NON-MULTIPLICATIVITY OF FOURIER COEFFICIENTS OF MODULAR FORMS FOR NON-ARITHMETIC GROUPS: A THEOREM OF A. VENKATESH

E. KOWALSKI

1. Introduction

In this note, we give the proof (due to A. Venkatesh) of a result which states, informally, that the Fourier coefficients occuring in the Fourier expansion of a cusp form for a *non-arithmetic* fuchsian group with finite covolume in $\mathrm{SL}_2(\mathbf{R})$ are not multiplicative.

The motivation for this result is as a comparison with the fact that coefficients of cusp forms for arithmetic groups, say congruence subgroups $\Gamma_0(q)$, where $q \geqslant 1$ is an integer, are multiplicative, or short linear combinations of multiplicative functions. This property is of course a consequence of the theory of Hecke operators and primitive forms (Atkin-Lehner theory). On the other hand, Duke and Iwaniec prove in [1] that if f is a holomorphic cusp form of *half-integral* weight $k = \ell + \frac{1}{2}$ with $\ell \geqslant 2$ an integer, with Fourier expansion

$$f(z) = \sum_{n \geqslant 1} \lambda(n) n^{(k-1)/2} e(nz),$$

then the function $n \mapsto \lambda(n)$ is *very far* from multiplicative. Indeed, they prove a bilinear form estimate

$$\sum_{n \leqslant X} \sideset{}{^\flat}\sum_{m \leqslant Y} a_m b_n \lambda(mn) \ll (X^{1/2} + X^{1/4} Y)(XY)^{\varepsilon} \|a\| \|b\|$$

for any complex numbers (a_m), (b_n), X, $Y \geqslant 2$, $\varepsilon > 0$, the implied constant depending only on f and ε. This is quite strong: indeed, they deduce ([1, (9)]) from it an estimate for sums over primes

$$\sum_{p \leqslant X} \lambda(p) \ll X^{155/156+\varepsilon}$$

which, if it were known for a non-zero integral weight modular form on a congruence group, would imply the existence of a zero-free strip (of width 1/156) for its Hecke L-function.

The lack of Hecke theory for non-arithmetic groups strongly suggests that such a result should hold for a cusp form for such a group, but the method in [1], though it looks applicable (it is based on the Peterson formula and properties of Kloosterman sums) does not, as of yet, lead to the result.

However, after I mentioned the problem in a lecture in Montréal in 2004, A. Venkatesh quickly found a proof, based on methods of ergodic theory, for the following *qualitative* form of non-multiplicativity:

Theorem 1 (Venkatesh)**.** *Let $\Gamma \subset \mathrm{SL}_2(\mathbf{R})$ be a cofinite, non co-compact, subgroup of $\mathrm{SL}_2(\mathbf{R})$, and assume Γ is not arithmetic. Let $f \neq 0$ be a holomorphic cusp form of weight k for Γ. Let $\mathfrak{a}$ be a cusp of Γ and $\lambda_{\mathfrak{a}}(n)$ the Fourier coefficients in the expansion of f around the cusp $\mathfrak{a}$.*

(1) *We have*

$$\sum_n |\lambda_{\mathfrak{a}}(n)|^2 e^{-n/X} \gg X^k, \tag{1}$$

for $X \geqslant 1$, the implied constant depending on f.

(2) *For all integers $d \geqslant 2$, we have*

$$\lim_{X\to+\infty} \frac{1}{X^k} \sum_n \lambda_{\mathfrak{a}}(n)\overline{\lambda_{\mathfrak{a}}(nd)} e^{-n/X} = 0.$$

Remark 2. (1) The lower bound is well-known (it also follows from the method of proof). It shows clearly that the conclusion indicates that $\lambda_{\mathfrak{a}}(nd)$ is very far from being usually proportional to $\lambda_{\mathfrak{a}}(n)$ (unless if the proportionality factor was zero, i.e., unless $\lambda_{\mathfrak{a}}(nd)$ is usually extremely small).

(2) It would be very interesting to find a quantitative version of the theorem, say for non-arithmetic Hecke triangle groups, either by adapting the method of [1], or by quantifying the proof of Venkatesh.

(3) Obviously, any mistake or misunderstanding in the statement or proof are only mine, and not due to Venkatesh.

2. Proof of the theorem

First of all, fix a scaling matrix $\sigma_{\mathfrak{a}}$ so that the Fourier expansion of f at $\mathfrak{a}$ takes the form

$$f(z) = \sum_{n\geqslant 1} \lambda_{\mathfrak{a}}(n) e(n\sigma_{\mathfrak{a}} z).$$

where $e(z) = e^{2i\pi z}$ for $z \in \mathbf{C}$. Up to replacing Γ with $\sigma_{\mathfrak{a}}\Gamma\sigma_{\mathfrak{a}}^{-1}$, we may assume that the cusp is at infinity with width 1, and that the Fourier expansion is

$$f(z) = \sum_{n\geqslant 1} \lambda_{\mathfrak{a}}(n) e(nz).$$

Fix an integer $d \geqslant 2$ and denote

$$S(X) = \sum_{n\geqslant 1} \lambda_{\mathfrak{a}}(n)\overline{\lambda_{\mathfrak{a}}(nd)} e^{-n/X}.$$

From the Fourier expansion, we derive the integral representation

$$S(X) = \int_0^1 f\Bigl(t + \frac{i}{4\pi X}\Bigr)\overline{f\Bigl(dt + \frac{i}{4\pi X}\Bigr)}dt.$$

of the sum.

Now the crucial step is to express this in group-theoretic terms using unipotent orbits, in order to be able to apply the understanding of such actions arising from ergodic theory (Ratner theory).

Denote as usual $j(g, z) = (cz + d)$ for a matrix g in $\mathrm{SL}_2(\mathbf{R})$ with second row $(c\ d)$. Let $G = \mathrm{SL}_2(\mathbf{R}) \times \mathrm{SL}_2(\mathbf{R})$ and $\Gamma_2 = \Gamma \times \Gamma \subset G$. We consider the function φ defined on G by

$$\varphi(g_1, g_2) = f(g_1 \cdot i)j(g_1, i)^{-k}\overline{f(g_2 \cdot i)j(g_2, i)^{-k}}$$

or

$$\varphi\Bigl(\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1\end{pmatrix}, \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2\end{pmatrix}\Bigr) = f\Bigl(\frac{a_1 i + b_1}{c_1 i + d_1}\Bigr)\overline{f\Bigl(\frac{a_2 i + b_2}{c_2 i + d_2}\Bigr)}(c_1 i + d)^{-k}\overline{(c_2 i + d)^{-k}}.$$

The fact that f is a modular form of weight k implies that φ is a function on $\Gamma_2\backslash G$. It is continuous and bounded (because f is a cusp form). Because f is cuspidal, we also have

$$\int_{\Gamma_2\backslash G} \varphi(g_1, g_2)d\mu = 0$$

where μ is the product of Haar measures on $\Gamma_2\backslash G$.

Furthermore, for any $t \in [0, 1]$ and $X > 0$, we have

$$f\Big(t + \frac{i}{X}\Big)\overline{f\Big(dt + \frac{i}{X}\Big)} = f(g_1 \cdot i, g_2 \cdot i)$$

where

$$g_1 = \begin{pmatrix} X^{-1/2} & tX^{1/2} \\ 0 & X^{1/2} \end{pmatrix}, \qquad g_2 = \begin{pmatrix} X^{-1/2} & dtX^{1/2} \\ 0 & X^{1/2} \end{pmatrix}.$$

Hence we find

$$f\Big(t + \frac{i}{X}\Big)\overline{f\Big(dt + \frac{i}{X}\Big)} = X^k\varphi(g_1, g_2).$$

Observe now that

$$(g_1, g_2) = (x_t, a_d x_t a_d^{-1})m(X)$$

where

$$m(X) = \left(\begin{pmatrix} 1/\sqrt{X} & 0 \\ 0 & \sqrt{X} \end{pmatrix}, \begin{pmatrix} 1/\sqrt{X} & 0 \\ 0 & \sqrt{X} \end{pmatrix}\right)$$

$$x_t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad a_d = \begin{pmatrix} d & 0 \\ 0 & 1 \end{pmatrix}.$$

Therefore we can express the sum as

$$S(X) = X^k \int_0^1 \varphi((x_t, a_d x_t a_d^{-1})m(4\pi X))dt.$$

The point is that we have expressed $S(X)$ as an integral of an automorphic function on $\Gamma_2\backslash G$ along a one-parameter unipotent subgroup. Since the normalization X^k has appeared naturally, the goal is now simply to prove that

$$\lim_{X\to+\infty} \int_0^1 \varphi((x_t, a_d x_t a_d^{-1})m(4\pi X))dt = 0 \tag{2}$$

for integers coprime to some fixed $q \geqslant 1$.

Define measures μ_X on $\Gamma_2\backslash G$ by

$$\int f d\mu_X = \int_0^1 f((x_t, a_d x_t a_d^{-1})m(X))dt.$$

Since the integral of φ over $\Gamma_2\backslash G$ is 0, the limit (2) follows from the convergence of μ_X to the Haar measure as $X \to +\infty$. It follows from results of Shah and Ratner (see [3, Th. 1.4], as exploited similarly and explained by Strömbergsson [4, §6], in the case where a_d is replaced by some unipotent matrix) that this is the case, provided a_d does not belong to the commensurator

$$C(\bar{\Gamma}) = \{g \in \mathrm{PSL}_2(\mathbf{R}) \mid [\bar{\Gamma} : \bar{\Gamma} \cap g\bar{\Gamma}g^{-1}] < +\infty, \quad [g\bar{\Gamma}g^{-1} : \bar{\Gamma} \cap g\bar{\Gamma}g^{-1}] < +\infty\}$$

of $\bar{\Gamma} = \Gamma/\{\pm 1\}$ in $\mathrm{PSL}_2(\mathbf{R})$. Since we assumed that Γ is non-arithmetic, a famous result of Margulis implies that $\bar{\Gamma}$ has finite index in $C(\bar{\Gamma})$ (see [2, Ch. IX, Th. 1.16] or [5, Th. 6.2.5]).

We now use a simple lemma:

Lemma 3. *Let Γ be a fuchsian group with finite covolume and $\mathfrak{a}$ a cusp of Γ. Then Γ does not contain a non-identity hyperbolic matrix g of $\mathrm{SL}_2(\mathbf{R})$ fixing $\mathfrak{a}$.*

Proof. We may assume that $\mathfrak{a} = \infty$ and that the width is 1. There exist $y_0 > 0$ such that the projection

$$\{z = x + iy \in \mathbf{H} \mid y > y_0 \text{ and } |x| < 1/2\} \to \Gamma\backslash\mathbf{H}$$

is injective. A non-identity hyperbolic matrix g of $\mathrm{SL}_2(\mathbf{R})$ fixing ∞ has the form

$$g = \pm \begin{pmatrix} u^{1/2} & v \\ 0 & u^{-1/2} \end{pmatrix}$$

for some $u > 0$ and $v \in \mathbf{R}$. If there is such an element in Γ that is not the identity, then up to inverting it, we may assume that it satisfies $u > 1$. This element g then acts on $\mathbf{H}$ by $z \mapsto uz + \sqrt{u}v$. So it identifies iy with $iuy + \sqrt{u}v$ for $y > 0$. If y is large enough, then composing with a power of $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, it identifies iy with $iuy + w$ where $|w| < 1/2$. This is incompatible with the injectivity statement. $\square$

Applied to $C(\bar{\Gamma})$, this implies that $a_d \in C(\bar{\Gamma})$ if and only if $d = 1$ and concludes the proof.

References

[1] W. Duke and H. Iwaniec: *Bilinear forms in the Fourier coefficients of half-integral weight cusp forms and sums over primes*, Math. Ann. 286 (1990), 783–802.

[2] G. Margulis: *Discrete subgroups of semisimple Lie groups*, Ergebnisse der Math. (3) 17, Springer Verlag 1991.

[3] N. Shah: *Limit distribution of expanding translates of certain orbits on homogeneous spaces*, Proc. Indian Acad. Sci. (Math. Sci.) 106 (1996), 105–125.

[4] A. Strömbergsson: *On the uniform equidistribution of long closed horocycles*, Duke Math. J. 123 (2004), 507–547.

[5] R. Zimmer: *Ergodic theory and semisimple groups*, Monographs in Math., Birkhäuser, 1984.

ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland

E-mail address: kowalski@math.ethz.ch