exp(tw) : t \in \mathbb{R}\big\} < G$ be a one-parameter unipotent subgroup — here w is a nilpotent element of the Lie algebra of G. Then instead of the above we consider the pieces of orbits $xu_{[0,T]} = \{xu_t : 0 \le t \le T\}$ for points $x \in X$. For this Ratner [15] has shown that the normalized image of the Lebesgue measure on $xu_{[0,T]}$ converges to a natural measure on X — as before we say the orbit equidistributes with respect to this measure. This natural invariant measure on X is for many points[7] the Haar measure on X, but the theorem applies to any point as follows. For a given x Ratner proves [15] that the orbit closure $\overline{xU} \subset X$ is of the form xL for some closed connected subgroup $L < G$ and that this orbit supports an L-invariant probability measure, the Haar measure m_{xL}, this is known as Raghunathan's conjecture. Then the measure on $xu_{[0,T]}$ converges in the weak* topology to m_{xL}. However, the problem of estimating the error in this theorem and in this generality is wide open.

A special case of the above setup is given by $U = \left\{u_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : t \in \mathbb{R}\right\}$ acting on $X = \Gamma\backslash \operatorname{SL}_2(\mathbb{R})$, i.e. the horocycle flow on the unit tangent bundle of a hyperbolic surface. If $X = \Gamma\backslash \operatorname{SL}(2,\mathbb{R})$ is compact, then the equidistribution of orbits has been

[7] Unlike the abstract ergodic theorem Ratner's theorem establishes precisely for which points this is true.

established by Furstenberg [9] already much earlier. Moreover, in this case error rates are known:

$$|\frac{1}{T}\int f(xu_t)dt - \int f dm_X\Big| \ll S(f)T^{-\delta},$$

where m_X is the Haar measure on X, $S(f)$ is a Sobolev norm of the function f and $\delta > 0$ is a constant which depends on the spectral properties of X. In particular, this error is independent of the starting point x. We refer the reader to [3], [17, Sect. 9.3.1], and [8] for more details. If X is noncompact with finite volume, e.g. $X = \mathrm{SL}(2,\mathbb{Z})\backslash\mathrm{SL}(2,\mathbb{R})$, then the above error is again more delicate. This is because, in X there are periodic orbits for the action of U. Even assuming that x is not periodic, x could in fact be very close to a periodic orbit for U which makes it impossible to give an error that is independent of x.

3.2. **Equidistribution of closed orbits.** A problem related to the distribution of pieces of the orbit is the distribution of closed orbits. Here a toy problem is the effective distribution properties of rational lines in the two-dimensional torus (which is a special case of the discussion in §2.3).

3.2.1. *Periodic horocycles.* A more interesting case concerns the distribution properties of closed horocyle orbits on noncompact quotients. Here an error rate has been established by Sarnak [16]. We now describe this result for $\mathrm{SL}(2,\mathbb{Z})\backslash\mathrm{SL}(2,\mathbb{R})$ and outline the argument from [17, Sect. 9] which establishes a slightly weaker form of the effective equidistribution.

We start by recalling that periodic orbits of U are easily visualized using the unit tangent bundle of the hyperbolic plane $\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. Here the fundamental domain for $\mathrm{SL}(2,\mathbb{Z})$ is the triangle bounded by $\mathrm{Re}(z) = \pm\frac{1}{2}$ and the unit circle. In this picture the horocycle transports vectors along the horocycle normal to the vector, and horocycles are horizontal lines and circles touching the real axis. In particular, we can visualize periodic orbits for the horocycle flow as horizontal line segments cutting through the fundamental domain with the arrows pointing up. Let y be the y-coordinate of the points in the orbit and write P_y for the periodic orbit. Going up inside the fundamental domain (i.e. for $y \to \infty$) the length of the periodic orbit, which equals $\frac{1}{y}$ goes to zero and the orbit escapes to infinity.

However, as $y \to 0$ we can still draw periodic orbits as horizontal lines outside the standard fundamental domain. If we draw P_y for small y inside the fundamental domain (applying the appropriate isometries from $\mathrm{SL}(2,\mathbb{Z})$) the orbit will look much more complicated, but will be periodic of large length $\frac{1}{y}$. In fact, the orbit P_y becomes equidistributed in X as $y \to 0$. Moreover, as Sarnak showed (in greater generality and with more information regarding δ) this can be made effective, i.e.

$$|\int_{P_y} f - \int_X f dm_X| \ll y^\delta S(f)$$

for any $f \in C_c^\infty(X)$. Here $\int_{P_y} f$ denotes the normalized integral over the periodic orbit P_y.

3.2.2. *Outline of a proof.* The geodesic flow is hyperbolic, i.e. inside the 3-dimensional space X there are three special directions:

(0) the orbit direction of the geodesic flow,

(s) the horocycle direction (corresponding to U) which is contracted by the geodesic flow in forward time (the stable direction), and

(u) the opposite horocycle direction (corresponding to $\begin{pmatrix} 1 & 0 \\ * & 1 \end{pmatrix}$) which is expanded (the unstable direction).

Now let B be a small **box** (using the above three directions as "directions for the sides") around the periodic orbit P_1 (the periodic orbit for U going through i), then

$$\langle f, g_t \cdot \chi_B \rangle \to m_X(B) \int_X f dm_X \text{ as } t \to \pm\infty$$

by the Howe-Moore theorem on vanishing of matrix coefficients or (equivalently) the mixing property of the geodesic flow. Here $\langle \cdot, \cdot \rangle$ is the inner product in $L^2(X)$, χ_B is the characteristic function, $g_t \cdot$ denotes the unitary action of the geodesic flow on $L^2(X)$, and m_X denotes the Haar measure on X. However, $g_t \cdot \chi_B$ equals the characteristic function of $B_t = B \begin{pmatrix} e^{-\frac{t}{2}} & \\ & e^{\frac{t}{2}} \end{pmatrix}$, which we should think of as a box around the periodic orbit $P_{e^{-t}}$. The distance of points in this new box to the new periodic orbit in the direction of geodesic flow is unchanged, and in the direction of the opposite horocycle flow has decreased exponentially. Hence for $f \in C_c(X)$ and large enough t we have (by uniform continuity and the careful construction of a thin enough box)

$$\int_{P_{e^{-t}}} f \approx \frac{1}{m_X(B)} \int_{B_t} f dm_X = \frac{1}{m_X(B)} \langle f, g_t \cdot \chi_B \rangle \approx \int_X f,$$

which can be made more precise to give a proof of the (noneffective) claim.

Using $f \in C_c^\infty$ one can use the same argument as above (the method in [16] is different and gives a better constant for δ), replacing the box with a smooth box-like function, and replacing the Howe-Moore theorem with the **effective decay of matrix coefficients** as discussed in §4.

3.2.3. *More general closed orbits.* More generally, one may ask about the distribution properties of closed, finite volume orbits xH of closed subgroups $H \subset G$ on quotients $X = \Gamma \backslash G$. If H is generated by unipotent subgroups, a theorem of Mozes and Shah [13] describes limits of such finite volume orbits — the limit measure is again a Haar measure m_{xL} just as in Ratner's theorem. However, also just as in Ratner's equidistribution theorem for individual orbits, the problem of establishing an error rate in this generality is wide open. We will discuss in §6 a special case where an error rate has been obtained in joint work with Margulis and Venkatesh [7].

4. Effective decay of matrix coefficients and spectral gap

We assume G is a closed linear semisimple group. We say we have *effective decay of matrix coefficients* for $X = \Gamma \backslash G$ if there exists some $\delta > 0$ such that

$$\left| \langle g \cdot f_1 - \textstyle\int f_1 dm_X, f_2 - \textstyle\int f_2 dm_X \rangle \right| \ll \|g\|^{-\delta} S(f_1) S(f_2), \tag{6}$$

where $g \in G$, $f_1, f_2 \in C_c^\infty(X)$, and $\|g\|$ denotes the maximum of the matrix entries of g. As before $S(f)$ denotes a Sobolev norm of f.

Also we say that the action of G on X has a *spectral gap* if there exists some nonnegative $\chi \in C_c(G)$ with $\int_G \chi(g) dm_G(g) = 1$ such that for any $f \in L^2(X)$ with $\int_X f dm_X = 0$ we have $\|\chi * f\|_2 \leq \theta \|f\|_2$ for some fixed $\theta < 1$. Here

$$\chi * f(x) = \int \chi(g) f(xg) \text{ for } x \in X$$

may be thought of as the average of the images $g \cdot f$ of f under the unitary transformation induced by right multiplication by g on X with respect to the weights $\chi(g)$. As $\chi * 1 = 1$ the assumption that $\theta < 1$ amounts to having a gap in the spectrum of the operator $\chi *$.

Both of the above notions generalize to more general unitary representations of G, in both notions we restrict ourself to representations without fixed vectors (or equivalently the orthogonal complement of the space of vectors fixed under G). The existence of a spectral gap $\theta < 1$ that is independent of the unitary representation is the well-known property (T) of the group G. We recall that $\mathrm{SL}(3, \mathbb{R})$ has property (T), but that $\mathrm{SL}(2, \mathbb{R})$ does not have property (T).

From representation theory one knows that (6) is equivalent to spectral gap for the G-action on $L^2(X)$ (where δ and the gap $1 - \theta$ are related). We refer to [7, Sect. 6] and the references there for a discussion of this equivalence.

4.1. **Effective decay for** $\mathrm{SL}(3, \mathbb{R})$**.** Spectral gap, in the form of effective decay of matrix coefficients, is an essential input for establishing effective equidistribution for homogeneous spaces, and so we would like to discuss where it comes from. However, instead of describing the general argument for establishing spectral gap and effective decay of matrix coefficients on "congruence" quotients, which would be quite hard in these short notes, we will give a direct proof of the effective decay of matrix coefficients for unitary representations of $\mathrm{SL}(3, \mathbb{R})$. I.e. we will prove (6) by showing

$$\big|\langle g \cdot v, w \rangle\big| \ll \|g\|^{-\frac{3}{8}} S(v) S(w), \tag{7}$$

whenever v, w are smooth vectors belonging to a Hilbert space $\mathcal{H}$ which has a unitary action of $\mathrm{SL}(3, \mathbb{R})$ on it and does not contain any $\mathrm{SL}(3, \mathbb{R})$-fixed vectors. E.g. this will apply to the subspace of $L^2(\Gamma \backslash \mathrm{SL}(3, \mathbb{R}))$ of functions of integral zero for any lattice Γ. Again we will use a Sobolev norm $S(v)$ for smooth vectors $v \in \mathcal{H}$ which we define below. The argument we present is an effectivization of the proof that $\mathrm{SL}(3, \mathbb{R})$ has property (T) and is likely well known to experts of the field.

4.1.1. *Smooth vectors and the Sobolev norm.* Let π be a unitary representation of $\mathrm{SL}(3, \mathbb{R})$ on a Hilbert space $\mathcal{H}$, for which we will also write $\pi(g)v = g \cdot v$ for $g \in \mathrm{SL}(3, \mathbb{R})$ and $v \in \mathcal{H}$. A vector $v \in \mathcal{H}$ is called *smooth* if all partial derivates of $g \mapsto \pi(g)v$ as a map from G to $\mathcal{H}$ exist. Taking a basis $e_1, \ldots, e_8$ of the Lie algebra $\mathfrak{sl}_3$ of $\mathrm{SL}(3, \mathbb{R})$ we can define the *Sobolev norm* (of degree one) by $S(v)^2 = \|v\|^2 + \sum_{j=1}^8 \big\| \big(\frac{\partial}{\partial t} \exp(t e_j) \cdot v\big)\big|_{t=0} \big\|^2$ where the sum is over all partial derivatives corresponding to the basis elements.

4.2. **Spectral measures.** We will also be needing some basic properties of the spectral measures which we recall next. Let π be a unitary representation of $\mathbb{R}^2$. Then $\mathbf{t} \to \langle \pi(\mathbf{t}) v, v \rangle$ is a positive definite function and so equals $\int_{\mathbb{R}^2} \exp(2\pi i \mathbf{t} \cdot \mathbf{s}) d\mu_{v,v}(\mathbf{s})$ for some finite measure $\mu_{v,v}$ on $\mathbb{R}^2$ by Bochner's theorem. We will refer to $\mu_{v,v}$ as the *spectral measure* of v. These are used in the theory of unitary representations of $\mathbb{R}^2$ to define the projection-valued measure E_B on $\mathcal{H}$ for any Borel subset $B \subset \mathbb{R}^2$ which have the property that the spectral measure $\mu_{E_B v, E_B v}$ is the restriction of $\mu_{v,v}$ to B. The map E_B is an orthogonal projection commuting with $\pi(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^2$, satisfies that $E_{\mathbb{R}^2}$ is the identity and that $E_{B_1 \cup B_2} = E_{B_1} + E_{B_2}$ whenever $B_1, B_2 \subset \mathbb{R}^2$ are disjoint. In particular, if $\mathcal{H}$ does not contain any vectors fixed under $\mathbb{R}^2$, then $\mu_{v,v}(\{0\}) = 0$. Finally, we note that if $v, w \in \mathcal{H}$ have singular spectral measures, then v and w are orthogonal.

We will be using these tools for the restriction of the unitary representation of $\mathrm{SL}(3,\mathbb{R})$ to the subgroup

$$U = \left\{ \begin{pmatrix} 1 & 0 & t_1 \\ 0 & 1 & t_2 \\ 0 & 0 & 1 \end{pmatrix} : \mathbf{t} \in \mathbb{R}^2 \right\}.$$

Note that by the Mautner phenomenon we have no $\mathbb{R}^2$-fixed vectors in $\mathcal{H}$ as we assumed that there are no $\mathrm{SL}(3,\mathbb{R})$-invariant vectors.

We note that the subgroup $\mathrm{SL}(2,\mathbb{R})$ embedded into the upper left corner of $\mathrm{SL}(3,\mathbb{R})$ normalizes the subgroup U. This leads to a relationship of the spectral measure of v and of $g \cdot v$ for $g \in \mathrm{SL}(2,\mathbb{R})$. In fact, we claim that $\mu_{g\cdot v, g\cdot v} = (g^{-1})^T_* \mu_{v,v}$. This follows from uniqueness of the measure in Bochner's theorem and the equation

$$\begin{aligned} \langle \pi(\mathbf{t}) g \cdot v, g \cdot v \rangle = \langle \pi(g^{-1}\mathbf{t}) v, v \rangle &= \int_{\mathbb{R}^2} \exp(2\pi i (g^{-1}\mathbf{t}) \cdot \mathbf{s}) d\mu_{v,v}(\mathbf{s}) \\ &= \int_{\mathbb{R}^2} \exp(2\pi i \mathbf{t} \cdot ((g^{-1})^T \mathbf{s}) d\mu_{v,v}(\mathbf{s}). \end{aligned}$$

Here we used that $\begin{pmatrix} g^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & t_1 \\ 0 & 1 & t_2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} g & 0 \\ 0 & 1 \end{pmatrix}$ belongs to U and is the element corresponding to $g^{-1}\mathbf{t}$.

4.3. **Eigenfunctions of** SO(2) **first.** We assume first that $v, w \in \mathcal{H}$ are eigenfunctions of SO(2), i.e. that for the matrix $k_\theta \in \mathrm{SO}(2)$ corresponding to a rotation by angle θ we have $k_\theta \cdot v = \exp(i\theta n) v$ and $k_\theta \cdot w = \exp(i\theta m) v$ for some $n, m \in \mathbb{Z}$. In this case we have $\langle \pi(\mathbf{t}) k_\theta \cdot v, k_\theta \cdot v \rangle = \langle \pi(\mathbf{t}) v, v \rangle$ which shows that the spectral measures of v and $k_\theta \cdot v$ are the same. This implies that the spectral measure of v, and similarly for w, is invariant under SO(2).

We claim that for such eigenfunctions v, w we have

$$|\langle a_r \cdot v, w \rangle| \ll e^{-|r|/2} \|v\| \|w\| \text{ where } a_r = \begin{pmatrix} e^{-r} & 0 & 0 \\ 0 & e^r & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{8}$$

where as before we assume $\mathcal{H}$ does not contain any $\mathrm{SL}(3,\mathbb{R})$-invariant vectors. We assume that $r > 0$, the argument for the other case is similar. We show this by splitting both v and w into two components $v = v_{\mathrm{main}} + v_{\mathrm{vertical}}$ and $w = w_{\mathrm{main}} + w_{\mathrm{horizontal}}$. Here v_{vertical} is defined as the image of v under the orthogonal projection defined by the set $\left\{(t_2, t_1) : \left|\frac{t_2}{t_1}\right| \geq e^r\right\}$ which is a sector shaped neighborhoods of the t_2-axis of angle $\ll e^{-r}$. Hence by invariance of the spectral measure under SO(2) we get $\|v_{\mathrm{vertical}}\|^2 \ll e^{-r} \|v\|^2$. Similary, $w_{\mathrm{horizontal}}$ is defined as the image of w under the orthogonal projection defined by $\left\{(t_2, t_1) : \left|\frac{t_2}{t_1}\right| \leq e^{-r}\right\}$ which also has

$\|w_{\mathrm{horizontal}}\| \ll e^{-r}\|w\|$. The other two vectors v_{main} and w_{main} are defined as the projections w.r.t. the complements of these sets. Recall that the spectral measure of v_{main} is supported on $\left\{(t_2,t_1) : \left|\frac{t_2}{t_1}\right| < e^r\right\}$ and that the spectral measure of $a_r \cdot v_{\mathrm{main}}$ is the push forward of the spectral measure of v_{main} under $(a_r^{-1})^T = a_r^{-1}$. This shows that the spectral measure of $a_r \cdot v_{\mathrm{main}}$ is supported on $\left\{(t_2,t_1) : \left|\frac{t_2}{t_1}\right| < e^{-r}\right\}$ and so $a_r \cdot v_{\mathrm{main}}$ is orthogonal to w_{main} as their spectral measures are supported on disjoint sets. Applying this to

$$\begin{aligned}|\langle a_r \cdot v, w\rangle| \leq |\langle a_r \cdot v_{\mathrm{main}}, w_{\mathrm{main}}\rangle| &+ |\langle a_r \cdot v_{\mathrm{main}}, w_{\mathrm{horizontal}}\rangle| \\ &+ |\langle a_r \cdot v_{\mathrm{vertical}}, w_{\mathrm{main}}\rangle| + |\langle a_r \cdot v_{\mathrm{vertical}}, w_{\mathrm{horizontal}}\rangle|,\end{aligned}$$

we get that the first term is zero, and the other are bounded by $\ll e^{-r/2}\|v\|\|w\|$ which gives (8).

4.4. **Bootstrapping to general vectors and general group elements.** We first extend (8) to any diagonal matrix

$$a = \begin{pmatrix} e^{r_1} & 0 & 0 \\ 0 & e^{r_2} & 0 \\ 0 & 0 & e^{r_3} \end{pmatrix}$$

with $r_1 + r_2 + r_3 = 0$ and any two smooth vectors $v, w \in \mathcal{H}$ to say

$$|\langle a \cdot v, w\rangle| \ll e^{-\frac{1}{4}|r_2 - r_1|} S(v) S(w). \tag{9}$$

To obtain this we decompose $v = \sum_{n\in\mathbb{Z}} v_n$ and $w = \sum_{m\in\mathbb{Z}} w_m$ into eigenfunctions for SO(2) — by smoothness these sums converge absolutely. Next notice that

$$a = \begin{pmatrix} e^{r_1+\frac{1}{2}r_3} & 0 & 0 \\ 0 & e^{r_2+\frac{1}{2}r_3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e^{-\frac{1}{2}r_3} & 0 & 0 \\ 0 & e^{-\frac{1}{2}r_3} & 0 \\ 0 & 0 & e^{r_3} \end{pmatrix} = a_{r_2+\frac{1}{2}r_3} c$$

where $a_{r_2+\frac{1}{2}r_3} = a_{\frac{1}{2}(r_2-r_1)}$ is as in (8) and c commutes with SO(2). The latter implies that v_n is mapped under c again to eigenfunctions of SO(2). Therefore, we may apply (8) to each $c \cdot v_n$ and w_m to get

$$|\langle a \cdot v, w\rangle| \leq \sum_{m,n\in\mathbb{Z}} |\langle a_{\frac{1}{2}(r_2-r_1)} c \cdot v_n, w_m\rangle| \ll e^{-\frac{1}{4}|r_2-r_1|} \sum_{m,n\in\mathbb{Z}} \|v_n\|\|w_m\|.$$

However, the last sum on the right may be written as the product of $\sum_{n\in\mathbb{Z}} \|v_n\|$ and the corresponding sum for w_m. Notice that the derivative of v_n along some element r of the Lie algebra of SO(2) equals

$$\big(\frac{\partial}{\partial t} \exp(tr) \cdot v_n\big)\big|_{t=0} = n v_n$$

and so $\left(\frac{\partial}{\partial t} \exp(tr) \cdot v\right)\big|_{t=0} = \sum_{n\in\mathbb{Z}} n v_n$, and that the terms in the last sum are all orthogonal to each other. Hence Cauchy-Schwarz gives

$$\sum_{n\in\mathbb{Z}} \|v_n\| = \|v_0\| + \sum_{n\in\mathbb{Z}\setminus\{0\}} \frac{1}{n}\|n v_n\| \ll \|v_0\| + \big\|\big(\frac{\partial}{\partial t}\exp(tr)\cdot v\big)\big|_{t=0}\big\| \ll S(v),$$

where we used that r can be expressed as a linear combination of the basis elements $e_1, \ldots, e_8 \in \mathfrak{sl}_3$ that we used to define $S(v)$. This gives (9).

To make (9) closer to (7) we notice that in (9) we could have proven the same statement with either $e^{-\frac{1}{4}|r_3-r_2|}S(v)S(w)$ or with $e^{-\frac{1}{4}|r_3-r_1|}S(v)S(w)$ on the right. We claim that

$$\min(e^{-\frac{1}{4}|r_2-r_1|}, e^{-\frac{1}{4}|r_3-r_2|}, e^{-\frac{1}{4}|r_3-r_1|}) \leq \|a\|^{-\frac{3}{8}},$$

which then shows that (7) holds for all diagonal matrices. To prove the above, assume that $r_3 \geq r_2 \geq r_1$. Then $\|a\| = e^{r_3}$ and $e^{r_1} \leq e^{\frac{1}{2}(r_1+r_2)}$ which together with $r_1 + r_2 + r_3 = 0$ gives

$$e^{r_1-r_3} \leq e^{\frac{1}{2}(r_1+r_2+r_3)}e^{-\frac{3}{2}r_3} = e^{-\frac{3}{2}r_3} = \|g\|^{-\frac{3}{2}},$$

which proves the claim and so (7) in this case.

To prove (7) for all $g \in \mathrm{SL}(3,\mathbb{R})$ we recall that $g = k_1ak_2$ for some $k_1, k_2 \in \mathrm{SO}(3)$ and some diagonal matrix a by the Cartan decomposition of g in $\mathrm{SL}(3,\mathbb{R})$. As $\mathrm{SO}(3)$ is compact, $\|g\|$ and $\|a\|$ are bounded by some multiplies of each others. Similarly, $S(k_2 \cdot v) \ll S(v)$ and $S(k_1^{-1}w) \ll S(w)$. Together this gives using (9)

$$|\langle g \cdot v, w\rangle| = |\langle a \cdot (k_2 \cdot v), k_1^{-1} \cdot w\rangle| \ll \|g\|^{-\frac{3}{8}}S(v)S(w),$$

which proves (7).

4.5. **Groups without property (T).** As we mentioned before the above argument is the effectivization of the proof that $\mathrm{SL}(3,\mathbb{R})$ has property (T). However, e.g. $\mathrm{SL}(2,\mathbb{R})$ and $\mathrm{SU(m,1)}(\mathbb{R})$ do not have property (T). For these groups spectral gap (respective effective decay of matrix coefficients) is not an automatic property for any unitary representation. However, Selberg showed that the $\mathrm{SL}(2,\mathbb{R})$-action on congruence quotients $\Gamma\backslash \mathrm{SL}(2,\mathbb{R})$ has a spectral gap — in fact there is a uniform spectral gap that is independent of Γ. In this case and in similar cases the spectral gap is a property of the space and not of the group.

5. An effective pointwise ergodic theorem

For the main theorem of [7], which we will discuss in §6, a pointwise ergodic theorem was needed and also proven in [7, Prop. 9.2]. As we outline now this is a consequence of the effective decay of matrix coefficients (6) discussed earlier (but we we will refer to [7] for the last step of the argument).

There are two basic types of non-compact one-parameter subgroups of semisimple groups $G \subseteq \mathrm{SL}(n,\mathbb{R})$: Diagonalizable subgroups and unipotent subgroups. The estimate in (6) can be used to establish an effective ergodic theorem for both of them, but as the unipotent case may seem a bit more delicate and at the same time is the case that will be used later, let us focus on that case. Hence suppose $u_t = \exp(tp)$ is a unipotent one-parameter subgroup defined by some nilpotent element p in the Lie algebra of G. Then we notice that $t \ll \|u_t\| \ll t^n$ as the entries of the matrix u_t are polynomials in t and so (6) is the statement that matrix coefficients decay at a polynomial rate with respect to the time parameter of the subgroup. (For diagonalizable subgroups (6) would be exponential decay of matrix coefficients.)

5.1. **A single function and a given time first.** For $f \in C_c^\infty(X)$ and $T > 0$ we define the *discrepancy* at x by

$$D_T(f)(x) = \frac{1}{T}\int_0^T f(xu_t)dt - \int_X f dm_X,$$

it measures how far the time average over $[0,T]$ is away from the expected value. Using Fubini's theorem several times as well as that $u_t \in G$ preserves m_X we get

$$\int |D_T(f)|^2 dm_X = \frac{1}{T^2}\int_0^T \int_0^T \int_X f(xu_{t_1})\overline{f}(xu_{t_2}) dm_X dt_2 dt_1 -$$
$$2\,\mathrm{Re}\frac{1}{\mathrm{T}}\int_0^{\mathrm{T}} \int_X f(xu_t) dm_X dt \int_X \overline{f} dm_X + \Big|\int_X f dm_X\Big|^2 =$$
$$\frac{1}{T^2}\int_0^T \int_0^T \Big(\langle u_{t_1-t_2}\cdot f, f\rangle - \big|\int_X f dm_X\big|^2\Big) dt_2 dt_1.$$

Now notice that for most $(t_1,t_2) \in [0,T]^2$ we have that $|t_1-t_2|$ is quite big, and so the expression within the last integral is quite small. More precisely, if $|t_1-t_2| > T^{\frac{1}{2}}$, then by (6) we have that

$$\Big|\langle u_{t_1-t_2}\cdot f, f\rangle - \big|\int_X f dm_X\big|^2\Big| \ll T^{-\frac{1}{2}\delta} S(f)^2$$

while the integral over the part $|t_1-t_2| \le T^{\frac{1}{2}}$ is bounded by the area $\le TT^{\frac{1}{2}}$ times the trivial estimate $\ll \|f\|_\infty^2$ of the integrand. Together this gives

$$\int |D_T(f)|^2 dm_X \ll T^{-\frac{1}{2}\delta} S(f)^2 + T^{-\frac{1}{2}}\|f\|_\infty^2.$$

We can simplify this as follows. If we modify our notion of Sobolev norm, we can make sure that

$$\|f\|_\infty \ll S(f),$$

see [7, Lemma 5.1.1]. This is not entirely trivial, because in fact, we claim that one can modify the norm in such a way that $S(f)$ is still the norm of a pre-Hilbert-space strucure (i.e. is an Hermitian norm) on $C_c^\infty(X)$ — this is not important right now, but will be for the last step of the argument. If additionally we also assume w.l.o.g. that $\delta \le 1$ then we have

$$\int |D_T(f)|^2 dm_X \ll T^{-\frac{1}{2}\delta} S(f)^2.$$

This shows that

$$W^2 m_X\big(\{x \in X : D_T(f)(x) \ge W\}\big) \ll T^{-\frac{1}{2}\delta} S(f)^2 \tag{10}$$

for any $W > 0$. We still have some freedom in W – asking for a better estimate, i.e. a smaller value of W, will make the estimate of the set worse. To achieve a reasonable estimate on the set, we set $W = T^{-\frac{1}{6}\delta} S(f)$ which makes the above into

$$m_X\big(\{x \in X : D_T(f)(x) \ge T^{-\frac{1}{6}\delta} S(f)\}\big) \ll T^{-\frac{1}{6}\delta}. \tag{11}$$

This is already an effective version of the pointwise ergodic theorem: For a given $f \in C_c^\infty(X)$ and $T > 0$ we know that the average of f over the $[0,T]$-orbit of x is $T^{-\frac{1}{6}\delta} S(f)$ close to $\int_X f dm_X$ except possibly for a set of points x of measure $\ll T^{-\frac{1}{6}\delta}$. However, this is not yet very satisfactory as the exceptional set is still allowed to depend on f and on T.

5.2. **A single function with large enough times.** It is relatively easy to obtain the following strengthening of the above: There exists some $\epsilon > 0$ such that for a given $f \in C_c^\infty(X)$ we have that for any T_0 the average of f over the $[0,T]$-orbit of x is $T^{-\epsilon}S(f)$ close to $\int_X f dm_X$ for all $T \geq T_0$ except possible for a set of measure $\ll T_0^{-\epsilon}$. I.e. at some cost in the exponents we can make the set independent of the particular time interval $[0,T]$ that we use to average and still get a very good estimate on the measure of the exceptional points if we only restrict ourself to large enough times $T \geq T_0$.

To prove the above let $M = 12\frac{1}{\delta}$. Then we may apply (11) for $T_n = n^M$ which gives

$$m_X\big(\{x \in X : D_T(f)(x) \geq T_n^{-\frac{1}{6}\delta} S(f)\}\big) \ll n^{-2}.$$

Call this exceptional set E_n, then $m_X(\bigcup_{n \geq n_0} E_n) \ll n_0^{-1}$ for any real $n_0 > 0$.

Now choose some T_0 and define $n_0 = T_0^{\frac{1}{M}}$. Now let $T \geq T_0$ and let $n = \lceil T^{\frac{1}{M}} \rceil \geq n_0$. Then $|n^M - T| \ll n^{M-1} \ll T^{\frac{M-1}{M}} = T^{1-\frac{1}{M}}$ and from this it is easy to see that

$$\Big|\frac{1}{n^M}\int_0^{n^M} f(xu_t)dt - \frac{1}{T}\int_0^T f(xu_t)dt\Big| \ll \|f\|_\infty T^{-\frac{1}{M}}.$$

This gives for $x \notin E_n$ that

$$\Big|\frac{1}{T}\int_0^T f(xu_t)dt - \int_X f dm_X\Big| \ll (T^{-\frac{1}{M}} + T^{-\frac{1}{6}\delta})S(f).$$

Setting $\epsilon = \min(\frac{1}{M}, \frac{1}{6}\delta) = \frac{1}{12}\delta$ gives the desired estimate.

5.3. **Bootstrapping to all functions** $f \in C_c^\infty(X)$**.** We fix some $\epsilon > 0$. We say a point $x \in X$ is (T_0, ϵ)-*generic* if

$$\Big|\frac{1}{T}\int_0^T f(xu_t)dt - \int_X f dm_X\Big| \ll T^{-\epsilon} S'(f) \tag{12}$$

for all $T \geq T_0$ and all $f \in C_c^\infty(X)$. Then an even stronger effective version of the pointwise ergodic theorem would be that there exists a choice of ϵ for which

$$m_X\big(\{x : x \text{ is not } (T_0,\epsilon)\text{-generic}\}\big) \ll T_0^{-\epsilon}.$$

This can be obtained by the argument in [7, Sect. 9]. The hidden cost is that in (12) a different notion of Sobolev norm S' (defined using more derivatives) is used than in (10). Allowing for that, gives us the possibility of making W in (10) also depend on $\frac{S'(f)}{S(f)}$. The argument is in some way then similar to §2.3.2 and §4.4. Using different Sobolev norms one can find an orthonormal basis $f_1, \ldots, f_k, \ldots$ w.r.t. $S'(\cdot)$ such that $\sum_{n=1}^\infty S(f_n)$ is finite. This uses some ideas (relative traces of Hermitian norms) of Bernstein and Reznikov [1]

6. Effective equidistribution for semisimple subgroups

We shall assume that:

- There is a semisimple $\mathbb{Q}$-group $\mathbf{G}$ so that $G = \mathbf{G}(\mathbb{R})^\circ$ and Γ is a congruence subgroup of $\mathbf{G}(\mathbb{Q})$.
- H is a connected semisimple subgroup without compact factors.

We note that in this context an H-orbit $x_0H \subset X = \Gamma\backslash G$ is closed **if and only if** it has finite volume.

Two examples of this setup are $H = \mathrm{SO}(2,1)(\mathbb{R})^\circ$ acting on $\mathrm{SL}(3,\mathbb{Z})\backslash \mathrm{SL}(3,\mathbb{R})$, and $H = \mathrm{SL}(k,\mathbb{R})$ embedded diagonally in $\mathrm{SL}(k,\mathbb{R})\times\mathrm{SL}(k,\mathbb{R})$ acting on $\mathrm{SL}(k,\mathbb{Z})\times \mathrm{SL}(k,\mathbb{Z})\backslash \mathrm{SL}(k,\mathbb{R}) \times \mathrm{SL}(k,\mathbb{R})$ for $k \geq 2$. In fact in both of these examples H is a maximal subgroup, where a subgroup $H \subset G$ is called *maximal* if there is no subgroup $S \subset G$ containing H with dimension strictly between the dimensions of G and H.

6.1. **Maximal subgroup theorem.** In joint work with Margulis and Venkatesh we proved last year [7] the following theorem[8].

Theorem 1 ([7], simpler form)**.** *Let $\Gamma, H \subset G$ be as above. Assume that H is a maximal subgroup of G.*

There exists $\delta > 0$ depending only on G, H so that the Haar measure m_{x_0H} on a closed orbit x_0H is $\mathrm{Vol}^{-\delta}$-close to m_X, i.e. for any $f \in C_c^\infty(X)$ we have

$$\left|\int_{x_0H} f - \int_X f\right| \ll \mathrm{Vol}^{-\delta}\, S(f),$$

where Vol *denotes the volume*[9] *of the orbit x_0H .*

Crucial input: This theorem has as the major input the **spectral gap** for the **H-action on** $L^2(x_0H)$ in a **uniform** way for all possible closed orbits x_0H, i.e. δ and the implicit constant as in the discussion of effective decay of matrix coefficients (6) are not allowed to depend on x_0.

- If H has property (T) as e.g. for $H = \mathrm{SL}(3,\mathbb{R})$, this holds always.
- If H does not have (T) as e.g. for $H = \mathrm{SO}(2,1)(\mathbb{R})^\circ$, the required statement is property (τ) as established by Clozel [5] (building on work of Burger and Sarnak [4]). This is where the congruence assumption on Γ is crucial, see [7, Sect. 6].

6.2. **A comment about the proof.** Our proof has little to do with the outline for the horocycle flow in §3.2.2, instead may be viewed as an effective version of the measure classification theorem by Ratner and the limiting distribution theorem due to Mozes and Shah (in the semisimple case considered here). It uses a version of the effective ergodic theorem discussed in §5 (where the measure m_X is replaced by m_{x_0H}). The difference of the effective ergodic theorem in [7, Prop. 9.2] and what we discussed in §5 is that in the former the average is not taken over initial intervals $[0,T]$ but rather over long intervals very far away from the origin. More precisely, in [7, Prop. 9.2] an error is obtained for the average of f over the interval $[T^M,(T+1)^M]$ which roughly speaking has length T^{M-1}, and this error holds for all points but those in a set of small measure. This is desirable, as the divergence of two nearby points under a unipotent one-parameter subgroup in H is determined by a polynomial. If this polynomial is unifomly bounded on $[0,(T+1)^M]$ then it is nearly constant on the interval $[T^M,(T+1)^M]$. This allows the effectivization of a

[8] The first simpler version of the theorem was presented by Margulis in several talks before our joint work and may also be approachable by other methods. In fact most of the work in [7] goes into the discussion of possible intermediate subgroups where the argument becomes more involved, see Theorem 2.

[9] The volume Vol is calculated in comparison with a fixed Haar measure on H, but the Haar measure m_{x_0H} is normalized to be a probability measure.

particular argument that appears in Ratner's work (the combination of the ergodic theorem and polynomial divergence for unipotent orbits, see [14] and [15, pg. 244]), we refer to [6] or [7, Sect. 2] for the ineffective argument in precisely the context we need here.

6.3. More general version.

The more general version of our theorem is the following.

Theorem 2 ([7], current form)**.** *Let $\Gamma, H \subset G$ be as above. Assume that H has finite centralizer in G. There exists $\delta > 0$ depending only on G, H and $V_0 > 0$ depending only on Γ, G, H so that, for any $V \geq V_0$ and any closed orbit $x_0 H$ there exists an intermediate subgroup $H \subseteq S \subseteq G$ for which*

- *$x_0 S$ is a closed S-orbit with volume $< V$, and*
- *the Haar measure on $x_0 H$ is $V^{-\delta}$-close to the Haar measure on $x_0 S$, i.e. for any $f \in C_c^\infty(X)$ we have*

$$\left| \int_{x_0 H} f - \int_{x_0 S} f \right| < V^{-\delta} S(f).$$

One may read this statement as follows: If $V = \mathrm{Vol}(x_0 H)$ is very large we may apply the theorem to this parameter and obtain some bigger group $S \supsetneq H$. The orbit $x_0 S$ of the higher dimensional group S has finite volume $V' = \mathrm{Vol}(x_0 S)$ (w.r.t. to a Haar measure on S) and should be thought of as being less complicated since $V' < V$. However, V' may still be large ($x_0 S$ may still be complicated), so that one may want to apply the theorem to the parameter V' to obtain a different group $S' \supsetneq S$ whose orbit $x_0 S'$ has smaller volume (is less complicated) at the cost of obtaining a worse error statement. This may be continued until the volume of the orbit of some group becomes less than V_0 (e.g. if $S'' = G$).

6.3.1. *Visualization on $\mathbb{T}^3$.*

A toy model for this problem of intermediate orbits is the image of long rational line $L \subset \mathbb{R}^3$ in a 3-dimensional torus $L/\mathbb{Z}^3 \subset \mathbb{T}^3$. It is determined $L = \mathbb{R}\mathbf{n}$ by a single primitive vector $\mathbf{n} \in \mathbb{Z}^3$ and the length of the closed circle $L/\mathbb{Z}^3$ is precisely $\|\mathbf{n}\|$. As we mentioned in §3.2 it is quite easy to establish an effective error for the distribution properties of a rational torus in $\mathbb{T}^2$. However, unlike the case of a rational line in $\mathbb{T}^2$ the circle $L/\mathbb{Z}^3$ is contained in rational planes $P \subset \mathbb{R}^3$. A rational plane is determined by a primitive orthogonal vector $\mathbf{v} \in \mathbb{Z}^3$, and one may check that $\|\mathbf{v}\|$ equals the area of the image torus $P/\mathbb{Z}^3$ — we will also think of $\|\mathbf{v}\|$ as a measure of the complexity of $P/\mathbb{Z}^3$ inside $\mathbb{T}^3$. If $\|\mathbf{v}\|$ is much smaller than $\|\mathbf{n}\|$ for some choice of the plane, then an effective error with an error determined by $\|\mathbf{n}\|$ can only be given if we compare the Lebesgue measure on the circle $L/\mathbb{Z}^3$ to the Lebesgue measure on the two-dimensional subtorus $P/\mathbb{Z}^3$. If $\|\mathbf{v}\|$ is also big (for all rational planes containing L), then one can also compare the Lebesgue measure on the circle $L/\mathbb{Z}^3$ to the Lebesgue measure on $\mathbb{T}^3$ but the error would be expressed in terms of the smallest $\|\mathbf{v}\|$.

7. Transportation of spectral gap

7.1. Hecke correspondences.

The above theorem (in fact the maximal case in Theorem 1) may be used in the context of $G = \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SL}_2(\mathbb{R})$ with Γ equal to the product of $\mathrm{SL}_2(\mathbb{Z})$ with itself. Then the Hecke correspondence T_p^n (roughly speaking) corresponds to big volume orbits inside $X = \Gamma\backslash G$ with respect to the diagonal subgroup isomorphic to $\mathrm{SL}_2(\mathbb{R})$, i.e. $H_\Delta = \{(g, g) : g \in \mathrm{SL}_2(\mathbb{R})\}$. These orbits are isomorphic to congruence quotients of $\mathrm{SL}_2(\mathbb{R})$. The uniform effective decay of matrix coefficients (which comes from Selberg's theorem) for the action of H_Δ then implies a bound for the eigenvalue of the Hecke operator T_p. In that sense, the theorem allows us to **transport** the spectral gap from one place to another (in this case from ∞ to p).

7.2. **General setup.** Another instance of this **transportation of spectral gap** can be set up as follows.

Let G_1, G_2 be simple groups, and suppose G_2 has (T) but G_1 has not. Let Γ be an irreducible lattice in $G = G_1 \times G_2$, e.g. this is possible for $G_1 = \mathrm{SU}(2,1)(\mathbb{R})$ and $G_2 = \mathrm{SL}_3(\mathbb{R})$. As we discussed in §4 (resp. §4.1 in the case of $\mathrm{SL}(3,\mathbb{R})$) we then have effective decay of matrix coefficients for the action of G_2.

We wish to bound the matrix coefficients of G_1 acting on $X = \Gamma\backslash G$. Let $H_\Delta = \{(g,g) : g \in G\}$. Notice that the diagonal orbit $\Gamma \times \Gamma H_\Delta \subset X \times X$ is 'responsible' for the inner product in the sense that the integral of $f_1 \otimes \bar{f}_2$ over this orbit equals the inner product $\langle f_1, f_2\rangle$. In the same sense is the deformed orbit $\Gamma \times \Gamma H_\Delta(g,e)$ responsible for the matrix coefficients of g, i.e.

$$\int f_1 \otimes \bar{f}_2 dm_{\Gamma\times\Gamma H_\Delta(g,e)} = \int_X f_1(xg)\bar{f}_2(x) d_{m_X}(x) = \langle g \cdot f_1, f_2\rangle.$$

The volume of the deformed orbit $\Gamma \times \Gamma H_\Delta(g,e)$ is roughly speaking a power of $\|g\|$, more precisely bounded from above and below by multiples of powers of $\|g\|$. Hence effective equidistribution of the Haar measures on these orbits to the Haar measure on $X \times X$ gives effective decay of matrix coefficients.

However, notice that the theorem does not apply as the group giving the closed orbit has been conjugated and does not remain fixed. (In the theorem the rate of equidistribution is allowed to depend on the group H, which is changing in this case.)

On a positive side, if $g = (g_1, e)$ then the simple factor of H_Δ corresponding to G_2 remains (as a subgroup of $G \times G$) fixed and this is the part with known effective decay. In this case the method behind the proof of the theorem can be used to show effective equidistribution and so decay of matrix coefficients for the G_1-action. In all of this, the rate (i.e. the δ appearing in the discussion) of decay of matrix coefficients for G_1 only depends on the spectral gap for G_2 (but not on Γ).

7.3. **Effective equidistribution implies a weak form of (τ).** Using the above construction for a p-adic group G_2, one can prove a weaker version of property (τ) for all simple algebraic groups of absolute higher rank (i.e. all groups except forms of SL_2 for which property (τ) has been known much longer).

So let $\mathbf{G}$ be a simple, simply connected algebraic $\mathbb{Q}$-group of absolute rank ≥ 2. Let $G_1 = \mathbf{G}(\mathbb{R})$, $G_2 = \mathbf{G}(\mathbb{Q}_p)$, and let Γ be commensurable with $\mathbf{G}(\mathbb{Z}[\frac{1}{p}])$, then $L^2(\Gamma_1\backslash G)$ (with $\Gamma_1 = \Gamma \cap \mathbf{G}(\mathbb{Z}_p)$) is contained in $L^2(\Gamma\backslash G_1 \times G_2)$. We choose p such that G_2 has $\mathbb{Q}_p$-rank ≥ 2. This gives that G_2 has property (T), and so also effective decay of matrix coefficients. The latter is the only input to the method which establishes the result.

Hence the spectral gap of the G_2-action and its independence from Γ gives also some spectral gap of the G_1-action on $\Gamma_1\backslash G_1$ and in a uniform way (as long as the lattice in G_1 can be obtained from a lattice in $G_1 \times G_2$ by intersection which is always possible for congruence subgroups). We then obtain a proof of uniform spectral gap of the action of G_1, a version of property (τ). We note that the gap hereby obtained is probably quite bad in comparison to what Clozel obtained in [5]. This is part of an ongoing joint work with Margulis and Venkatesh.

References

[1] J. Bernstein and A. Reznikov. *Sobolev norms of automorphic functionals.* Int. Math. Res. Not. 2002, no. 40, 2155–2174.

[2] M. Björklund and A. Fish. *Equidistribution of Dilations of Polynomial Curves in Nilmanifolds.* To appear in Proc. AMS.

[3] M. Burger. Horocycle flow on geometrically finite surfaces. Duke Math. J. 61 (1990), no. 3, 779–803.

[4] M. Burger and P. Sarnak. Ramanujan duals. II. *Invent. Math.*, 106(1):1–11, 1991.

[5] L. Clozel. Démonstration de la conjecture τ. *Invent. Math.*, 151(2):297–328, 2003.

[6] M. Einsiedler. Ratner's theorem on $\mathrm{SL}(2,\mathbb{R})$-invariant measures. *Jahresber. Deutsch. Math.-Verein.* 108 (2006), no. 3, 143–164.

[7] M. Einsiedler, G. Margulis and A. Venkatesh. *Effective results for closed orbits of semisimple groups on homogeneous spaces.* To appear in Invent. Math.

[8] L. Flaminio and G. Forni. *Invariant distributions and time averages for horocycle flows.* Duke Math. Journal, 119: 465526, 2003.

[9] H. Furstenberg *The Unique Ergodicity of the Horocycle Flow.* pp. 95115, Lecture Notes, 318. Springer Berlin, 1972.

[10] L. Grafakos. *Classical and Modern Fourier Analysis.* Pearson, Prentice Hall, 2004.

[11] B. Green and T. Tao. *The quantitative behaviour of polynomial orbits on nilmanifolds.* Preprint 2008.

[12] D. Kleinbock and G. Margulis. *Flows on homogeneous spaces and Diophantine approximation on manifolds.* Ann. Math. 148 (1998), 339360.

[13] S. Mozes, N. Shah. *On the space of ergodic invariant measures of unipotent flows.* Ergodic Theory Dynam. Systems 15 (1995), no. 1, 149–159.

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

[15] M. Ratner. *Raghunathan's topological conjecture and distributions of unipotent flows.* Duke Math. J. 63 (1991), no. 1, 235–280.

[16] P. Sarnak. *Asymptotic behavior of periodic orbits of the horocycle flow and Eisenstein series.* Comm. Pure Appl. Math. 34 (1981), 719–739.

[17] A. Venkatesh. Sparse equidistribution problems, period bounds, and subconvexity. preprint, arxiv:math/0506224.