Automorphic L-functions, Geometric Invariants, and Dynamics

Thesis by
Alexandre Perozim de Faveri

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
Defended May 16, 2022

Alexandre Perozim de Faveri
ORCID: 0000-0001-7180-9382

ACKNOWLEDGEMENTS

This thesis would not have been possible without the outstanding mentorship of my advisor Maksym Radziwiłł. I thank him for the generosity with his time, for the valuable advice and guidance, for a vast number of inspiring conversations, and for introducing me to so many interesting problems and ideas.

I am immensely grateful to my teachers Ilca Brisante, Carlos Shine, Régis Barbosa, Edmilson Motta, and Peter Sarnak, without whom I would not be here.

I would also like to thank Alex Dunn, Philippe Michel, and Dinakar Ramakrishnan for being part of my thesis committee.

Thanks to my Caltech friends for making my time in Pasadena so enjoyable and exciting. You were responsible for so many of my favorite memories of the last years, and I am extremely happy to have you in my life!

Finally, I would like to thank my parents and sister for their unconditional love and support, which transcends distance, defies description, and constantly propels me forward. I dedicate this thesis to them.

ABSTRACT

We address three different problems in analytic number theory.

In the first part, we show that the completed L-function of a modular form has $\Omega(T^\delta)$ simple zeros with imaginary part in $[-T, T]$, for any $\delta < \frac{2}{27}$. This is the first power bound for forms with non-trivial level in this problem, where previously the best result was $\Omega(\log\log\log T)$. Along the way, we also improve the corresponding bound in the case of trivial level, and sharpen a certain zero-density result.

In the second part, we study the variance for the distribution of closed geodesics in random balls on the modular surface. A probabilistic model in which closed geodesics are modeled using random geodesic segments is proposed, and we rigorously analyze this model using mixing of the geodesic flow. This leads to a conjecture for the asymptotic behavior of the variance, and we prove this conjecture for sufficiently small balls.

In the third part, we prove Sarnak's Möbius disjointness conjecture for $C^{1+\varepsilon}$ skew products on the 2-torus over a rotation of the circle.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] A. de Faveri. Simple zeros of GL(2) L-functions. *arXiv:2109.15311*, 2021.

[2] A. de Faveri. The variance of closed geodesics in balls and annuli on the modular surface. *Adv. Math.*, 403, 2022. DOI: 10.1016/j.aim.2022.108390.

[3] A. de Faveri. Möbius disjointness for $C^{1+\varepsilon}$ skew products. *Int. Math. Res. Not. IMRN*, (4):2513–2531, 2022. DOI: 10.1093/imrn/rnaa185.

Alexandre de Faveri conducted all of the research and authored each of the manuscripts listed above.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

1.1 General remarks

This thesis addresses three different problems in the intersection of analytic number theory with adjacent areas, including harmonic analysis, automorphic forms, spectral theory, and dynamical systems.

The first two chapters are focused on the study of automorphic L-functions through analytic methods, and on subsequent applications. This includes understanding properties of their zeros in Chapter 2 and Appendix A, as well as examining their large values through moments and subconvexity and applying such results to arithmetic questions in Chapter 3.

Results along these lines are arguably interesting for their own sake, due to their close connection to central open problems in analytic number theory such as the Riemann Hypothesis. Furthermore, they also often have important applications in a wide array of topics, ranging from fine questions about the distribution of prime numbers to problems in the interface of number theory with geometry and spectral theory (e.g. equidistribution of geometric invariants, quantum unique ergodicity, etc).

Chapter 4 investigates an instance of the Möbius disjointness conjecture of Sarnak. This conjecture, which combines multiplicative number theory and dynamical systems, has seen spectacular progress in recent years. Major results include the Matomäki-Radziwiłł theorem on multiplicative functions in short intervals [82], Tao's subsequent resolution of the two-point logarithmic Chowla conjecture [103] and of the Erdös discrepancy problem [102], and the work of Frantzikinakis and Host establishing the logarithmic Sarnak conjecture for uniquely ergodic systems [37].

We now describe the results of each chapter in more detail and provide some background on each of the questions investigated, referring to the introduction of the corresponding chapters for precise statements.

1.2 Context and description of results

Chapter 2: Simple zeros

The study of zeros of automorphic L-functions is one of the central areas of analytic number theory. While the horizontal distribution of such zeros is (conjecturally) given by the Grand Riemann Hypothesis (GRH), the vertical distribution of zeros was only more recently understood, and has been a topic of intense research in the last 30 years. Inspired by the Pair Correlation Conjecture of Montgomery [88], a rich correspondence with random matrix theory has been developed by many authors [67, 66, 94]. This has led to very precise conjectures for the distribution of zeros (and also for moments [20]) of L-functions, which are supported by both theoretical [95] and computational [90] evidence.

In a landmark result, Selberg [99] pioneered the use of mollifiers to show that a (small) positive proportion of the zeros of ζ have real part $\frac{1}{2}$. When it comes to the vertical distribution of zeros, one of the simplest observations one may hope to prove is the Grand Simplicity Hypothesis (GSH), which says that all the zeros should be simple, apart from at most one exception (connected to the BSD conjecture). The first unconditional result in that direction is due to Levinson [76] (combined with observations of Selberg and Heath-Brown [49]), who showed that at least one third of the zeros of ζ are both simple and on the central line. There have been many improvements on the proportion obtained through Levinson's method, using tools from harmonic analysis and spectral theory [92, 35, 19, 18]. The result can also be generalized to Dirichlet L-functions [4].

The situation is different already for degree 2 automorphic L-functions. While both the methods of Selberg [45, 46, 47] and of Levinson [2] generalize to show that a positive proportion of the zeros satisfy GRH, they are not able to deal with simple zeros. It is only known that a positive proportion of the zeros have order at most 3 [34], and even under GRH it is an open problem to obtain a positive proportion of simple zeros [84].

Let $f \in S_k(\Gamma_1(N))$ be a primitive holomorphic form, and let $N_f^s(T)$ denote the number of simple zeros of the completed L-function of f with imaginary part in $[-T, T]$. In 1988, Conrey and Ghosh [21] developed a new method for detecting simple zeros, and were able to show that if $f = \Delta$ is the Ramanujan function, then $N_f^s(T) = \Omega(T^{\frac{1}{6}-\varepsilon})$ for any $\varepsilon > 0$. Their method applies to any

f of level $N = 1$, as long as one assumes the existence of at least one simple zero (which they verified for $f = \Delta$).

It was only after a breakthrough of Booker [7] in 2012 that the existence of a simple zero for arbitrary $f \in S_k(\Gamma_1(N))$ was established (in fact he obtains $N_f^s(T) \to \infty$ as $T \to \infty$). We remark that Booker's method has applications to other important problems. For instance, related ideas were used to show that the Artin conjecture for a given 2-dimensional Galois representation over $\mathbb{Q}$ implies the Langlands modularity conjecture for the corresponding L-function [6], and also to strengthen the converse theorem [9, 11, 10].

While Booker's result combined with the work of Conrey and Ghosh gives $N_f^s(T) = \Omega(T^{\frac{1}{6}-\varepsilon})$ for f of level $N = 1$, in general one runs into issues related to the level N that are reminiscent of the difficulties in extending the Hecke converse theorem to non-trivial level. Booker, Milinovich, and Ng [12] made Booker's result quantitative for N odd, obtaining $N_f^s(T) = \Omega(\log\log\log T)$.

In Chapter 2 we remove the parity restriction on N, and overcome the limitations coming from the level, leading to the first power bound for the number of simple zeros when f has non-trivial level. More precisely, we show that if $f \in S_k(\Gamma_1(N))$ is primitive of arbitrary weight k and level N, then $N_f^s(T) = \Omega\left(T^{\delta}\right)$ for any $\delta < \frac{2}{27}$.

The main novel ingredient is the use of zero-density estimates for the family of character twists of f, in order to control certain pole cancellations. Philosophically, one may interpret this extra control over twists as precisely the kind of ingredient that allows Weil to generalize the converse theorem to general level, and something analogous is true for simple zeros (though the mechanisms are somewhat different). Finally, we also improve the exponent in the result of Conrey and Ghosh from $\frac{1}{6}$ to $\frac{1}{5}$ for $f \in S_k(\Gamma_1(1))$.

The proofs of the two results described above lead to new applications for some open problems that may be in reach of current techniques, including a certain non-vanishing problem for twists of a fixed form, and a generalization of the sixth moment bound of Jutila [63] for general level (the corresponding subconvex bound has been obtained recently [13], but the moment bound remains open). It should also be possible to extend the results to the case of Maass forms, along the lines of work of Booker, Cho, and Kim [8], since the proofs do not rely on the Ramanujan conjecture.

Chapter 3: Closed geodesics

The study of statistical properties of closed geodesics in negatively curved surfaces lies at the intersection of geometry, ergodic theory, and spectral theory. Striking results have been obtained in the last two decades (for instance through work of Mirzakhani [85, 86]), and the field continues to be an active area of research [72, 33, 62]. The modern theory of automorphic L-functions is – somewhat surprisingly – a powerful tool in that direction, and has important consequences for the topic in the case of arithmetic manifolds, where one can prove very precise results.

There is a vast literature in analytic number theory connecting subconvex bounds for L-functions to equidistribution of various geometric invariants [29, 30, 48, 53, 65]. The prototypical example concerns the equidistribution of the integral solutions to $a^2 + b^2 + c^2 = D$ on the surface of the unit sphere (after projection), for large fundamental D. This was established independently by Duke [28] and Golubeva-Fomenko [41], after a breakthrough of Iwaniec [57] which can be interpreted as a subconvex bound for quadratic twists of a fixed half-integral weight form. This kind of equidistribution theorem has also been obtained for many other geometric invariants, such as those associated to the quadratic form $b^2 - 4ac = D$ (namely Heegner points and closed geodesics on the modular surface $\Gamma\backslash\mathbb{H}$, depending on the sign of D).

Let Λ_D denote the set of closed geodesics of discriminant $D > 0$ (which we assume to be squarefree and fundamental from now on). Equidistribution in balls B_R means in this case that for any fixed $w \in \Gamma\backslash\mathbb{H}$, we have

$$\sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C} \cap B_R(w)) \sim \frac{\mu(B_R)}{\mu(\Gamma\backslash\mathbb{H})} \sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C}) \tag{1.1}$$

as $D \to \infty$, where μ is the hyperbolic measure and ℓ denotes the hyperbolic length. Such a result holds if $D^{-\frac{1}{18}+\varepsilon} \ll R \ll 1$ by work of Humphries [53] and Young [106], and it is conjectured that $\frac{1}{18}$ can be replaced with $\frac{1}{2}$.

If one is interested on a result for almost every ball, it is natural to vary $w \in \Gamma\backslash\mathbb{H}$ randomly according to μ, and consider the random variable given by the LHS of (1.1). It is tautological that the expected value is equal to the RHS of the same equation. One is then led to consider the variance

$$\mathrm{Var}(R;\Lambda_D) := \frac{1}{\mu(\Gamma\backslash\mathbb{H})} \int_{\Gamma\backslash\mathbb{H}} \left(\sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C} \cap B_R(w)) - \frac{\mu(B_R)}{\mu(\Gamma\backslash\mathbb{H})} \sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C}) \right)^2 d\mu(w).$$

Such an expression was first studied by Bourgain, Rudnick, and Sarnak [16] in the context of lattice points in spheres. Based on probabilistic considerations, they conjectured that if the radii satisfy certain mild conditions, then the variance should be asymptotically equal to the corresponding expected value of the underlying random variable.

Humphries and Radziwiłł [54] were able to unconditionally prove this conjecture when one replaces the balls B_R by very thin annuli, both in the case of lattice points in spheres and of Heegner points in $\mathbb{H}$. In the case of closed geodesics, they obtained equidistribution for almost all balls in the full range $D^{-\frac{1}{2}+\varepsilon} \ll R \ll 1$, but did not compute the variance.

While for Heegner points and lattice points on spheres there is a very simple probabilistic model which allows one to easily conjecture the correct asymptotics for the variance, this is not the case for closed geodesics, and a priori it is not clear what one should expect. In Chapter 3 we propose a probabilistic model based on geodesic segments of the appropriate length taken at random according to the Liouville measure in the unit tangent bundle of $\Gamma\backslash\mathbb{H}$. Using a quantitative version [81] of Ratner's theorem on exponential mixing of the geodesic flow for the modular surface [93], we rigorously analyze this random model.

This leads to a conjecture for the asymptotics of $\mathrm{Var}(R;\Lambda_D)$, which is no longer equal to the expected value. Finally, the main result of Chapter 3 is a proof of this conjecture for sufficiently small balls, using the methods of Humphries and Radziwiłł. Namely, we show that if $0 < R \leq D^{-\frac{5}{12}-\varepsilon}$, then

$$\mathrm{Var}(R;\Lambda_D) \sim \frac{64\sqrt{D}L(1,\chi_D)R^3}{\pi}$$

as $D \to \infty$ through squarefree fundamental discriminants.

We also obtain the asymptotics for a wide class of annuli in place of the balls B_R, and interestingly the variance depends on the shape of the annulus, and not only on its area (since a certain special function appears in the asymptotics, exactly matching the value predicted by the random model).

The exponent $\frac{5}{12}$ in the range for the variance asymptotics is limited by the best bound available for a certain first moment of twisted GL(2) L-functions, ultimately coming from the work of Young [106], based on a landmark result of Conrey and Iwaniec [22].

Chapter 4: Möbius disjointness

The Möbius disjointness conjecture of Sarnak [96, 97] roughly predicts that if $f : \mathbb{N} \to \mathbb{C}$ is a function of "low complexity", then it should not correlate with the Möbius function, in the sense that

$$\sum_{n=1}^{N} f(n)\mu(n) = o(N).$$

Here the appropriate meaning of "low complexity" turns out to be dynamical. Namely, it is that there is a topological dynamical system (X, T) of entropy zero, a point $x \in X$, and a continuous function $g : X \to \mathbb{C}$ such that $f(n) = g(T^n x)$.

The Sarnak conjecture follows from the Chowla conjecture, and its importance stems from the fact that while it seems hard to make progress on the latter, one may use various tools (including measure classification, exponential sums, combinatorial methods, etc) to deal with specific classes of systems and make progress on the former. We also point out that establishing the conjecture in specific cases often leads to applications, such as in the work of Green and Tao [44].

Sarnak's conjecture has been proved for many dynamical systems, and a common feature of many such results is that the underlying system is regular. The first class of systems with irregular dynamics [38] for which Möbius disjointness was established were the skew products $(\mathbb{T}^2, T_{\alpha,\phi})$, where $\mathbb{T} := \mathbb{R}/\mathbb{Z}$, $\alpha \in \mathbb{R}$, $\phi : \mathbb{T} \to \mathbb{T}$ is a continuous map, and the transformation is given by

$$T_{\alpha,\phi}(x, y) := (x + \alpha, y + \phi(x))$$

for all $(x, y) \in \mathbb{T}^2$. These are the building blocks of the important class of distal flows [39], and Möbius disjointness was obtained, assuming ϕ is analytic and satisfies a certain mild property, by Liu and Sarnak [78] (see also the work of Kułaga-Przymus and Lemańczyk [71]). The restrictions on ϕ were relaxed, to analytic by Wang [104], to C^∞ by Huang, Wang, and Ye [52], and finally to $C^{2+\varepsilon}$ (with a small extra condition) by Kanigowski, Lemańczyk, and Radziwiłł [64]. In Chapter 4, we use a more refined analysis of the diophantine properties of α to further lower the smoothness condition to $C^{1+\varepsilon}$, providing the best disjointness result to date for this class of dynamical systems.

The basic method to obtain such a result is the same as that of Kanigowski, Lemańczyk, and Radziwiłł [64], which is heavily based on the breakthrough

of Matomäki and Radziwiłł [82] on multiplicative functions in short intervals. We also show that the exponent $1+\varepsilon$ is the limit of such an argument, which is puzzling since disjointness should hold just with the assumption of continuity for ϕ. The bottleneck here is a generalization of the Matomäki-Radziwiłł theorem to short arithmetic progressions, where an issue arises if for instance the modulus is a primorial (the same kind of limitation also appears in the work of Klurman, Mangerel, and Teräväinen [68]).

Chapter 2

SIMPLE ZEROS OF GL(2) L-FUNCTIONS

2.1 Introduction

Discussion

Let π be a cuspidal automorphic representation of $\mathrm{GL}(n, \mathbb{A}_{\mathbb{Q}})$ with completed L-function Λ_π. It is conjectured that all the zeros of $\Lambda_\pi(s)$ are on the critical line $\Re(s) = \frac{1}{2}$ and, apart from at most one multiple zero of algebraic origin, are all simple. For degree $n = 1$ (Dirichlet L-functions), Levinson's method [76, 49, 4, 105] shows that a positive proportion of the zeros are simultaneously simple and on the critical line. An adaptation of that method for degree $n = 2$ also implies that a positive proportion of the zeros are on the critical line [2], but already cannot tackle simple zeros and only shows that a positive proportion of the zeros are of order at most three [34].

In this chapter we consider the problem of obtaining lower bounds for the number of simple zeros in the case of degree $n = 2$. Let $f \in S_k(\Gamma_1(N))$ be a primitive form (i.e. a normalized Hecke newform) of arbitrary weight k and level N. The first challenge is to show that Λ_f has any simple zeros at all. While for a given f this can be checked computationally, the problem was only completely solved in 2012, after a breakthrough of Booker [7], who in fact showed that Λ_f has infinitely many simple zeros. The argument relies on simple zeros of local factors of Λ_f, thus differentiating it from counterexamples such as the square of a degree one L-function. Another key ingredient in Booker's method is non-vanishing of automorphic L-functions on the line $\Re(s) = 1$, more specifically applied to multiplicative twists of f, foreshadowing an important obstruction in the method.

With Booker's result in hand, the next challenge is to obtain quantitative bounds on the number of simple zeros of Λ_f. Here one runs into issues related to the level N that are somewhat reminiscent of the difficulties in extending the Hecke converse theorem to general level. As in Weil's generalization of the converse theorem, an important tool are the twists of f by multiplicative characters. However, in our case an obstruction remains. It roughly consists of the possibility that $\Lambda_f(s)$ has simple zeros arbitrarily close to the line $\Re(s) = 1$,

and in addition that a certain conspiracy between additive twists of f happens at those simple zeros — namely that (2.4) below doesn't have a pole at any of those simple zeros, for any choice of $\alpha \in \mathbb{Q}^\times$.

Let

$$N_f^s(T) := |\{\rho \in \mathbb{C} : |\Im(\rho)| \leq T \text{ and } \rho \text{ is a simple zero of } \Lambda_f\}|$$

denote the number of simple zeros of Λ_f with imaginary part in $[-T, T]$. For the case of full level ($N = 1$), it is easy to directly check that no widespread pole cancellation in (2.4) can happen. In a paper from 1988 which introduced ideas used in most subsequent works on this topic, Conrey and Ghosh [21] showed that if $f = \Delta$ is the Ramanujan function, then $N_f^s(T) = \Omega(T^{\frac{1}{6}-\varepsilon})$ for any $\varepsilon > 0$. Their method applies to any f of level $N = 1$, as long as one assumes the existence of at least one simple zero for Λ_f (which they verified for $f = \Delta$, and is now known to hold in general due to Booker's work).

For general level N, Booker, Milinovich, and Ng [12] recently showed that there exists an unspecified Dirichlet character χ, possibly depending on f, such that $N_{f\otimes\chi}^s(T) = \Omega(T^{\frac{1}{6}-\varepsilon})$ for any $\varepsilon > 0$ (see also [23] for a strong result on simple zeros of twists of f). In the same paper, the authors also used the zero-free region of Λ_f to slightly limit where pole cancellations in (2.4) can happen. As a result, they made Booker's result quantitative for f of odd level, showing that $N_f^s(T) = \Omega(\log\log\log T)$. The restriction $2 \nmid N$ comes from the prominent role played by certain additive twists by $1/2$ in their argument (the use of such twists dates back to the work of Conrey and Ghosh), relying on the fact that there are no non-trivial Dirichlet characters modulo 2.

Results

Our main result removes the parity restriction on the level, and rules out complete pole cancellation in (2.4) on a wide strip, leading to the first power bound for the number of simple zeros of Λ_f when f has non-trivial level.

Theorem 1 (Power bound for arbitrary level)**.** *Let $f \in S_k(\Gamma_0(N), \xi)$ be a primitive holomorphic modular form of arbitrary weight k, level N, and nebentypus ξ. Then*

$$N_f^s(T) = \Omega\left(T^\delta\right)$$

for any $\delta < \frac{2}{27}$.

We obtain a power bound by showing that the aforementioned complete pole cancellation in (2.4) at a simple zero ρ of Λ_f would imply that $\Lambda_{f\otimes\chi}(\rho) = 0$ for a large number of characters χ. Such an amount of vanishing can then be ruled out at points ρ close to the line $\Re(s) = 1$, using zero-density results. In order to remove the parity restriction on the level, we get the method started by producing a pole for a certain Dirichlet series via ideas of Booker [7], instead of relying on the special nature of twists by 1/2 to do so. See Section 2.1 for a sketch of both arguments.

In Appendix A, we use standard Dirichlet polynomial methods [87, 89, 55] to obtain a zero-density bound in degree two which is better in the twist aspect (hence for the application at hand) than other general results from the literature [69, 75]. It is likely that the exponent in Theorem 1 can be improved by refining this zero-density result, or better yet by dealing directly with non-vanishing at an arbitrary (but fixed) point ρ. To be more precise, the problem is to show that the number of primitive characters $\chi \pmod q$ with $q \leq Q$ such that $\Lambda_{f\otimes\chi}(\rho) = 0$ is $o(Q/\log Q)$. Using Proposition 11 we obtain this result as long as $\Re(\rho) > \frac{7}{9}$, and the challenge is to enlarge such a half-plane (for instance, the density hypothesis for the family of twists of f would allow one to replace $\frac{7}{9}$ with $\frac{3}{4}$). This type of non-vanishing problem for families has received considerable attention at the central point [101, 70, 60], but much less seems to be known in general, and we hope that providing an application will lead to further study. An important feature is that we require more than a 100% rate of non-vanishing, and in fact wish to rule out a thin set of zeros, of size less than the square-root of the size of the family.

Finally, we also improve the exponent in the result of Conrey and Ghosh [21] from 1/6 to 1/5.

Theorem 2 (Improved exponent for full level)**.** *Let $f \in S_k(\Gamma_0(1))$ be a primitive holomorphic modular form of arbitrary weight k for the full modular group. Then*

$$N_f^s(T) = \Omega\left(T^\nu\right)$$

for any $\nu < \frac{1}{5}$.

Theorem 2 comes from a simple modification of the last step of their original argument (or its reformulated version in the language of this chapter, as presented in Section 2.5). Instead of using Weyl subconvexity for Λ_f, we input

Jutila's sixth moment bound [63]. Analogous improvements in the exponent of Theorem 1 would also follow if one had a similar moment bound for f of general level, which may be accessible with current tools but we do not pursue here.

It seems likely that the methods of this chapter would apply to Maass forms as well, along the lines of work of Booker, Cho, and Kim [8]. Indeed, while we do use the Ramanujan conjecture for convenience, the argument only really requires information which is already provided by Rankin-Selberg. We restrict ourselves to holomorphic forms for simplicity.

Sketch of the argument

Let us describe the obstructions that arise when the level is non-trivial. First we must give an overview of the general method, but we shall be somewhat imprecise and use standard notations that will be familiar to the experts without further explanation, postponing the definitions until Section 2.2. The fundamental object is the Dirichlet series

$$D_f(s) := L_f(s)\left(\frac{L_f'}{L_f}\right)'(s) = \sum_{n=1}^{\infty} c_f(n)n^{-s} \qquad \text{for } \Re(s) > 1, \tag{2.1}$$

which has meromorphic continuation to $\mathbb{C}$ with poles exactly at the simple zeros of $L_f(s)$ (the incomplete L-function of f), including the trivial ones at $s = \frac{1-k}{2} - n$, for $n \in \mathbb{Z}_{\geq 0}$. It is convenient to work with the completed version $\Delta_f(s) := \Gamma_{\mathbb{C}}\left(s + \frac{k-1}{2}\right) D_f(s)$, which satisfies a certain functional equation coming from that of Λ_f.

The way we obtain information about simple zeros is using the inverse Mellin transform

$$F_f(z) := 2\sum_{n=1}^{\infty} c_f(n) n^{\frac{k-1}{2}} e(nz) = \frac{1}{2\pi i} \int\limits_{\Re(s)=2} \Delta_f(s)(-iz)^{-s-\frac{k-1}{2}}\, ds,$$

for $z \in \mathbb{H}$. Indeed, shifting the line of integration to the left of the critical strip and returning to the right via the functional equation of Δ_f, we pick up poles of Δ_f and obtain a relation of the form

$$F_f(z) = (*) \cdot F_{\overline{f}}\left(-\frac{1}{Nz}\right) + S_f(z) + (**) \tag{2.2}$$

for certain factors $(*)$ and $(**)$ that we brush aside for now. Here the poles contribute

$$S_f(z) := -\sum_{\rho} \Lambda_f'(\rho)(-iz)^{-\rho-\frac{k-1}{2}}, \tag{2.3}$$

where ρ runs over the simple zeros of Λ_f.

Understanding the size of S_f gives information about the simple zeros of Λ_f. To do so we apply a Mellin transform to (2.2) along the half-line $\Re(z) = \alpha \in \mathbb{Q}^\times$. This gives rise to additive twists of Δ_f, and in the end one obtains a relation between the Mellin transform of S_f and an expression of the form

$$\Delta_f(s,\alpha) - (***)\cdot \Delta_{\overline{f}}\left(s, -\frac{1}{N\alpha}\right) \tag{2.4}$$

for some non-vanishing factor $(***)$ which we ignore in this sketch.

The goal now becomes to show that (2.4) has a pole with large real part (i.e. at least $1/2$) for some $\alpha \in \mathbb{Q}^\times$, since then this pole gets transferred to the Mellin transform of S_f and we get a lower bound for simple zeros. As an aside, the reason why the method produces omega results is that we obtain only minimal information about the pole structure of the Mellin transform of S_f (which makes the application of Tauberian theorems difficult), as opposed to bounds for S_f itself.

Since additive twists of Δ_f are not so well-behaved, we expand them into multiplicative twists instead to understand their poles. At least for $\alpha = \frac{a}{q}$, with $q \nmid N$ a prime, we obtain

$$\Delta_f\left(s, \frac{a}{q}\right) = \Delta_f(s) + b_{\chi_0,a}\cdot \Delta_f(s,\chi_0) + \sum_{\substack{\chi \pmod q \\ \chi \neq \chi_0}} b_{\chi,a}\cdot \Delta_{f\otimes\chi}(s) \tag{2.5}$$

for certain coefficients $b_{\chi,a}$, where $\chi_0 \pmod q$ denotes the trivial character. A key point is that the term $\Delta_f(s) + b_{\chi_0,a}\cdot\Delta_f(s,\chi_0)$ has the same poles as $\Delta_f(s)$ in the interior of the critical strip.

Here it becomes clear why the case $N = 1$ is special: one may simply plug $\alpha = \frac{1}{2}$ into (2.4). Applying (2.5) and using the fact that there are no non-trivial characters modulo 2, one checks that (2.4) has the same poles as $\Delta_f(s)$ inside the critical strip (hence by the aforementioned result of Booker [7] it has at least one pole with real part greater or equal to $1/2$, and one recovers the bound of Conrey and Ghosh).

For non-trivial level, as was pointed out in [12], one encounters obstacles that are reminiscent of the difficulties in extending Hecke's converse theorem to arbitrary level. However, Booker, Milinovich, and Ng are still able to obtain a result for N odd, using not only the special nature of the choice $\alpha = \frac{1}{2}$, but also adding an extra additive twist in the outset of the problem and leveraging various choices of α against each other.

The improvements of the present chapter are twofold, and in essentially disjoint parts of the argument sketched above. To obtain a result for f of any level (without parity restrictions), instead of using twists by $1/2$ we provide in Section 2.3 a new unified way of verifying that (2.4) has poles with real part greater or equal to $1/2$ for some choice of $\alpha \in \mathbb{Q}^\times$. The idea is that it is possible to construct a linear combination of certain terms of the form (2.4) that equals

$$\Delta_f\left(s, \frac{1}{p}\right) - \Delta_f\left(s, -\frac{\overline{N}}{p}\right) \tag{2.6}$$

for a certain prime p. Then one may use techniques of Booker [7] to show that (2.6) has a pole inside the critical strip, ultimately coming from the simple zeros of local factors of Λ_f.

To upgrade such a pole inside the critical strip to one with real part greater or equal to $1/2$, we use the important feature that (2.6) was constructed specifically to satisfy a certain functional equation relating s to $1 - s$ (reminiscent of Voronoi summation). Thus the poles of (2.6) inside the critical strip are invariant under reflection through the central point, which gives the desired pole with real part at least $1/2$ and makes the method applicable to all N.

We now turn to the second improvement, which is what allows us to obtain a power bound. Observe from (2.3) that the contribution to S_f of each simple zero ρ is weighted by a factor that becomes larger with $\Re(\rho)$, so in its current form the result is poorer if $\Lambda_f(s)$ has simple zeros close to $\Re(s) = 1$. If all the simple zeros ρ satisfy $\Re(\rho) \leq \frac{7}{9}$, then we simply use the argument above and obtain a power bound for the number of simple zeros of Λ_f. Otherwise, if ρ is a simple zero with $\Re(\rho) > \frac{7}{9}$, we will show that there exists $\alpha \in \mathbb{Q}^\times$ such that (2.4) also has a pole at ρ (in [12] the key to control this scenario is using the zero-free region of Λ_f to limit $\Re(\rho)$, which is why the resulting bound is of logarithmic quality). The pole of (2.4) at ρ ultimately also gives a (rather good) power bound, so either way we obtain the desired result.

Let ρ be a simple zero of Λ_f (therefore a pole of Δ_f) with $\Re(\rho) > \frac{7}{9}$. To rule out pole cancellations in (2.4) for every $\alpha \in \mathbb{Q}^\times$, we introduce a new number-theoretic input into the argument, namely a zero-density estimate for twists of f. This is done by observing that for any prime $p \equiv 1 \pmod{N}$ there is a linear combination of terms of the form (2.4) that gives

$$p^{1-2s}\Delta_f(s) - \Delta_f\left(s, \frac{1}{p}\right). \tag{2.7}$$

One can use (2.5) to understand (2.7), concluding that it is equal (modulo a term that is holomorphic at $s = \rho$) to

$$b_{f,\chi_0}(s) \cdot \Delta_f(s) + \sum_{\substack{\chi \pmod{p} \\ \chi \neq \chi_0}} b_\chi \cdot \Delta_{f\otimes\chi}(s) \tag{2.8}$$

for some factor $b_{f,\chi_0}(s)$ which is non-vanishing inside the critical strip (hence at $s = \rho$).

If (2.7) does not have a pole at $s = \rho$, then the pole of $\Delta_f(s)$ there must be cancelled in (2.8), so $\Delta_{f\otimes\chi}(s)$ must have a pole at $s = \rho$ for at least one non-trivial character $\chi \pmod{p}$. This implies that $\Lambda_{f\otimes\chi}(\rho) = 0$ for at least one non-trivial character χ modulo every prime $p \equiv 1 \pmod{N}$. However, since $\Re(\rho) > \frac{7}{9}$, we can rule this out via zero-density estimates for twists of f. Therefore we show that (2.4) has a pole at $s = \rho$ for some $\alpha \in \mathbb{Q}^\times$, which implies a power bound for the number of simple zeros of Λ_f.

Acknowledgments

Thanks to Andrew Booker for his helpful comments and correspondence.

2.2 The setup

Definitions and background

Let $f \in S_k(\Gamma_0(N), \xi)$ be a primitive form (i.e. a normalized holomorphic Hecke cusp newform) of arbitrary weight k, level N, and nebentypus character $\xi \pmod{N}$. Writing the Fourier expansion

$$f(z) = \sum_{n=1}^{\infty} \lambda_f(n) n^{\frac{k-1}{2}} e(nz)$$

for $z \in \mathbb{H}$, where $\lambda_f(1) = 1$, we have Deligne's bound $|\lambda_f(n)| \leq d(n)$. Associate to f the usual completed L-function $\Lambda_f(s) := \Gamma_{\mathbb{C}}\left(s + \frac{k-1}{2}\right) L_f(s)$, which is

entire, where $\Gamma_{\mathbb{C}}(s) := 2(2\pi)^{-s}\Gamma(s)$ and

$$L_f(s) := \sum_{n=1}^{\infty} \lambda_f(n) n^{-s} = \prod_{p \text{ prime}} \left(1 - \lambda_f(p)p^{-s} + \xi(p)p^{-2s}\right)^{-1} \qquad \text{for } \Re(s) > 1.$$

Then we have the functional equation $\Lambda_f(s) = \epsilon_f N^{\frac{1}{2}-s} \Lambda_{\overline{f}}(1-s)$, where $\overline{f} \in S_k(\Gamma_0(N), \overline{\xi})$ is the dual of f, with Fourier coefficients $\lambda_{\overline{f}}(n) = \overline{\lambda_f(n)}$, and $\epsilon_f \in \mathbb{C}$ is the root number of f, with $|\epsilon_f| = 1$.

Let D_f be as in (2.1). For $\alpha \in \mathbb{Q}^{\times}$, χ a Dirichlet character, and $\Re(s) > 1$, we define the additive twists

$$L_f(s, \alpha) := \sum_{n=1}^{\infty} \lambda_f(n) e(n\alpha) n^{-s} \qquad \text{and} \qquad D_f(s, \alpha) := \sum_{n=1}^{\infty} c_f(n) e(n\alpha) n^{-s},$$

and the multiplicative twists

$$L_f(s, \chi) := \sum_{n=1}^{\infty} \lambda_f(n) \chi(n) n^{-s} \qquad \text{and} \qquad D_f(s, \chi) := \sum_{n=1}^{\infty} c_f(n) \chi(n) n^{-s}.$$

Denote

$$Q(N) := \{1\} \cup \{p \text{ prime} : p \nmid N\}.$$

For each Dirichlet character $\chi \pmod q$, there is a unique primitive form $f \otimes \chi$ such that $\lambda_{f\otimes\chi}(n) = \lambda_f(n)\chi(n)$ for every n with $(n, q) = 1$, by [3, Theorem 3.2]. If $q \in Q(N)$ and χ is non-trivial, then in fact $L_f(s, \chi) = L_{f\otimes\chi}(s)$ and therefore this multiplicative twist has analytic continuation to $\mathbb{C}$. This shows that $D_f(s, \chi) = L_f(s, \chi)\left(\frac{L_f'(s,\chi)}{L_f(s,\chi)}\right)' = D_{f\otimes\chi}(s)$ has meromorphic continuation to $\mathbb{C}$.

Similar results hold for the additive twists as well. Indeed, if $q \in Q(N)$, then we can expand our additive characters with multiplicative ones using

$$e\left(\frac{n}{q}\right) = \frac{q-1}{\phi(q)} + \frac{q}{\phi(q)}\tau(\chi_0)\chi_0(n) + \frac{1}{\phi(q)} \sum_{\substack{\chi \pmod q \\ \chi \neq \chi_0}} \tau(\overline{\chi})\chi(n), \tag{2.9}$$

where $\chi_0 \pmod q$ is the trivial character, the sum ranges over every non-trivial $\chi \pmod q$, and τ denotes the Gauss sum (observe that $\tau(\chi_0) = 1$ if $q = 1$ and $\tau(\chi_0) = -1$ otherwise). For any $a \in \mathbb{Z}$, this implies that $L_f\left(s, \frac{a}{q}\right)$ is entire, and $D_f\left(s, \frac{a}{q}\right)$ extends meromorphically to $\mathbb{C}$.

To be more precise, for $q \in Q(N)$, consider the rational functions

$$P_{f,q}(x) := \begin{cases} 1 & \text{if } q = 1, \\ 1 - \lambda_f(q)x + \xi(q)x^2 & \text{otherwise,} \end{cases}$$

and

$$R_{f,q}(x) := \begin{cases} 0 & \text{if } q = 1, \\ \frac{q \log^2 q}{\phi(q)} \frac{x(\lambda_f(q) - 4\xi(q)x + \lambda_f(q)\xi(q)x^2)}{P_{f,q}(x)} & \text{otherwise.} \end{cases}$$

Then (2.9) gives

$$\begin{aligned} D_f\left(s, \frac{a}{q}\right) = \frac{q-1}{\phi(q)} D_f(s) + \frac{q}{\phi(q)} \tau(\chi_0)\chi_0(a) D_f(s, \chi_0) \\ + \frac{1}{\phi(q)} \sum_{\substack{\chi \pmod q \\ \chi \neq \chi_0}} \tau(\overline{\chi})\chi(a) D_f(s, \chi). \end{aligned} \tag{2.10}$$

We have seen before that if $\chi \pmod q$ is non-trivial then $D_f(s, \chi) = D_{f \otimes \chi}(s)$ extends meromorphically to $\mathbb{C}$, but also from $D_f(s, \chi_0) = L_f(s, \chi_0) \left(\frac{L_f'(s,\chi_0)}{L_f(s,\chi_0)}\right)'$ and $L_f(s, \chi_0) = P_{f,q}(q^{-s}) L_f(s)$ (coming from the Euler product of L_f) we get

$$D_f(s, \chi_0) = P_{f,q}(q^{-s}) D_f(s) - \frac{\phi(q)}{q} R_{f,q}(q^{-s}) L_f(s), \tag{2.11}$$

and this provides the meromorphic continuation of $D_f\left(s, \frac{a}{q}\right)$ to $\mathbb{C}$. The analytic continuation of $L_f\left(s, \frac{a}{q}\right)$ to $\mathbb{C}$ follows in the same way.

For $(a, q) = 1$, it will be convenient to work with

$$D_{f,a,q}(s) := D_f\left(s, \frac{a}{q}\right) - R_{f,q}(q^{-s}) L_f(s) = \sum_{n=1}^{\infty} c_{f,a,q}(n) n^{-s} \qquad \text{for } \Re(s) > 1,$$

where the Dirichlet series expansion follows from (2.10), (2.11), and (2.1). Clearly $D_{f,a,q}(s)$ extends meromorphically to $\mathbb{C}$. We then define additive and multiplicative twists of $D_{f,a,q}(s)$. Namely, if χ is a Dirichlet character and $\alpha \in \mathbb{Q}^\times$, then for $\Re(s) > 1$ we let

$$D_{f,a,q}(s, \chi) := \sum_{n=1}^{\infty} c_{f,a,q}(n)\chi(n) n^{-s} \quad \text{and} \quad D_{f,a,q}(s, \alpha) := \sum_{n=1}^{\infty} c_{f,a,q}(n) e(n\alpha) n^{-s}.$$

Finally, associate to each of $L_f, D_f, D_{f,a,q}$ and their (additive or multiplicative) twists the completed versions $\Lambda_f, \Delta_f, \Delta_{f,a,q}$, respectively, obtained by multiplying by $\Gamma_{\mathbb{C}}\left(s + \frac{k-1}{2}\right)$.

Functional equations

If $q \in Q(N)$ and $\chi \pmod q$ is non-trivial, the functional equation for $f \otimes \chi$ gives

$$\Lambda_f(s,\chi) = \epsilon_f \xi(q)\chi(N)\frac{\tau(\chi)^2}{q}(Nq^2)^{\frac{1}{2}-s}\Lambda_{\overline{f}}(1-s,\overline{\chi}),$$

and as a consequence we obtain the corresponding functional equation for $\Delta_f(s,\chi) = \Delta_{f\otimes\chi}(s)$, given by

$$\begin{aligned}&\Delta_f(s,\chi) - \epsilon_f \xi(q)\chi(N)\frac{\tau(\chi)^2}{q}(Nq^2)^{\frac{1}{2}-s}\Delta_{\overline{f}}(1-s,\overline{\chi})\\ &= \Lambda_f(s,\chi)\left(\psi'\left(\frac{k+1}{2}-s\right) - \psi'\left(s+\frac{k-1}{2}\right)\right),\end{aligned}$$

where $\psi(s) := \frac{\Gamma'}{\Gamma}(s)$. Combining that with the relation

$$\begin{aligned}\Delta_{f,a,q}(s) = &\left(\frac{q-1}{\phi(q)} + \frac{q}{\phi(q)}\tau(\chi_0)P_{f,q}(q^{-s})\right)\Delta_f(s)\\ &+ \frac{1}{\phi(q)}\sum_{\substack{\chi \pmod q\\ \chi\neq\chi_0}}\tau(\overline{\chi})\chi(a)\Delta_f(s,\chi)\end{aligned} \tag{2.12}$$

for $q \in Q(N)$ and $(a,q)=1$, which follows from (2.10) and (2.11), we obtain a functional equation for additive twists of Δ_f.

Proposition 1 (Functional equation for $\Delta_{f,a,q}$ [12, Proposition 2.1])**.** *Let $f \in S_k(\Gamma_0(N),\xi)$ be a primitive form, $q \in Q(N)$, and $a \in \mathbb{Z}$ coprime to q. Then*

$$\begin{aligned}&\Delta_{f,a,q}(s) - \epsilon_f\xi(q)(Nq^2)^{\frac{1}{2}-s}\Delta_{\overline{f},-\overline{Na},q}(1-s)\\ &= \Lambda_f\left(s,\frac{a}{q}\right)\left(\psi'\left(\frac{k+1}{2}-s\right) - \psi'\left(s+\frac{k-1}{2}\right)\right).\end{aligned}$$

The detection mechanism for simple zeros

We give a brief account of the techniques of [12], since they will be relevant in what follows, but refer to that paper for details. The main idea originates in [21], and is developed in greater generality in [7]. The starting point is to study the poles of Δ_f by relating them to the inverse Mellin transform of Δ_f, via a contour integral. We develop the notation in the more general case of $\Delta_{f,a,q}$ for future reference. For $z \in \mathbb{H}$, let

$$F_{f,a,q}(z) := 2\sum_{n=1}^{\infty} c_{f,a,q}(n)n^{\frac{k-1}{2}}e(nz),$$

$$S_{f,a,q}(z) := \sum_{\Re(\rho)\in(0,1)} \operatorname*{Res}_{s=\rho} \Delta_{f,a,q}(s)(-iz)^{-s-\frac{k-1}{2}},$$

$$A_{f,a,q}(z) := \frac{1}{2\pi i} \int\limits_{\Re(s)=\frac{k}{2}} \Lambda_f\left(s, \frac{a}{q}\right) \times \left(\psi'\left(s+\frac{k-1}{2}\right)+\psi'\left(s-\frac{k-1}{2}\right)\right)(-iz)^{-s-\frac{k-1}{2}}\, ds,$$

and

$$B_{f,a,q}(z) := \frac{1}{2\pi i} \int\limits_{\Re(s)=\frac{k}{2}} \Lambda_f\left(s, \frac{a}{q}\right) \frac{\pi^2}{\sin^2\left(\pi\left(s+\frac{k-1}{2}\right)\right)}(-iz)^{-s-\frac{k-1}{2}}\, ds,$$

where $(-iz)^{-s-\frac{k-1}{2}}$ is defined in terms of the principal branch of $\log(-iz)$.

Taking the inverse Mellin transform of $\Delta_{f,a,q}$ (evaluated at $-iz$), shifting the line of integration to the left of the critical strip — where we pick up the factor $S_{f,a,q}$ corresponding to the poles — and using the functional equation (Proposition 1) to return to the right of the critical strip, we obtain (see [12, Lemma 2.3] for details) the relation

$$S_{f,a,q}(z) = F_{f,a,q}(z) - \frac{\epsilon_f \xi(q)}{(-i\sqrt{N}qz)^k} F_{\overline{f},-\overline{Na},q}\left(-\frac{1}{Nq^2 z}\right) + A_{f,a,q}(z) - B_{f,a,q}(z).$$

The next step is to take the Mellin transform for $z \in \mathbb{H}$ along a vertical line in the relation above. Such a procedure along the line $\Re(z) = 0$ would essentially bring us back to the previous step, but we instead integrate along $\Re(z) = \alpha$ for some $\alpha \in \mathbb{Q}^\times$ and obtain additive twists. The final result is the following.

Proposition 2 (Detecting poles of $\Delta_{f,a,q}$ via further additive twists [12, Proposition 2.2])**.** *Define*

$$H_{f,a,q,\alpha}(s) := \Delta_{f,a,q}(s,\alpha) - \epsilon_f \xi(q)(i\operatorname{sgn}(\alpha))^k (Nq^2\alpha^2)^{s-\frac{1}{2}} \Delta_{\overline{f},-\overline{Na},q}\left(s, -\frac{1}{Nq^2\alpha}\right)$$

and

$$I_{f,a,q,\alpha}(s) := \int_0^{\frac{|\alpha|}{4}} S_{f,a,q}(\alpha+iy) y^{s+\frac{k-1}{2}} \frac{dy}{y}.$$

Then $I_{f,a,q,\alpha}(s) - H_{f,a,q,\alpha}(s)$ has analytic continuation to $\Re(s) > 0$. Therefore, if

$$\int_0^{\frac{|\alpha|}{4}} |S_{f,a,q}(\alpha+iy)| y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} < \infty \tag{2.13}$$

for some $\sigma \geq 0$, then $H_{f,a,q,\alpha}(s)$ is holomorphic for $\Re(s) > \sigma$.

We will use only the special case $a = q = 1$ (i.e. detecting poles of Δ_f, or equivalently simple zeros of L_f) of Proposition 2, but the method of proof used for the general case $\Delta_{f,a,q}$ will be the key for showing that $H_{f,1,1,\alpha}$ has a pole in the critical strip, for some $\alpha \in \mathbb{Q}^\times$. For convenience, from now on we denote $H_{f,\alpha} := H_{f,1,1,\alpha}$.

2.3 Existence of poles of $H_{f,\alpha}$

Outline of the method

To establish an abundance of simple zeros of L_f (i.e. poles of Δ_f), we will use the poles of $H_{f,\alpha}$ in the critical strip, since through (2.13) their existence would imply that $S_{f,1,1}$ cannot be always small. However, showing that even a single such pole of $H_{f,\alpha}$ exists turns out to be difficult, since one needs to rule out a cancellation of poles between the two terms of $H_{f,\alpha}$. The purpose of this section is to establish such a result.

In [12] the authors circumvent this issue in the case $2 \nmid N$ by exploring the relations between the $H_{f,a,q,\alpha}$ for various choices of parameters (a, q, α). The limitation on the level N comes from the key role played by twists by $1/2$ (which also play an important role in [21]), since the poles of $\Delta_{f,1,2}$ are easily understood in terms of those of Δ_f, due to (2.12) and the fact that there are no non-trivial characters modulo 2. The issue is that this line of argument requires the case $q = 2$ of Proposition 2, which is not available if $2 \mid N$ since the functional equation in Proposition 1 no longer holds, as the local factor for a prime dividing the level has different, more problematic properties.

We will follow a different approach based on the methods of [7], where a significant difficulty is showing that Δ_f has even a single pole in the critical strip, and this is reminiscent of our situation for $H_{f,\alpha}$. The argument in the reference proceeds by contradiction, and the critical input is that $\Delta_f(s, \alpha)$ has poles in the line $\Re(s) = 0$ coming from simple zeros of local factors of L_f. We apply this argument for a certain difference of L-functions related to $H_{f,\alpha}$, instead of for Δ_f, and our issue of ruling out cancellations of poles in $H_{f,\alpha}$ at unknown locations inside the critical strip reduces to the simpler task of ruling out such cancellations at the simple zeros of certain local factors, where this can be explicitly done.

Implementation

Observe that

$$H_{f,1}(s) = \Delta_f(s) - \epsilon_f i^k N^{s-\frac{1}{2}} \Delta_{\overline{f}}\left(s, -\frac{1}{N}\right),$$

and if p is a prime satisfying $p \equiv 1 \pmod{N}$ then $\Delta_{\overline{f}}\left(s, -\frac{p}{N}\right) = \Delta_{\overline{f}}\left(s, -\frac{1}{N}\right)$, so

$$H_{f,\frac{1}{p}}(s) = \Delta_f\left(s, \frac{1}{p}\right) - \epsilon_f i^k \left(\frac{N}{p^2}\right)^{s-\frac{1}{2}} \Delta_{\overline{f}}\left(s, -\frac{1}{N}\right).$$

Therefore,

$$\begin{aligned} p^{1-2s} H_{f,1}(s) - H_{f,\frac{1}{p}}(s) &= p^{1-2s}\Delta_f(s) - \Delta_f\left(s, \frac{1}{p}\right) \\ &= p^{1-2s}\Delta_f(s) - \Delta_{f,1,p}(s) + R_{f,p}(p^{-s})\Lambda_f(s). \end{aligned} \tag{2.14}$$

Similarly, if we let $d := \frac{p-1}{N} \in \mathbb{Z}_{>0}$ then

$$H_{f,d}(s) = \Delta_f(s) - \epsilon_f i^k (Nd^2)^{s-\frac{1}{2}} \Delta_{\overline{f}}\left(s, -\frac{1}{Nd}\right),$$

and $\Delta_{\overline{f}}\left(s, -\frac{p}{Nd}\right) = \Delta_{\overline{f}}\left(s, -\frac{1}{Nd}\right)$, so

$$H_{f,\frac{d}{p}}(s) = \Delta_f\left(s, \frac{d}{p}\right) - \epsilon_f i^k \left(\frac{Nd^2}{p^2}\right)^{s-\frac{1}{2}} \Delta_{\overline{f}}\left(s, -\frac{1}{Nd}\right).$$

Therefore, since $d \equiv -\overline{N} \pmod{p}$,

$$\begin{aligned} p^{1-2s} H_{f,d}(s) - H_{f,\frac{d}{p}}(s) &= p^{1-2s}\Delta_f(s) - \Delta_f\left(s, \frac{d}{p}\right) \\ &= p^{1-2s}\Delta_f(s) - \Delta_{f,-\overline{N},p}(s) + R_{f,p}(p^{-s})\Lambda_f(s). \end{aligned} \tag{2.15}$$

Subtracting (2.14) from (2.15), we conclude that

$$p^{1-2s} H_{f,d}(s) - H_{f,\frac{d}{p}}(s) - p^{1-2s} H_{f,1}(s) + H_{f,\frac{1}{p}}(s) = \Delta_{f,1,p}(s) - \Delta_{f,-\overline{N},p}(s). \tag{2.16}$$

We will be able to show the existence of useful poles for at least one of $H_{f,1}(s)$, $H_{f,\frac{1}{p}}(s)$, $H_{f,d}(s)$, or $H_{f,\frac{d}{p}}(s)$ using the key proposition below.

Proposition 3 (Ruling out complete cancellation of poles)**.** *For any prime $p \neq N+1$ such that $p \equiv 1 \pmod{N}$, the meromorphic function*

$$G_{f,p}(s) := \Delta_{f,1,p}(s) - \Delta_{f,-\overline{N},p}(s)$$

has at least one pole in $\Re(s) \in (0,1)$.

Remark 1. *Our proof of Proposition 3 can easily be adapted to obtain infinitely many poles of $G_{f,p}(s)$ in $\Re(s) \in (0,1)$. Such a result has the same strength for our application as the existence of a single pole, so for simplicity we stick with the current statement.*

Assuming Proposition 3, we have the following consequence which will be the starting point in the course of our subsequent analysis.

Proposition 4 (Existence of poles with large real part)**.** *There exists $\alpha_f \in \mathbb{Q}^\times$ such that at least one of $H_{f,\alpha_f}(s)$ or $H_{\overline{f},\alpha_f}(s)$ has a pole in $\Re(s) \in \left[\frac{1}{2}, 1\right)$.*

Proof. For any prime $p \neq N+1$ such that $p \equiv 1 \pmod{N}$, from the functional equation in Proposition 1 we have that

$$\begin{aligned} &G_{f,p}(s) + \epsilon_f (Np^2)^{\frac{1}{2}-s} G_{\overline{f},p}(1-s) \\ &= \left(\Lambda_f\left(s, \frac{1}{p}\right) - \Lambda_f\left(s, -\frac{\overline{N}}{p}\right)\right)\left(\psi'\left(\frac{k+1}{2} - s\right) - \psi'\left(s + \frac{k-1}{2}\right)\right), \end{aligned}$$

as $\xi(p) = 1$. Since $\Lambda_f\left(s, \frac{1}{p}\right)$ and $\Lambda_f\left(s, -\frac{\overline{N}}{p}\right)$ are both entire, as easily follows from expanding into characters (see (2.17) below for details), and the poles of $\psi'(s)$ coincide with the poles of $\Gamma(s)$, we conclude that $G_{f,p}(s)$ and $G_{\overline{f},p}(1-s)$ have the same poles in $\Re(s) \in (0,1)$, as the RHS of the equation above is holomorphic in that region.

Combining this with Proposition 3, we get that at least one of $G_{f,p}(s)$ or $G_{\overline{f},p}(s)$ has a pole in $\Re(s) \in \left[\frac{1}{2}, 1\right)$, so (2.16) shows that the desired result holds for some $\alpha_f \in \left\{d, \frac{d}{p}, 1, \frac{1}{p}\right\}$, where $d = \frac{p-1}{N}$ as before.

□

Preliminary results

Before proceeding to the proof of Proposition 3, we take note of certain computations essentially contained in [12] that will be relevant for our argument. Those are reproduced in the auxiliary results below for ease of reference.

Lemma 1 (Inverse Mellin transform computations)**.** *Let $0 < \eta < 1/2$. Then for $z \in \mathbb{H}$ we have*

$$I^R_{f,a,q}(z) := \frac{1}{2\pi i} \int\limits_{\Re(s)=1+\eta} \Delta_{f,a,q}(s)(-iz)^{-s-\frac{k-1}{2}}\, ds = F_{f,a,q}(z)$$

and

$$I^L_{f,a,q}(z) := \frac{1}{2\pi i}\int\limits_{\Re(s)=-\eta} \Delta_{f,a,q}(s)(-iz)^{-s-\frac{k-1}{2}}\,ds$$
$$= \frac{\epsilon_f\xi(q)}{(-i\sqrt{N}qz)^k}F_{\overline{f},-\overline{Na},q}\left(-\frac{1}{Nq^2z}\right) - A_{f,a,q}(z) + B_{f,a,q}(z) - \operatorname*{Res}_{s=0}\Delta_{f,a,q}(s).$$

Proof. This follows from the functional equation in Proposition 1 (for the case of $I^L_{f,a,q}(z)$) and a computation of inverse Mellin transforms. The details are contained in the proof of [12, Lemma 2.3] — see in particular equations (2.9) and (2.12) there, and keep in mind that the residue at $s = 0$ only contributes if $k = 1$. Our statement above corrects a small typo in the computation of this residue at the last display of page 382 of the reference, where the term $\Delta_{f,a,q}\left(s, \frac{a}{q}\right)$ should be replaced by $\Delta_{f,a,q}(s)$, according to the functional equation.

□

Lemma 2 (Auxiliary analytic continuations)**.** *Let $\alpha \in \mathbb{Q}^\times$. Then for any $M \in \mathbb{Z}_{\geq 0}$,*

$$\int_0^{\frac{|\alpha|}{4}} (-i\sqrt{N}q(\alpha+iy))^{-k}F_{\overline{f},-\overline{Na},q}\left(-\frac{1}{Nq^2(\alpha+iy)}\right)y^{s+\frac{k-1}{2}}\frac{dy}{y} - (i\,\mathrm{sgn}(\alpha))^k$$
$$\times\sum_{m=0}^{M-1}(-i\alpha)^{-m}\binom{s+m-\frac{k+1}{2}}{m}(Nq^2\alpha^2)^{s-\frac{1}{2}+m}\Delta_{\overline{f},-\overline{Na},q}\left(s+m,-\frac{1}{Nq^2\alpha}\right)$$

continues to a holomorphic function in $\{s \in \mathbb{C} : \Re(s) > 1-M\}$. Furthermore, each of

$$\int_0^{\frac{|\alpha|}{4}} F_{f,a,q}(\alpha+iy)y^{s+\frac{k-1}{2}}\frac{dy}{y} - \Delta_{f,a,q}(s,\alpha),$$

$$\Gamma_{\mathbb{C}}(s)^{-1}\int_0^{\frac{|\alpha|}{4}} A_{f,a,q}(\alpha+iy)y^s\frac{dy}{y},$$

and

$$\Gamma_{\mathbb{C}}(s)^{-1}\int_0^{\frac{|\alpha|}{4}} B_{f,a,q}(\alpha+iy)y^s\frac{dy}{y}$$

continues to an entire function of s.

Proof. Those are precisely [12, Lemmas 2.4, 2.5, 2.6, and 2.7] in our notation. The first result is the hardest to establish, and it follows from Taylor

expanding the phases in $F_{\overline{f},-\overline{Na},q}$ and carefully analyzing the ensuing Mellin transforms. The idea is that as $z := \alpha + iy \in \mathbb{H}$ ranges over the vertical half-line $\Re(z) = \alpha$, $w := -\frac{1}{Nq^2 z} \in \mathbb{H}$ ranges over a semicircle centered in the x-axis with an endpoint at $-\frac{1}{Nq^2\alpha}$, so to a first approximation the input w of $F_{\overline{f},-\overline{Na},q}$ in the first integral can be considered to range over the vertical half-line $\Re(w) = -\frac{1}{Nq^2\alpha}$, which by Mellin inversion gives rise to a term of the form $\Delta_{\overline{f},-\overline{Na},q}\left(s, -\frac{1}{Nq^2\alpha}\right)$. The other terms arise from lower order components of the aforementioned Taylor expansion.

□

Lemma 3 (Analytic continuation of Mellin transforms)**.** *Let $\alpha \in \mathbb{Q}^\times$ and $M \in \mathbb{Z}_{\geq 0}$. Then*

$$\int_0^{\frac{|\alpha|}{4}} I^R_{f,a,q}(\alpha + iy) y^{s+\frac{k-1}{2}} \frac{dy}{y} - \Delta_{f,a,q}(s,\alpha)$$

continues to an entire function of s, and

$$\int_0^{\frac{|\alpha|}{4}} \left(I^L_{f,a,q}(\alpha+iy) + \operatorname*{Res}_{s=0} \Delta_{f,a,q}(s)\right) y^{s+\frac{k-1}{2}} \frac{dy}{y} - \epsilon_f \xi(q)(i\operatorname{sgn}(\alpha))^k$$
$$\times \sum_{m=0}^{M-1} (-i\alpha)^{-m} \binom{s+m-\frac{k+1}{2}}{m} (Nq^2\alpha^2)^{s-\frac{1}{2}+m} \Delta_{\overline{f},-\overline{Na},q}\left(s+m, -\frac{1}{Nq^2\alpha}\right)$$

continues to a meromorphic function in $\{s \in \mathbb{C} : \Re(s) > 1 - M\}$ whose only possible poles in that region must be at $s = \frac{1-k}{2} - n$, for $n \in \mathbb{Z}_{\geq 0}$.

Proof. Follows directly from plugging the equations in Lemma 1 into the integrals above and using Lemma 2 for each term that arises. The only possible poles come from the integral terms corresponding to $A_{f,a,q}$ and $B_{f,a,q}$, whose poles must be poles of $\Gamma_{\mathbb{C}}\left(s + \frac{k-1}{2}\right)$.

□

The next two results determine the locations of the poles of $\Delta_{f,a,q}(s)$ and some of its additive twists. Lemma 4 is essentially contained in [12, Proposition 2.2], while Lemma 5 requires a more careful analysis.

Lemma 4 (No exotic poles for $\Delta_{f,a,q}$)**.** *The poles of $\Delta_{f,a,q}(s)$ satisfy $\Re(s) \in (0,1)$ or $s = \frac{1-k}{2} - n$ for some $n \in \mathbb{Z}_{\geq 0}$.*

Proof. First observe that $\Delta_{f,a,q}(s)$ has no poles with $\Re(s) \geq 1$. Indeed, this follows from (2.12) and the fact that for non-trivial $\chi \pmod q$ the poles of $\Delta_f(s)$ and $\Delta_f(s,\chi) = \Delta_{f\otimes\chi}(s)$ are at simple zeros of $L_f(s)$ and $L_{f\otimes\chi}(s)$, respectively, but there are no such zeros with $\Re(s) \geq 1$ by non-vanishing for automorphic L-functions [61]. As a consequence, we can also understand the poles of $\Delta_{f,a,q}(s)$ with $\Re(s) \leq 0$, through the functional equation. Using (2.9) and $\Lambda_f(s,\chi_0) = P_{f,q}(q^{-s})\Lambda_f(s)$, since $(a,q)=1$ we get

$$\begin{aligned}\Lambda_f\left(s,\frac{a}{q}\right) = &\left(\frac{q-1}{\phi(q)} + \frac{q}{\phi(q)}\tau(\chi_0)P_{f,q}(q^{-s})\right)\Lambda_f(s) \\ &+ \frac{1}{\phi(q)}\sum_{\substack{\chi \pmod q \\ \chi\neq\chi_0}} \tau(\overline{\chi})\chi(a)\Lambda_{f\otimes\chi}(s),\end{aligned} \tag{2.17}$$

so $\Lambda_f\left(s,\frac{a}{q}\right)$ is entire. The poles of $\psi'(s)$ coincide with the poles of $\Gamma(s)$, so Proposition 1 shows that $\Delta_{f,a,q}(s)$ has no poles with $\Re(s) \leq 0$, except possibly for $s = \frac{1-k}{2} - n$, for some $n \in \mathbb{Z}_{\geq 0}$.

□

Lemma 5 (Location of exotic poles for additive twists of $\Delta_{f,a,p}$)**.** *Let $p, q \in Q(N)$ with $p \neq q$, and let $a, b \in \mathbb{Z}$ with $(a,p) = (b,q) = 1$. If we let $\chi_0 \pmod q$ and $\psi_0 \pmod p$ denote the trivial characters, then*

$$\begin{aligned}&\Delta_{f,a,p}\left(s,\frac{b}{q}\right) + \tau(\chi_0)\left(\frac{p-1}{\phi(p)} + \frac{p}{\phi(p)}\tau(\psi_0)P_{f,p}(p^{-s})\right)R_{f,q}(q^{-s})\Lambda_f(s) \\ &\quad + \frac{\tau(\chi_0)}{\phi(p)}\sum_{\substack{\psi \pmod p \\ \psi\neq\psi_0}} \tau\left(\overline{\psi}\right)\psi(a)R_{f\otimes\psi,q}(q^{-s})\Lambda_{f\otimes\psi}(s)\end{aligned}$$

continues to a holomorphic function in $\{s \in \mathbb{C} : \Re(s) \leq 0\} \setminus \frac{1}{2}\mathbb{Z}$.

Proof. By (2.9) we have

$$\begin{aligned}\Delta_{f,a,p}\left(s,\frac{b}{q}\right) = &\frac{q-1}{\phi(q)}\Delta_{f,a,p}(s) + \frac{q}{\phi(q)}\tau(\chi_0)\Delta_{f,a,p}(s,\chi_0) \\ &+ \frac{1}{\phi(q)}\sum_{\substack{\chi \pmod q \\ \chi\neq\chi_0}} \tau\left(\overline{\chi}\right)\chi(b)\Delta_{f,a,p}(s,\chi).\end{aligned}$$

For non-trivial $\chi \pmod q$, we can twist (2.12) by χ to get

$$\begin{aligned}\Delta_{f,a,p}(s,\chi) &= \left(\frac{p-1}{\phi(p)} + \frac{p}{\phi(p)}\tau(\psi_0)P_{f,p}(p^{-s}\chi(p))\right)\Delta_f(s,\chi) \\ &\qquad + \frac{1}{\phi(p)}\sum_{\substack{\psi \pmod p \\ \psi\neq\psi_0}} \tau\left(\overline{\psi}\right)\psi(a)\Delta_f(s,\psi\chi) \\ &= \left(\frac{p-1}{\phi(p)} + \frac{p}{\phi(p)}\tau(\psi_0)P_{f\otimes\chi,p}(p^{-s})\right)\Delta_{f\otimes\chi}(s) \\ &\qquad + \frac{1}{\phi(p)}\sum_{\substack{\psi \pmod p \\ \psi\neq\psi_0}} \tau\left(\overline{\psi}\right)\psi(a)\Delta_{f\otimes\psi\chi}(s),\end{aligned}$$

as $\chi \pmod q$ and $\psi\chi \pmod{pq}$ are primitive characters. This shows that $\Delta_{f,a,p}(s,\chi)$ is holomorphic in $\{s\in\mathbb{C} : \Re(s)\leq 0\}\setminus\frac{1}{2}\mathbb{Z}$, since this is the case for each of $\Delta_{f\otimes\chi}(s)$ and $\Delta_{f\otimes\psi\chi}(s)$ due to Lemma 4. The same property also holds for $\Delta_{f,a,p}(s)$ by the same lemma, so we are left with analyzing $\Delta_{f,a,p}(s,\chi_0)$. Since $\chi_0(p)=1$, once again by (2.12) we have

$$\begin{aligned}\Delta_{f,a,p}(s,\chi_0) &= \left(\frac{p-1}{\phi(p)} + \frac{p}{\phi(p)}\tau(\psi_0)P_{f,p}(p^{-s})\right)\Delta_f(s,\chi_0) \\ &\qquad + \frac{1}{\phi(p)}\sum_{\substack{\psi \pmod p \\ \psi\neq\psi_0}} \tau\left(\overline{\psi}\right)\psi(a)\Delta_{f\otimes\psi}(s,\chi_0).\end{aligned}$$

Now, (2.11) gives

$$\Delta_f(s,\chi_0) = P_{f,q}(q^{-s})\Delta_f(s) - \frac{\phi(q)}{q}R_{f,q}(q^{-s})\Lambda_f(s),$$

and analogously, since for non-trivial $\psi \pmod p$ the primitive form $f\otimes\psi$ has level Np^2 and $q\in Q(Np^2)$,

$$\Delta_{f\otimes\psi}(s,\chi_0) = P_{f\otimes\psi,q}(q^{-s})\Delta_{f\otimes\psi}(s) - \frac{\phi(q)}{q}R_{f\otimes\psi,q}(q^{-s})\Lambda_{f\otimes\psi}(s).$$

However, $\Delta_f(s)$ and $\Delta_{f\otimes\psi}(s)$ are both holomorphic in $\{s\in\mathbb{C} : \Re(s)\leq 0\}\setminus\frac{1}{2}\mathbb{Z}$, so the only remaining terms are the ones with the factors $R_{f,q}(q^{-s})$ and $R_{f\otimes\psi,q}(q^{-s})$. Plugging those back along our sequence of equations, we obtain the desired result.

□

Producing poles

We are now ready for the proof of the main result in this section.

Proof of Proposition 3. Assume by contradiction that $G_{f,p}(s)$ has no poles in $\Re(s) \in (0,1)$. Then by Lemma 4 the only possible pole of $G_{f,p}(s)$ with $\Re(s) > -1/2$ is $s = 0$, which can only occur if $k = 1$.

Let $0 < \eta < 1/2$. For $z \in \mathbb{H}$, define

$$\mathcal{I}^R(z) := \frac{1}{2\pi i} \int\limits_{\Re(s)=1+\eta} G_{f,p}(s)(-iz)^{-s-\frac{k-1}{2}}\, ds$$

and

$$\mathcal{I}^L(z) := \frac{1}{2\pi i} \int\limits_{\Re(s)=-\eta} G_{f,p}(s)(-iz)^{-s-\frac{k-1}{2}}\, ds.$$

By Stirling's formula, the decomposition (2.12), and the PhragménLindelöf principle, we see that $G_{f,p}(s)$ is rapidly decaying in vertical strips, so we can shift contours. Since we are assuming that $G_{f,p}(s)$ has no poles in $\Re(s) \in (0,1)$, and it has a pole at $s = 0$ only if $k = 1$, we get

$$\mathcal{I}^L(z) + \operatorname*{Res}_{s=0} G_{f,p}(s) = \mathcal{I}^R(z). \tag{2.18}$$

Observe that

$$\begin{aligned}\mathcal{I}^R(z) &= \frac{1}{2\pi i} \int\limits_{\Re(s)=1+\eta} \left(\Delta_{f,1,p}(s) - \Delta_{f,-\overline{N},p}(s)\right)(-iz)^{-s-\frac{k-1}{2}}\, ds \\ &= I^R_{f,1,p}(z) - I^R_{f,-\overline{N},p}(z)\end{aligned}$$

in the notation of Lemma 1. Similarly, we have

$$\begin{aligned}&\mathcal{I}^L(z) + \operatorname*{Res}_{s=0} G_{f,p}(s) \\ &= \frac{1}{2\pi i} \int\limits_{\Re(s)=-\eta} \left(\Delta_{f,1,p}(s) - \Delta_{f,-\overline{N},p}(s)\right)(-iz)^{-s-\frac{k-1}{2}}\, ds + \operatorname*{Res}_{s=0} G_{f,p}(s) \\ &= \left(I^L_{f,1,p}(z) + \operatorname*{Res}_{s=0} \Delta_{f,1,p}(s)\right) - \left(I^L_{f,-\overline{N},p}(z) + \operatorname*{Res}_{s=0} \Delta_{f,-\overline{N},p}(s)\right).\end{aligned}$$

Therefore, (2.18) becomes

$$\begin{aligned}&\left(I^L_{f,1,p}(z) + \operatorname*{Res}_{s=0} \Delta_{f,1,p}(s)\right) - \left(I^L_{f,-\overline{N},p}(z) + \operatorname*{Res}_{s=0} \Delta_{f,-\overline{N},p}(s)\right) \\ &= I^R_{f,1,p}(z) - I^R_{f,-\overline{N},p}(z).\end{aligned} \tag{2.19}$$

We now set $z = \alpha + iy$, with $\alpha \in \mathbb{Q}^\times$ and $y > 0$, and perform a truncated Mellin transform along y. More precisely, consider

$$\mathcal{R}(s) := \int_0^{\frac{|\alpha|}{4}} \left(I^R_{f,1,p}(\alpha + iy) - I^R_{f,-\overline{N},p}(\alpha + iy) \right) y^{s+\frac{k-1}{2}} \frac{dy}{y}.$$

Applying Lemma 3 we conclude that

$$\mathcal{R}(s) - \left(\Delta_{f,1,p}(s,\alpha) - \Delta_{f,-\overline{N},p}(s,\alpha) \right) \tag{2.20}$$

continues to an entire function of s. Similarly, let

$$\begin{aligned}\mathcal{L}(s) := \int_0^{\frac{|\alpha|}{4}} \Big(\Big(I^L_{f,1,p}(\alpha + iy) + \operatorname*{Res}_{s=0} \Delta_{f,1,p}(s) \Big) \\ - \Big(I^L_{f,-\overline{N},p}(\alpha + iy) + \operatorname*{Res}_{s=0} \Delta_{f,-\overline{N},p}(s) \Big) \Big) y^{s+\frac{k-1}{2}} \frac{dy}{y}.\end{aligned}$$

By Lemma 3 we conclude that, for any $M \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned}&\mathcal{L}(s) - \epsilon_f (i \operatorname{sgn}(\alpha))^k (Np^2\alpha^2)^{s-\frac{1}{2}} \sum_{m=0}^{M-1} (iNp^2\alpha)^m \binom{s+m-\frac{k+1}{2}}{m} \\ &\times \left(\Delta_{\overline{f},-\overline{N},p}\left(s+m, -\frac{1}{Np^2\alpha} \right) - \Delta_{\overline{f},1,p}\left(s+m, -\frac{1}{Np^2\alpha} \right) \right)\end{aligned} \tag{2.21}$$

continues to a meromorphic function in $\{s \in \mathbb{C} : \Re(s) > 1 - M\}$ whose only possible poles in that region must be at $s = \frac{1-k}{2} - n$, for $n \in \mathbb{Z}_{\geq 0}$. Here we used the fact that $\xi(p) = 1$, as $p \equiv 1 \pmod{N}$.

Since $\mathcal{L}(s) = \mathcal{R}(s)$ due to (2.19), we conclude from (2.20) and (2.21) that

$$\begin{aligned}&\Delta_{f,1,p}(s,\alpha) - \Delta_{f,-\overline{N},p}(s,\alpha) - \epsilon_f (i \operatorname{sgn}(\alpha))^k (Np^2\alpha^2)^{s-\frac{1}{2}} \sum_{m=0}^{M-1} (iNp^2\alpha)^m \\ &\times \binom{s+m-\frac{k+1}{2}}{m} \left(\Delta_{\overline{f},-\overline{N},p}\left(s+m, -\frac{1}{Np^2\alpha} \right) - \Delta_{\overline{f},1,p}\left(s+m, -\frac{1}{Np^2\alpha} \right) \right)\end{aligned} \tag{2.22}$$

continues to a holomorphic function in $\{s \in \mathbb{C} : \Re(s) > 1 - M\} \setminus \frac{1}{2}\mathbb{Z}$.

Fix $b \in (\mathbb{Z}/Np^2\mathbb{Z})^\times$. Let $q_1, q_2, \ldots, q_M$ be distinct primes satisfying $q_j \equiv b \pmod{Np^2}$ for all $1 \leq j \leq M$, and let m_0 be an integer satisfying $0 \leq m_0 \leq M - 1$. Setting $\alpha = \frac{1}{q_j}$, (2.22) shows that

$$\begin{aligned}&\left(\frac{Np^2}{q_j^2} \right)^{\frac{1}{2}-s} \left(\Delta_{f,1,p}\left(s, \frac{1}{q_j} \right) - \Delta_{f,-\overline{N},p}\left(s, \frac{1}{q_j} \right) \right) - \epsilon_f i^k \sum_{m=0}^{M-1} \left(\frac{iNp^2}{q_j} \right)^m \\ &\times \binom{s+m-\frac{k+1}{2}}{m} \left(\Delta_{\overline{f},-\overline{N},p}\left(s+m, -\frac{b}{Np^2} \right) - \Delta_{\overline{f},1,p}\left(s+m, -\frac{b}{Np^2} \right) \right)\end{aligned} \tag{2.23}$$

continues to a holomorphic function in $\{s \in \mathbb{C} : \Re(s) > 1 - M\} \setminus \frac{1}{2}\mathbb{Z}$. By the Vandermonde determinant, we can find $c_1, c_2, \ldots, c_M \in \mathbb{Q}$ such that for every $m \in \mathbb{Z}$ with $0 \leq m \leq M-1$,

$$\sum_{j=1}^{M} c_j q_j^{-m} = \begin{cases} 1 & \text{if } m = m_0, \\ 0 & \text{if } m \neq m_0. \end{cases}$$

Summing (2.23) for each q_j with weight c_j, for $1 \leq j \leq M$, it follows that

$$\begin{aligned} &\epsilon_f i^k (iNp^2)^{m_0} \binom{s + m_0 - \frac{k+1}{2}}{m_0} \\ &\qquad \times \left(\Delta_{\overline{f},-\overline{N},p}\left(s + m_0, -\frac{b}{Np^2}\right) - \Delta_{\overline{f},1,p}\left(s + m_0, -\frac{b}{Np^2}\right)\right) \\ &- \sum_{j=1}^{M} c_j \left(\frac{Np^2}{q_j^2}\right)^{\frac{1}{2}-s} \left(\Delta_{f,1,p}\left(s, \frac{1}{q_j}\right) - \Delta_{f,-\overline{N},p}\left(s, \frac{1}{q_j}\right)\right) \end{aligned} \tag{2.24}$$

continues to a holomorphic function in $\{s \in \mathbb{C} : \Re(s) > 1 - M\} \setminus \frac{1}{2}\mathbb{Z}$.

Now, observe that both $\Delta_{f,1,p}\left(s, \frac{1}{q_j}\right)$ and $\Delta_{f,-\overline{N},p}\left(s, \frac{1}{q_j}\right)$ are holomorphic in $\{s \in \mathbb{C} : \Re(s) < 0\} \setminus \frac{1}{2}\mathbb{Z}$. Indeed, this follows from Lemma 5 and the fact that for a non-trivial character $\psi \pmod p$, the poles of $R_{f,q_j}(q_j^{-s})$ and $R_{f\otimes\psi,q_j}(q_j^{-s})$ satisfy $\Re(s) = 0$, since $\lambda_{f\otimes\psi}(q_j) = \lambda_f(q_j)\psi(q_j)$ and $|\lambda_f(q_j)| \leq 2$ by Deligne's bound. Therefore, (2.24) implies that

$$\Delta_{\overline{f},-\overline{N},p}\left(s, -\frac{b}{Np^2}\right) - \Delta_{\overline{f},1,p}\left(s, -\frac{b}{Np^2}\right)$$

continues to a holomorphic function in $\{s \in \mathbb{C} : 1 - M + m_0 < \Re(s) < m_0\} \setminus \frac{1}{2}\mathbb{Z}$. Since $M \in \mathbb{Z}_{\geq 0}$ and $0 \leq m_0 \leq M-1$ are arbitrary, we conclude that it indeed continues to a holomorphic function in $\mathbb{C} \setminus \frac{1}{2}\mathbb{Z}$. Finally, this in conjunction with (2.23) shows that

$$\Delta_{f,1,p}\left(s, \frac{1}{q}\right) - \Delta_{f,-\overline{N},p}\left(s, \frac{1}{q}\right) \tag{2.25}$$

continues to a holomorphic function in $\mathbb{C}\setminus\frac{1}{2}\mathbb{Z}$, for any prime $q \equiv b \pmod{Np^2}$. Since we can choose the congruence class $b \in (\mathbb{Z}/Np^2\mathbb{Z})^\times$ arbitrarily, the result holds for any prime $q \nmid Np$.

Let $\chi_0 \pmod q$ and $\psi_0 \pmod p$ denote the trivial characters, and observe that $\tau(\chi_0) = \tau(\psi_0) = -1$. Applying Lemma 5 to each term of (2.25) we verify

that

$$\frac{1}{p-1}\sum_{\substack{\psi \pmod p\\ \psi\neq\psi_0}} \tau\left(\overline{\psi}\right)\left(\psi(1)-\psi\left(-\overline{N}\right)\right) R_{f\otimes\psi,q}(q^{-s})\Lambda_{f\otimes\psi}(s) \tag{2.26}$$

continues to a holomorphic function in $\{s\in\mathbb{C}:\Re(s)\leq 0\}\setminus\frac{1}{2}\mathbb{Z}$, so in particular it has no poles $s\neq 0$ with $\Re(s)=0$.

Observe that for any $c\in(\mathbb{Z}/p\mathbb{Z})^\times$ we have

$$\sum_{\substack{r\leq x \text{ prime}\\ r\equiv c \pmod p}} |\lambda_f(r)|^2 \sim \frac{1}{\phi(p)}\frac{x}{\log x} \qquad \text{as } x\to\infty \tag{2.27}$$

by Rankin-Selberg (see for instance [73, Lemma 1] for details when f has trivial nebentypus), as $f\otimes\psi$ is orthogonal to f for each non-trivial $\psi \pmod p$, since it is a primitive form of level Np^2. From now on we assume that $q\in\mathcal{Q}_{f,p}:=\{r \text{ prime}: r\equiv 1 \pmod p, r\nmid N, \text{ and } |\lambda_f(r)|<2\}$. Observe that $\mathcal{Q}_{f,p}$ is an infinite set, by (2.27).

Since $q\equiv 1 \pmod p$, for any non-trivial $\psi \pmod p$ we have $R_{f\otimes\psi,q}(q^{-s}) = R_{f,q}(q^{-s}\psi(q)) = R_{f,q}(q^{-s})$. But

$$R_{f,q}(q^{-s}) = -\frac{q}{\phi(q)}P_{f,q}(q^{-s})\left(\frac{(P_{f,q}(q^{-s}))'}{P_{f,q}(q^{-s})}\right)', \tag{2.28}$$

where the derivatives are with respect to s, so the poles of $R_{f,q}(q^{-s})$ are precisely at the simple zeros of $P_{f,q}(q^{-s}) = 1-\lambda_f(q)q^{-s}+\xi(q)q^{-2s} =: (1-\alpha_f(q)q^{-s})(1-\beta_f(q)q^{-s})$. We chose q with $|\lambda_f(q)|<2$, so $|\alpha_f(q)|=|\beta_f(q)|=1$ and $\alpha_f(q)\neq\beta_f(q)$. Therefore, all the zeros of $P_{f,q}(q^{-s})$ are simple and satisfy $\Re(s)=0$.

Choose $t\in\mathbb{R}^\times$ such that $q^{it}=\alpha_f(q)$, so $P_{f,q}(q^{-it})=0$ and (2.28) gives

$$\begin{aligned}\operatorname*{Res}_{s=it} R_{f,q}(q^{-s}) &= \frac{q}{\phi(q)}(P_{f,q}(q^{-s}))'\Big|_{s=it} = \frac{q^{1-it}\log q}{q-1}\left(\lambda_f(q)-2\xi(q)q^{-it}\right)\\ &= \frac{q\log q}{q-1}\overline{\alpha_f(q)}\left(\alpha_f(q)-\beta_f(q)\right)\neq 0,\end{aligned}$$

as $\alpha_f(q)\beta_f(q)=\xi(q)$. We now take residues of (2.26) at $s=it\neq 0$ to obtain

$$\frac{1}{p-1}\sum_{\substack{\psi \pmod p\\ \psi\neq\psi_0}} \tau\left(\overline{\psi}\right)\left(\psi(1)-\psi\left(-\overline{N}\right)\right)\Lambda_{f\otimes\psi}(it)\cdot\operatorname*{Res}_{s=it} R_{f,q}(q^{-s}) = 0.$$

But $\operatorname{Res}_{s=it} R_{f,q}(q^{-s}) \neq 0$ as we saw above, so in fact using (2.17) we get

$$\begin{aligned} 0 &= \frac{1}{p-1} \sum_{\substack{\psi \pmod p \\ \psi \neq \psi_0}} \tau\left(\overline{\psi}\right)\left(\psi(1) - \psi\left(-\overline{N}\right)\right) \Lambda_{f\otimes\psi}(it) \\ &= \Lambda_f\left(it, \frac{1}{p}\right) - \Lambda_f\left(it, -\frac{\overline{N}}{p}\right) \end{aligned} \tag{2.29}$$

for any $t \in \mathcal{T}_{f,q} := \left\{\frac{\theta_f(q)+2\pi n}{\log q} : n \in \mathbb{Z}\right\} \setminus \{0\}$, where $\theta_f(q) \in [0, 2\pi)$ is defined by $\alpha_f(q) = e^{i\theta_f(q)}$. Since this holds for any $q \in \mathcal{Q}_{f,p}$ and $\bigcup_{q\in\mathcal{Q}_{f,p}} \mathcal{T}_{f,q}$ is dense in $\mathbb{R}$ (as $\mathcal{Q}_{f,p}$ is infinite), we conclude by analytic continuation that

$$\Lambda_f\left(s, \frac{1}{p}\right) = \Lambda_f\left(s, -\frac{\overline{N}}{p}\right)$$

for every $s \in \mathbb{C}$. This is a contradiction, as we can compare the coefficients of the respective Dirichlet series expansions in $\Re(s) > 1$ and they do not match. For instance, (2.27) shows that there is a prime $r \equiv 1 \pmod p$ such that $\lambda_f(r) \neq 0$, hence the r-th coefficients in the corresponding Dirichlet series expansions are $\lambda_f(r)e\left(1/p\right)$ and $\lambda_f(r)e\left(-\overline{N}/p\right)$, which are distinct since $-\overline{N} \not\equiv 1 \pmod p$, as $p > N+1$ by hypothesis. A standard argument using Perron's formula then gives the desired contradiction, so we conclude that $G_{f,p}(s)$ has at least one pole in $\Re(s) \in (0,1)$, as desired.

□

2.4 Location of poles of $H_{f,\alpha}$

In this section we will show that if $H_{f,\alpha}(s)$ has a pole in $\Re(s) \in \left[\frac{1}{2}, 1\right)$ for some $\alpha \in \mathbb{Q}^\times$, then Λ_f must have many simple zeros. This will be enough to prove our main results, since Proposition 4 guarantees the existence of such a pole for some α in the case of either f or $\overline{f}$, but Λ_f and $\Lambda_{\overline{f}}$ have the same number of simple zeros, by the functional equation.

From poles of $H_{f,\alpha}$ to simple zeros of Λ_f

Denote $S_f(z) := S_{f,1,1}(z)$ for $z \in \mathbb{H}$, as in the introduction. As we have described before, the basic mechanism uses (2.13) to show that S_f cannot be always small if $H_{f,\alpha}$ has a pole of large real part. The next lemma provides a more direct link between the quantity in (2.13) and simple zeros of Λ_f (i.e. poles of Δ_f in the critical strip). It is essentially contained in [12, Lemma 3.2], but we provide a proof for completeness.

Lemma 6 (Bounding the truncated Mellin transform of S_f)**.** *Let $\eta > 0$ be fixed. For any $\sigma \in [\eta, 2]$ and $\alpha \in \mathbb{Q}^\times$,*

$$\int_0^{\frac{|\alpha|}{4}} |S_f(\alpha+iy)| y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} \ll_{f,\alpha,\eta} \sum_{\substack{\rho=\beta+i\gamma \\ \text{a pole of } \Delta_f \\ \text{with } \beta>0}} \left|\Lambda_f'(\rho)\right| e^{\frac{\pi|\gamma|}{2}} (1+|\gamma|)^{-\sigma-\frac{k-1}{2}}.$$

Proof. We have

$$\begin{aligned} &\int_0^{\frac{|\alpha|}{4}} |S_f(\alpha+iy)| y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} \\ &\leq \sum_{\Re(\rho)\in(0,1)} \int_0^{\frac{|\alpha|}{4}} \left|\operatorname*{Res}_{s=\rho} \Delta_f(s)\right| \cdot \left|(y-i\alpha)^{-\rho-\frac{k-1}{2}}\right| y^{\sigma+\frac{k-1}{2}} \frac{dy}{y}, \end{aligned} \tag{2.30}$$

where we can exchange the order of summation and integration by Tonelli's theorem. Let $\rho = \beta+i\gamma$ be a pole of Δ_f with $\beta \in (0,1)$, and denote $\tau := 1+|\gamma|$. Then since $\Lambda_f(\rho) = 0$, observe that

$$\operatorname*{Res}_{s=\rho} \Delta_f(s) = -\Gamma_{\mathbb{C}}\left(\rho + \frac{k-1}{2}\right) L_f'(\rho) = -\Lambda_f'(\rho),$$

and for $0 \leq y \leq \frac{|\alpha|}{4}$,

$$\begin{aligned} \left|(y-i\alpha)^{-\rho-\frac{k-1}{2}}\right| &= |y-i\alpha|^{-\beta-\frac{k-1}{2}} e^{\gamma \arctan\left(-\frac{\alpha}{y}\right)} \\ &\ll_{f,\alpha} e^{-\gamma \arctan\left(\frac{\alpha}{y}\right)} = e^{\gamma \operatorname{sgn}(\alpha)\left(\arctan\left(\frac{y}{|\alpha|}\right)-\frac{\pi}{2}\right)}. \end{aligned}$$

Therefore, we conclude that the RHS of (2.30) is

$$\ll_{f,\alpha} \sum_{\substack{\rho=\beta+i\gamma \\ \text{a pole of } \Delta_f \\ \text{with } \beta\in(0,1)}} \left|\Lambda_f'(\rho)\right| \cdot \int_0^{\frac{|\alpha|}{4}} e^{\gamma \operatorname{sgn}(\alpha)\left(\arctan\left(\frac{y}{|\alpha|}\right)-\frac{\pi}{2}\right)} y^{\sigma+\frac{k-1}{2}} \frac{dy}{y}.$$

Using $\gamma \operatorname{sgn}(\alpha)\left(\arctan\left(\frac{y}{|\alpha|}\right)-\frac{\pi}{2}\right) \leq -|\gamma|\left(\arctan\left(\frac{y}{|\alpha|}\right)-\frac{\pi}{2}\right) \leq -\frac{|\gamma|y}{2|\alpha|} + \frac{\pi|\gamma|}{2}$, since $\arctan(x) \geq \frac{x}{2}$ for $0 \leq x \leq \frac{1}{4}$, we have

$$\begin{aligned} \int_0^{\frac{|\alpha|}{4}} e^{\gamma \operatorname{sgn}(\alpha)\left(\arctan\left(\frac{y}{|\alpha|}\right)-\frac{\pi}{2}\right)} y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} &\leq e^{\frac{\pi|\gamma|}{2}} \int_0^{\frac{|\alpha|}{4}} e^{-\frac{|\gamma|y}{2|\alpha|}} y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} \\ &\ll e^{\frac{\pi|\gamma|}{2}} \int_0^{\frac{|\alpha|}{4}} e^{-\frac{\tau y}{2|\alpha|}} y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} \leq e^{\frac{\pi|\gamma|}{2}} \int_0^{\infty} e^{-\frac{\tau y}{2|\alpha|}} y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} \\ &= e^{\frac{\pi|\gamma|}{2}} \left(\frac{2|\alpha|}{\tau}\right)^{\sigma+\frac{k-1}{2}} \Gamma\left(\sigma+\frac{k-1}{2}\right) \ll_{f,\alpha,\eta} e^{\frac{\pi|\gamma|}{2}} \tau^{-\sigma-\frac{k-1}{2}}, \end{aligned}$$

so the desired result follows.

□

For the case of general level, we will apply Lemma 6 in conjunction with a pointwise bound coming from subconvexity (we will see how to improve this for $N = 1$ in the next section).

Lemma 7 (Weyl subconvexity for L'_f [12, Lemma 3.1])**.** *If $\rho = \beta + i\gamma$ is a zero of Λ_f, then*

$$\Lambda'_f(\rho) \ll_{f,\varepsilon} (1 + |\gamma|)^{\frac{k}{2}+\frac{1}{3}|\beta-\frac{1}{2}|-\frac{1}{6}+\varepsilon} e^{-\frac{\pi|\gamma|}{2}}$$

for any $\varepsilon > 0$.

Proof (sketch). Follows from the Weyl subconvex bound $L_f\left(\frac{1}{2} + it\right) \ll_{f,\varepsilon} (1 + |t|)^{\frac{1}{3}+\varepsilon}$ of [13], and a standard argument using Cauchy's formula combined with the PhragménLindelöf principle, the functional equation, and Stirling's formula. See the reference for details.

□

Remark 2. *If $\mu \in \left[0, \frac{1}{2}\right]$ and we had a subconvexity bound of the form $L_f\left(\frac{1}{2} + it\right) \ll_{f,\varepsilon} (1 + |t|)^{\mu+\varepsilon}$ for all $\varepsilon > 0$, then Lemma 7 would become $\Lambda'_f(\rho) \ll_{f,\varepsilon} (1 + |\gamma|)^{\frac{k}{2}+(1-2\mu)(|\beta-\frac{1}{2}|-\frac{1}{2})+\varepsilon} e^{-\frac{\pi|\gamma|}{2}}$ for any $\varepsilon > 0$. The given result corresponds to $\mu = \frac{1}{3}$.*

For a meromorphic function h on $\{s \in \mathbb{C} : \Re(s) > 1\}$, let

$$\Theta(h) := \inf\{\theta \geq 0 : h \text{ continues analytically to } \{s \in \mathbb{C} : \Re(s) > \theta\}\}.$$

Furthermore, let

$$\begin{aligned} \theta_f :=& \sup\left(\{0\} \cup \{\Re(\rho), 1 - \Re(\rho) : \rho \text{ is a pole of } \Delta_f\}\right) \\ =& \sup\left(\{0\} \cup \left\{\Re(\rho) : \rho \text{ is a simple zero of } \Lambda_f \text{ or } \Lambda_{\overline{f}}\right\}\right). \end{aligned}$$

Then Lemma 6 and Lemma 7 can be combined into the following result, which is a particular case of [12, Proposition 3.3]. We again provide the proof for completeness.

Proposition 5 (General bound for N_f^s)**.** *Let $\alpha \in \mathbb{Q}^\times$. If $\Theta(H_{f,\alpha}) > 0$, then $\theta_f \geq \frac{1}{2}$ and*

$$N_f^s(T) = \Omega\left(T^{\frac{1}{3}(1-\theta_f)+\Theta(H_{f,\alpha})-\frac{1}{2}-\varepsilon}\right)$$

for any $\varepsilon > 0$.

Proof. Let $\beta_n + i\gamma_n$ run through the poles of Δ_f with $\beta_n > 0$, in increasing order of $|\gamma_n|$. For $\sigma \in (0, 1]$, Lemma 6 and Lemma 7 give

$$\int_0^{\frac{|\alpha|}{4}} |S_f(\alpha + iy)| y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} \ll_{f,\alpha,\sigma,\varepsilon} \sum_{n=1}^{\infty} (1 + |\gamma_n|)^{\frac{1}{3}|\beta_n - \frac{1}{2}| + \frac{1}{3} - \sigma + \varepsilon} \tag{2.31}$$

for any $\varepsilon > 0$. If $\Theta(H_{f,\alpha}) > 0$, set $\sigma = \Theta(H_{f,\alpha}) - \varepsilon$, where $0 < \varepsilon < \Theta(H_{f,\alpha})$ is arbitrary. Then Proposition 2 implies that (2.31) diverges, so in particular Δ_f has infinitely many poles $\beta_n + i\gamma_n$ with $\beta_n > 0$, and therefore $\theta_f \geq \frac{1}{2}$.

Now assume by contradiction that $N_f^s(T) = o\left(T^{\frac{1}{3}(1-\theta_f)+\Theta(H_{f,\alpha})-\frac{1}{2}-3\varepsilon}\right)$ for some $0 < \varepsilon < \Theta(H_{f,\alpha})$. Then by (2.31), since $\left|\beta_n - \frac{1}{2}\right| \leq \theta_f - \frac{1}{2}$, we get

$$\begin{aligned}
\infty = \int_0^{\frac{|\alpha|}{4}} |S_f(\alpha + iy)| y^{\sigma+\frac{k-1}{2}} \frac{dy}{y} &\ll_{f,\alpha,\varepsilon} \sum_{n=1}^{\infty} (1 + |\gamma_n|)^{\frac{1}{3}\theta_f + \frac{1}{6} - \Theta(H_{f,\alpha}) + 2\varepsilon} \\
&\ll_f 1 + \int_1^{\infty} t^{\frac{1}{3}\theta_f + \frac{1}{6} - \Theta(H_{f,\alpha}) + 2\varepsilon} \, dN_f^s(t) \\
&\ll_f 1 + \int_1^{\infty} t^{\frac{1}{3}\theta_f + \frac{1}{6} - \Theta(H_{f,\alpha}) + 2\varepsilon - 1} N_f^s(t) \, dt \\
&= 1 + \int_1^{\infty} o\left(t^{-1-\varepsilon}\right) \, dt < \infty,
\end{aligned}$$

which is a contradiction.

□

Remark 3. *Assuming a subconvexity exponent $\mu \in \left[0, \frac{1}{2}\right]$ as in Remark 2, the result of Proposition 5 becomes $N_f^s(T) = \Omega\left(T^{(1-2\mu)(1-\theta_f)+\Theta(H_{f,\alpha})-\frac{1}{2}-\varepsilon}\right)$ for any $\varepsilon > 0$.*

Observe that Proposition 5 fails to give a power of T (even if the subconvexity exponent were to be improved) if $\theta_f = 1$ and $\Theta(H_{f,\alpha}) = \frac{1}{2}$, which cannot be ruled out with what we have done so far. However, we will use this proposition for the case of θ_f sufficiently far from 1, where it gives a good bound.

Corollary 1 (Main result for θ_f away from 1)**.** *We have $\theta_f \geq \frac{1}{2}$ and*

$$N_f^s(T) = \Omega\left(T^{\frac{1}{3}(1-\theta_f)-\varepsilon}\right)$$

for any $\varepsilon > 0$.

Proof. By the functional equation $\Lambda_f(s) = \epsilon_f N^{\frac{1}{2}-s}\Lambda_{\overline{f}}(1-s)$, we have $N_f^s(T) = N_{\overline{f}}^s(T)$ and $\theta_f = \theta_{\overline{f}}$. By Proposition 4, there is $\alpha_f \in \mathbb{Q}^\times$ such that

$$\max\left\{\Theta(H_{f,\alpha_f}), \Theta(H_{\overline{f},\alpha_f})\right\} \geq \frac{1}{2}.$$

Then applying Proposition 5 to either f or $\overline{f}$ gives the desired result.

□

Improvements for θ_f close to 1

If θ_f is close to 1, then either Δ_f or $\Delta_{\overline{f}}$ must have a pole ρ with real part close to 1. We will show that if for instance that is the case for Δ_f, then there exists $\alpha \in \mathbb{Q}^\times$ such that $H_{f,\alpha}$ also has a pole at ρ, so Proposition 5 gives a much stronger result than before. The main tool for showing such a pole transference will be a certain zero density estimate, which we introduce now.

For a primitive form $g \in S_k(\Gamma_1(M))$, $\beta \in \mathbb{R}$, and $T \geq 0$, let

$$N_g(\beta, T) := \left|\{s \in \mathbb{C} : \Re(s) \geq \beta, |\Im(s)| \leq T, \text{ and } L_g(s) = 0\}\right|, \tag{2.32}$$

where the zeros are counted with multiplicity.

Lemma 8 (Zero density for twists close to the line 1)**.** *Let $f \in S_k(\Gamma_1(N))$ be a primitive form. For each prime $p \equiv 1 \pmod N$, let $\psi_p \pmod p$ be an arbitrary non-trivial character modulo p. Then for any $T \geq 2$, $X \geq 2$, $\varepsilon > 0$, and $\frac{3}{4} \leq \beta \leq 1$, we have*

$$\sum_{\substack{p \leq X \text{ prime} \\ p \equiv 1 \pmod N}} N_{f\otimes\psi_p}(\beta, T) \ll_{f,\varepsilon,T} X^{4(1-\beta)+\varepsilon} + X^{\frac{6(1-\beta)}{3\beta-1}+\varepsilon}.$$

Proof. This follows directly from the more general result of Proposition 11 in Appendix A.

□

Now, let κ be such that $\frac{6(1-\kappa)}{3\kappa-1} = 1$, i.e. $\kappa = \frac{7}{9}$. The important point is that $\kappa < 1$.

Proposition 6 (Ruling out pole cancellation in $H_{f,\alpha}$ via zero density)**.** *If Δ_f has a pole $\rho = \beta + i\gamma$ with $\beta > \kappa$, then there exists some $\alpha \in \mathbb{Q}^\times$ (depending on f and ρ) such that ρ is also a pole of $H_{f,\alpha}$.*

Proof. We will show that there exists a prime p satisfying $p \equiv 1 \pmod{N}$ such that ρ is a pole of either $H_{f,1}$ or $H_{f,\frac{1}{p}}$, so we will be able to pick $\alpha = 1$ or $\alpha = \frac{1}{p}$.

Suppose by contradiction that ρ is not a pole of either $H_{f,1}$ or $H_{f,\frac{1}{p}}$ for any prime $p \equiv 1 \pmod{N}$. By (2.14) we have

$$p^{1-2s}H_{f,1}(s) - H_{f,\frac{1}{p}}(s) = p^{1-2s}\Delta_f(s) - \Delta_{f,1,p}(s) + R_{f,p}(p^{-s})\Lambda_f(s),$$

and by assumption this meromorphic function does not have a pole at $s = \rho$. Observe that since $|\lambda_f(p)| \leq 2$ by Deligne's bound, the poles of $R_{f,p}(p^{-s})\Lambda_f(s)$ all satisfy $\Re(s) = 0$, so ρ is not a pole of $R_{f,p}(p^{-s})\Lambda_f(s)$ (as $\kappa > 0$). Hence it also cannot be a pole of

$$\begin{aligned} p^{1-2s}\Delta_f(s) - \Delta_{f,1,p}(s) = &\left(p^{1-2s} - 1 + \frac{p}{p-1}P_{f,p}(p^{-s})\right)\Delta_f(s) \\ &- \frac{1}{p-1}\sum_{\substack{\psi \pmod{p} \\ \psi \neq \psi_0}} \tau\left(\overline{\psi}\right)\Delta_{f\otimes\psi}(s), \end{aligned}$$

where $\psi_0 \pmod{p}$ denotes the trivial character, and we have used (2.12).

Since $\xi(p) = 1$, a direct computation gives

$$p^{1-2s} - 1 + \frac{p}{p-1}P_{f,p}(p^{-s}) = \frac{p^{2-2s} - \lambda_f(p)p^{1-s} + 1}{p-1} = \frac{1}{p-1}P_{f,p}(p^{1-s}).$$

Observe that $P_{f,p}(p^{1-\rho}) \neq 0$, as $\Re(1-\rho) = 1 - \beta \neq 0$, since $\beta < 1$ by [61]. Furthermore,

$$\operatorname*{Res}_{s=\rho}\Delta_f(s) = -\Gamma_{\mathbb{C}}\left(\rho + \frac{k-1}{2}\right)L_f'(\rho) = -\Lambda_f'(\rho) \neq 0,$$

as ρ is a simple zero of Λ_f. We conclude that

$$\sum_{\substack{\psi \pmod{p} \\ \psi \neq \psi_0}} \tau\left(\overline{\psi}\right) \cdot \operatorname*{Res}_{s=\rho}\Delta_{f\otimes\psi}(s) = \operatorname*{Res}_{s=\rho}P_{f,p}(p^{1-s})\Delta_f(s) = -P_{f,p}(p^{1-\rho})\Lambda_f'(\rho) \neq 0,$$

so there is at least one non-trivial character $\psi_p \pmod p$ such that $\Delta_{f\otimes\psi_p}$ has a pole at ρ, or in other words $\Lambda_{f\otimes\psi_p}$ has a simple zero at $\rho = \beta + i\gamma$. This holds for every prime $p \equiv 1 \pmod N$, so it follows that

$$\sum_{\substack{p\leq X \text{ prime} \\ p\equiv 1 \pmod N}} N_{f\otimes\psi_p}(\beta, 2+|\gamma|) \geq \pi(X; N, 1) \gg_f \frac{X}{\log X} \tag{2.33}$$

for every X sufficiently large (in terms of N). However, applying Lemma 8 we conclude that

$$\sum_{\substack{p\leq X \text{ prime} \\ p\equiv 1 \pmod N}} N_{f\otimes\psi_p}(\beta, 2+|\gamma|) \ll_{f,\varepsilon,\rho} X^{4(1-\beta)+\varepsilon} + X^{\frac{6(1-\beta)}{3\beta-1}+\varepsilon} \tag{2.34}$$

for every $X \geq 2$ and $\varepsilon > 0$. Observe that both $\frac{6(1-x)}{3x-1}$ and $4(1-x)$ are strictly decreasing for $\frac{3}{4} \leq x \leq 1$, so since $\beta > \kappa$, $\frac{6(1-\kappa)}{3\kappa-1} = 1$, and $4(1-\kappa) = \frac{8}{9} < 1$, we conclude that $4(1-\beta)+\varepsilon < 1$ and $\frac{6(1-\beta)}{3\beta-1} + \varepsilon < 1$ for $\varepsilon > 0$ sufficiently small. But this is a contradiction, as (2.33) and (2.34) imply

$$X^{4(1-\beta)+\varepsilon} + X^{\frac{6(1-\beta)}{3\beta-1}+\varepsilon} \gg_{f,\varepsilon,\rho} \frac{X}{\log X}$$

for every $\varepsilon > 0$ and X sufficiently large, which cannot hold for small $\varepsilon > 0$ when $X \to \infty$. Therefore, the desired result follows by contradiction.

□

Corollary 2 (Main result for θ_f close to 1)**.** *If $\theta_f > \kappa$, then*

$$N_f^s(T) = \Omega\left(T^{\frac{2}{3}\theta_f - \frac{1}{6} - \varepsilon}\right)$$

for any $\varepsilon > 0$.

Proof. If $\theta_f > \kappa$, then for any given $0 < \varepsilon < \theta_f - \kappa$, either Δ_f or $\Delta_{\overline{f}}$ must have a pole $\rho = \beta + i\gamma$ with $\beta > \theta_f - \varepsilon$. Since $\theta_f - \varepsilon > \kappa$, by Proposition 6 there exists some $\alpha \in \mathbb{Q}^\times$ (depending on f and ρ) such that ρ is also a pole of either $H_{f,\alpha}$ or $H_{\overline{f},\alpha}$. Therefore,

$$\max\left\{\Theta(H_{f,\alpha}), \Theta(H_{\overline{f},\alpha})\right\} \geq \beta > \theta_f - \varepsilon.$$

Then we can use the relations $N_f^s(T) = N_{\overline{f}}^s(T)$ and $\theta_f = \theta_{\overline{f}}$ to get the desired result after applying Proposition 5 to either f or $\overline{f}$, since $\varepsilon > 0$ can be chosen arbitrarily small.

□

Obtaining a power bound

The proof of our main theorem easily follows from what we have done so far.

Proof of Theorem 1. If $\theta_f > \kappa$, we apply Corollary 2 and observe that $\frac{2}{3}\theta_f - \frac{1}{6} > \frac{2}{3}\kappa - \frac{1}{6} = \frac{19}{54}$ to get

$$N_f^s(T) = \Omega\left(T^{\frac{2}{3}\theta_f - \frac{1}{6} - \varepsilon}\right) = \Omega\left(T^{\frac{19}{54} - \varepsilon}\right)$$

for any $\varepsilon > 0$, so in this case we have a rather strong bound.

Otherwise, if $\theta_f \leq \kappa$, we apply Corollary 1 and observe that $\frac{1}{3}(1 - \theta_f) \geq \frac{1}{3}(1 - \kappa) = \frac{2}{27}$ to get

$$N_f^s(T) = \Omega\left(T^{\frac{1}{3}(1-\theta_f) - \varepsilon}\right) = \Omega\left(T^{\frac{2}{27} - \varepsilon}\right)$$

for any $\varepsilon > 0$. In either case, we obtain the desired result.

□

2.5 An improved estimate for f of level 1

If f has level $N = 1$, then we will easily see that there is $\alpha \in \mathbb{Q}^\times$ with $\Theta(H_{f,\alpha}) \geq \theta_f$, so Proposition 5 gives $N_f^s(T) = \Omega\left(T^{\frac{2}{3}\theta_f - \frac{1}{6} - \varepsilon}\right) = \Omega\left(T^{\frac{1}{6} - \varepsilon}\right)$ for any $\varepsilon > 0$, as was proved in [21]. We improve this result by using the sixth moment bound of Jutila [63] instead of subconvexity in the last step of the argument. An improvement in the exponent for the case of general level N would also follow by the same reasoning, except that at present a sixth moment bound does not seem to be in the literature in such generality.

To begin, we convert pointwise values of our L-function into moments via the following standard lemma.

Lemma 9 (Pointwise values to short moments)**.** *Let $f \in S_k(\Gamma_1(N))$ be a primitive form and $T \geq 2$. For $\rho = \beta + i\gamma$ with $\beta \geq \frac{1}{2}$ and $|\gamma| \leq T$, we have*

$$L_f'(\rho) \ll_f \log^4 T + \log^5 T \cdot \int_{-\log^2 T}^{\log^2 T} \left| L_{\overline{f}}\left(\frac{1}{2} - i(\gamma + x)\right)\right| dx.$$

Proof. Let $c = \frac{1}{100 \log T}$. Observe that

$$\frac{1}{2\pi i}\int_{1-i\infty}^{1+i\infty} L_f(\rho + w)\Gamma(w)^2\, dw \ll 1,$$

as $\Re(\rho+w)\geq\frac{3}{2}$. Shifting the line of integration to $\Re(w)=\frac{1}{2}-\beta-c$, we pick up a pole at $w=0$ with residue $L_f'(\rho)$. By Stirling's formula we have the rough bound

$$\Gamma\left(\frac{1}{2}-\beta-c+it\right)\ll e^{-|t|}\left(\left|\frac{1}{2}-\beta-c\right|+|t|\right)^{-1}\ll e^{-|t|}\left(c+|t|\right)^{-1},$$

so we get

$$L_f'(\rho)\ll 1+\int_{-\infty}^{\infty}\left|L_f\left(\frac{1}{2}-c+i(\gamma+t)\right)\right|e^{-2|t|}(c+|t|)^{-2}\,dt.$$

By convexity,

$$\int_{\pm\frac{1}{2}\log^2 T}^{\pm\infty}\left|L_f\left(\frac{1}{2}-c+i(\gamma+t)\right)\right|e^{-2|t|}(c+|t|)^{-2}\,dt\ll_f\int_{\frac{1}{2}\log^2 T}^{\infty}e^{-t}\,dt\ll 1,$$

therefore

$$L_f'(\rho)\ll_f 1+\log^2 T\cdot\int_{-\frac{1}{2}\log^2 T}^{\frac{1}{2}\log^2 T}\left|L_f\left(\frac{1}{2}-c+i(\gamma+t)\right)\right|dt. \tag{2.35}$$

The functional equation combined with Stirling's formula gives

$$\begin{aligned}L_f\left(\frac{1}{2}-c+i(\gamma+t)\right)&\ll_f\frac{\Gamma_{\mathbb{C}}\left(\frac{k}{2}+c-i(\gamma+t)\right)}{\Gamma_{\mathbb{C}}\left(\frac{k}{2}-c+i(\gamma+t)\right)}L_{\overline{f}}\left(\frac{1}{2}+c-i(\gamma+t)\right)\\&\ll L_{\overline{f}}\left(\frac{1}{2}+c-i(\gamma+t)\right).\end{aligned}$$

Now we use an argument similar to the one above. For $\vartheta=\frac{1}{2}+c-i(\gamma+t)$ we have

$$\frac{1}{2\pi i}\int_{1-i\infty}^{1+i\infty}L_{\overline{f}}(\vartheta+w)\Gamma(w)\,dw\ll 1,$$

and shifting the line of integration to $\Re(w)=-c$, picking up a pole at $w=0$ with residue $L_{\overline{f}}(\vartheta)$, and using $\Gamma(-c+iv)\ll e^{-|v|}(c+|v|)^{-1}$, we argue as before to get

$$L_{\overline{f}}(\vartheta)\ll_f 1+\log T\cdot\int_{-\frac{1}{2}\log^2 T}^{\frac{1}{2}\log^2 T}\left|L_{\overline{f}}\left(\frac{1}{2}-i(\gamma+t-v)\right)\right|dv. \tag{2.36}$$

Inserting (2.36) into (2.35) then gives the desired result.

□

The key new input is the lemma below.

Lemma 10 (Hölder against sixth moment)**.** *Let $f \in S_k(\Gamma_0(1))$ be a primitive form and $T \geq 2$. If $\rho_n = \beta_n + i\gamma_n$ runs through the simple zeros of Λ_f in increasing order of $|\gamma_n|$, then*

$$\sum_{\substack{\beta_n \geq \frac{1}{2} \\ |\gamma_n| \leq T}} |L_f'(\rho_n)| \ll_{f,\varepsilon} N_f^s(T)^{\frac{5}{6}} T^{\frac{1}{3}+\varepsilon}$$

for any $\varepsilon > 0$.

Proof. Denote $K = T + \log^2 T$. By Lemma 9,

$$\sum_{\substack{\beta_n \geq \frac{1}{2} \\ |\gamma_n| \leq T}} |L_f'(\rho_n)| \ll_f N_f^s(T) \log^4 T + \log^5 T \cdot \int_{-K}^{K} \left| L_{\overline{f}}\left(\frac{1}{2} - it\right) \right| \cdot \sum_{\substack{\beta_n \geq \frac{1}{2} \\ |\gamma_n| \leq T}} \mathbb{1}_{|t-\gamma_n| \leq \log^2 T} \, dt.$$

Observe that $N_f^s(x+1) - N_f^s(x) \ll_f \log(2 + |x|)$ by standard zero-density results, so we have the bounds $N_f^s(T) \log^4 T \ll_f N_f^s(T)^{\frac{5}{6}} T^{\frac{1}{3}}$ and

$$\sum_{|\gamma_n| \leq T} \mathbb{1}_{|t-\gamma_n| \leq \log^2 T} \ll_f \log^3 T$$

for any $t \in \mathbb{R}$. Therefore, using Hölder's inequality twice we get

$$\begin{aligned}
&\int_{-K}^{K} \left| L_{\overline{f}}\left(\frac{1}{2} - it\right) \right| \cdot \sum_{\substack{\beta_n \geq \frac{1}{2} \\ |\gamma_n| \leq T}} \mathbb{1}_{|t-\gamma_n| \leq \log^2 T} \, dt \\
&\leq \left(\int_{-K}^{K} \left| L_{\overline{f}}\left(\frac{1}{2} - it\right) \right|^6 dt \right)^{\frac{1}{6}} \left(\int_{-K}^{K} \left(\sum_{|\gamma_n| \leq T} \mathbb{1}_{|t-\gamma_n| \leq \log^2 T} \right)^{\frac{6}{5}} dt \right)^{\frac{5}{6}} \\
&\ll_f \left(\int_{-K}^{K} \left| L_{\overline{f}}\left(\frac{1}{2} - it\right) \right|^6 dt \right)^{\frac{1}{6}} \left((\log^3 T)^{\frac{1}{5}} N_f^s(T) \log^2 T \right)^{\frac{5}{6}}.
\end{aligned}$$

Then Jutila's sixth moment bound [63, Theorem 4.7] gives

$$\int_{-K}^{K} \left| L_{\overline{f}}\left(\frac{1}{2} - it\right) \right|^6 dt \ll_{f,\varepsilon} K^{2+\varepsilon} \ll_{\varepsilon} T^{2+\varepsilon}$$

for any $\varepsilon > 0$, and the lemma follows.

□

We are now ready to obtain the desired bound for $N_f^s(T)$.

Proof of Theorem 2. For $f \in S_k(\Gamma_0(1))$ a primitive form, we can apply (2.14) with $p = 2$ to get

$$\begin{aligned} 2^{1-2s}H_{f,1}(s) - H_{f,\frac{1}{2}}(s) &= 2^{1-2s}\Delta_f(s) - \Delta_{f,1,2}(s) + R_{f,2}(2^{-s})\Lambda_f(s) \\ &= \left(2^{1-2s} - 1 + 2P_{f,2}(2^{-s})\right)\Delta_f(s) + R_{f,2}(2^{-s})\Lambda_f(s) \\ &= P_{f,2}(2^{1-s})\Delta_f(s) + R_{f,2}(2^{-s})\Lambda_f(s), \end{aligned}$$

where we have used (2.12). Observe that $P_{f,2}(2^{1-s}) \neq 0$ and $R_{f,2}(2^{-s})$ is holomorphic for $0 < \Re(s) < 1$, so the function above has the same poles as Δ_f in this region. We conclude that

$$\max\left\{\Theta(H_{f,1}), \Theta(H_{f,\frac{1}{2}})\right\} \geq \theta_f.$$

Let $\alpha = 1$ or $\frac{1}{2}$ be such that $\Theta(H_{f,\alpha}) \geq \theta_f$. Also let $0 < \varepsilon < \theta_f$ (recall that $\theta_f \geq \frac{1}{2}$ by Corollary 1) and $\sigma = \theta_f - \varepsilon$. Then since $0 < \sigma < \Theta(H_{f,\alpha})$, Lemma 6 and Proposition 2 give

$$\sum_{\substack{\rho=\beta+i\gamma \\ \text{a pole of } \Delta_f \\ \text{with } \beta>0}} \left|\Lambda_f'(\rho)\right| e^{\frac{\pi|\gamma|}{2}}(1+|\gamma|)^{-\sigma-\frac{k-1}{2}} = \infty. \tag{2.37}$$

By the functional equation, $\Lambda_f'(\rho) = -\epsilon_f N^{\frac{1}{2}-\rho}\Lambda_{\overline{f}}'(1-\rho) \ll_f \Lambda_{\overline{f}}'(1-\rho)$, so the LHS of (2.37) is

$$\ll_f \sum_{\substack{\rho=\beta+i\gamma \\ \text{a pole of } \Delta_f \\ \text{with } \beta\geq\frac{1}{2}}} \left|\Lambda_f'(\rho)\right| e^{\frac{\pi|\gamma|}{2}}(1+|\gamma|)^{-\sigma-\frac{k-1}{2}} + \sum_{\substack{\rho=\beta+i\gamma \\ \text{a pole of } \Delta_{\overline{f}} \\ \text{with } \beta\geq\frac{1}{2}}} \left|\Lambda_{\overline{f}}'(\rho)\right| e^{\frac{\pi|\gamma|}{2}}(1+|\gamma|)^{-\sigma-\frac{k-1}{2}}. \tag{2.38}$$

Applying Stirling's bound $\Gamma\left(\rho + \frac{k-1}{2}\right) \ll (1+|\gamma|)^{\beta+\frac{k}{2}-1}e^{-\frac{\pi|\gamma|}{2}}$, valid for $\beta \geq \frac{1}{2}$, we have

$$\sum_{\substack{\rho=\beta+i\gamma \\ \text{a pole of } \Delta_f \\ \text{with } \beta\geq\frac{1}{2}}} \left|\Lambda_f'(\rho)\right| e^{\frac{\pi|\gamma|}{2}}(1+|\gamma|)^{-\sigma-\frac{k-1}{2}} \ll \sum_{\beta_n\geq\frac{1}{2}} \left|L_f'(\rho_n)\right| (1+|\gamma_n|)^{\beta_n-\sigma-\frac{1}{2}},$$

where we use the notation of Lemma 10. Observing that $\beta_n \leq \theta_f$ and applying Lemma 10, we obtain

$$\sum_{\beta_n \geq \frac{1}{2}} \left|L_f'(\rho_n)\right| (1+|\gamma_n|)^{\beta_n - \sigma - \frac{1}{2}} \ll_f 1 + \sum_{k=1}^{\infty} \sum_{T=2^k} T^{-\frac{1}{2}+\varepsilon} \sum_{\substack{\beta_n \geq \frac{1}{2} \\ \frac{T}{2} < |\gamma_n| \leq T}} |L_f'(\rho_n)|$$

$$\ll_{f,\varepsilon} 1 + \sum_{k=1}^{\infty} \sum_{T=2^k} N_f^s(T)^{\frac{5}{6}} T^{-\frac{1}{6}+2\varepsilon}$$

for any sufficiently small $\varepsilon > 0$.

Now suppose by contradiction that $N_f^s(T) = o\left(T^{\frac{1}{5}-6\varepsilon}\right)$. Then

$$\sum_{\beta_n \geq \frac{1}{2}} \left|L_f'(\rho_n)\right| (1+|\gamma_n|)^{\beta_n - \sigma - \frac{1}{2}} \ll_{f,\varepsilon} 1 + \sum_{k=1}^{\infty} \sum_{T=2^k} o\left(T^{-3\varepsilon}\right) < \infty.$$

The same argument, exchanging f with $\overline{f}$ (and observing that $\theta_f = \theta_{\overline{f}}$), shows that the second term of (2.38) is also finite. This contradicts (2.37), so we conclude that

$$N_f^s(T) = \Omega\left(T^{\frac{1}{5}-\varepsilon}\right)$$

for any $\varepsilon > 0$, as desired.

□

Chapter 3

THE VARIANCE OF CLOSED GEODESICS IN BALLS AND ANNULI ON THE MODULAR SURFACE

3.1 Introduction

Let $\Gamma := \mathrm{PSL}_2(\mathbb{Z})$ denote the modular group and let $D > 0$ be a fundamental discriminant, meaning that D is the discriminant of the real quadratic field $\mathbb{Q}(\sqrt{D})$. There is a well-known correspondence between narrow ideal classes in the narrow class group Cl_D^+ of $\mathbb{Q}(\sqrt{D})$ and Γ-orbits of primitive irreducible integral binary quadratic forms $ax^2 + bxy + cy^2$ of discriminant $b^2 - 4ac = D$. Those, in turn, can also be associated to Γ-orbits of geodesics on the upper half-plane $\mathbb{H}$ with endpoints $\frac{-b\pm\sqrt{D}}{2a}$, or equivalently to the corresponding closed geodesics on the modular surface $\Gamma\backslash\mathbb{H}$.

Denote the set of such closed geodesics of discriminant D by Λ_D. Then $|\Lambda_D| = |\,\mathrm{Cl}_D^+\,| =: h_D^+$, and each closed geodesic in Λ_D has length $2\log\varepsilon_D^+$, where $\varepsilon_D^+ > 1$ is the smallest unit of positive norm in $\mathbb{Q}(\sqrt{D})$. The class number formula then gives

$$\sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C}) = h_D^+ \cdot 2\log\varepsilon_D^+ = 2\sqrt{D}L(1,\chi_D),$$

where χ_D is the primitive quadratic character modulo D and $\ell(\mathcal{C}) := \int_{\mathcal{C}} ds$ denotes the length in $\mathbb{H}$, which is equipped with the hyperbolic metric and corresponding hyperbolic measure given respectively by

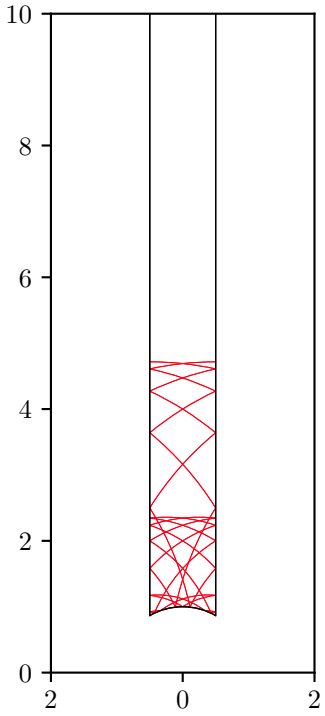
$$ds^2 := \frac{dx^2 + dy^2}{y^2} \quad \text{and} \quad d\mu(z) := \frac{dx\,dy}{y^2}$$

for $z = x + iy$. The bounds $D^{-\varepsilon} \ll_\varepsilon L(1,\chi_D) \ll \log D$ allow us to understand the total length quite well.

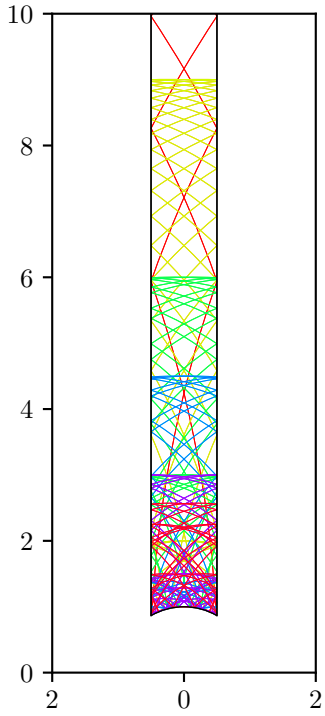
The elements of Λ_D are expected to behave "randomly" in various senses (we will make this more precise below). In that direction, it is known that they become equidistributed in shrinking balls B_R: if we fix $\delta > 0$ and $w \in \Gamma\backslash\mathbb{H}$, then for $D^{-\frac{1}{18}+\delta} \ll R \ll 1$ we have

$$\sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C}\cap B_R(w)) \sim \frac{\mu(B_R)}{\mu(\Gamma\backslash\mathbb{H})} \sum_{\mathcal{C}\in\Lambda_D} \ell(\mathcal{C}) \tag{3.1}$$

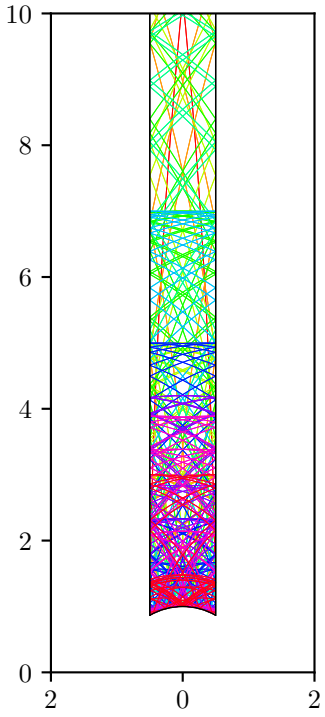
as $D \to \infty$ through squarefree fundamental discriminants.



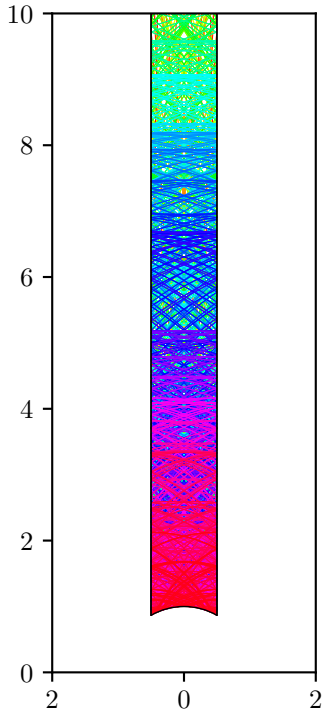
(a) $D = 89$, $h_D^+ = 1$



(b) $D = 1297$, $h_D^+ = 11$



(c) $D = 44101$, $h_D^+ = 19$



(d) $D = 1032257$, $h_D^+ = 80$

Figure 3.1: Closed geodesics in Λ_D

Under the generalized Lindelöf hypothesis we may replace the exponent $1/18$ by $1/6$, and equidistribution is expected to hold for exponents up to $1/2$. Such a result was first proved for fixed R and with a congruence condition on D by Skubenko [100], using Linnik's ergodic method [77, Chapter VI]. The congruence condition was only removed almost 30 years later by Duke [28], following a breakthrough of Iwaniec [57] (see [31] for a history of the problem). The result for shrinking R mentioned above is given by Humphries [53, Theorem 1.24], based on work of Young [106]. Analogous results are also available for geometric invariants in other contexts, such as Heegner points in $\mathbb{H}$ (corresponding to $D < 0$) and lattice points in spheres [41, 30], but we will restrict our attention to closed geodesics.

If one does not require equidistribution for every ball but instead is satisfied with a result covering almost all balls, then it is possible to go further. Considering a random variable given by the LHS of (3.1), where w is distributed according to (a normalized version of) the measure μ, it is tautological that the expected value is equal to the RHS of the same equation. One is then naturally led to consider the variance, which in the more general context of annuli $A_{r,R}(w)$ centered at $w \in \Gamma\backslash\mathbb{H}$, with inner radius r and outer radius R, is given by

$$\mathrm{Var}(r, R; \Lambda_D) := \frac{1}{\mu(\Gamma\backslash\mathbb{H})} \int_{\Gamma\backslash\mathbb{H}} \left(\sum_{\mathcal{C} \in \Lambda_D} \ell(\mathcal{C} \cap A_{r,R}(w)) - \frac{\mu(A_{r,R})}{\mu(\Gamma\backslash\mathbb{H})} \sum_{\mathcal{C} \in \Lambda_D} \ell(\mathcal{C}) \right)^2 d\mu(w). \tag{3.2}$$

Such an expression was first studied by Bourgain, Rudnick, and Sarnak [16] in the context of lattice points in spheres. Based on probabilistic considerations, they conjectured that if the radii satisfy certain mild conditions, then the variance should be asymptotically equal to the corresponding expected value of the underlying random variable. An upper bound was then obtained assuming the generalized Lindelöf hypothesis.

Humphries and Radziwiłł [54] were able to unconditionally prove the conjecture for certain very thin annuli, both in the case of lattice points in spheres and of Heegner points in $\mathbb{H}$. Furthermore, in the case of closed geodesics they obtained equidistribution for almost all annuli by showing that if $0 \leq r < R \ll 1$

and $D^{-1+\delta} \ll \mu(A_{r,R}) \ll 1$ for some fixed $\delta > 0$, then for any fixed $c > 0$,

$$\mu\left(\left\{w \in \Gamma\backslash\mathbb{H} : \left|\frac{\mu(\Gamma\backslash\mathbb{H})}{\mu(A_{r,R})}\frac{\sum_{\mathcal{C}\in\Lambda_D}\ell(\mathcal{C}\cap A_{r,R}(w))}{\sum_{\mathcal{C}\in\Lambda_D}\ell(\mathcal{C})} - 1\right| > c\right\}\right) = o(1)$$

as $D \to \infty$ through squarefree fundamental discriminants. They did so by obtaining the bound $\mathrm{Var}(r, R; \Lambda_D) = o((\mu(A_{r,R})\sqrt{D}L(1, \chi_D))^2)$ in this range, and indeed a careful examination of their method gives in particular

$$\mathrm{Var}(0, R; \Lambda_D) \ll_\varepsilon D^{\frac{1}{2}}R^{3-\varepsilon}$$

for $R \ll D^{-\frac{5}{12}}$. This shows that the variance is not asymptotically equal to the expected value (which for balls is $D^{\frac{1}{2}+o(1)}R^2$, since $\mu(B_R) \asymp R^2$ for $R \ll 1$), as was the case for Heegner points in $\mathbb{H}$ and lattice points in spheres. A deviation of this kind is somewhat unexpected, since it implies better than "square-root cancellation" in (3.2). However, in retrospect such a result is quite reasonable, since the geometric invariants have codimension 1 in the case of closed geodesics, but 2 in the other cases mentioned.

Given the discussion above, it is not completely clear what one should expect for the behavior of $\mathrm{Var}(r, R; \Lambda_D)$, and the purpose of this chapter is to tackle this question. We start by proposing a probabilistic model, using geodesic segments of the appropriate length $2\log\varepsilon_D^+$ taken at random according to the Liouville measure in the unit tangent bundle of $\Gamma\backslash\mathbb{H}$, to model the elements of Λ_D (see Section 3.4 for details). A rigorous analysis of this model turns out to be considerably more complicated than that for the geometric invariants of codimension 2. We make critical use of a quantitative bound on the rate of mixing for the geodesic flow on the modular surface, combined with basic hyperbolic lattice point counting and some elementary hyperbolic geometry, to arrive at an asymptotic formula for the variance in the context of our probabilistic model.

The main result in that direction is Theorem 3, where we show — in the case of balls — that for a single random geodesic segment of length L in $\Gamma\backslash\mathbb{H}$, under mild conditions, the corresponding expression for the variance is $\sim \frac{16LR^3}{\pi}$. For annuli, a certain special function $\mathbf{G}$ appears in the asymptotics (see Lemma 13 for its definition and key properties). We also refer to Section 3.4 for a heuristic explanation of why the factor R^3 (instead of R^2) and the constant $\frac{16}{\pi}$ emerge in the asymptotics for this problem. Finally, it is worth pointing out that Luo

and Sarnak [79] have computed the quantum variance for the geodesic flow. In its classical incarnation, this variance is related to the spectral decomposition of our random model.

Using the analysis of the probabilistic model above, we are able to predict the asymptotic behavior of the variance for closed geodesics. In particular, in the case of balls we conjecture that if $0 < R \leq D^{-\delta}$ for some fixed $\delta > 0$, then

$$\mathrm{Var}(0, R; \Lambda_D) \sim \frac{64\sqrt{D}L(1, \chi_D)R^3}{\pi}$$

as $D \to \infty$ through squarefree fundamental discriminants (see Conjecture 1 for the general case of annuli). Finally, our main result shows that the conjecture is true for balls of small radius.

Corollary 3. *Let $\delta > 0$ be given. If $0 < R \leq D^{-\frac{5}{12}-\delta}$, then as $D \to \infty$ through squarefree fundamental discriminants,*

$$\mathrm{Var}(0, R; \Lambda_D) \sim \frac{64\sqrt{D}L(1, \chi_D)R^3}{\pi}.$$

Indeed, Corollary 3 is a particular case of Theorem 4, where we treat a wide class of annuli and the special function $\mathbf{G}$ appears, as expected. An interesting feature of the result is that the variance depends on the shape of the annulus, and not only on its area. The significance of the exponent $5/12$ and the obstacles towards extending the range of R for which Corollary 3 holds are discussed in Section 3.6.

The proof of Theorem 4 follows a completely different path than that of Theorem 3, and we instead apply the methods of [54] to the case of closed geodesics. What allows us to prove a result for balls in this case is the presence of a different weight function than the one for Heegner points, due to the fact that the Gamma factors that arise when one expresses the relevant Weyl sums in terms of L-functions depend on the sign of D. The fact that the weight function decays faster is also a source of complications, since in our case the main contribution to the variance comes from forms with spectral parameter of size roughly between $1/R$ and $1/(R-r)$, as opposed to just around $1/(R-r)$ for Heegner points. This forces us to deal with the transition range $|x| \asymp 1$ for the Bessel function $J_0(x)$, where clear asymptotics are not available (see Remark 4). Thus instead of approximating with trigonometric functions, we carry the Bessel factors throughout the argument, and after certain integral

transforms they are ultimately what gives rise to the special function **G** mentioned before in the asymptotics for the variance.

Acknowledgments

Thanks to Valentin Blomer, Peter Humphries, Steve Lester, Carlos Matheus, and Zeév Rudnick for helpful comments and suggestions.

3.2 Background and notation

Geometry of the upper half-plane

The distance function $\rho : \mathbb{H} \times \mathbb{H} \to \mathbb{R}_{\geq 0}$ and its more convenient proxy $u : \mathbb{H} \times \mathbb{H} \to \mathbb{R}_{\geq 0}$ are given by

$$\rho(z,w) := \log\left(\frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|}\right)$$

and

$$u(z,w) := \frac{|z-w|^2}{4\Im(z)\Im(w)} = \sinh^2\left(\frac{\rho(z,w)}{2}\right).$$

The group of isometries is $G := \mathrm{PSL}_2(\mathbb{R})$, which acts transitively through fractional linear transformations. The stabilizer of i is $K := \mathrm{PSO}_2(\mathbb{R})$, so $gK \mapsto gi$ gives an identification $G/K \simeq \mathbb{H}$.

Moreover, the corresponding action of G on the unit tangent bundle $T^1(\mathbb{H})$ (through the derivative map) is simply transitive, so if $v \in T^1(\mathbb{H})$ denotes the unit tangent vector pointing up at i then $g \mapsto gv$ gives an identification $G \simeq T^1(\mathbb{H})$. More concretely, we can use the Iwasawa decomposition $G = NAK$, where

$$N := \left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : t \in \mathbb{R} \right\} \quad \text{and} \quad A := \left\{ \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix} : a \in \mathbb{R}_{>0} \right\},$$

to describe this identification as

$$\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y^{1/2} & 0 \\ 0 & y^{-1/2} \end{pmatrix} \begin{pmatrix} \cos(\frac{\theta}{2}) & \sin(\frac{\theta}{2}) \\ -\sin(\frac{\theta}{2}) & \cos(\frac{\theta}{2}) \end{pmatrix} \longleftrightarrow (z,\theta),$$

where θ is the angle with the unit tangent vector pointing up at $z = x + iy$. The derivative action of G on $T^1(\mathbb{H})$ becomes left multiplication in G under the map described above, and the Liouville measure

$$d\nu(z,\theta) := \frac{dx\,dy}{y^2}\frac{d\theta}{2\pi}$$

on $T^1(\mathbb{H})$ is invariant under this action of G, i.e. corresponds (up to a constant multiple) to the left-invariant Haar measure in G under our identification. Furthermore since the group G is unimodular, ν is also right-invariant.

Geometry of the modular surface

Let $X := \Gamma\backslash\mathbb{H}$ denote the modular surface, so that our previous identification quotients out to $X \simeq \Gamma\backslash G/K$, and similarly for the unit tangent bundle[1] identification $T^1(X) \simeq \Gamma\backslash G$. The metric space structure of X is obtained from the distance function

$$\widetilde{\rho}(\Gamma z, \Gamma w) := \min_{\gamma\in\Gamma} \rho(z, \gamma w).$$

Considering the usual (closure of a) fundamental domain

$$\mathcal{F} := \left\{ z \in \mathbb{H} : |\Re(z)| \leq \frac{1}{2} \quad \text{and} \quad |z| \geq 1 \right\},$$

we can define measures $\widetilde{\mu}$ and $\widetilde{\nu}$ in X and $T^1(X)$, respectively, by

$$\widetilde{\mu}(\Gamma A) := \mu\left(\bigcup_{\gamma\in\Gamma} \gamma A \cap \mathcal{F}\right) \quad \text{and} \quad \widetilde{\nu}(\Gamma B) := \nu\left(\bigcup_{\gamma\in\Gamma} \gamma B \cap \pi^{-1}(\mathcal{F})\right)$$

for measurable $A \subset \mathbb{H}$ and $B \subset T^1(\mathbb{H})$, where $\pi : T^1(\mathbb{H}) \to \mathbb{H}$ is the projection map. In particular, $\widetilde{\nu}(T^1(X)) = \nu(\pi^{-1}(\mathcal{F})) = \pi/3 = \mu(\mathcal{F}) = \widetilde{\mu}(X)$. Both measures are G-invariant under multiplication on the right, since the particular choice of fundamental domain turns out to be immaterial.

Geodesic flow

Given $t \in \mathbb{R}$, the geodesic flow $\mathcal{G}_t : T^1(\mathbb{H}) \to T^1(\mathbb{H})$ is

$$\mathcal{G}_t(g) := g \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$$

for $g \in G \simeq T^1(\mathbb{H})$, and in geometric terms it amounts to parallel transport along the geodesic with starting point and direction given by the element of $T^1(\mathbb{H})$ corresponding to g, for (hyperbolic) signed length t. The right-invariance of the Liouville measure ν implies that it is preserved by $\mathcal{G}_t$.

The geodesic flow clearly commutes with left multiplication by G (and in particular by Γ), so it descends to a well-defined map $\widetilde{\mathcal{G}}_t : T^1(X) \to T^1(X)$ given by

$$\widetilde{\mathcal{G}}_t(\Gamma g) := \Gamma g \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$$

[1]Technically the modular surface has singularities at i and $\frac{1+i\sqrt{3}}{2}$, since these points have nontrivial stabilizer in Γ. To correctly interpret the unit tangent bundle $T^1(X)$ we need to consider the orbifold structure of X, but this minor issue can be safely ignored for our purposes.

for $\Gamma g \in T^1(X) \simeq \Gamma\backslash G$. Once again, $\widetilde{\mathcal{G}}_t$ preserves $\widetilde{\nu}$ and amounts to parallel transport by (hyperbolic) signed length t along the corresponding geodesic in X.

3.3 Estimates for the Selberg–Harish-Chandra transform

Definitions

We follow [54] with some minor modifications.

Let $k_{r,R}(u(z,w))$ be the identity function of the annulus

$$\begin{aligned} A_{r,R}(w) &:= \{z \in \mathbb{H} : r \leq \rho(z,w) \leq R\} \\ &= \left\{ z \in \mathbb{H} : \sinh^2\left(\frac{r}{2}\right) \leq u(z,w) \leq \sinh^2\left(\frac{R}{2}\right) \right\} \end{aligned}$$

of hyperbolic volume

$$\mu(A_{r,R}) := \mu(A_{r,R}(w)) = 4\pi\left(\sinh^2\left(\frac{R}{2}\right) - \sinh^2\left(\frac{r}{2}\right)\right),$$

that is,

$$k_{r,R}(t) := \begin{cases} 1 & \text{if } \sinh^2\left(\frac{r}{2}\right) \leq t \leq \sinh^2\left(\frac{R}{2}\right), \\ 0 & \text{otherwise.} \end{cases}$$

Observe that we use a different normalization from [54] both here and in what follows below. Since $k_{r,R}(u(z,w))$ is a point-pair invariant, we can define the automorphic kernel $K_{r,R} : X \times X \to \mathbb{R}_{\geq 0}$ given by

$$K_{r,R}(z,w) := \sum_{\gamma \in \Gamma} k_{r,R}(u(z,\gamma w)).$$

The spectral expansion of this kernel involves the Selberg–Harish-Chandra transform $h_{r,R}$ of $k_{r,R}$, which is given by

$$\begin{aligned} h_{r,R}(t) &:= 2\pi \int_0^\infty P_{-\frac{1}{2}+it}(\cosh\rho) k_{r,R}\left(\sinh^2\left(\frac{\rho}{2}\right)\right) \sinh\rho \, d\rho \\ &= 2\pi \int_r^R P_{-\frac{1}{2}+it}(\cosh\rho) \sinh\rho \, d\rho, \end{aligned} \tag{3.3}$$

where P_λ is the Legendre function of the first kind.

Bounds and asymptotics for $h_{r,R}$

To understand the behavior of $h_{r,R}$ we express $P_{-\frac{1}{2}+it}$ in terms of Bessel functions, which will be more convenient to evaluate under the various integral transforms that will arise later.

Lemma 11 (Hilb's formula [54, Lemma 2.24])**.** *Fix $\varepsilon > 0$. For $t \in \mathbb{R}$ and $0 < \rho < 1/\varepsilon$,*

$$P_{-\frac{1}{2}+it}(\cosh \rho) = \sqrt{\frac{\rho}{\sinh \rho}} J_0(\rho t) + \begin{cases} O(\rho^2) & \text{for } |t| \le \frac{1}{\rho}, \\ O_\varepsilon\left(\frac{\sqrt{\rho}}{|t|^{3/2}}\right) & \text{for } |t| \ge \frac{1}{\rho} \ge \varepsilon. \end{cases}$$

With this in mind, an asymptotic formula for $h_{r,R}$ easily follows. We restrict our attention to the case $R - r \gg R$, which will be relevant to us, but a similar statement also holds in the complementary case.

Lemma 12. *Suppose that $0 \le r < R \ll 1$ satisfy $R - r \gg R$, and $t \in \mathbb{R}$. Then*

$$h_{r,R}(t) = 2\pi \frac{R \cdot J_1(Rt) - r \cdot J_1(rt)}{t} + \begin{cases} O(R^4) & \text{for } |t| \le \frac{1}{R}, \\ O\left(\frac{R^{7/2}}{\sqrt{|t|}}\right) & \text{for } |t| \ge \frac{1}{R}. \end{cases}$$

Furthermore,

$$h_{r,R}(t) \ll \begin{cases} R^2 & \text{for } |t| \le \frac{1}{R}, \\ \frac{\sqrt{R}}{|t|^{3/2}} & \text{for } |t| \ge \frac{1}{R}. \end{cases}$$

Proof. Plugging Lemma 11 into (3.3) gives

$$h_{r,R}(t) = 2\pi \int_r^R \sqrt{\rho \sinh \rho} \cdot J_0(\rho t)\, d\rho + \begin{cases} O\left(R^4\right) & \text{for } |t| \le \frac{1}{R}, \\ O\left(\frac{R^{5/2}}{|t|^{3/2}}\right) & \text{for } |t| \ge \frac{1}{R}. \end{cases}$$

Using $\sinh \rho \ll \rho$, the bounds

$$J_0(x) = \begin{cases} 1 + O(x^2) & \text{for } |x| \le 1, \\ \sqrt{\frac{2}{\pi |x|}} \cos\left(|x| - \frac{\pi}{4}\right) + O\left(\frac{1}{|x|^{3/2}}\right) & \text{for } |x| \ge 1 \end{cases} \tag{3.4}$$

for $x \in \mathbb{R}$ [42, 8.411.1 and 8.451.1], and integrating by parts in the case $|t| \ge 1/R$ (antidiferentiating the cosine term) gives the desired upper bound for $h_{r,R}$, as in [54, Lemma 2.33]. For the first asymptotic statement we use instead $\sinh \rho = \rho + O(\rho^3)$ combined with (3.4) to get

$$h_{r,R}(t) = 2\pi \int_r^R \rho \cdot J_0(\rho t)\, d\rho + \begin{cases} O\left(R^4\right) & \text{for } |t| \le \frac{1}{R}, \\ O\left(\frac{R^{7/2}}{\sqrt{|t|}}\right) & \text{for } |t| \ge \frac{1}{R}. \end{cases}$$

We can directly evaluate the remaining integral, since [42, page 8.472.1] yields $(xJ_1(x))' = xJ_0(x)$, and the result follows.

□

Remark 4. *The reason we keep an expression with Bessel functions in the result above, instead of using* (3.4) *as in [54, Lemma 2.27] to write it in terms of simpler trigonometric functions, is that the main term in our variance computation will come roughly from* $|t| \asymp 1/R$*. This can be seen from the ranges of integration for the main term in* (3.39)*, as defined in* (3.35)*. In the case* $R - r \ll R$ *that range would be roughly* $1/R \ll |t| \ll 1/(R-r)$*, so either way we must deal with the transition range* $|x| \asymp 1$ *for* $J_0(x)$*, and* (3.4) *is not good enough to obtain asymptotics there.*

In contrast, the main term in [54, (7.18)] — with relevant ranges defined in [54, (7.11)] — turns out to come roughly from $|t| \asymp 1/(R-r)$*, which is much larger than* $1/R$ *(with the assumptions present there), so one still obtains an asymptotic for the Bessel function in the most important range. The main difference between the two cases is the presence of the extra weight* $H(t)$ *given by* (3.22) *in the spectral expansion of the variance for closed geodesics, which is not present in the case of Heegner points considered by Humphries and Radziwiłł (see [54, Lemma 2.13] for a comparison of the two weight functions).*

3.4 Variance for random geodesic segments

Since the closed geodesics in Λ_D are expected to behave in many aspects like "random geodesics", we will model them using uniformly distributed geodesic segments in X, so first we must understand the variance in that case.

By a geodesic segment of length L in X, we mean a curve in X of the form $\widetilde{\pi} \circ \widetilde{\mathcal{G}}_t(g)$ for $0 \leq t \leq L$ (observe that it is parametrized by hyperbolic arc length), where $g \in T^1(X)$ and $\widetilde{\pi} : T^1(X) \to X$ denotes the projection map. Uniform distribution means that the initial condition $g \in T^1(X)$ is distributed (up to normalization) according to the Liouville measure $\widetilde{\nu}$.

Given the discussion above, the random variable given by the length of the intersection between a random geodesic segment of length L in X with a random annulus $A_{r,R}$ in X (with center distributed independently of the geodesic

segment and according to the normalized measure in X) has variance

$$\begin{aligned}&\mathrm{Var}(r,R;L):=\\&\int_{T^1(X)}\int_X\left(\int_0^L K_{r,R}(\widetilde{\pi}\circ\widetilde{\mathcal{G}}_t(g),w)\,dt-L\frac{\mu(A_{r,R})}{\widetilde{\mu}(X)}\right)^2\frac{d\widetilde{\mu}(w)}{\widetilde{\mu}(X)}\frac{d\widetilde{\nu}(g)}{\widetilde{\nu}(T^1(X))}=\\&\int_{\pi^{-1}(\mathcal{F})}\int_{\mathcal{F}}\left(\int_0^L\sum_{\gamma\in\Gamma}k_{r,R}(u(\pi\circ\mathcal{G}_t(g),\gamma w))\,dt-L\frac{\mu(A_{r,R})}{\mu(\mathcal{F})}\right)^2\frac{d\mu(w)}{\mu(\mathcal{F})}\frac{d\nu(g)}{\mu(\mathcal{F})}.\end{aligned}\tag{3.5}$$

Heuristics

This discussion is partially inspired by the heuristics in [72].

For simplicity, we consider only the case of balls B_R and assume $R=o(1)$. Let $\mathcal{S}$ be a random geodesic segment of length $L\gg 1$ on the modular surface X, and let $Y=Y(w):=\ell(\mathcal{S}\cap B_R(w))$, where w is uniformly distributed in X. Also denote by $\mathcal{S}^R$ the "tube" of radius R around $\mathcal{S}$. We wish to compute $\mathrm{Var}(Y)$, and will think of $\mathcal{S}$ as fixed but "generic".

First suppose that $LR=o(1)$. Observe that $Y(w)=0$ precisely for $w\notin\mathcal{S}^R$, while $Y(w)\asymp R$ for most $w\in\mathcal{S}^R$ — certainly for a typical $w\in\mathcal{S}^{\frac{R}{2}}$, since we expect this tube to have few self-intersections, as it has area $\asymp LR=o(1)$. Therefore, $\mathbb{E}(Y^2)\asymp LR^3$ and $\mathbb{E}(Y)^2\asymp L^2R^4=o(LR^3)$, so $\mathrm{Var}(Y)\asymp LR^3$. Furthermore, if for instance $\mathcal{S}^R$ has no self-intersections, then we can unfold it to $\mathbb{H}$ and obtain asymptotics for the variance using elementary hyperbolic geometry.

For the complementary case, suppose (say) that $LR\gg R^{\frac{1}{10}}$. We let $T\asymp R^{-\frac{9}{10}}$ and split $\mathcal{S}$ into $\frac{L}{T}\gg 1$ pieces $\mathcal{S}_1,\mathcal{S}_2,\ldots,\mathcal{S}_{\frac{L}{T}}$ of length T. Since the geodesic flow is mixing (of all orders) and $T\to\infty$, we expect these segments to essentially behave independently. Let $Y_i=Y_i(w):=\ell(\mathcal{S}_i\cap B_R(w))$. Observe that $TR=o(1)$, so by the previous case we should have $\mathrm{Var}(Y_i)\asymp TR^3$. Then independence gives $\mathrm{Var}(Y)=\mathrm{Var}\left(\sum_{i=1}^{\frac{L}{T}}Y_i\right)\approx\sum_{i=1}^{\frac{L}{T}}\mathrm{Var}(Y_i)\asymp LR^3$, and it is reasonable to expect asymptotics for $\mathrm{Var}(Y)$ if we could obtain those for each $\mathrm{Var}(Y_i)$.

In fact, one may heuristically determine the constant in $\mathrm{Var}(Y)\sim\frac{16LR^3}{\pi}$ as follows. It suffices to consider the case $LR=o(1)$, by the argument using independence from the previous paragraph. The tube $\mathcal{S}^R$ has few self-intersections in that case, and the geometry of the problem is essentially Euclidean (as we

work at the scale $R = o(1)$). Therefore, our situation can be modeled by the toy problem where the geodesic segment $\mathcal{S}$ is replaced by a straight line segment of length L in $\mathbb{R}^2$, and the the point w is randomized over some region of the appropriate area $\mu(X) = \frac{\pi}{3}$ that contains the (now Euclidean) tube $\mathcal{S}^R$. Since $\mathbb{E}(Y^2 \mid w \notin \mathcal{S}^{\mathcal{R}}) = 0$, we get $\mathbb{E}(Y^2) = \mathbb{E}(Y^2 \mid w \in \mathcal{S}^R) \cdot \mathbb{P}(w \in \mathcal{S}^R)$. Away from the endpoints of $\mathcal{S}$, $Y = 2\sqrt{R^2 - x^2}$ depends only on the (signed) distance $x = x(w)$ from w to the line that contains the segment $\mathcal{S}$, so we obtain

$$\begin{aligned}\mathbb{E}(Y^2 \mid w \in \mathcal{S}^R) \cdot \mathbb{P}(w \in \mathcal{S}^R) &\approx \frac{1}{2R} \int_{-R}^{R} \left(2\sqrt{R^2 - x^2}\right)^2 \, dx \cdot \frac{2RL}{\mu(X)} \\ &= \frac{8R^2}{3} \cdot \frac{6RL}{\pi} = \frac{16LR^3}{\pi}.\end{aligned}$$

These heuristics provide a good intuition for the upcoming arguments in this section, but we will have to do something more complicated to effectively deal with self-intersections of $\mathcal{S}^R$.

The cuspidal contribution

Before delving into the variance computation, we need to make a small technical modification to (3.5), since with the current definition it turns out that $\mathrm{Var}(r, R; L) = \infty$ for $L \gg 1$. This is essentially due to the fact that the automorphic kernel $K_{r,R}(z, w)$ becomes quite large as z and w go towards the cusp together, so it is in particular not in $L^2(\mathcal{F} \times \mathcal{F})$. This is the same issue that gives rise to continuous spectrum in the spectral resolution of the Laplacian in X.

For simplicity, consider the case of balls B_R, so $r = 0$. For $k \in \mathbb{Z}$ we have $\rho(w, w+k) \ll \sinh\left(\frac{\rho(w,w+k)}{2}\right) = \sqrt{u(w, w+k)} = \frac{|k|}{2\Im(w)}$, so for $\gg R\Im(w)$ values of $k \in \mathbb{Z}$ we have $\rho(w, w+k) \leq \frac{R}{4}$. Therefore, for all g with $\pi(g) \in B_{\frac{R}{4}}(w)$ there are $\gg R\Im(w)$ values of $k \in \mathbb{Z}$ such that $\rho(\pi(g), w+k) \leq \rho(\pi(g), w) + \rho(w, w+k) \leq \frac{R}{2}$. Also observe that if $\rho(\pi(g), w+k) \leq \frac{R}{2}$ then $\pi(\mathcal{G}_t(g)) \in B_R(w+k)$ for all $0 \leq t \leq \frac{R}{2}$.

We conclude that if $L \gg 1 \gg R$ then

$$\int_{\pi^{-1}(\mathcal{F})} \int_{\mathcal{F}} \left(\int_0^L \sum_{\gamma \in \Gamma} k_{0,R}(u(\pi \circ \mathcal{G}_t(g), \gamma w)) \, dt \right)^2 d\mu(w) \, d\nu(g) \gg$$

$$\int_{\mathcal{F}} \int_{\mathcal{F} \cap B_{\frac{R}{4}}(w)} \left(\frac{R}{2} R\Im(w)\right)^2 d\mu(z) \, d\mu(w) \gg \int_{\frac{1}{R}}^{\infty} \frac{R}{y} (R^2 y)^2 \frac{dy}{y^2} = \infty$$

and this gives $\mathrm{Var}(0, R; T) = \infty$.

The truncated variance

In view of the necessity to exclude the contribution from the cusp, we let

$$\mathcal{F}_A := \{z \in \mathcal{F} : \Im(z) \le A\}$$

and consider averaging over annuli $A_{r,R}(w)$ only for $w \in \mathcal{F}_A$ instead of $w \in \mathcal{F}$, so that the relevant expression for the variance is

$$\mathrm{Var}_A(r, R; L) := \int_{\pi^{-1}(\mathcal{F})} \int_{\mathcal{F}_A} \left(\int_0^L \sum_{\gamma \in \Gamma} k_{r,R}(u(\pi \circ \mathcal{G}_t(g), \gamma w))\, dt - L \frac{\mu(A_{r,R})}{\mu(\mathcal{F})} \right)^2 \frac{d\mu(w)}{\mu(\mathcal{F}_A)} \frac{d\nu(g)}{\mu(\mathcal{F})}.$$

The asymptotic behavior of the expression above will involve a special function, so now we define it and express its key properties.

Lemma 13 (Basic asymptotic properties of $\mathbf{G}$)**.** *For $0 \le w < 1$, let*

$$\mathbf{G}(w) := 1 + w^3 + (1 - w^2)\mathbf{K}(w) - (1 + w^2)\mathbf{E}(w),$$

where $\mathbf{K}$ and $\mathbf{E}$ are the complete elliptic integrals of the first and second kinds, respectively. Then

$$\mathbf{G}(0) = 1, \tag{3.6}$$

$$\mathbf{G}(w) = \frac{3}{4}(1 - w)^2 \log\left(\frac{2}{1 - w}\right) + O\left((1 - w)^2\right), \tag{3.7}$$

$$\mathbf{G}(w) \gg (1 - w)^2 \log\left(\frac{2}{1 - w}\right), \tag{3.8}$$

and

$$\mathbf{G}'(w) \ll (1 - w) \log\left(\frac{2}{1 - w}\right). \tag{3.9}$$

Proof. The definitions of $\mathbf{K}$ and $\mathbf{E}$ [42, page 8.112] give (3.6), while (3.7) and (3.9) follow from [42, 8.113.3 and 8.114.3]. Indeed, for $u := \sqrt{1 - w^2}$, those give

$$\mathbf{K}(w) = \log\left(\frac{4}{u}\right) + \frac{u^2}{4} \log\left(\frac{4}{u}\right) + O(u^2)$$

and

$$\mathbf{E}(w) = 1 + \frac{u^2}{2} \log\left(\frac{4}{u}\right) - \frac{u^2}{4} + \frac{3u^4}{16} \log\left(\frac{4}{u}\right) + O(u^4),$$

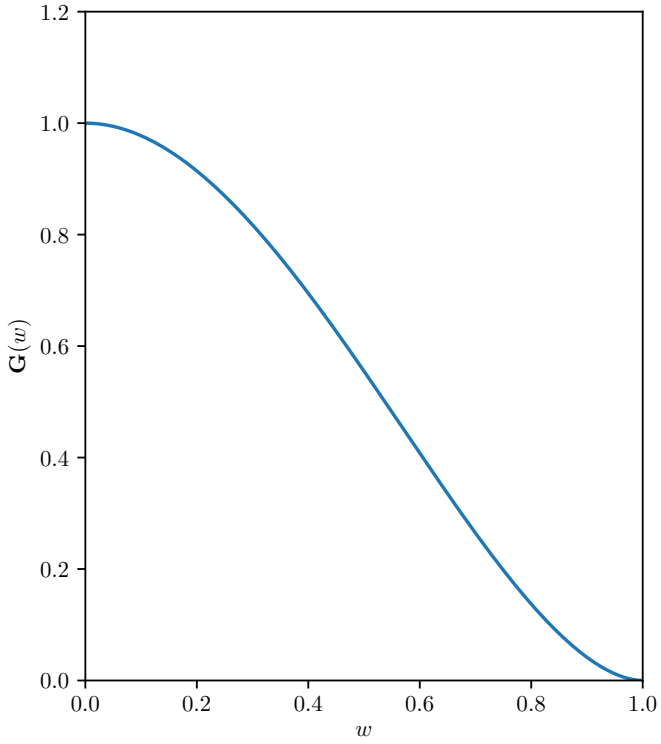


Figure 3.2: Graph of $\mathbf{G}(w)$

so

$$\begin{aligned}\mathbf{G}(w) &= 1 + w^3 + u^2\mathbf{K}(w) - (2-u^2)\mathbf{E}(w) \\ &= -1 + w^3 + \frac{3u^2}{2} + \frac{3u^4}{8}\log\left(\frac{4}{u}\right) + O(u^4).\end{aligned}$$

Changing variables to $v := 1-w$, so $-1+w^3 = v^3+3v^2-3v$ and $u^2 = -v^2+2v$, we get

$$\begin{aligned}\mathbf{G}(w) &= v^3 + \frac{3}{2}v^2 + \frac{3}{8}\left(-v^2+2v\right)^2\log\left(\frac{4}{\sqrt{v(1+w)}}\right) + O(v^2) \\ &= \frac{3}{4}v^2\log\left(\frac{1}{v}\right) + O(v^2),\end{aligned}$$

which gives (3.7). Similarly, the identity $\mathbf{G}'(w) = 3w(w-\mathbf{E}(w))$, which follows from [42, page 8.123], gives

$$\mathbf{G}'(w) \ll w - \mathbf{E}(w) \ll v\log\left(\frac{1}{v}\right) + O(v)$$

and we obtain (3.9). This identity also shows that $\mathbf{G}'(w) \leq 0$, since $\mathbf{E}(w) \geq 1 > w$, so (3.8) follows from (3.7) after choosing an appropriate cutoff.

□

Remark 5. *Perhaps the simplest way to understand the function* $\mathbf{G}(w)$ *geometrically is to describe it as follows: let* $t \in [-1,1]$ *be uniformly distributed, and consider the (Euclidean) annulus* $A'_{w,1}(it) := \{z \in \mathbb{C} : w \leq |z - it| \leq 1\}$, *for* $w \in [0,1]$. *Let* Y_w *be the (Euclidean) length of the intersection of* $A'_{w,1}(it)$ *with the real axis. Then an explicit computation shows that*

$$\mathbf{G}(w) = \frac{\mathbb{E}(Y_w^2)}{\mathbb{E}(Y_0^2)} = \frac{3}{8}\mathbb{E}(Y_w^2) = \frac{3}{8} \cdot \frac{1}{2} \int_{-1}^{1} \left(2\sqrt{1-t^2} - \mathbb{1}_{|t|\leq w} \cdot 2\sqrt{w^2-t^2}\right)^2 dt.$$

We are ready to state the main result of this section, omitting the dependence on a parameter t that governs all the asymptotic statements below (meaning that the quantities L, r, R, A are all functions of t, and asymptotic notations such as $o(\dots)$ or $\sim$ should be interpreted in the limit as $t \to \infty$).

Theorem 3. *Suppose that* $L \gg 1$, $0 \leq r < R = o(1)$, *and* $1 \ll \log A = o\left(R^{-1} \log^{-1}\left(\frac{1}{R-r}\right)\right)$, *so in particular we require* $\log\left(\frac{1}{R-r}\right) = o\left(\frac{1}{R}\right)$. *Then*

$$\mathrm{Var}_A(r, R; L) \sim \frac{16LR^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right).$$

In particular, for balls we get

$$\mathrm{Var}_A(0, R; L) \sim \frac{16LR^3}{\pi},$$

and for thin annuli (i.e. such that $R - r = o(R)$*) satisfying the restrictions above we get*

$$\mathrm{Var}_A(r, R; L) \sim \frac{12LR(R-r)^2}{\pi} \log\left(\frac{R}{R-r}\right).$$

Auxiliary results

An important ingredient for Theorem 3 will be the fact that the geodesic flow is mixing, and in fact it is so with an exponential rate, due to a theorem of M. Ratner [93]. We will use the following effective version of Ratner's result, due to C. Matheus and adapted here to the modular group Γ.

Lemma 14 (Exponential mixing for the geodesic flow [81, Corollary 2.1])**.** *Let* $\phi, \psi \in L^2(X)$ *be such that* $\int_X \phi \, d\mu = \int_X \psi \, d\mu = 0$. *Then*

$$\int_{T^1(X)} \phi(\widetilde{\pi}(g)) \cdot \psi(\widetilde{\pi} \circ \widetilde{\mathcal{G}}_t(g)) \, d\widetilde{\nu}(g) \ll \|\phi\|_{L^2(X)} \cdot \|\psi\|_{L^2(X)} \cdot (|t|+1) e^{-\frac{|t|}{2}}.$$

In order to deal with the problem of self-intersections alluded to in our heuristic discussion, we will need two basic observations regarding the distribution of orbits of Γ acting on $\mathbb{H}$. The first one is the following standard density estimate.

Lemma 15 (Density of hyperbolic lattice points [58, Lemma 2.11])**.** *If $z \in \mathbb{H}$, $w \in \mathcal{F}$, and $\delta > 0$ then*

$$|\{\gamma \in \Gamma : u(z, \gamma w) \leq \delta\}| \ll \sqrt{\delta(\delta+1)}\Im(w) + 1.$$

The second observation about the orbits of Γ in $\mathbb{H}$ deals with the minimum spacing between distinct points in such an orbit. It formalizes the idea that the spacing can only be small if either the orbit comes close to a point of $\mathbb{H}$ with nontrivial stabilizer in Γ, or if it has a point very high up towards the cusp.

Lemma 16 (Minimum spacing of hyperbolic lattice points)**.** *If $w \in \mathcal{F}$ then*

$$\min_{1 \neq \gamma \in \Gamma} \rho(w, \gamma w) \gg \min\left\{\rho(w,i), \rho(w,j), \rho(w,j'), \frac{1}{\Im(w)}\right\}$$

where $j := \frac{1+i\sqrt{3}}{2}$ and $j' := \frac{-1+i\sqrt{3}}{2}$.

Proof. Since the minimum is $\leq \rho(w, w+1) \ll 1$, it suffices to show that

$$\min_{1 \neq \gamma \in \Gamma} u(w, \gamma w) \gg \min\left\{u(w,i), u(w,j), u(w,j'), \frac{1}{\Im(w)^2}\right\}.$$

Write $w = x + iy$, $1 \neq \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ for an element that attains the minimum, and $U^2 := u(w, \gamma w)$ with $1 \gg U > 0$.

Case 1: $y \geq 2$.

If $c \neq 0$ we have $\Im(\gamma w) = \frac{y}{|cw+d|^2} \leq \frac{y}{(cy)^2} \leq \frac{1}{2}$, which gives $U \gg 1$. If $c = 0$ we have $\gamma w = w \pm b$ and $b \neq 0$, which gives $U \gg \frac{1}{\Im(w)}$. In any case, the result holds for $y \geq 2$.

Case 2: $y < 2$.

Observe that since $w \in \mathcal{F}$ we have $y \geq \frac{\sqrt{3}}{2}$. If $|c| \geq 2$ then as before $\Im(\gamma w) = \frac{y}{|cw+d|^2} \leq \frac{1}{c^2 y} \leq \frac{1}{2\sqrt{3}}$, so $U \gg 1$. If $c = 0$ we also get $U \gg 1$ as

before, so $|c| = 1$ and we can assume that $c = 1$ since the entries of γ are only determined up to flipping all the signs.

Now, $\gamma = \begin{pmatrix} a & ad-1 \\ 1 & d \end{pmatrix}$ and

$$U^2 = \frac{|w-\gamma w|^2}{4y\Im(\gamma w)} = \frac{|w-\gamma w|^2|w+d|^2}{4y^2}$$
$$= \frac{|w(w+d)-(aw+ad-1)|^2}{4y^2} \gg |z|^2$$

for $z := w(w+d)-(aw+ad-1)$, so that

$$\Re(z) = x^2 - y^2 + (d-a)x - ad + 1 \qquad \text{and} \qquad \Im(z) = 2xy + (d-a)y.$$

If $|d-a| \geq 2$ then since $|x| \leq \frac{1}{2}$, as $w \in \mathcal{F}$, we get $|\Im(z)| \geq y \gg 1$ and therefore $U \gg 1$.

If $d = a \neq 0$ then we have $\Re(z) = x^2 - y^2 - a^2 + 1 \leq \frac{1}{4} - \frac{3}{4} = -\frac{1}{2}$, so $U \gg 1$.

If $d = a = 0$ then $U^2 \gg \Im(z)^2 \gg x^2$ and $U^2 \gg \Re(z)^2 = (x^2 - y^2 + 1)^2$, which gives $|x| \ll U$ and $y^2 = 1 + x^2 + O(U) = 1 + O(U)$ since $U \ll 1$, so that $|y-1| \ll U$. We conclude that $u(w,i) \asymp x^2 + (y-1)^2 \ll U^2$, as desired.

If $d - a = 1$ then we can assume that $d = 0$ or 1, otherwise $ad \geq 2$ and $\Re(z) = x^2 - y^2 + x - ad + 1 \leq \frac{1}{4} - \frac{3}{4} + \frac{1}{2} - 2 + 1 = -1$, so $U \gg 1$. For $d = 0$ we get $U \gg |\Im(z)| \gg \left|x + \frac{1}{2}\right|$, so $x = -\frac{1}{2} + O(U)$ and then looking at the real part we get $y^2 = 1 + x + x^2 + O(U) = \frac{3}{4} + O(U)$, once again since $U \ll 1$. This gives $y = \frac{\sqrt{3}}{2} + O(U)$ and therefore $u(w,j) \asymp \left(x + \frac{1}{2}\right)^2 + \left(y - \frac{\sqrt{3}}{2}\right)^2 \ll U^2$, as desired. For $d = 1$ the exact same reasoning shows that $u(w,j) \ll U^2$.

Finally, if $d - a = -1$ then an argument analogous to the previous paragraph, but exchanging x with $-x$, gives $u(w,j') \ll U^2$, so we have covered all possibilities and the result follows.

□

The last ingredients necessary to prove Theorem 3 are bounds and asymptotics for integral expressions that measure the lengths of intersections between geodesics and annuli in $\mathbb{H}$, averaged over various parameters. We deal with those geometric quantities in the next two lemmas, and emphasize that the

results are analogous (except for large distances) to those for the Euclidean version of the problem, in which straight lines intersect Euclidean annuli.

Lemma 17 (Average intersection of geodesics through z with $A_{r,R}(w)$)**.** *For any $z, w \in \mathbb{H}$, and $0 \le r < R \ll 1$, if we denote $D := \rho(z,w)$ then*

$$\Theta_{r,R}(z,w) := \int_0^{2\pi} \int_0^{\infty} k_{r,R}(u(\pi \circ \mathcal{G}_t(z,\theta), w))\, dt\, d\theta$$
$$\ll \begin{cases} (R-r)\log\left(\frac{2R}{R-r}\right) & \text{if } D < 2R, \\ \frac{R(R-r)}{D} & \text{if } 2R \le D < 1, \\ \frac{R(R-r)}{e^D} & \text{if } 1 \le D. \end{cases}$$

Proof. Since the geodesic flow is parametrized by arc length, the system of coordinates (t,θ) corresponds to geodesic polar coordinates centered at z, therefore the hyperbolic measure becomes $d\mu = 2\sinh t\, dt\, d\theta$.

If $D \ge 2R$, then observing that the integrand is simply the indicator function of the annulus $A_{r,R}(w)$ and that $\sinh t \ge \sinh(D-R) \gg \sinh D$ for all points (t,θ) inside it (since by the triangle inequality $t = \rho(z,(t,\theta)) \ge \rho(z,w) - \rho((t,\theta),w) \ge D - R$), we get

$$\Theta_{r,R}(z,w) \ll \int_0^{2\pi} \int_0^{\infty} k_{r,R}(u(\pi \circ \mathcal{G}_t(z,\theta), w)) \frac{\sinh t}{\sinh D}\, dt\, d\theta \ll \frac{\mu(A_{r,R})}{\sinh D}$$

and the result follows. If $R - r \gg R$ then the result for $D < 2R$ is trivial since the integral over t is always $\le 2R$ by the triangle inequality. Therefore we can assume that $r \ge \frac{R}{2}$, and then a slight modification of the argument above also takes care of $D \le \frac{R}{4}$, since $\sinh t \gg \sinh R$ in that case.

We are left with the trickiest case $r \ge \frac{R}{2}$ and $\frac{R}{4} \le D \le 2R$. The issue here is that the intersection of each geodesic with the annulus $A_{r,R}(w)$ no longer has length $\ll R - r$ when the latter is thin — in fact the length can be $\gg \sqrt{R(R-r)}$. In what follows it is worth keeping in mind that since $R \ll 1$ the geometry is roughly Euclidean.

First let us change variables, shifting θ so that it corresponds to the angle with the geodesic from z to w (instead of with the vertical line). Since $D \ll R$, we can choose a sufficiently small (absolute) $\varepsilon > 0$ such that if $|\sin\theta| < \varepsilon$ then each θ contributes $\ll R - r$. Indeed, the integrand for each θ is now the length of the intersection of the one-sided geodesic determined by (z,θ) with

the annulus $A_{r,R}(w)$. If d is the (orthogonal) distance between that geodesic and w, then the hyperbolic law of sines gives $\sinh d = \sinh D \sin\theta$, which implies $d \ll R\sin\theta$. The length of the intersection of the corresponding two-sided geodesic with $A_{r,R}(w)$ is $2(\ell_R - \ell_r)$, where we define $2\ell_R$ as the length of the intersection of that two-sided geodesic with $B_R(w)$, and similarly for $2\ell_r$ (both intersections will be non-empty for sufficiently small $\varepsilon > 0$, as we assume $D \asymp R \asymp r$). By the hyperbolic law of cosines we have (see Figure 3.3)

$$\cosh \ell_R = \frac{\cosh R}{\cosh d} \qquad \text{and} \qquad \cosh \ell_r = \frac{\cosh r}{\cosh d}.$$

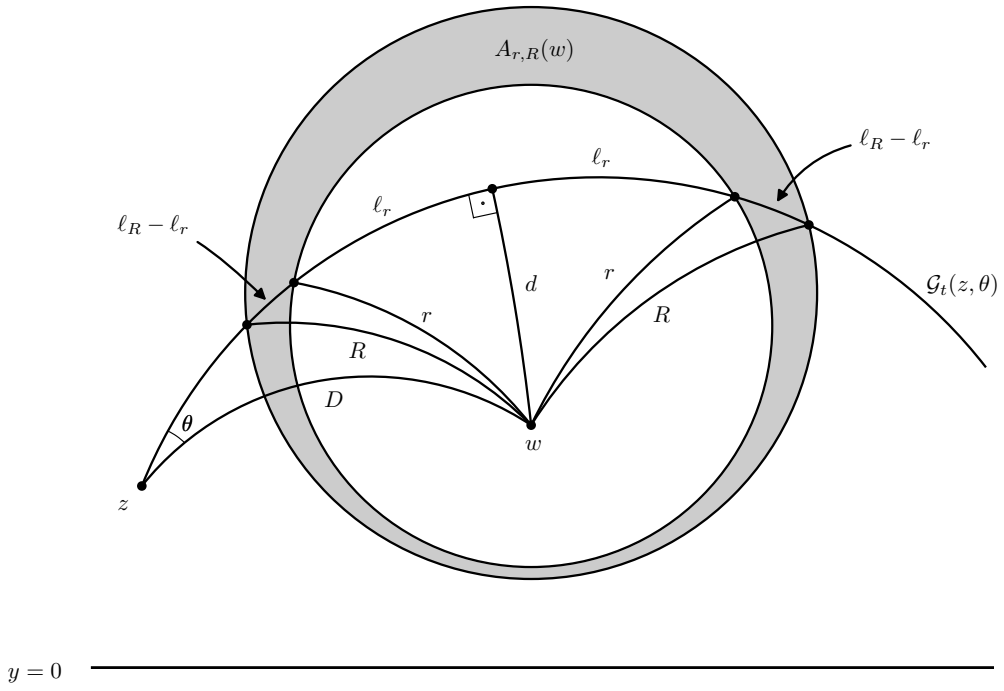


Figure 3.3: Lengths in intersection of $\mathcal{G}_t(z,\theta)$ with $A_{r,R}(w)$

Choosing $\varepsilon > 0$ sufficiently small so that $d \le \frac{R}{4} \le \frac{r}{2}$ we get $\ell_r \gg R$ and therefore by the MVT

$$\ell_R - \ell_r \ll \frac{\cosh \ell_R - \cosh \ell_r}{\sinh \ell_r} \ll \frac{\cosh R - \cosh r}{R \cosh d} \ll R - r.$$

Now, for the remaining angles θ satisfying $|\sin\theta| \ge \varepsilon$ we will fix the radius t and evaluate the angular contribution. The values of t for which $\partial B_t(z)$ intersects only one of $\partial B_R(w)$ or $\partial B_r(w)$ contribute $\ll R - r$ (bounding the

integral over θ trivially), so they may be excluded and we can assume that both circles are intersected. Let $0 \leq \theta_r < \theta_R \leq \pi$ be the angles corresponding to the intersection points in the upper half, so the integral over θ with no restrictions contributes $2(\theta_R - \theta_r)$ and by the hyperbolic law of cosines we have (see Figure 3.4)

$$\cos\theta_R = \frac{\cosh t \cosh D - \cosh R}{\sinh t \sinh D} \qquad \text{and} \qquad \cos\theta_r = \frac{\cosh t \cosh D - \cosh r}{\sinh t \sinh D}.$$

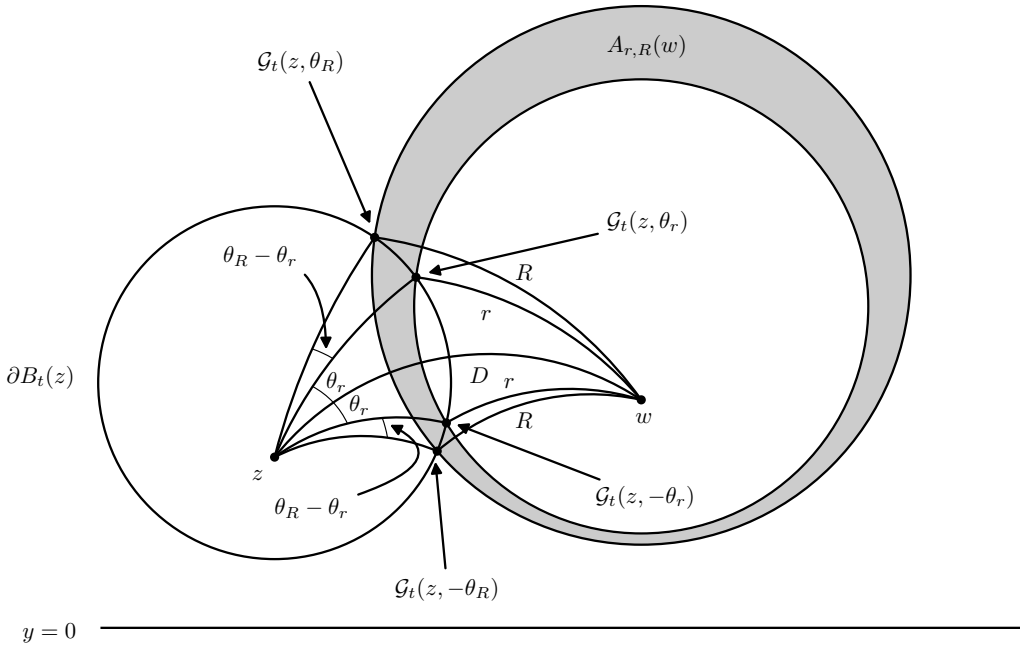


Figure 3.4: Angles in intersection of $\partial B_t(z)$ with $A_{r,R}(w)$

If $[\theta'_r, \theta'_R] := [\theta_r, \theta_R] \cap [\arcsin\varepsilon, \pi - \arcsin\varepsilon]$, then since we have already excluded angles θ with $|\sin\theta| < \varepsilon$ by bounding the corresponding terms as in the previous part of the argument, the contribution of the remaining terms corresponding to t is actually just $2(\theta'_R - \theta'_r)$. But since the sine of both angles is bounded away from zero we can use the MVT to get

$$\theta'_R - \theta'_r \ll \frac{\cos\theta'_r - \cos\theta'_R}{\sin\varepsilon} \ll \cos\theta_r - \cos\theta_R = \frac{\cosh R - \cosh r}{\sinh t \sinh D} \ll \frac{R - r}{t},$$

as $D \gg R$.

Finally, we integrate over $t \leq D+R \ll R$, where we can assume that $t \geq R-r$ since those radii trivially contribute $\ll R-r$. In conclusion,

$$\Theta_{r,R}(z,w) \ll R-r+\int_{R-r}^{D+R} \frac{R-r}{t}\,dt \ll (R-r)\log\left(\frac{2R}{R-r}\right),$$

as desired.

□

Lemma 18 (Main term computation)**.** *For any $g \in T^1(\mathbb{H})$ and $0 \leq r < R = o(1)$,*

$$\int_{\mathbb{H}} k_{r,R}(u(\pi(g),w))\left(\int_{-\infty}^{\infty} k_{r,R}(u(\pi\circ\mathcal{G}_t(g),w))\,dt\right)\,d\mu(w) \sim \frac{16R^3}{3}\mathbf{G}\left(\frac{r}{R}\right).$$

Proof. Observe that the LHS is independent of g, since the integral over w is invariant under isometries, so denoting the whole expression by $I_{r,R}$ we can average over the geodesic segment of length $S>0$ (which we will choose to be sufficiently large later) starting at $(i,0) \in T^1(\mathbb{H})$ to get

$$I_{r,R} = \frac{1}{S}\int_{\mathbb{H}}\left(\int_0^S k_{r,R}(u(ie^s,w))\,ds\right)\left(\int_{-\infty}^{\infty} k_{r,R}(u(ie^t,w))\,dt\right)\,d\mu(w).$$

Given $D>0$, let

$$A(D,S) := \left\{re^{i\delta} \in \mathbb{H} : 1 \leq r \leq e^S \quad \text{and} \quad \frac{\pi}{2}-\theta_D \leq \delta \leq \frac{\pi}{2}+\theta_D\right\},$$

for $0 \leq \theta_D \leq \frac{\pi}{2}$ defined by $\sin\theta_D = \tanh D$ (see Figure 3.5). It will be important to compute $\mu(A(D,S))$ for the computation of $I_{r,R}$ that follows below, so we do that now and come back to the integral afterwards.

The locus of points $z \in \mathbb{H}$ with (orthogonal) distance to the line $y=0$ equal to D is the pair of straight half-lines through the origin with angles θ_D and $-\theta_D$ with the vertical. The (hyperbolic) arc length parametrization of the half-line corresponding to θ_D is $z(s) = e^{s\cos\theta_D}e^{i(\frac{\pi}{2}-\theta_D)}$. A computation shows that its geodesic curvature is constant equal to $\sin\theta_D = \tanh D$ (see for instance the discussion after [26, Corollary 4], where our situation corresponds to $a=1$ and $\alpha = \frac{\pi}{2}-\theta_D$). The region $A(D,S)$ has as boundaries two geodesics (Euclidean circles with center at the origin) and the two straight lines through the origin with angles θ_D and $-\theta_D$ with the vertical line $y=0$. Examining the arc

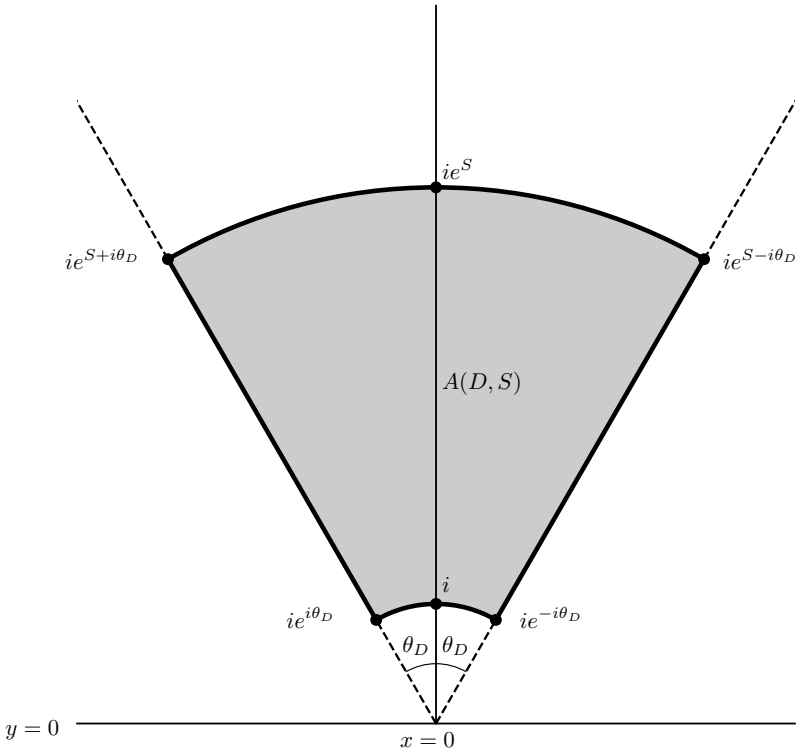


Figure 3.5: The region $A(D,S)$

length parametrization we see that each of those has length $\frac{S}{\cos\theta_D}$, so the total geodesic curvature along the boundary of $A(D,S)$ in the positive direction is

$$2S\tan\theta_D = 2S\frac{\tanh D}{\sqrt{1-\tanh^2 D}} = 2S\sinh D.$$

Since the four external angles of $A(D,S)$ are equal to $\frac{\pi}{2}$, denoting by $K=-1$ the Gaussian curvature of $A(D,S)$ and by k_g the geodesic curvature of $\partial A(D,S)$, the Gauss-Bonnet theorem gives

$$\mu(A(D,S)) = -\int_{A(D,S)} K\,d\mu = -2\pi + 4\frac{\pi}{2} + \int_{\partial A(D,S)} k_g\,ds = 2S\sinh D.$$

Observe that $\int_0^S k_{r,R}(u(ie^s,w))\,ds$ is nonzero only for $w\in A(R,S)\cup B_R(i)\cup B_R(ie^S)$, since this is the locus of points within distance $\leq R$ from the geodesic segment between the points i and ie^S. Furthermore, if $w\notin B_R(i)\cup B_R(ie^S)$ then $\int_0^S k_{r,R}(u(ie^s,w))\,ds$ is equal to $\int_{-\infty}^{\infty} k_{r,R}(u(ie^s,w))\,ds$, which is the length of the intersection of the vertical line $y=0$ with the annulus $A_{r,R}(w)$ (and therefore trivially $\ll R$). As discussed in the proof of Lemma 17, the hyperbolic law of cosines shows that for points at distance D from the vertical this

length is $2(\ell_R - \ell_r)$ if $0 \le D \le r$ and $2\ell_R$ for $r \le D \le R$, where

$$\cosh \ell_R = \frac{\cosh R}{\cosh D} \qquad \text{and} \qquad \cosh \ell_r = \frac{\cosh r}{\cosh D}.$$

We conclude that

$$I_{r,R} = \frac{1}{S} \int_{A(R,S)} \left(\int_{-\infty}^{\infty} k_{r,R}(u(ie^t, w))\, dt \right)^2 d\mu(w) + O\left(\frac{R^4}{S}\right).$$

Choosing say $S \gg (R-r)^{-2}$ we see that the error term is $o(R(R-r)^2)$ and will be negligible. The remaining term can be written as

$$\begin{aligned}
&\frac{4}{S} \int_0^R \left(\operatorname{arccosh}\left(\frac{\cosh R}{\cosh D}\right) - \mathbb{1}_{D \le r} \cdot \operatorname{arccosh}\left(\frac{\cosh r}{\cosh D}\right) \right)^2 d\left(\mu(A(D,S))\right) \\
&= 8 \int_0^R \left(\operatorname{arccosh}\left(\frac{\cosh R}{\cosh D}\right) - \mathbb{1}_{D \le r} \cdot \operatorname{arccosh}\left(\frac{\cosh r}{\cosh D}\right) \right)^2 \cosh D\, dD \\
&\sim 8 \int_0^r \log^2 \left(\frac{\cosh R + \sqrt{\cosh^2 R - \cosh^2 D}}{\cosh r + \sqrt{\cosh^2 r - \cosh^2 D}} \right) dD \\
&\qquad\qquad + 8 \int_r^R \log^2 \left(\frac{\cosh R + \sqrt{\cosh^2 R - \cosh^2 D}}{\cosh D} \right) dD
\end{aligned} \tag{3.10}$$

using $\operatorname{arccosh} x = \log\left(x + \sqrt{x^2 - 1}\right)$ and $\cosh D = 1 + O(R^2)$.

Denoting

$$\begin{aligned}
f_{r,R}(D) &:= \sqrt{\cosh^2 R - \cosh^2 D} - \sqrt{\cosh^2 r - \cosh^2 D} \\
&= \sqrt{\sinh^2 R - \sinh^2 D} - \sqrt{\sinh^2 r - \sinh^2 D}
\end{aligned}$$

for $0 \le D \le r$, we see that it is increasing and therefore the MVT gives

$$R - r \ll f_{r,R}(D) \ll \sqrt{R(R-r)}. \tag{3.11}$$

Similarly, $\cosh r + \sqrt{\cosh^2 r - \cosh^2 D} = 1 + O(R)$ and $\cosh R - \cosh r = O(R(R-r))$, so

$$\frac{\cosh R + \sqrt{\cosh^2 R - \cosh^2 D}}{\cosh r + \sqrt{\cosh^2 r - \cosh^2 D}} = 1 + f_{r,R}(D)(1 + O(R)) + O(R(R-r))$$

and therefore with the aid of (3.11) we obtain

$$\log^2 \left(\frac{\cosh R + \sqrt{\cosh^2 R - \cosh^2 D}}{\cosh r + \sqrt{\cosh^2 r - \cosh^2 D}} \right) = f_{r,R}(D)^2 (1 + O(R)) + O(R^2(R-r)^2).$$

The same sort of analysis for $g_R(D) := \sqrt{\cosh^2 R - \cosh^2 D} \ll \sqrt{R(R-r)}$ when $r \le D \le R$, separating into cases depending on whether $g_R(D)$ is larger than $R - r$ or not, gives

$$\log^2\left(\frac{\cosh R + \sqrt{\cosh^2 R - \cosh^2 D}}{\cosh D}\right) = g_R(D)^2(1 + O(R)) + O(R(R-r)^2).$$

Plugging those into (3.10) we get an error term $\ll R^2(R-r)^2 = o(R(R-r)^2)$ and an integral term

$$\begin{aligned}
&\sim 8\int_0^r f_{r,R}(D)^2\,dD + 8\int_r^R g_R(D)^2\,dD \\
&\sim 8\int_0^{\sinh r} f_{r,R}(\operatorname{arcsinh} x)^2\,dx + 8\int_{\sinh r}^{\sinh R} g_R(\operatorname{arcsinh} x)^2\,dx,
\end{aligned}$$

changing variables to $x = \sinh D$ and observing that $\cosh D = 1 + O(R^2)$. The result can be written, by [42, page 3.155.8], as

$$\begin{aligned}
&\frac{16}{3}\left(\sinh^3 R + \sinh^3 r\right) - 16\int_0^{\sinh r} \sqrt{(\sinh^2 R - x^2)(\sinh^2 r - x^2)}\,dx \\
&= \frac{16}{3}\left(\sinh^3 R + \sinh^3 r\right) - \frac{16}{3}\sinh R\left(\left(\sinh^2 R + \sinh^2 r\right)\mathbf{E}\left(\frac{\sinh r}{\sinh R}\right)\right. \\
&\qquad\qquad \left. - \left(\sinh^2 R - \sinh^2 r\right)\mathbf{K}\left(\frac{\sinh r}{\sinh R}\right)\right) \\
&= \frac{16\sinh^3 R}{3}\mathbf{G}\left(\frac{\sinh r}{\sinh R}\right) \sim \frac{16R^3}{3}\mathbf{G}\left(\frac{\sinh r}{\sinh R}\right),
\end{aligned}$$

where we use the notation of Lemma 13. A computation with Taylor series gives

$$\frac{\sinh r}{\sinh R} = \frac{r}{R} + O\left(R(R-r)\right),$$

so (3.9) and the MVT give

$$\mathbf{G}\left(\frac{\sinh r}{\sinh R}\right) = \mathbf{G}\left(\frac{r}{R}\right) + O\left((R-r)^2 \log\left(\frac{2R}{R-r}\right)\right).$$

We conclude that

$$I_{r,R} \sim \frac{16R^3}{3}\mathbf{G}\left(\frac{r}{R}\right) + o(R(R-r)^2) \sim \frac{16R^3}{3}\mathbf{G}\left(\frac{r}{R}\right)$$

by (3.8), as desired.

□

Putting it all together: the proof of Theorem 3

Proof of Theorem 3. By absolute convergence, we can freely exchange the order of integration and write

$$\mathrm{Var}_A(r,R;L) = \int_{\mathcal{F}_A}\int_{T^1(X)}\left(\int_0^L K_{r,R}(\widetilde{\pi}\circ\widetilde{\mathcal{G}}_t(g),w)\,dt - L\frac{\mu(A_{r,R})}{\mu(\mathcal{F})}\right)^2 \frac{d\widetilde{\nu}(g)}{\mu(\mathcal{F})}\frac{d\mu(w)}{\mu(\mathcal{F}_A)}.$$

The inner integral over $g\in T^1(X)$ is, after changing variables, equal to

$$\int_0^L\int_0^L\int_{T^1(X)}\phi(\widetilde{\pi}(g))\cdot\phi(\widetilde{\pi}\circ\widetilde{\mathcal{G}}_{t-t'}(g))\frac{d\widetilde{\nu}(g)}{\mu(\mathcal{F})}\,dt'\,dt \tag{3.12}$$

for $\phi\in L^2(X)$ given by $\phi(z):=K_{r,R}(z,w)-\frac{\mu(A_{r,R})}{\mu(\mathcal{F})}$, so that $\int_X\phi\,d\mu=0$ and

$$\begin{aligned}\|\phi\|^2_{L^2(X)} &= \int_{\mathcal{F}}K^2_{r,R}(z,w)\,d\mu(z) - \left(\frac{\mu(A_{r,R})}{\mu(\mathcal{F})}\right)^2 < \int_{\mathcal{F}}K^2_{r,R}(z,w)\,d\mu(z)\\ &\le \max_{\xi\in\mathcal{F}}\left|\left\{\gamma\in\Gamma : u(\xi,\gamma w)\le\sinh^2\left(\frac{R}{2}\right)\right\}\right|\cdot\int_{\mathcal{F}}K_{r,R}(z,w)\,d\mu(z) \qquad (3.13)\\ &\ll (R\Im(w)+1)\frac{\mu(A_{r,R})}{\mu(\mathcal{F})} \ll (R\Im(w)+1)R(R-r)\end{aligned}$$

by Lemma 15. Let $L\ge T\gg 1$. If $|t-t'|>T$ we can use Lemma 14 in the inner integral of (3.12), inputting the bound (3.13), to conclude that the contribution of all such terms to $\mathrm{Var}_A(r,R;L)$ is

$$\begin{aligned}&\ll LR(R-r)\int_{\mathcal{F}_A}(R\Im(w)+1)\,d\mu(w)\int_T^\infty (x+1)e^{-\frac{x}{2}}\,dx\\ &\ll LR(R-r)Te^{-\frac{T}{2}}(R\log A+1)\ll LR(R-r)e^{-\frac{T}{4}}\end{aligned} \tag{3.14}$$

if $T<L$, and it is $=0$ if we choose $T=L$ (since no such terms exist in that case).

The remaining set of $|t-t'|\le T$ has measure $\ll LT$, so replacing $\phi(z)$ with $K_{r,R}(z,w)$ in (3.12) we pick up an error term

$$\ll LTR^2(R-r)^2, \tag{3.15}$$

and what remains is

$$\begin{aligned}&\int_0^L\int_{\max\{t'-T,0\}}^{\min\{t'+T,L\}}\int_{T^1(X)}K_{r,R}(\widetilde{\pi}(g),w)\cdot K_{r,R}(\widetilde{\pi}\circ\widetilde{\mathcal{G}}_{t-t'}(g),w)\frac{d\widetilde{\nu}(g)}{\mu(\mathcal{F})}\,dt\,dt'\\ &\quad=\int_{T^1(X)}K_{r,R}(\widetilde{\pi}(g),w)\left(\int_{-T}^T(L-|t|)\cdot K_{r,R}(\widetilde{\pi}\circ\widetilde{\mathcal{G}}_t(g),w)\,dt\right)\frac{d\widetilde{\nu}(g)}{\mu(\mathcal{F})}.\end{aligned}$$

Inserting the integral above into the expression for the variance and expanding the automorphic kernel, we are left with

$$\int_{\mathcal{F}_A}\int_{\pi^{-1}(\mathcal{F})}\sum_{\gamma'\in\Gamma}k_{r,R}(u(\pi(g),\gamma' w))$$
$$\times\left(\int_{-T}^{T}(L-|t|)\sum_{\gamma\in\Gamma}k_{r,R}(u(\pi\circ\mathcal{G}_t(g),\gamma w))\,dt\right)\frac{d\nu(g)}{\mu(\mathcal{F})}\frac{d\mu(w)}{\mu(\mathcal{F}_A)}=$$
$$\int_{\mathcal{F}_A}\sum_{\gamma'\in\Gamma}\int_{\pi^{-1}(\mathcal{F}\cap A_{r,R}(\gamma' w))}\sum_{\gamma\in\Gamma}\int_{-T}^{T}(L-|t|)\cdot k_{r,R}(u(\pi\circ\mathcal{G}_t(g),\gamma w))\,dt\frac{d\nu(g)}{\mu(\mathcal{F})}\frac{d\mu(w)}{\mu(\mathcal{F}_A)}. \tag{3.16}$$

Now, let $\mathcal{M}$ denote the terms corresponding to $\gamma=\gamma'$, which is where the main contribution will come from, and let $\mathcal{E}$ denote all other terms.

Decomposing $g=(z,\theta)$ for $z\in\mathcal{F}\cap A_{r,R}(\gamma' w)$ and $0\le\theta<2\pi$, what is left in (3.16) corresponding to the terms in $\mathcal{E}$ is

$$\ll L\int_{\mathcal{F}_A}\sum_{\gamma'\in\Gamma}\int_{\mathcal{F}\cap A_{r,R}(\gamma' w)}\int_0^{2\pi}\sum_{\gamma'\neq\gamma\in\Gamma}\int_{-T}^{T}k_{r,R}(u(\pi\circ\mathcal{G}_t(z,\theta),\gamma w))\,dt\,d\theta\,d\mu(z)\,d\mu(w). \tag{3.17}$$

For given $w\in\mathcal{F}_A$, $\gamma'\in\Gamma$, and $z\in\mathcal{F}\cap A_{r,R}(\gamma' w)$, we can use the notation of Lemma 17 to bound

$$\int_0^{2\pi}\sum_{\gamma'\neq\gamma\in\Gamma}\int_{-T}^{T}k_{r,R}(u(\pi\circ\mathcal{G}_t(z,\theta),\gamma w))\,dt\,d\theta\ll\sum_{\gamma'\neq\gamma\in\Gamma}\mathbb{1}_{\rho(z,\gamma w)<T+R}\cdot\Theta_{r,R}(z,\gamma w). \tag{3.18}$$

Denote

$$h(\gamma',w,z;D):=\left|\{\gamma'\neq\gamma\in\Gamma:\rho(z,\gamma w)\le D\}\right|$$

and

$$f(w;D):=\left|\{1\neq\gamma\in\Gamma:\rho(w,\gamma w)\le D\}\right|.$$

The fact that $z\in A_{r,R}(\gamma' w)$ implies $h(\gamma',w,z;D)\le f(w;D+R)$. The next step is to apply Lemma 17 in the equation below, where in order to simplify the notation we keep a term $(R-r)^{-1}\log^{-1}\left(\frac{2R}{R-r}\right)$ in the left and adjust the bounds in each range so that the boundary terms cancel out after integration

by parts. This gives

$$\begin{aligned}
&(R-r)^{-1}\log^{-1}\left(\frac{2R}{R-r}\right)\sum_{\gamma'\neq\gamma\in\Gamma}\mathbb{1}_{\rho(z,\gamma w)<T+R}\cdot\Theta_{r,R}(z,\gamma w)\\
&\ll\int_{0^-}^{R} dh(\gamma',w,z;D)+\int_{R}^{1}\frac{R}{D}\,dh(\gamma',w,z;D)+\int_{1}^{T+R}\frac{R}{e^{D-1}}\,dh(\gamma',w,z;D)\\
&=\int_{R}^{1}\frac{R}{D^2}h(\gamma',w,z;D)\,dD+\int_{1}^{T+R}\frac{R}{e^{D-1}}h(\gamma',w,z;D)\,dD\\
&\qquad\qquad+\frac{R}{e^{T+R-1}}h(\gamma',w,z;T+R)\\
&\ll\int_{R}^{1}\frac{R}{D^2}f(w;D+R)\,dD+\int_{1}^{T+R}\frac{R}{e^{D}}f(w;D+R)\,dD+\frac{R}{e^{T}}f(w;T+2R).
\end{aligned}\tag{3.19}$$

We now denote

$$m(w):=\min_{1\neq\gamma\in\Gamma}\rho(w,\gamma w)=\min\{D\in\mathbb{R}:f(w;D)>0\}$$

and consider two different cases for $w\in\mathcal{F}_A$:

Case 1: $w\in\mathcal{F}_A\cap B_{1/100}(q)$ for some $q\in\{i,j,j'\}$.

In this case Lemma 15 and Lemma 16 give respectively

$$f(w;D)\ll\begin{cases}1 & \text{for } D\leq 1,\\ e^{D} & \text{for } D>1,\end{cases}\quad\text{and}\quad m(w)\gg\rho(w,q),$$

so that (3.19) is

$$\ll\int_{\max\{m(w),R\}}^{1}\frac{R}{D^2}\,dD+\int_{1}^{T+R}R\,dD+R\ll\frac{R}{\max\{\rho(w,q),R\}}+TR.$$

Plugging this into (3.18) and then (3.17), we see that the total contribution

to $\mathcal{E}$ in this case is

$$\begin{aligned}
&\ll L(R-r)\log\left(\frac{2R}{R-r}\right)\int_{\mathcal{F}_A\cap B_{1/100}(q)}\sum_{\gamma'\in\Gamma}\\
&\qquad\times\int_{\mathcal{F}\cap A_{r,R}(\gamma' w)}\left(\frac{R}{\max\{\rho(w,q),R\}}+TR\right)d\mu(z)\,d\mu(w)\\
&\ll L(R-r)\log\left(\frac{2R}{R-r}\right)\int\limits_{B_{1/100}(q)}\left(\frac{R}{\max\{\rho(w,q),R\}}+TR\right)\\
&\qquad\times\int_{\mathcal{F}}K_{r,R}(z,w)\,d\mu(z)\,d\mu(w)\\
&\ll LR(R-r)^2\log\left(\frac{2R}{R-r}\right)(R^2+R+TR)\\
&\ll LTR^2(R-r)^2\log\left(\frac{2R}{R-r}\right).
\end{aligned}\tag{3.20}$$

Case 2: $w\in\mathcal{F}_A$ but $w\notin B_{1/100}(q)$ for any $q\in\{i,j,j'\}$.

In this case Lemma 15 and Lemma 16 give respectively

$$f(w;D)\ll\begin{cases}D\Im(w)+1 & \text{for } D\le 1,\\ e^D\Im(w) & \text{for } D>1,\end{cases}\quad\text{and}\quad m(w)\gg\frac{1}{\Im(w)},$$

so that (3.19) is

$$\begin{aligned}
&\ll\int_{\max\{m(w),R\}}^1\frac{R}{D^2}(D\Im(w)+1)\,dD+\int_1^{T+R}R\Im(w)\,dD+R\Im(w)\\
&\ll R\Im(w)\log\left(\frac{1}{R}\right)+RT\Im(w).
\end{aligned}$$

Denote $\mathcal{F}_A^*:=\mathcal{F}_A\setminus\bigcup_{q\in\{i,j,j'\}}B_{1/100}(q)$. Plugging the bound above into (3.18) and then (3.17), we see that the total contribution to $\mathcal{E}$ in this case is

$$\begin{aligned}
&\ll L(R-r)\log\left(\frac{2R}{R-r}\right)\int_{\mathcal{F}_A^*}\sum_{\gamma'\in\Gamma}\\
&\qquad\times\int_{\mathcal{F}\cap A_{r,R}(\gamma' w)}\left(R\Im(w)\log\left(\frac{1}{R}\right)+RT\Im(w)\right)d\mu(z)\,d\mu(w)\\
&\ll L(R-r)\log\left(\frac{2R}{R-r}\right)\int_{\mathcal{F}_A}\left(R\Im(w)\log\left(\frac{1}{R}\right)+RT\Im(w)\right)\\
&\qquad\times\int_{\mathcal{F}}K_{r,R}(z,w)\,d\mu(z)\,d\mu(w)\\
&\ll LR^2(R-r)^2\log\left(\frac{2R}{R-r}\right)\left(\log\left(\frac{1}{R}\right)+T\right)\log A.
\end{aligned}\tag{3.21}$$

Collecting the error terms (3.14) and (3.15), and the estimates for $\mathcal{E}$ in (3.20) and (3.21), we conclude that

$$\begin{aligned}\mathrm{Var}_A(r,R;L) &= \mathcal{M}+\mathcal{E}+O\left(LTR^2(R-r)^2+\mathbb{1}_{T\neq L}\cdot LR(R-r)e^{-\frac{T}{4}}\right)\\ &= \mathcal{M}+O\left(LR^2(R-r)^2\log\left(\frac{2R}{R-r}\right)\left(\log\left(\frac{1}{R}\right)+T\right)\log A\right)\\ &\quad +O\left(\mathbb{1}_{T\neq L}\cdot LR(R-r)e^{-\frac{T}{4}}\right).\end{aligned}$$

Choosing $T=\min\left\{8\log\left(\frac{1}{R-r}\right),L\right\}$, recalling that $\log A=o\left(R^{-1}\log^{-1}\left(\frac{1}{R-r}\right)\right)$ we see that the error term is $o\left(LR(R-r)^2\log\left(\frac{2R}{R-r}\right)\right)$.

We are left with computing the main term

$$\mathcal{M}:=\int_{\mathcal{F}_A}\sum_{\gamma'\in\Gamma}\int_{\pi^{-1}(\mathcal{F}\cap A_{r,R}(\gamma' w))}\int_{-T}^{T}(L-|t|)\\ \times k_{r,R}(u(\pi\circ\mathcal{G}_t(g),\gamma' w))\,dt\frac{d\nu(g)}{\mu(\mathcal{F})}\frac{d\mu(w)}{\mu(\mathcal{F}_A)}.$$

Observe that since $\pi(g)\in A_{r,R}(\gamma' w)$ we can restrict the integral over t to $[-2R,2R]$, as the intersection of a geodesic with an annulus $A_{r,R}$ in $\mathbb{H}$ is contained in a segment of the geodesic of length $\leq 2R$. Then $L-|t|=L-O(R)\sim L$. Therefore,

$$\mathcal{M}\sim L\int_{\mathcal{F}_A}\sum_{\gamma'\in\Gamma}\int_{\pi^{-1}(\mathcal{F})}k_{r,R}(u(\pi(g),\gamma' w))\\ \times\left(\int_{-2R}^{2R}k_{r,R}(u(\pi\circ\mathcal{G}_t(g),\gamma' w))\,dt\right)\frac{d\nu(g)}{\mu(\mathcal{F})}\frac{d\mu(w)}{\mu(\mathcal{F}_A)}.$$

We have $u(\pi\circ\mathcal{G}_t(g),\gamma' w)=u(\gamma'^{-1}\pi\circ\mathcal{G}_t(g),w)=u(\pi\circ\mathcal{G}_t(\gamma'^{-1}g),w)$, and the measure ν is (left) G-invariant, so it is possible to unfold the integral over $g\in\pi^{-1}(\mathcal{F})$ to get

$$\mathcal{M}\sim\frac{L}{\mu(\mathcal{F})\mu(\mathcal{F}_A)}\int_{\mathcal{F}_A}\int_G k_{r,R}(u(\pi(g),w))\\ \times\left(\int_{-2R}^{2R}k_{r,R}(u(\pi\circ\mathcal{G}_t(g),w))\,dt\right)d\nu(g)\,d\mu(w).$$

By the same argument via G-invariance, the integral over $g\in G$ is independent of w, so we can replace the domain $\mathcal{F}_A$ with $\mathcal{F}$ in the display above, multiplying by $\frac{\mu(\mathcal{F}_A)}{\mu(\mathcal{F})}$, to get

$$\frac{L}{\mu(\mathcal{F})^2}\int_{\mathcal{F}}\int_G k_{r,R}(u(\pi(g),w))\left(\int_{-2R}^{2R}k_{r,R}(u(\pi\circ\mathcal{G}_t(g),w))\,dt\right)d\nu(g)\,d\mu(w).$$

Now we revert the unfolding process in the integral over $g \in G$, obtaining

$$\mathcal{M} \sim \frac{L}{\mu(\mathcal{F})^2} \int_{\mathcal{F}} \sum_{\gamma \in \Gamma} \int_{\pi^{-1}(\mathcal{F})} k_{r,R}(u(\pi(g), \gamma w)) \\ \times \left(\int_{-2R}^{2R} k_{r,R}(u(\pi \circ \mathcal{G}_t(g), \gamma w))\, dt \right) d\nu(g)\, d\mu(w).$$

Next we can complete the inner integral to $t \in (-\infty, \infty)$ with no loss, as was previously discussed, and after unfolding (this time the integral over $w \in \mathcal{F}$) we are left with

$$\mathcal{M} \sim \frac{L}{\mu(\mathcal{F})^2} \int_{\pi^{-1}(\mathcal{F})} \int_{\mathbb{H}} k_{r,R}(u(\pi(g), w)) \\ \times \left(\int_{-\infty}^{\infty} k_{r,R}(u(\pi \circ \mathcal{G}_t(g), w))\, dt \right) d\mu(w)\, d\nu(g).$$

Applying Lemma 18 and (3.8), we conclude that

$$\mathcal{M} \sim \frac{16LR^3}{3\mu(\mathcal{F})} \mathbf{G}\left(\frac{r}{R}\right) = \frac{16LR^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right) \gg LR(R-r)^2 \log\left(\frac{2R}{R-r}\right).$$

Therefore,

$$\mathrm{Var}_A(r, R; L) = \mathcal{M} + o\left(LR(R-r)^2 \log\left(\frac{2R}{R-r}\right)\right) \sim \frac{16LR^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right),$$

as desired. Combining this with (3.6) and (3.7) finishes the proof of the theorem.

□

Remark 6. *It may be possible to remove the technical condition* $\log\left(\frac{1}{R-r}\right) = o\left(\frac{1}{R}\right)$ *in Theorem 3, extending the result to all* $L \gg 1$, $0 \le r < R = o(1)$, *and* $1 \ll \log A = o\left(R^{-1} \log^{-1}\left(\frac{1}{R}\right)\right)$. *That is because we use a somewhat simplified mixing estimate, for functions on* X *instead of on* $T^1(X)$. *Similar estimates for the latter are available [81, Theorem 2], and it would be natural to use those to express an analogous version of* (3.12) *but with* $\phi \in L^2(T^1(X))$ *given by*

$$\phi(g) := \int_0^T K_{r,R}(\widetilde{\pi} \circ \widetilde{\mathcal{G}}_t(g), w)\, dt - T\frac{\mu(A_{r,R})}{\mu(\mathcal{F})}.$$

The L^2*-norm estimate is essentially the rest of the proof of Theorem 3, and one would be able to gain an extra factor of* $(R-r)\log\left(\frac{2R}{R-r}\right)$ *in the error term*

coming from cutting the geodesic into small pieces of length T. The issue is that [81, Theorem 2] requires estimates for Lie derivatives of order 3 in the angular direction, which would add a lot of complexity at diminishing returns.

Instead we are satisfied with the mild restriction on $R - r$, which is already enough to accommodate for instance $R - r \geq \exp(-R^{-1+\delta})$ for any fixed $\delta > 0$.

3.5 Variance for closed geodesics

Predictions from the random model

We can rewrite (3.2) as

$$\mathrm{Var}(r, R; \Lambda_D) := \int_X \left(\sum_{\mathcal{C} \in \Lambda_D} \ell(\mathcal{C} \cap A_{r,R}(w)) - \frac{\mu(A_{r,R})}{\widetilde{\mu}(X)} \sum_{\mathcal{C} \in \Lambda_D} \ell(\mathcal{C}) \right)^2 \frac{d\widetilde{\mu}(w)}{\widetilde{\mu}(X)}.$$

If $J := (\sqrt{D}) \in \mathrm{Cl}_D^+$, then the closed geodesic corresponding to any $B \in \mathrm{Cl}_D^+$ is the same as that corresponding to JB^{-1}, but with opposite orientation. If $B^2 = J$, they are the same and correspond to a (so-called reciprocal [98]) closed geodesic that goes through its image twice, once in each orientation. Let $a_D := |\{B \in \mathrm{Cl}_D^+ : B^2 = J\}|$, so $h_D^+ = a_D + 2b_D$. The images of the closed geodesics from Λ_D in $\Gamma\backslash\mathbb{H}$ correspond to a_D geodesic segments of length $\log \varepsilon_D^+$ with multiplicity 2, and b_D geodesic segments of length $2 \log \varepsilon_D^+$ also with multiplicity 2. Furthermore, the height of each of those closed geodesics in the fundamental domain is $\leq \sqrt{D}/2$ [31, Proposition 3.1]. Therefore, if we model each of those geodesic segments using independent (except for the multiplicities) random geodesic segments in X, with a cutoff $A > \sqrt{D}/2$, we may expect

$$\begin{aligned}
\mathrm{Var}(r, R; \Lambda_D) &\approx 4 \left(a_D \, \mathrm{Var}_A(r, R; \log \varepsilon_D^+) + b_D \, \mathrm{Var}_A(r, R; 2 \log \varepsilon_D^+) \right) \\
&\sim 4 \left(a_D \frac{16 \cdot \log \varepsilon_D^+ \cdot R^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right) + b_D \frac{16 \cdot 2 \log \varepsilon_D^+ \cdot R^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right) \right) \\
&= \frac{64 \sqrt{D} L(1, \chi_D) R^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right),
\end{aligned}$$

at least for $\log\left(\frac{\sqrt{D}}{2}\right) < \log A = o\left(R^{-1} \log^{-1}\left(\frac{1}{R}\right)\right)$ (taking Remark 6 into account, but already from Theorem 3 for thick annuli with $R - r \gg R$). It would suffice to restrict to $R \leq (\log D)^{-1-\delta}$ for any fixed $\delta > 0$. Therefore, being a bit conservative, this leads to the conjecture below.

Conjecture 1. *Let $\delta > 0$ be given. If $0 \le r < R \le D^{-\delta}$, then as $D \to \infty$ through squarefree fundamental discriminants,*

$$\mathrm{Var}(r, R; \Lambda_D) \sim \frac{64\sqrt{D}L(1, \chi_D)R^3}{\pi}\mathbf{G}\left(\frac{r}{R}\right).$$

Our main result confirms the conjecture for sufficiently small annuli that are not too thin, and in particular for small balls. Observe that the allowed range of radii intersects the regime where one would expect equidistribution, i.e. $\mu(A_{r,R}) \ge D^{-1+\delta}$.

Theorem 4. *Let $\delta > 0$ be given. If $0 \le r < R \le D^{-\frac{5}{12}-\delta}$ and $R - r \gg R$, then as $D \to \infty$ through squarefree fundamental discriminants,*

$$\mathrm{Var}(r, R; \Lambda_D) \sim \frac{64\sqrt{D}L(1, \chi_D)R^3}{\pi}\mathbf{G}\left(\frac{r}{R}\right).$$

Remark 7. *The restriction $R - r \gg R$ in Theorem 4 is mostly technical in nature, due to the fact that the behavior of the weight function $h_{r,R}(t)$ changes when $R - r \ll R$. We stick with it for simplicity, since it is enough to cover the most interesting case of balls ($r = 0$).*

Remark 8. *The proof of Theorem 4 actually gives a power-saving error term of the form $O_\varepsilon(D^{\frac{1}{2}}R^{3+\varepsilon})$ for any $\varepsilon > 0$ sufficiently small (depending on δ).*

Spectral expansion and automorphic transformations

Let $D > 0$ be a squarefree fundamental discriminant, χ_D be the primitive quadratic character modulo D, and $\mathcal{B}_0(\Gamma)$ be an orthonormal basis of the space of Maaß cusp forms for the modular group Γ, which we may choose to consist of HeckeMaaß cusp forms.

Expressing the variance in terms of the automorphic kernel $K_{r,R}$, performing a spectral expansion, and using the work of Duke–Imamoḡlu–Tóth [29] to compute the resulting Weyl sums, we are left with L-functions.

Lemma 19 (Spectral expansion of the variance [54, Lemma 2.20])**.** *We have*

$$\begin{aligned}\mathrm{Var}(r, R; \Lambda_D) = \frac{\sqrt{D}}{2\widetilde{\mu}(X)}\sum_{f \in \mathcal{B}_0(\Gamma)}\frac{L\left(\frac{1}{2}, f\right)L\left(\frac{1}{2}, f \otimes \chi_D\right)}{L\left(1, \mathrm{sym}^2 f\right)}H(t_f)\left|h_{r,R}(t_f)\right|^2 \\ + \frac{\sqrt{D}}{4\pi\widetilde{\mu}(X)}\int_{-\infty}^{\infty}\left|\frac{\zeta\left(\frac{1}{2}+it\right)L\left(\frac{1}{2}+it, \chi_D\right)}{\zeta(1+2it)}\right|^2 H(t)\left|h_{r,R}(t)\right|^2 dt,\end{aligned}$$

where

$$H(t) := \frac{\Gamma\left(\frac{1}{4}+\frac{it}{2}\right)^2\Gamma\left(\frac{1}{4}-\frac{it}{2}\right)^2}{\Gamma\left(\frac{1}{2}+it\right)\Gamma\left(\frac{1}{2}-it\right)} = \frac{4\pi}{|t|+1} + O\left(\frac{1}{(|t|+1)^2}\right). \tag{3.22}$$

Before the next lemma we need to establish some notation. Recall that the Mellin transform $\widehat{W}$ of a function $W : (0,\infty) \to \mathbb{C}$ is given by

$$\widehat{W}(s) := \int_0^\infty W(x) x^s \frac{dx}{x}$$

for $s \in \mathbb{C}$ for which the integral is absolutely convergent, and conversely the inverse Mellin transform $\widecheck{\mathcal{W}}$ of a holomorphic function $\mathcal{W} : \{s \in \mathbb{C} : a < \Im(s) < b\} \to \mathbb{C}$ is given by

$$\widecheck{\mathcal{W}}(x) := \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \mathcal{W}(s) x^{-s}\, ds$$

for $a < \sigma < b$ and $x \in (0,\infty)$ for which the integral converges absolutely.

Lemma 20 (Automorphic transformations [54, Corollary 5.7])**.** *Let $h(t)$ be an even holomorphic function in the strip $-2M < \Im(t) < 2M$ for some $M \geq 20$ with zeroes at $\pm\left(n-\frac{1}{2}\right)i$ for $n \in \{1,2,\dots,2M\}$ and satisfying $h(t) \ll (|t|+1)^{-2M}$ in this region. Then the moment*

$$\sum_{f\in\mathcal{B}_0(\Gamma)} \frac{L\left(\frac{1}{2},f\right)L\left(\frac{1}{2},f\otimes\chi_D\right)}{L\left(1,\operatorname{sym}^2 f\right)} h(t_f) + \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\frac{\zeta\left(\frac{1}{2}+it\right)L\left(\frac{1}{2}+it,\chi_D\right)}{\zeta(1+2it)}\right|^2 h(t)\,dt$$

is equal to the sum of the main term

$$2L(1,\chi_D)\int_{-\infty}^{\infty} h(t)\, d_{\mathrm{spec}}t \tag{3.23}$$

and the shifted convolution

$$\begin{aligned}\frac{2}{\sqrt{D}}\sum_{\pm}\sum_{D_1D_2=D}\sum_{\substack{m=1\\ m\neq\mp D_2}}^{\infty} &\chi_1(\operatorname{sgn}(m\pm D_2))\lambda_{\chi_1,\chi_2}(m,0)\lambda_{\chi_1,\chi_2}(|m\pm D_2|,0)\\ &\times\frac{1}{2\pi i}\int_{\sigma_1-i\infty}^{\sigma_1+i\infty}\widehat{\mathscr{K}^- h}(s)\widehat{\mathcal{J}_0^{\pm}}(1-s)\left(\frac{m}{D_2}\right)^{\frac{s-1}{2}} ds,\end{aligned} \tag{3.24}$$

where $1-M < \sigma_1 < -1$,

$$\left(\mathscr{K}^- h\right)(x) := \int_{-\infty}^{\infty} h(t)\mathcal{J}_t^-(x)\, d_{\mathrm{spec}}t, \qquad d_{\mathrm{spec}}t := \frac{1}{2\pi^2}t\tanh(\pi t)\, dt,$$

$$\mathcal{J}_t^-(x) := 4\cosh(\pi t)\, K_{2it}(4\pi x), \qquad \mathcal{J}_0^+(x) := -2\pi Y_0(4\pi x),$$

the decomposition $\chi_D = \chi_1\chi_2$ *corresponds to* $D = D_1D_2$*, and*

$$\lambda_{\chi_1,\chi_2}(m,0) := \sum_{ab=m} \chi_1(a)\chi_2(b).$$

Combining this with work of M. Young [106] leads to the following bound for moments of L-functions, which will be useful for bounding some of our error terms later on.

Lemma 21 (Dyadic moment bound [54, Proposition 2.35 (1)])**.** *For* $T \geq 1$*, we have*

$$\sum_{\substack{f\in\mathcal{B}_0(\Gamma)\\ T\leq t_f\leq 2T}} \frac{L\left(\frac{1}{2},f\right)L\left(\frac{1}{2},f\otimes\chi_D\right)}{L(1,\mathrm{sym}^2 f)} + \frac{1}{2\pi}\int\limits_{T\leq|t|\leq 2T}\left|\frac{\zeta\left(\frac{1}{2}+it\right)L\left(\frac{1}{2}+it,\chi_D\right)}{\zeta(1+2it)}\right|^2 dt$$
$$\ll_\varepsilon \begin{cases} D^{\frac{1}{3}+\varepsilon}T^{2+\varepsilon} & \text{for } T \ll D^{\frac{1}{12}},\\ D^{\frac{1}{2}+\varepsilon} & \text{for } D^{\frac{1}{12}} \ll T \ll D^{\frac{1}{4}},\\ D^{\varepsilon}T^{2+\varepsilon} & \text{for } T \gg D^{\frac{1}{4}}. \end{cases}$$

Choice of test function

To prove Theorem 4, we will start with the expression in Lemma 19 and approximate the weights $H(t)|h_{r,R}(t)|^2$ by a function $h(t)$ satisfying the conditions of Lemma 20. The error terms coming from switching from one set of weights to the other may be bounded using Lemma 21, and the problem will be reduced to evaluating the main term (3.23) and the error term (3.24). We once again follow [54], adapting their construction to our context.

The conditions of Lemma 20 require that $h(t)$ be even, extend holomorphically to $|\Im(t)| < 2M$, have zeros at $\pm\left(n-\frac{1}{2}\right)i$ for $n \in \{1,2,\ldots,2M\}$, and satisfy $h(t) \ll (|t|+1)^{-2M}$ for some integer $M \geq 20$. From now on, fix a sufficiently large constant $M \in \mathbb{N}$.

First we localize $h(t)$ to the region $[-T_2,-T_1]\cup[T_1,T_2]$, where $T_1 := R^{-1+\alpha}$ and $T_2 := R^{-1-\alpha}$ for a sufficiently small fixed constant $\alpha > 0$. This is because the main contribution to $\mathrm{Var}(r,R;\Lambda_D)$ will come from this range when $R-r \gg R$. To achieve this localization, let

$$h_1(t) := e^{-\left(\frac{t}{T_2}\right)^{2M}}\left(1-e^{-\left(\frac{t}{T_1}\right)^{2M}}\right),$$

which is even and for $|\Im(t)| < 2M$ satisfies

$$h_1(t) = \begin{cases} O\left(\left(\frac{|\Re(t)|+1}{T_1}\right)^{2M}\right) & \text{for } |\Re(t)| \le T_1, \\ 1 + O\left(\left(\frac{\Re(t)}{T_2}\right)^{2M} + e^{-\left(\frac{\Re(t)}{T_1}\right)^{2M}}\right) & \text{for } T_1 \le |\Re(t)| \le T_2, \\ O\left(e^{-\left(\frac{\Re(t)}{T_2}\right)^{2M}}\right) & \text{for } |\Re(t)| \ge T_2. \end{cases} \tag{3.25}$$

Moreover, for $j \in \{1, 2, \dots, 2M\}$ and $t \in \mathbb{R}$,

$$h_1^{(j)}(t) \ll_j \begin{cases} \frac{|t|^{2M-j}}{T_1^{2M}} & \text{for } |t| \le T_1, \\ \frac{|t|^{2M-j}}{T_2^{2M}} + \frac{|t|^{(2M-1)j}}{T_1^{2Mj}} e^{-\left(\frac{t}{T_1}\right)^{2M}} & \text{for } T_1 \le |t| \le T_2, \\ \frac{|t|^{(2M-1)j}}{T_2^{2Mj}} e^{-\left(\frac{t}{T_2}\right)^{2M}} & \text{for } |t| \ge T_2. \end{cases} \tag{3.26}$$

Next, ignoring the Bessel factors for now, we see from Lemma 12 and (3.22) that a factor asymptotic to $16\pi^3/|t|^3$ arises. Therefore consider

$$h_2(t) := 2(2\pi)^{-4M+1}(4M+3)^{-3} \frac{\Gamma\left(\frac{2M}{4M+3} + \frac{it}{4M+3}\right)^{4M+3} \Gamma\left(\frac{2M}{4M+3} - \frac{it}{4M+3}\right)^{4M+3}}{\Gamma\left(\frac{1}{2} + it\right)\Gamma\left(\frac{1}{2} - it\right)},$$

which is even and holomorphic in the strip $|\Im(t)| < 2M$, where it has zeros at $\pm\left(n - \frac{1}{2}\right) i$ for $n \in \{1, 2, \dots, 2M\}$ and satisfies

$$h_2(t) = \frac{16\pi^3}{(|t|+1)^3} + O\left(\frac{1}{(|t|+1)^4}\right), \tag{3.27}$$

by Stirling's formula. Furthermore, for $j \in \mathbb{Z}_{\ge 0}$ and $t \in \mathbb{R}$,

$$h_2^{(j)}(t) \ll_j (|t|+1)^{-j-3}. \tag{3.28}$$

Finally, let

$$h_3(t) := (R \cdot J_1(Rt) - r \cdot J_1(rt))^2,$$

which is entire (as this is the case for J_1) and even (as J_1 is odd [42, page 8.476.1]). Using the crude bound

$$J_1(z) \ll \begin{cases} |z| & \text{for } |z| \le 1, \\ \frac{e^{|\Im(z)|}}{\sqrt{|z|}} & \text{for } |z| \ge 1 \end{cases} \tag{3.29}$$

[42, 8.411.3 and 8.451.1], for $|\Im(t)| < 2M$ we get

$$h_3(t) \ll \begin{cases} R^4|t|^2 & \text{for } |t| \le \frac{1}{R}, \\ \frac{R}{|t|} & \text{for } |t| \ge \frac{1}{R}. \end{cases} \tag{3.30}$$

Also, for $j \in \mathbb{Z}_{\geq 0}$ and $y \in \mathbb{R}$ we have

$$J_1^{(j)}(y) \ll_j \begin{cases} |y| & \text{for } |y| \leq 1 \text{ and } j \text{ even}, \\ 1 & \text{for } |y| \leq 1 \text{ and } j \text{ odd}, \\ \frac{1}{\sqrt{|y|}} & \text{for } |y| \geq 1 \end{cases}$$

[42, 8.471.2, 8.411.4, and 8.451.1], which for $t \in \mathbb{R}$ gives

$$h_3^{(j)}(t) \ll_j \begin{cases} R^4|t|^{2-j} & \text{for } |t| \leq \frac{1}{R} \text{ and } j \in \{0,1\}, \\ R^{2+j} & \text{for } |t| \leq \frac{1}{R} \text{ and } j \geq 2, \\ \frac{R^{1+j}}{|t|} & \text{for } |t| \geq \frac{1}{R}. \end{cases} \tag{3.31}$$

We choose the test function

$$h(t) := h_1(t)h_2(t)h_3(t), \tag{3.32}$$

so that combining (3.25), (3.27), and (3.30) gives the following upper bounds and asymptotics for h.

Lemma 22. *For $|\Im(t)| < 2M$,*

$$h(t) \ll \begin{cases} \frac{R^4(|\Re(t)|+1)^{2M-1}}{T_1^{2M}} & \text{for } |t| \leq T_1, \\ \frac{R^4}{|\Re(t)|} & \text{for } T_1 \leq |t| \leq \frac{1}{R}, \\ \frac{R}{|\Re(t)|^4} & \text{for } \frac{1}{R} \leq |t| \leq T_2, \\ \frac{R}{|\Re(t)|^4} e^{-\left(\frac{\Re(t)}{T_2}\right)^{2M}} & \text{for } |t| \geq T_2. \end{cases} \tag{3.33}$$

Furthermore, if $t \in \mathbb{R}$ then

$$\begin{aligned} h(t) = \frac{4\pi}{|t|}\left(2\pi\frac{R\cdot J_1(Rt) - r\cdot J_1(rt)}{t}\right)^2 \\ + \begin{cases} O\left(\frac{R^4}{|t|^2} + \frac{R^4|t|^{2M-1}}{T_2^{2M}} + \frac{R^4}{|t|}e^{-\left(\frac{t}{T_1}\right)^{2M}}\right) & \text{for } T_1 \leq |t| \leq \frac{1}{R}, \\ O\left(\frac{R}{|t|^5} + \frac{R|t|^{2M-4}}{T_2^{2M}}\right) & \text{for } \frac{1}{R} \leq |t| \leq T_2. \end{cases} \end{aligned} \tag{3.34}$$

We record the following important definitions and bounds for future reference:

$$R \ll D^{-\frac{5}{12}-\delta}, \qquad R - r \gg R, \qquad T_1 = R^{-1+\alpha}, \qquad T_2 = R^{-1-\alpha}, \tag{3.35}$$

where $\alpha, \delta > 0$ are sufficiently small fixed constants and $M \in \mathbb{N}$ is a sufficiently large fixed constant.

Change of test function for the variance

Lemma 23. *Under the assumptions of* (3.35) *and for* $h(t)$ *as in* (3.32)*, we have*

$$\begin{aligned}\mathrm{Var}(r,R;\Lambda_D) = \frac{\sqrt{D}}{2\widetilde{\mu}(X)}\sum_{f\in\mathcal{B}_0(\Gamma)}\frac{L\left(\frac{1}{2},f\right)L\left(\frac{1}{2},f\otimes\chi_D\right)}{L\left(1,\mathrm{sym}^2 f\right)}h(t_f)\\ +\frac{\sqrt{D}}{4\pi\widetilde{\mu}(X)}\int_{-\infty}^{\infty}\left|\frac{\zeta\left(\frac{1}{2}+it\right)L\left(\frac{1}{2}+it,\chi_D\right)}{\zeta(1+2it)}\right|^2 h(t)\,dt\\ +O_\varepsilon\left(D^{\frac{1}{2}+\varepsilon}R^{3-\varepsilon}\left(RD^{\frac{5}{12}}+RT_1+\frac{RD^{\frac{1}{2}}}{T_1}+\frac{1}{(RT_2)^2}\right)\right).\end{aligned} \tag{3.36}$$

Proof. Follows from the spectral expansion in Lemma 19 and a change of test function. The error term is estimated using the bounds and asymptotics in Lemma 12, (3.22), and Lemma 22, considering each of the ranges separately. More specifically, if we denote $\Delta(t) := H(t)|h_{r,R}(t)|^2 - h(t)$, then putting those bounds together yields, for $t\in\mathbb{R}$,

$$\Delta(t)\ll\begin{cases}\frac{R^4}{1+|t|} & \text{for } |t|\le T_1,\\ \frac{R^4}{|t|^2}+\frac{R^4|t|^{2M-1}}{T_2^{2M}}+\frac{R^4}{|t|}e^{-\left(\frac{t}{T_1}\right)^{2M}} & \text{for } T_1\le|t|\le\frac{1}{R},\\ \frac{R}{|t|^5}+\frac{R|t|^{2M-4}}{T_2^{2M}} & \text{for } \frac{1}{R}\le|t|\le T_2,\\ \frac{R}{|t|^4} & \text{for } |t|\ge T_2.\end{cases}$$

We used the fact that $\alpha>0$ is small in the inequality above.

Therefore, combining this with Lemma 21 gives, for $T \geq 1$,

$$\sum_{\substack{f \in \mathcal{B}_0(\Gamma) \\ T \leq t_f \leq 2T}} \frac{L\left(\frac{1}{2}, f\right) L\left(\frac{1}{2}, f \otimes \chi_D\right)}{L\left(1, \operatorname{sym}^2 f\right)} |\Delta(t_f)|$$

$$+ \frac{1}{2\pi} \int\limits_{T \leq |t| \leq 2T} \left| \frac{\zeta\left(\frac{1}{2} + it\right) L\left(\frac{1}{2} + it, \chi_D\right)}{\zeta(1 + 2it)} \right|^2 |\Delta(t)| \, dt$$

$$\ll_\varepsilon \begin{cases} D^{\frac{1}{3}+\varepsilon} T^{1+\varepsilon} R^4 & \text{for } 1 \leq T \ll D^{\frac{1}{12}}, \\ D^{\frac{1}{2}+\varepsilon} \frac{R^4}{T} & \text{for } D^{\frac{1}{12}} \ll T \ll D^{\frac{1}{4}}, \\ D^{\varepsilon} T^{1+\varepsilon} R^4 & \text{for } D^{\frac{1}{4}} \ll T \leq T_1, \\ D^{\varepsilon} T^{\varepsilon} \left(R^4 + R^4 T \left(\frac{T}{T_2}\right)^{2M} + R^4 T e^{-\left(\frac{T}{T_1}\right)^{2M}} \right) & \text{for } T_1 \leq T \leq \frac{1}{R}, \\ D^{\varepsilon} T^{\varepsilon} \left(\frac{R}{T^3} + \frac{R}{T^2} \left(\frac{T}{T_2}\right)^{2M} \right) & \text{for } \frac{1}{R} \leq T \leq T_2, \\ D^{\varepsilon} T^{\varepsilon} \frac{R}{T^2} & \text{for } T \geq T_2. \end{cases}$$

Here we recall that $T_1 = R^{-1+\alpha} \gg D^{(1-\alpha)(\frac{5}{12}+\delta)} \geq D^{\frac{1}{4}}$ for sufficiently small $\alpha > 0$. Multiplying by $\sqrt{D}$ and summing over $T = 2^k$ for $k \in \mathbb{Z}_{\geq 0}$ gives the claimed error term.

□

Observe that by (3.35) the error term is $O_\varepsilon(D^{\frac{1}{2}} R^{3+\varepsilon})$ for $\varepsilon > 0$ sufficiently small, and therefore it is asymptotically smaller than the main term of Theorem 4, as $L(1, \chi_D) \gg_\varepsilon D^{-\varepsilon}$.

Remark 9. *The error term $O_\varepsilon(D^{\frac{11}{12}+\varepsilon} R^{4-\varepsilon})$ in* (3.36) *is the only point in the proof of Theorem 4 where the range $R \leq D^{-\frac{5}{12}-\delta}$ is tight. Instead of using the bound $\ll_\varepsilon D^{\frac{1}{3}+\varepsilon} T^{2+\varepsilon}$ (coming from Young's work [106]) for the range $T \ll D^{\frac{1}{12}}$ of Lemma 21, we could have tried to use the weaker bound $\ll_\varepsilon D^{\frac{1}{2}+\varepsilon}$ (which holds in this range by the argument in Humphries-Radziwiłł [54, Proposition 2.35]). This would produce a corresponding error term of size $O_\varepsilon(D^{1+\varepsilon} R^4)$ in* (3.36)*, which is enough to obtain asymptotics for the variance if $R \leq D^{-\frac{1}{2}-\delta}$. Here it becomes clear that it is precisely the range of (conjectured) equidistribution, i.e. $R \geq D^{-\frac{1}{2}+\delta}$, which requires deeper arithmetic inputs.*

Applying Lemma 20 to the first two terms of (3.36), we obtain the main term

$$\frac{\sqrt{D} L(1, \chi_D)}{\widetilde{\mu}(X)} \int_{-\infty}^{\infty} h(t) \, d_{\text{spec}} t, \tag{3.37}$$

where $d_{\text{spec}}t := \frac{1}{2\pi^2} t \tanh(\pi t)\, dt$ as before, plus the shifted convolution

$$\frac{1}{\widetilde{\mu}(X)} \sum_{\pm} \sum_{D_1 D_2 = D} \sum_{\substack{m=1 \\ m \neq \mp D_2}}^{\infty} \chi_1(\operatorname{sgn}(m \pm D_2)) \lambda_{\chi_1,\chi_2}(m,0) \lambda_{\chi_1,\chi_2}(|m \pm D_2|, 0) \\ \times \frac{1}{2\pi i} \int_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} \widehat{\mathscr{K}^- h}(s) \widehat{\mathcal{J}_0^{\pm}}(1-s) \left(\frac{m}{D_2}\right)^{\frac{s-1}{2}} ds. \tag{3.38}$$

Asymptotics for main term

Lemma 24. *Under the assumptions of* (3.35) *and for* $h(t)$ *as in* (3.32)*, the main term* (3.37) *is equal to*

$$\frac{64\sqrt{D} L(1,\chi_D) R^3}{\pi} \mathbf{G}\left(\frac{r}{R}\right) + O_\varepsilon\left(D^{\frac{1}{2}+\varepsilon}\left(R^4 T_1 + \frac{R}{T_2^2}\right)\right).$$

Proof. Using the bounds and asymptotics of Lemma 22, combined with the fact that h is even and the bound $L(1,\chi_D) \ll \log D$, we see that (3.37) is equal to

$$\frac{16\pi\sqrt{D} L(1,\chi_D)}{\widetilde{\mu}(X)} \int_{T_1}^{T_2} \left(\frac{R \cdot J_1(Rt) - r \cdot J_1(rt)}{t}\right)^2 dt \\ + O_\varepsilon\left(D^{\frac{1}{2}+\varepsilon}\left(R^4 T_1 + \frac{R}{T_2^2}\right)\right). \tag{3.39}$$

Then (3.29) allows us to complete the integral to $(0,\infty)$ under the same error term as above.

From [42, page 6.574.2] it follows that

$$R^2 \int_0^\infty \frac{J_1(Rt)^2}{t^2}\, dt = \frac{R^3 \cdot \Gamma\left(\frac{1}{2}\right)}{4 \cdot \Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{5}{2}\right)\Gamma\left(\frac{3}{2}\right)} = \frac{4R^3}{3\pi}, \tag{3.40}$$

and similarly for the term corresponding to r. The cross-term can be evaluated using [42, page 6.574.3], which gives

$$2Rr \int_0^\infty \frac{J_1(Rt) J_1(rt)}{t^2}\, dt = Rr^2 \cdot {}_2F_1\left(\frac{1}{2}, -\frac{1}{2}; 2; \frac{r^2}{R^2}\right),$$

where ${}_2F_1$ denotes the ordinary hypergeometric function. By [42, 8.113.1, 8.114.1, and 9.137.14] we deduce that

$${}_2F_1\left(\frac{1}{2}, -\frac{1}{2}; 2; z^2\right) = \frac{4}{3\pi z^2}\left((1+z^2)\mathbf{E}(z) - (1-z^2)\mathbf{K}(z)\right).$$

Therefore,

$$2Rr\int_0^\infty \frac{J_1(Rt)J_1(rt)}{t^2}\,dt = \frac{4R^3}{3\pi}\left(\left(1+\frac{r^2}{R^2}\right)\mathbf{E}\left(\frac{r}{R}\right)-\left(1-\frac{r^2}{R^2}\right)\mathbf{K}\left(\frac{r}{R}\right)\right). \tag{3.41}$$

Combining (3.40) and (3.41), we conclude that

$$\int_0^\infty \left(\frac{R\cdot J_1(Rt)-r\cdot J_1(rt)}{t}\right)^2 dt = \frac{4R^3}{3\pi}\mathbf{G}\left(\frac{r}{R}\right),$$

which gives the desired result.

□

By (3.35), the error term in Lemma 24 is $O_\varepsilon(D^{\frac{1}{2}}R^{3+\varepsilon})$ for $\varepsilon>0$ sufficiently small, so it is once again asymptotically smaller than the main term of Theorem 4.

Bounds for shifted convolution

To finish the proof of Theorem 4, it suffices to show that the shifted convolution (3.38) is asymptotically smaller than the main term obtained in the previous subsection. This requires considerably more work and involves a more careful consideration of the oscillatory behavior of the test function $h(t)$. The final result is indicated in the lemma below.

Lemma 25. *Under the assumptions of* (3.35) *and for* $h(t)$ *as in* (3.32)*, the shifted convolution* (3.38) *is* $O_\varepsilon(D^{\frac{1}{2}}R^{3+\varepsilon})$ *for every* $\varepsilon>0$ *sufficiently small.*

Proof. We once again follow [54], with necessary modifications due to the fact that $D>0$ and also the presence of oscillations coming from a Bessel function, instead of a trigonometric function, in our choice of $h(t)$.

By Mellin inversion — where we use the convolution identity [56, (A.6)] — and the divisor bound, it suffices to show that

$$\sum_{\pm}\sum_{D_2|D}\sum_{m=1}^{\infty} m^\varepsilon \left|\int_0^\infty (\mathscr{K}^- h)(x)\mathcal{J}_0^\pm\left(\sqrt{\frac{m}{D_2}}x\right)\,dx\right| \tag{3.42}$$

is $O_\varepsilon(D^{\frac{1}{2}}R^{3+\varepsilon})$ for every $\varepsilon>0$ sufficiently small. We consider two different ranges for m.

Case 1: $m > \sqrt{D_2}$.

Via integration by parts and [42, 8.472.1 and 8.486.12], the integral in (3.42) is

$$\frac{c^{\pm} D_2}{16\pi^2 m} \int_0^{\infty} \frac{1}{x^2} \mathscr{L}(x) B_2^{\pm}\left(4\pi\sqrt{\frac{m}{D_2}}x\right) dx, \tag{3.43}$$

where

$$c^+ := -2\pi, \qquad c^- := 4, \qquad B_k^+(x) := Y_k(x), \qquad B_k^-(x) := K_k(x),$$

and

$$\mathscr{L}(x) := 3(\mathscr{K}^- h)(x) - 3x(\mathscr{K}^- h)'(x) + x^2(\mathscr{K}^- h)''(x).$$

We will split the integral in (3.43) into three different ranges for x and bound each one separately.

Sub-case 1a: $0 < x \leq 1$.

By [5, (A.2) and (A.4)],

$$\frac{d^j}{dx^j}\mathscr{J}_t^-(x) = \frac{(2\pi)^j \pi i}{\sinh(\pi t)} \sum_{n=0}^{j} \binom{j}{n} (I_{2it-j+2n}(4\pi x) - I_{-2it-j+2n}(4\pi x)).$$

Combining this with the bound

$$e^{-\pi|t|} I_{2it-j+2n}(4\pi x) \ll_{\Im(t),j} \frac{x^{-j+2(n-\Im(t))}}{(|\Re(t)|+1)^{\frac{1}{2}-j+2(n-\Im(t))}}$$

valid for $0 < x \ll \sqrt{|t|+1}$, which follows from a slight adaptation of [5, (A.6)], we can shift contours to obtain

$$x^j \frac{d^j}{dx^j}(\mathscr{K}^- h)(x) \ll_j \sum_{\pm} \sum_{n=0}^{j} x^{2(n-c_n)} \int\limits_{\Im(t)=\pm c_n} |h(t)|(|\Re(t)|+1)^{j-2(n-c_n)+\frac{1}{2}}\, dt$$

for any choice of integers $-2M < c_n < 2M$ (observe that the poles of $\cosh^{-1}(\pi t)$ are cancelled by the zeros of $h(t)$). Choose $c_n = n - 2M + 1$ and apply (3.33) to conclude that for $0 < x \leq 1$,

$$\mathscr{L}(x) \ll \frac{R^4 x^{4M-2}}{T_1^{2M}}. \tag{3.44}$$

For future reference, we note that if $k \in \mathbb{Z}_{\geq 0}$ and $x \in \mathbb{R}_{>0}$ then one has the general bound

$$B_k^{\pm}(x) \ll_{k,\varepsilon} \begin{cases} x^{-k-\varepsilon} & \text{for } 0 < x \leq 1, \\ \frac{1}{\sqrt{x}} & \text{for } x \geq 1 \end{cases} \tag{3.45}$$

for both Bessel functions in question [48, Proposition 9]. Therefore, we conclude that the contribution of $0 < x \leq 1$ to (3.43) is

$$\ll_\varepsilon \frac{R^4}{T_1^{2M}} \left(\left(\frac{D_2}{m} \right)^{2+\varepsilon} + \left(\frac{D_2}{m} \right)^{\frac{5}{4}} \right).$$

Summing over $m > \sqrt{D_2}$ and $D_2 | D$, this sub-case contributes $O_\varepsilon(R^4 T_1^{-2M} D^{\frac{3}{2}+\varepsilon})$ to (3.42), which is $O_\varepsilon(D^{\frac{1}{2}} R^{3+\varepsilon})$ by (3.35).

Sub-case 1b: $x \geq T_2 \log T_2$.

We use [5, (A.1)] to write

$$\frac{d^j}{dx^j} \mathcal{J}_t^-(x) = (-2\pi)^j \sum_{n=0}^{j} \binom{j}{n} 4 \cosh(\pi t) K_{2it-j+2n}(4\pi x)$$

and apply the uniform bound

$$\cosh(\pi t) K_{2it-j+2n}(4\pi x) \ll_{\Im(t),j} e^{\min\{0, -\pi(4x - |\Re(t)|)\}} \left(\frac{1 + |\Re(t)| + 4\pi x}{4\pi x} \right)^{|2\Im(t)+j-2n|+\frac{1}{10}}$$

valid for all $t \in \mathbb{C}$ [5, (A.3)]. Combining this with (3.33) gives, for $t \in \mathbb{R}$ and $x \geq T_2$,

$$h(t) t \frac{d^j}{dx^j} \mathcal{J}_t^-(x) \ll_j \begin{cases} \frac{R^4 (|t|+1)^{2M}}{T_1^{2M}} e^{\pi|t|} e^{-4\pi x} & \text{for } |t| \leq T_1, \\ R^4 e^{\pi|t|} e^{-4\pi x} & \text{for } T_1 \leq |t| \leq \frac{1}{R}, \\ \frac{R}{|t|^3} e^{\pi|t|} e^{-4\pi x} & \text{for } \frac{1}{R} \leq |t| \leq T_2, \\ \frac{R}{|t|^3} e^{-\left(\frac{t}{T_2}\right)^{2M}} e^{\pi|t|} e^{-4\pi x} & \text{for } T_2 \leq |t| \leq 4x, \\ \frac{R}{|t|^3} e^{-\left(\frac{t}{T_2}\right)^{2M}} \left(\frac{|t|}{x}\right)^{j+\frac{1}{10}} & \text{for } |t| \geq 4x. \end{cases}$$

Considering each range separately (and in fact dividing the fourth range into $|t| \leq 2x$ and $|t| \geq 2x$) leads to

$$\begin{aligned} \mathscr{L}(x) &\ll \sum_{j=0}^{2} x^j \int_{-\infty}^{\infty} \left| h(t) t \frac{d^j}{dx^j} \mathcal{J}_t^-(x) \right| dt \\ &\ll \sum_{j=0}^{2} x^j \left(R^4 e^{\pi T_2} e^{-4\pi x} + \frac{R}{T_2^2} e^{-2\pi x} + \frac{R}{x^3} e^{-\left(\frac{2x}{T_2}\right)^{2M}} + \frac{R T_2}{x^3} e^{-\left(\frac{4x}{T_2}\right)^{2M}} \right) \\ &\ll R^3 x^2 e^{-2\pi x} + \frac{R T_2}{x} e^{-\left(\frac{2x}{T_2}\right)^{2M}}. \end{aligned} \tag{3.46}$$

Therefore, using (3.45) once again, the contribution of $x \geq T_2 \log T_2$ to the integral (3.43) is $O_A(T_2^{-A}(D_2/m)^{\frac{5}{4}})$ for any $A > 0$, which easily gives the desired bound of $O_\varepsilon(D^{\frac{1}{2}}R^{3+\varepsilon})$ for the corresponding contribution to (3.42).

Sub-case 1c: $1 < x < T_2 \log T_2$.

We use the identity

$$(\mathscr{K}^- h)(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} e(2x\sinh(\pi u))\int_{-\infty}^{\infty} h(t)t\tanh(\pi t)e(-ut)\,dt\,du$$

from [5, (A.8)], which is valid due to the rapid decay of $h(t)$, following from (22). Then integrating by parts in u gives

$$\mathscr{L}(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} e(2x\sinh(\pi u))\int_{-\infty}^{\infty} \widetilde{h}(t)(c_0(u)+c_1(u)t+c_2(u)t^2)e(-ut)\,dt\,du, \tag{3.47}$$

where

$$\widetilde{h}(t) := h(t)t\tanh(\pi t)$$

and

$$\begin{aligned} c_0(u) &:= 8 - 8\tanh^2(\pi u) + 3\tanh^4(\pi u), \\ c_1(u) &:= -14i\tanh(\pi u) + 6i\tanh^3(\pi u), \\ c_2(u) &:= -4\tanh^2(\pi u). \end{aligned}$$

From $\frac{d}{dt}\tanh(\pi t) = \pi\operatorname{sech}(\pi t)^2$ and $\frac{d}{dt}\operatorname{sech}(\pi t) = -\pi\tanh(\pi t)\operatorname{sech}(\pi t)$ we can show by induction that for $j \geq 1$, there is a polynomial $Q_j(x,y)$ such that

$$\frac{d^j}{dt^j}\tanh(\pi t) = \operatorname{sech}(\pi t)^2 \cdot Q_j(\tanh(\pi t), \operatorname{sech}(\pi t)) \ll_j e^{-2\pi|t|},$$

which will be negligible in what follows. Combining such a bound with (3.26), (3.28), and (3.31) we conclude that for $j \in \{0, 1, \ldots, 2M\}$ and

$t \in \mathbb{R}$,

$$\widetilde{h}^{(j)}(t) \ll_j \begin{cases} \left(\frac{t}{T_1}\right)^{2M} \cdot \frac{1}{(|t|+1)^3} \cdot R^4|t|^2 \cdot |t| \cdot \frac{1}{|t|^j} & \text{for } |t| \le T_1, \\ \left(1 + \left(\frac{t}{T_2}\right)^{2M} + \left(\frac{t}{T_1}\right)^{2Mj} e^{-\left(\frac{t}{T_1}\right)^{2M}}\right) \cdot \frac{R^4}{|t|^j} & \text{for } T_1 \le |t| \le \frac{1}{R}, \\ \left(1 + \left(\frac{t}{T_2}\right)^{2M} + \left(\frac{t}{T_1}\right)^{2Mj} e^{-\left(\frac{t}{T_1}\right)^{2M}}\right) \cdot \frac{R^{j+1}}{|t|^3} & \text{for } \frac{1}{R} \le |t| \le T_2, \\ \left(\frac{t}{T_2}\right)^{2Mj} e^{-\left(\frac{t}{T_2}\right)^{2M}} \cdot \frac{1}{|t|^3} \cdot \frac{R}{|t|} \cdot |t| \cdot R^j & \text{for } |t| \ge T_2 \end{cases}$$

$$\ll_j \begin{cases} \frac{R^4(|t|+1)^{2M-j}}{T_1^{2M}} & \text{for } |t| \le T_1, \\ \frac{R^4}{T_1^j} & \text{for } T_1 \le |t| \le \frac{1}{R}, \\ \frac{R}{|t|^3 T_1^j} & \text{for } \frac{1}{R} \le |t| \le T_2, \\ \frac{R^{1+j}}{|t|^3} e^{-\frac{1}{2}\left(\frac{t}{T_2}\right)^{2M}} & \text{for } |t| \ge T_2. \end{cases} \tag{3.48}$$

We now bound (3.47) by dividing the integral over u into the ranges $|u| \le v$ and $|u| > v$, where $v \in (0,1)$ will be chosen later. In the case $|u| > v$, we estimate the integral over t by integrating by parts $2M$ times. Since $c_i(u) \ll |u|^i$ for $i \in \{0,1,2\}$, (3.48) shows that the contribution of this range to $\mathscr{L}(x)$ is

$$\ll R \log(1/R) T_1^{-2M} v^{-2M+1} (R+v)^2.$$

For $|u| \le v < 1$, we Taylor expand twice to get

$$e(2x \sinh(\pi u)) = e(2x(\pi u + O(u^3))) = e(2\pi x u) + O(x u^3),$$

as long as $xv^3 < 1$ (which will be the case for our choice of v). Plugging this into (3.47) and using (3.48), the error term is

$$\ll R \log(1/R) x v^4 (R+v)^2.$$

To make the two error terms collected so far match, we choose

$$v = T_1^{-1+\frac{3}{2M+3}} x^{-\frac{1}{2M+3}},$$

which satisfies the necessary restrictions since in the present sub-case $x < T_2 \log T_2$. In the remaining integral over $|u| \le v$ we use

$$\begin{aligned} c_0(u) &= 8 + O(u^2), \\ c_1(u) &= -14i\pi u + O(u^3), \\ c_2(u) &= -4\pi^2 u^2 + O(u^4). \end{aligned}$$

The contribution of these error terms to (3.47) is

$$\ll R\log(1/R)v^3(R+v)^2 \ll R\log(1/R)v^5,$$

as $1 < x < T_2\log T_2$. Finally, we can complete the integral over u to $(-\infty,\infty)$ under an error term

$$\begin{aligned}&\ll R\log(1/R)T_1^{-2M}v^{-2M+1}(R+v)^2\\&= R\log(1/R)xv^4(R+v)^2 \ll R\log(1/R)xv^6,\end{aligned}$$

by the argument via integration by parts from before. Therefore,

$$\begin{aligned}\mathscr{L}(x) &= \frac{1}{\pi}\int_{-\infty}^{\infty} e(2\pi xu)\int_{-\infty}^{\infty}\widetilde{h}(t)(8-14i\pi ut-4\pi^2u^2t^2)e(-ut)\,dt\,du\\&\qquad + O\left(R\log(1/R)v^5(1+xv)\right)\\&= \frac{1}{\pi}\left(3\widetilde{h}(2\pi x)-3x\frac{d}{dx}\left[\widetilde{h}(2\pi x)\right]+x^2\frac{d^2}{dx^2}\left[\widetilde{h}(2\pi x)\right]\right)\\&\qquad + O\left(R\log(1/R)v^5(1+xv)\right),\end{aligned} \tag{3.49}$$

where the double integral was evaluated via Fourier inversion. For $1 < x < T_2\log T_2$ we have

$$R\log(1/R)v^5(1+xv) = R\log(1/R)\left(\frac{x^{-\frac{5}{2M+3}}}{T_1^{5-\frac{15}{2M+3}}}+\frac{x^{1-\frac{5}{2M+3}}}{T_1^{6-\frac{18}{2M+3}}}\right)\ll\frac{R}{T_1^{\frac{9}{2}}}$$

due to (3.35). Applying this combined with (3.48) to (3.49), we obtain

$$\mathscr{L}(x)\ll\frac{R}{T_1^{\frac{9}{2}}}+\begin{cases}\frac{R^4x^2}{T_1^2} & \text{for } 1\le x\le\frac{1}{R},\\ \frac{R}{xT_1^2} & \text{for } \frac{1}{R}\le x\le T_2\log T_2.\end{cases} \tag{3.50}$$

Using the bound above and (3.45), the contribution of $1 < x < T_2\log T_2$ to (3.43) is

$$\begin{aligned}&\ll_\varepsilon \frac{D_2}{m}\int_1^{T_2\log T_2}\frac{1}{x^2}|\mathscr{L}(x)|\left(\left(\frac{D_2}{mx^2}\right)^{1+\varepsilon}+\left(\frac{D_2}{m}\right)^{\frac{1}{4}}\frac{1}{\sqrt{x}}\right)dx\\&\ll\frac{R}{T_1^{\frac{9}{2}}}\left(\left(\frac{D_2}{m}\right)^{2+\varepsilon}+\left(\frac{D_2}{m}\right)^{\frac{5}{4}}\right).\end{aligned}$$

Summing over $M > \sqrt{D_2}$ and $D_2|D$, this sub-case adds $O_\varepsilon(RT_1^{-\frac{9}{2}}D^{\frac{3}{2}+\varepsilon})$ to (3.42). This is the most delicate range, but from (3.35) we see that it contributes $O_\varepsilon(D^{\frac{1}{2}}R^{3+\varepsilon})$, as desired.

Case 2: $1 \leq m \leq \sqrt{D_2}$.

In this case we directly bound the integral from (3.42), which is

$$c^{\pm} \int_0^{\infty} (\mathscr{K}^- h)(x) B_0^{\pm}\left(4\pi\sqrt{\frac{m}{D_2}}x\right)\, dx. \tag{3.51}$$

The strategy is to divide it into the same three ranges for x, and observe that the bounds (3.44), (3.46), and (3.50) for $\mathscr{L}(x)$ are actually bounds for

$$\max_{j\in\{0,1,2\}} \left| x^j \frac{d^j}{dx^j}(\mathscr{K}^- h)(x)\right|,$$

so they apply verbatim to $(\mathscr{K}^- h)(x)$. We simply combine them with (3.45) to estimate (3.51).

Sub-case 2a: $0 < x \leq 1$.

We see from (3.44) that the contribution of $0 < x \leq 1$ to (3.51) is bounded by $O(R^4 T_1^{-2M}(D_2/m)^{\frac{1}{4}})$, so this corresponds to a term of size $O_\varepsilon(R^4 T_1^{-2M} D^{\frac{5}{8}+\varepsilon})$ in (3.42), which is acceptable.

Sub-case 2b: $x \geq T_2 \log T_2$.

From (3.46), the contribution of this sub-case to (3.51) is $O_A(T_2^{-A}(D_2/m)^{\frac{1}{4}})$ for any $A > 0$, and this easily leads to an acceptable error term of $O_{A,\varepsilon}(T_2^{-A} D^{\frac{5}{8}+\varepsilon})$ for (3.42).

Sub-case 2c: $1 < x < T_2 \log T_2$.

Finally, (3.50) shows that this final sub-case contributes $O(R^{\frac{3}{2}} T_1^{-2}(D_2/m)^{\frac{1}{4}})$ to (3.51), which translates to $O_\varepsilon(R^{\frac{3}{2}} T_1^{-2} D^{\frac{5}{8}+\varepsilon})$ in (3.42). This is $O_\varepsilon(D^{\frac{1}{2}} R^{3+\varepsilon})$ by (3.35), so we have exhausted all possible cases and the proof of Lemma 25 (and therefore also of Theorem 4) is complete.

□

3.6 Limitations and connections to subconvexity

As Lemma 19 shows and we use in the course of our argument, bounds towards subconvexity have implications to (at least upper bounds for) the variance $\mathrm{Var}(r, R; \Lambda_D)$. We remark that the opposite is also true, in the sense that upper bounds of the correct order of magnitude for the variance imply subconvexity for certain L-functions. This clarifies the obstacles for improving Theorem 4.

For simplicity consider the case of balls, $r = 0$. If one has an upper bound of the (expected) correct order of magnitude for the variance, i.e. $\mathrm{Var}(0, R; \Lambda_D) \ll$

$\sqrt{D}L(1,\chi_D)R^3$, then assuming $R = o(1)$ the argument in Lemma 12 shows that $h_{0,R}(t) \gg R^2$ for $|t| \leq \frac{1}{R}$, so by Lemma 19 and non-negativity of the terms we get

$$\begin{aligned}\sum_{\substack{f\in\mathcal{B}_0(\Gamma)\\ |t_f|\leq\frac{1}{R}}} \frac{L\left(\frac{1}{2},f\right)L\left(\frac{1}{2},f\otimes\chi_D\right)}{L\left(1,\mathrm{sym}^2 f\right)|t_f|} \\ + \frac{1}{2\pi}\int\limits_{|t|\leq\frac{1}{R}} \left|\frac{\zeta\left(\frac{1}{2}+it\right)L\left(\frac{1}{2}+it,\chi_D\right)}{\zeta(1+2it)}\right|^2 \frac{dt}{|t|+1} \ll \frac{L(1,\chi_D)}{R}\end{aligned} \tag{3.52}$$

for squarefree fundamental discriminants $D > 0$ (observe that $|t_f| \gg 1$ for Γ). As an aside, we note that here the significance of the exponent $5/12$ in Theorem 4 becomes clear. This is because the hardest range in (3.52) is $|t_f|, |t| \asymp D^{\frac{1}{12}}$, where the bounds of Lemma 21 intersect, and the best one can do is use Hölder's inequality against the third moment result of [106] and the large sieve, obtaining

$$\begin{aligned}\sum_{\substack{f\in\mathcal{B}_0(\Gamma)\\ |t_f|\asymp D^{\frac{1}{12}}}} \frac{L\left(\frac{1}{2},f\right)L\left(\frac{1}{2},f\otimes\chi_D\right)}{L\left(1,\mathrm{sym}^2 f\right)} \\ + \frac{1}{2\pi}\int\limits_{|t|\asymp D^{\frac{1}{12}}} \left|\frac{\zeta\left(\frac{1}{2}+it\right)L\left(\frac{1}{2}+it,\chi_D\right)}{\zeta(1+2it)}\right|^2 dt \ll_\varepsilon D^{\frac{1}{2}+\varepsilon}.\end{aligned} \tag{3.53}$$

An improvement in the first moment bound (3.53) is essentially equivalent to an extension of the range of R in Theorem 4.

Going back to our point about subconvexity, dropping all but one term in (3.52) and using the bound $L(1,\mathrm{sym}^2 f) \gg_\varepsilon |t_f|^{-\varepsilon}$ of [51] we get

$$L\left(\frac{1}{2},f\right)L\left(\frac{1}{2},f\otimes\chi_D\right) \ll_\varepsilon \frac{D^\varepsilon |t_f|^{1+\varepsilon}}{R}$$

for $f \in \mathcal{B}_0(\Gamma)$ with $|t_f| \leq \frac{1}{R}$. The conductor of the product of L-functions on the left is $\asymp D^2|t_f|^4$, so if $f \in \mathcal{B}_0(\Gamma)$ is fixed and $R \gg D^{-\frac{1}{3}+\delta}$ for a given $\delta > 0$, we would obtain sub-Weyl subconvexity for $f \otimes \chi_D$ in the twist aspect, which is currently an open problem.

In conclusion, improving the exponent $5/12$ of Theorem 4 requires a better bound for the first moment (3.53), and improving it to anything below $1/3$ seems especially difficult at present, as it implies a challenging case of sub-Weyl subconvexity.

Chapter 4

MÖBIUS DISJOINTNESS FOR $C^{1+\varepsilon}$ SKEW PRODUCTS

4.1 Introduction

Let (X, d) be a compact metric space and $T : X \to X$ be a homeomorphism. If the topological dynamical system (X, T) has (topological) entropy zero, then Sarnak's conjecture [96, 97] predicts that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n\leq N} f(T^n x)\mu(n) = 0$$

for any continuous $f : X \to \mathbb{R}$ and every $x \in X$. When this holds, we say that the system (X, T) is *Möbius disjoint*.

Sarnak's conjecture has been proved for a variety of dynamical systems: see for instance [14, 15, 43, 27, 32, 44, 83, 80, 91, 17]. A common feature of all the results listed is that the underlying system is *regular*, in the sense that for every $x \in X$ the sequence $\frac{1}{N}\sum_{n\leq N} \delta_{T^n(x)}$ converges in the weak-* topology to some T-invariant Borel probability measure on X.

Let $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ denote the circle. In this chapter we will deal with the so-called *Anzai skew products* $(\mathbb{T}^2, T_{\alpha,\phi})$, where $\alpha \in \mathbb{R}$, $\phi : \mathbb{T} \to \mathbb{T}$ is a continuous map and the transformation is given by

$$T_{\alpha,\phi}(x, y) := (x + \alpha, y + \phi(x))$$

for all $(x, y) \in \mathbb{T}^2$. We often denote the system simply by $T_{\alpha,\phi}$.

Observe that $T_{\alpha,\phi}$ is *distal*, so it has zero topological entropy and therefore we expect it to be Möbius disjoint. In fact, these skew products are the basic building blocks in Furstenberg's classification of minimal distal flows [39], so understanding them is the first step towards establishing Sarnak's conjecture for this important general case. The main novel dynamical challenge that arises when one deals with skew products is that they provide some of the simplest examples of irregular dynamics. Indeed, Furstenberg [38] showed that $T_{\alpha,\phi}$ is not regular for some α and some analytic ϕ.

Lifting $\phi : \mathbb{T} \to \mathbb{T}$ to the real line, we can write $\phi(x) = cx + \tilde{\phi}(x)$ for all $x \in \mathbb{T}$, where $c \in \mathbb{Z}$ is the *topological degree* of ϕ and $\tilde{\phi} : \mathbb{T} \to \mathbb{R}$ is a continuous 1-

periodic function, unique up to shifts by $\mathbb{Z}$ (we fix an arbitrary choice). Kułaga-Przymus and Lemańczyk [71] have shown that if $\phi \in C^{1+\varepsilon}$ for some $\varepsilon > 0$ then $T_{\alpha,\phi}$ is Möbius disjoint for a topologically generic set of α. Furthermore, they proved Möbius disjointness of $T_{\alpha,\phi}$ when $\alpha \in \mathbb{Q}$, assuming only continuity of ϕ [71, Proposition 2.3.1], so from now on we assume $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. A further consequence of their work [71, Remark 2.5.7] (see also [104, Corollary 2.6]) is that if ϕ is assumed to be Lipschitz continuous, then Sarnak's conjecture holds for $T_{\alpha,\phi}$ whenever $c \neq 0$. Therefore, with the underlying assumption on ϕ in mind, we can deal only with topological degree zero from now on, and with an abuse of notation we identify ϕ with $\tilde{\phi}$.

The first Möbius disjointness result for all α was established by Liu and Sarnak [78], who proved it for ϕ analytic and satisfying the technical condition $\widehat{\phi}(m) \gg e^{-\tau|m|}$ for some $\tau > 0$. This was the first time Sarnak's conjecture was proved for a system that is not regular (since Furstenberg's example satisfies the condition). A refinement of this result was recently obtained by Wang [104], who removed the need for a lower bound on Fourier coefficients, obtaining Möbius disjointness of $T_{\alpha,\phi}$ for all analytic ϕ. Huang, Wang, and Ye [52] later improved this to cover all $\phi \in C^{\infty}$. Finally, using the work of Matomäki and Radziwiłł [82] on the behavior of μ in short intervals, Kanigowski, Lemańczyk, and Radziwiłł [64] established Möbius disjointness of $T_{\alpha,\phi}$ for all $\phi \in C^{2+\varepsilon}$ subject to the condition $\widehat{\phi}(0) = 0$, where $\varepsilon > 0$ is arbitrary.

Our main result is a simultaneous improvement of the works of Kułaga-Przymus-Lemańczyk [71], Huang-Wang-Ye [52], and Kanigowski-Lemańczyk-Radziwiłł [64]:

Theorem 5. *Let $\varepsilon > 0$. For any $\alpha \in \mathbb{R}$ and $\phi : \mathbb{T} \to \mathbb{T}$ of class $C^{1+\varepsilon}$, the skew product $T_{\alpha,\phi}$ is Möbius disjoint.*

The proof follows the ideas laid out by Kanigowski-Lemańczyk-Radziwiłł in [64], but instead of aiming for a polynomial rate of convergence for $T_{\alpha,\phi}^{r_n} \to \mathrm{Id}$ in the uniform norm (along some unbounded sequence $\{r_n\}_{n\geq 1}$), we establish a polynomial rate of convergence for $T_{\alpha,\phi}^{r_n} \to \mathrm{Id}$ in the $L^2(\nu)$ norm, for each $T_{\alpha,\phi}$-invariant Borel probability measure ν. The difficulties in dealing with every such ν are overcome because they all project to the Lebesgue measure in the first coordinate. We also remove the condition $\widehat{\phi}(0) = 0$ present in [64] by slightly modifying their choice of the sequence $\{r_n\}_{n\geq 1}$.

Another important ingredient is better control of some sums related to the Fourier coefficients of ϕ, where the Diophantine properties of α play an important role. The idea here is that not many q's at a given scale can make $\|q\alpha\|$ small (i.e. be denominators of good rational approximations of α). Furthermore, the q's at a given scale that give rise to rational approximations of similar quality must be somewhat well-spaced. We apply the Denjoy-Koksma inequality to appropriately chosen functions in order to extract that information (see Section 4.3).

The smoothness exponent $1+\varepsilon$ seems to be the limit of this argument. Indeed, we prove in Section 4.5 that if one only assumes that $\phi \in C^1$ then, at least along the sequence of best rational approximations of the irrational α, the rate of rigidity of $T_{\alpha,\phi}$ can be logarithmic even when $\widehat{\phi}(0) = 0$.

In Section 4.6 we show that our ideas can be used to extend some general rigidity results so far only known for functions of mean zero to the general case. A modification of Lemma 26 to obtain uniform polynomial rates of rigidity in the case $\phi \in C^{1+\varepsilon}$ is also discussed.

Finally, in Section 4.7 we use our argument to deduce new Möbius disjointness results for flows in $\mathbb{T}^2$ and Rokhlin extensions.

Notations

For a topological dynamical system (X, T), let $M(X, T)$ be the set of T-invariant Borel probability measures on X. Write $\|\cdot\|$ for the distance to the nearest integer (which we use as the metric in $\mathbb{T}$), $d(\cdot,\cdot)$ for the product metric in $\mathbb{T}\times\mathbb{T}$, corresponding to $\|\cdot\|$ in each coordinate, and $\|\cdot\|_{L^2(\nu)}$ for the usual L^2 norm with respect to a measure ν. We also abbreviate $e(x) := e^{2\pi i x}$ and use the asymptotic notation $f(x) \ll g(x)$ (respectively $f(x) \ll_p g(x)$) to mean that there exists $C > 0$ absolute (respectively depending only on the parameter p) such that $|f(x)| \leq C|g(x)|$ for all x in the relevant range. Furthermore, $f(x) \asymp g(x)$ means $f(x) \ll g(x) \ll f(x)$.

Acknowledgments

Thanks to Adam Kanigowski and Mariusz Lemańczyk for pointing out a nice simplification to my initial proofs of Lemmas 27 and 28, and for providing valuable comments and references. I am also grateful to the American Institute of Mathematics (AIM) for their 2018 workshop on "Sarnak's Conjecture",

which played a role in motivating the work in this chapter.

4.2 Reduction of disjointness to a rigidity result

As previously outlined, we can assume that $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and $\deg(\phi) = 0$, so ϕ can be realized as a function from $\mathbb{T}$ to $\mathbb{R}$ of class $C^{1+\varepsilon}$, which by an abuse of notation we still denote by ϕ. Observe that ϕ is in particular Lipschitz continuous, so we have pointwise convergence of its Fourier series, and the smoothness condition gives

$$\phi(x) = \sum_{q \in \mathbb{Z}} c_q e(qx) \quad \text{with} \quad c_q \ll_\phi \frac{1}{|q|^{1+\varepsilon}} \text{ for } q \neq 0. \tag{4.1}$$

The key to the proof of Theorem 5 is the result below, which is motivated by [64].

Lemma 26. *Let $0 < \varepsilon < \frac{1}{100}$ and $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. If $\phi : \mathbb{T} \to \mathbb{R}$ is of class $C^{1+\varepsilon}$, then there exists an unbounded sequence of positive integers $\{r_n\}_{n \geq 1}$ such that*

$$\int_{\mathbb{T} \times \mathbb{T}} d(T_{\alpha,\phi}^{r_n}(x,y),(x,y))^2 \, d\nu(x,y) \ll_\phi r_n^{-\varepsilon/100}$$

for any $\nu \in M(\mathbb{T}^2, T_{\alpha,\phi})$, where the implied constant does not depend on ν.

Assuming Lemma 26, we can easily prove Theorem 5.

Proof of Theorem 5. Let $\{r_n\}_{n \geq 1}$ be the sequence from Lemma 26. For any $\nu \in M(\mathbb{T}^2, T_{\alpha,\phi})$, continuous $f : \mathbb{T}^2 \to \mathbb{R}$, and $k \in \mathbb{Z}$, the triangle inequality and the $T_{\alpha,\phi}$-invariance of ν imply

$$\|f \circ T_{\alpha,\phi}^{kr_n} - f\|_{L^2(\nu)}^2 \leq |k| \sum_{j=1}^{|k|} \|f \circ T_{\alpha,\phi}^{jr_n} - f \circ T_{\alpha,\phi}^{(j-1)r_n}\|_{L^2(\nu)}^2 = k^2 \cdot \|f \circ T_{\alpha,\phi}^{r_n} - f\|_{L^2(\nu)}^2. \tag{4.2}$$

If f is also Lipschitz continuous, then using Lemma 26 we get

$$\|f \circ T_{\alpha,\phi}^{r_n} - f\|_{L^2(\nu)}^2 \ll_f \int_{\mathbb{T} \times \mathbb{T}} d(T_{\alpha,\phi}^{r_n}(x,y),(x,y))^2 \, d\nu(x,y) \ll_\phi r_n^{-\varepsilon/100}. \tag{4.3}$$

Therefore, (4.2) and (4.3) together give

$$\lim_{n \to \infty} \sum_{|k| \leq r_n^{\varepsilon/400}} \|f \circ T_{\alpha,\phi}^{kr_n} - f\|_{L^2(\nu)}^2 = 0$$

for every $\nu \in M(\mathbb{T}^2, T_{\alpha,\phi})$, which is precisely the *PR rigidity* condition of [64] (using the linearly dense family $\mathcal{F}$ of Lipschitz continuous functions) for the system $(\mathbb{T}^2, T_{\alpha,\phi})$, so [64, Theorem 1.1] implies Möbius disjointness for this skew product, and Theorem 5 is proved.

□

4.3 Continued fractions and some arithmetic estimates

Before proceeding to the proof of Lemma 26, we recall some properties of continued fractions. Let $\frac{p_n}{q_n}$, with $q_n > 0$ and $(p_n, q_n) = 1$, be the n-th convergent of the continued fraction expansion $[a_0; a_1, a_2, \dots]$ of the irrational α, so that $a_i \geq 1$ for $i \neq 0$. Then

(P1) $q_0 = 1$, $q_1 = a_1$ and $q_{n+1} = a_{n+1}q_n + q_{n-1}$ for $n \geq 1$;

(P2) $\frac{1}{q_{n+1}+q_n} < \|q_n\alpha\| < \frac{1}{q_{n+1}}$;

(P3) If $0 < q < q_{n+1}$, then $\|q_n\alpha\| \leq \|q\alpha\|$.

The main technical tool that allows us to quickly explore the Diophantine properties of α through its continued fraction is the following inequality.

Proposition 7 (Denjoy-Koksma inequality)**.** *Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. If $f : \mathbb{T} \to \mathbb{R}$ is of bounded variation, which we denote by $\mathrm{Var}(f)$, then for any $n \geq 0$ and $x \in \mathbb{T}$ we have*

$$\left| \sum_{j=0}^{q_n-1} f(x + j\alpha) - q_n \int_{\mathbb{T}} f(z)\, dz \right| \leq \mathrm{Var}(f).$$

Proof. See [50, page VI.3.1].

□

The next two lemmas encapsulate estimates related to continued fractions that will be necessary to prove Lemma 26.

Lemma 27. *For any $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and $k \geq 2$,*

$$\sum_{0<|q|<q_k} \frac{1}{\|q\alpha\|^2} \asymp q_k^2.$$

Proof. The lower bound comes from positivity and the single term $q = q_{k-1}$, by (P2). The upper bound follows from [1, Lemma 2.5] (see also [74, Lemma 1] for a partial result). We give a quick proof for completeness.

Assume $0 < q < q_k$, as the sum over negative q is the same. Consider $f : \mathbb{T} \to \mathbb{R}$ given by

$$f(z) = \begin{cases} (2q_k)^2, & \text{if } \|z\| \le \frac{1}{2q_k} \\ \frac{1}{\|z\|^2}, & \text{if } \|z\| > \frac{1}{2q_k} \end{cases}.$$

Observe that by (P2) and (P3), $\|q\alpha\| > \frac{1}{2q_k}$ for all $0 < q < q_k$, so by the Denjoy-Koksma inequality we conclude that

$$\begin{aligned} \sum_{0<q<q_k} \frac{1}{\|q\alpha\|^2} = \sum_{q=1}^{q_k-1} f(q\alpha) &\le |f(0)| + q_k \left| \int_{\mathbb{T}} f(z)\, dz \right| + \mathrm{Var}(f) \\ &= 4q_k^2 + q_k(8q_k - 4) + (8q_k^2 - 8) \ll q_k^2, \end{aligned}$$

as desired.

□

Lemma 28. *For any $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, $k \ge 1$, and $1 \le c \le q_k$,*

$$\sum_{q_k \le |q| < q_{k+1}} \frac{1}{q^2} \min\left\{ \frac{1}{\|q\alpha\|^2}, c^2 \right\} \ll \frac{c}{q_k}.$$

Proof. We can assume $q_k \le q < q_{k+1}$ since the sum over negative q is the same.

Consider $f : \mathbb{T} \to \mathbb{R}$ given by

$$f(z) = \begin{cases} c^2, & \text{if } \|z\| \le \frac{1}{c} \\ \frac{1}{\|z\|^2}, & \text{if } \|z\| > \frac{1}{c} \end{cases}.$$

Observe that $f(q\alpha) = \min\left\{ \frac{1}{\|q\alpha\|^2}, c^2 \right\}$, so (P1) gives

$$\begin{aligned} &\sum_{q_k \le q < q_{k+1}} \frac{1}{q^2} \min\left\{ \frac{1}{\|q\alpha\|^2}, c^2 \right\} \\ &= \sum_{j=1}^{a_{k+1}-1} \sum_{jq_k \le q < (j+1)q_k} \frac{f(q\alpha)}{q^2} + \sum_{a_{k+1}q_k \le q < q_{k+1}} \frac{f(q\alpha)}{q^2} \\ &\le \sum_{j=1}^{a_{k+1}-1} \frac{1}{(jq_k)^2} \sum_{r=0}^{q_k-1} f(jq_k\alpha + r\alpha) + \frac{4}{q_{k+1}^2} \sum_{r=0}^{q_{k-1}-1} f(a_{k+1}q_k\alpha + r\alpha). \end{aligned}$$

Using the Denjoy-Koksma inequality for the sums over r as in the proof of Lemma 27, since $\int_{\mathbb{T}} f(z)\,dz \asymp c$ and $\mathrm{Var}(f) \ll c^2$ by direct computation, we conclude that the remaining expression is

$$\ll \sum_{j=1}^{\infty} \frac{q_k c + c^2}{(jq_k)^2} + \frac{q_{k-1}c + c^2}{q_{k+1}^2} \ll \frac{c}{q_k},$$

so we are done.

□

4.4 Polynomial rate of rigidity in $C^{1+\varepsilon}$

At last, we are ready to prove our main lemma.

Proof of Lemma 26. Denoting

$$S_r(g)(x) := g(x) + g(x+\alpha) + \cdots + g(x+(r-1)\alpha),$$

we have

$$T_{\alpha,\phi}^{r_n}(x,y) = (x + r_n\alpha, y + S_{r_n}(\phi)(x)),$$

so that

$$d(T_{\alpha,\phi}^{r_n}(x,y),(x,y))^2 \asymp \|r_n\alpha\|^2 + \|S_{r_n}(\phi)(x)\|^2.$$

Therefore,

$$\begin{aligned} D_n &:= \int_{\mathbb{T}\times\mathbb{T}} d(T_{\alpha,\phi}^{r_n}(x,y),(x,y))^2\,d\nu(x,y) \\ &\asymp \|r_n\alpha\|^2 + \int_{\mathbb{T}\times\mathbb{T}} \|S_{r_n}(\phi)(x)\|^2\,d\nu(x,y). \end{aligned} \tag{4.4}$$

Consider the projection map $\pi(x,y) = x$. Observe that the integrand in (4.4) is independent of the second coordinate, so we can rewrite the integral as

$$\int_{\mathbb{T}\times\mathbb{T}} \|S_{r_n}(\phi)(\pi(x,y))\|^2\,d\nu(x,y) = \int_{\mathbb{T}} \|S_{r_n}(\phi)(x)\|^2\,d(\pi_*\nu)(x). \tag{4.5}$$

Since $\pi : (\mathbb{T}^2, T_{\alpha,\phi}) \to (\mathbb{T}, R_\alpha)$ is a map of topological dynamical systems (where in the image the transformation is $R_\alpha(x) := x + \alpha$) and ν is $T_{\alpha,\phi}$-invariant, the Borel probability measure $\pi_*\nu$ is R_α-invariant. But α is irrational, so $(\mathbb{T}, R_\alpha)$ is uniquely ergodic and we conclude that $\pi_*\nu$ is the Lebesgue measure on $\mathbb{T}$.

Using the Fourier expansion of ϕ we get $S_{r_n}(\phi)(x) = \sum_{q\in\mathbb{Z}} c_q S_{r_n}(e_q)(x)$, where $e_q(x) := e(qx)$. A computation shows that

$$S_{r_n}(e_q)(x) = e(qx)\frac{1-e(qr_n\alpha)}{1-e(q\alpha)}$$

for $q \neq 0$ and $S_{r_n}(e_0)(x) = r_n$, so we can plug this into (4.5) and conclude, using the triangle inequality and replacing $\|\cdot\|$ by absolute values, that the integral there is bounded by a constant multiple of

$$\|c_0 r_n\|^2 + \int_{\mathbb{T}} \left|\sum_{q\neq 0} c_q \frac{1-e(qr_n\alpha)}{1-e(q\alpha)} e(qx)\right|^2 dx = \|c_0 r_n\|^2 + \sum_{q\neq 0} |c_q|^2 \left|\frac{1-e(qr_n\alpha)}{1-e(q\alpha)}\right|^2, \tag{4.6}$$

where we have used Parseval for $S_{r_n}(\phi) - c_0 r_n \in L^2(\mathbb{T})$.

Now, we make a preliminary choice of the sequence $\{r_n\}_{n\geq 1}$ by letting $r_n := \ell_n q_n$, where q_n is the denominator of the n-th convergent of the continued fraction expansion of α, as before, and $\ell_n \in \mathbb{Z}$ is chosen so that

$$0 < \ell_n \leq q_n^{\delta} \quad \text{and} \quad \|\ell_n q_n c_0\| < q_n^{-\delta},$$

where $\delta := \varepsilon/10$. Such ℓ_n exist for all n, by the Dirichlet approximation theorem.

Let $\lambda := \varepsilon/100$. In what follows it is worth keeping in mind the rough hierarchy "$\lambda \lll \delta \lll \varepsilon$" behind our choice of parameters. We wish to show that $D_n \ll_\phi r_n^{-\lambda}$. With our choice of $\{r_n\}_{n\geq 1}$ the first term in the RHS of (4.4) contributes at most

$$\ell_n^2 \cdot \|q_n\alpha\|^2 < q_n^{2\delta} q_{n+1}^{-2} < q_n^{2\delta-2} < q_n^{-\lambda(1+\delta)} \leq r_n^{-\lambda}, \tag{4.7}$$

so it is harmless. The first term of (4.6) contributes

$$\|c_0 \ell_n q_n\|^2 < q_n^{-2\delta} < q_n^{-\lambda(1+\delta)} \leq r_n^{-\lambda}, \tag{4.8}$$

and it is also harmless.

We break the remaining terms into two parts, corresponding to $0 < |q| < q_n$ and $|q| \geq q_n$. Observe that $|1-e(q\alpha)| \asymp \|q\alpha\|$ and $|1-e(qr_n\alpha)| \leq 2$.

Furthermore, $|S_{r_n}(e_q)(x)| \le r_n$ by a trivial bound, so

$$\begin{aligned}\sum_{|q|\ge q_n} |c_q|^2 \left|\frac{1-e(qr_n\alpha)}{1-e(q\alpha)}\right|^2 &\ll_\phi \sum_{|q|\ge q_n} \frac{1}{|q|^{2+2\varepsilon}} \min\left\{\frac{1}{\|q\alpha\|^2}, r_n^2\right\} \\ &< q_n^{-2\varepsilon}\ell_n^2 \sum_{k=n}^{\infty} \sum_{q_k\le|q|<q_{k+1}} \frac{1}{q^2}\min\left\{\frac{1}{\|q\alpha\|^2}, q_n^2\right\} \qquad (4.9)\\ &\ll q_n^{-2\varepsilon+2\delta}\sum_{k=n}^{\infty}\frac{q_n}{q_k} \ll q_n^{-2\varepsilon+2\delta} < q_n^{-\lambda(1+\delta)} \le r_n^{-\lambda},\end{aligned}$$

where we have used (4.1), Lemma 28 (for $c = q_n \le q_k$) and the fact that $q_{k+2} > 2q_k$ by (P1), so the q_k grow exponentially.

It remains to deal with $0 < |q| < q_n$. In this case, we use $|1-e(q\alpha)| \asymp \|q\alpha\|$ and $|1-e(qr_n\alpha)|^2 \asymp \|q\ell_n q_n\alpha\|^2 \le q^2\ell_n^2\cdot\|q_n\alpha\|^2 < q^2 q_n^{2\delta} q_{n+1}^{-2}$, so those terms contribute

$$\sum_{0<|q|<q_n} |c_q|^2 \left|\frac{1-e(qr_n\alpha)}{1-e(q\alpha)}\right|^2 \ll_\phi \frac{q_n^{2\delta}}{q_{n+1}^2}\sum_{0<|q|<q_n}\frac{1}{|q|^{2\varepsilon}}\frac{1}{\|q\alpha\|^2}. \qquad (4.10)$$

To deal with the sum over q we consider two cases.

Case 1: *There is a subsequence $\{q_{b_n}\}_{n\ge1}$ of $\{q_n\}_{n\ge1}$ such that $q_{b_n+1} \ge q_{b_n}^2$ for all $n \ge 1$.*

In this case we take the subsequence $\{r_{b_n}\}_{n\ge1}$ instead of the original sequence $\{r_n\}_{n\ge1}$. Observe that (4.7), (4.8) and (4.9) still hold along any subsequence. In (4.10) we can use the given condition and Lemma 27 to get the upper bound

$$\frac{q_{b_n}^{2\delta}}{q_{b_n}^4}\sum_{0<|q|<q_{b_n}}\frac{1}{\|q\alpha\|^2} \ll q_{b_n}^{2\delta-2} < q_{b_n}^{-\lambda(1+\delta)} \le r_{b_n}^{-\lambda},$$

and this finishes the proof.

Case 2: *For all sufficiently large n, we have $q_{n+1} < q_n^2$.*

In this case we stick with the original sequence $\{r_n\}_{n\ge1}$ and observe that for any $0 < k < n$ we can rewrite the sum in the RHS of (4.10) as

$$\begin{aligned}\left[\sum_{0<|q|<q_k} + \sum_{q_k\le|q|<q_n}\right]\frac{1}{|q|^{2\varepsilon}}\frac{1}{\|q\alpha\|^2} &< \sum_{0<|q|<q_k}\frac{1}{\|q\alpha\|^2} + q_k^{-2\varepsilon}\sum_{q_k\le|q|<q_n}\frac{1}{\|q\alpha\|^2} \\ &\ll q_k^2 + q_k^{-2\varepsilon}q_n^2,\end{aligned} \qquad (4.11)$$

where once again we have used Lemma 27.

Take $0 < k < n$ such that $q_k \in \left[q_n^{1/4}, q_n^{1/2}\right]$, which exists for all n sufficiently large since we can find such terms in any interval of the form $[a, a^2]$ for a sufficiently large, because of the given condition. Then the corresponding upper bound when we plug (4.11) into (4.10) is

$$\frac{q_n^{2\delta}}{q_{n+1}^2}\left(q_n + q_n^{2-\varepsilon/2}\right) \ll q_n^{2\delta-\varepsilon/2} < q_n^{-\lambda(1+\delta)} \leq r_n^{-\lambda},$$

which establishes the result of Lemma 26.

□

4.5 Counterexample to polynomial rate of rigidity in C^1

Lemma 26 raises the question of how low one can push the smoothness of ϕ and still have a polynomial rate of rigidity for $T_{\alpha,\phi}$. We show that, at least along the sequence $\{q_n\}_{n\geq 1}$ of denominators of best rational approximations for an irrational α, there is $\phi \in C^1$ with $\widehat{\phi}(0) = 0$ such that

$$\int_{\mathbb{T}\times\mathbb{T}} d(T_{\alpha,\phi}^{q_n}(x,y),(x,y))^2\, d\nu(x,y) \gg_\delta q_n^{-\delta}$$

for every $\delta > 0$, unlike what happens for $\phi \in C^{1+\varepsilon}$ with $\widehat{\phi}(0) = 0$ (observe that in that case $\ell_n = 1$ works in Lemma 26).

Indeed, let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and choose $\phi : \mathbb{T} \to \mathbb{R}$ given by

$$\phi(x) := \frac{1}{C}\sum_{k\geq 2} \frac{e(q_k x) + e(-q_k x)}{q_k (\log q_k)^2},$$

where $C > 0$ will be chosen to be sufficiently large. Since $q_k \geq 2^{(k-1)/2}$ by (P1), $\sum_{k\geq 2}(\log q_k)^{-2}$ is absolutely convergent and therefore $\phi \in C^1$. Take $C > 0$ large enough so that $\mathrm{Var}(\phi) < 1/2$. By the Denjoy-Koksma inequality, we have

$$\left|\sum_{j=0}^{q_n-1} \phi(x + j\alpha) - q_n \int_{\mathbb{T}} \phi(z)\, dz \right| \leq \mathrm{Var}(\phi) < \frac{1}{2}$$

for every $x \in \mathbb{T}$. Since $\widehat{\phi}(0) = 0$ we conclude that $|S_{q_n}(\phi)(x)| < 1/2$, so that $\|S_{q_n}(\phi)(x)\| = |S_{q_n}(\phi)(x)|$ for all $x \in \mathbb{T}$. Therefore, the beginning of the proof of Lemma 26 shows that

$$\int_{\mathbb{T}\times\mathbb{T}} d(T_{\alpha,\phi}^{q_n}(x,y),(x,y))^2\, d\nu(x,y) \asymp \|q_n\alpha\|^2 + \sum_{k\geq 2} \frac{1}{q_k^2(\log q_k)^4}\left|\frac{1 - e(q_n q_k \alpha)}{1 - e(q_k\alpha)}\right|^2. \tag{4.12}$$

If $q_n \cdot \|q_n\alpha\| < 1/2$ then $\|q_n^2\alpha\| = q_n \cdot \|q_n\alpha\|$, so

$$\frac{1}{q_n^2(\log q_n)^4}\left|\frac{1-e(q_n^2\alpha)}{1-e(q_n\alpha)}\right|^2 \asymp \frac{1}{q_n^2(\log q_n)^4}\frac{\|q_n^2\alpha\|^2}{\|q_n\alpha\|^2} = \frac{1}{(\log q_n)^4}.$$

If instead $q_n \cdot \|q_n\alpha\| > 1/2$ then from $\|q_n\alpha\| < 1/q_{n+1}$ (by (P2)) we get $q_{n+1} < 2q_n$. Since $q_{n+2} > 2q_n$ we have $q_n \cdot \|q_{n+1}\alpha\| < q_n/q_{n+2} < 1/2$, so $\|q_nq_{n+1}\alpha\| = q_n \cdot \|q_{n+1}\alpha\|$, and in conclusion

$$\frac{1}{q_{n+1}^2(\log q_{n+1})^4}\left|\frac{1-e(q_nq_{n+1}\alpha)}{1-e(q_{n+1}\alpha)}\right|^2 \asymp \frac{1}{q_n^2(\log q_n)^4}\frac{\|q_nq_{n+1}\alpha\|^2}{\|q_{n+1}\alpha\|^2} = \frac{1}{(\log q_n)^4}.$$

Taking respectively the terms corresponding to $k = n$ and $k = n+1$ in (4.12) and using positivity of the other terms we conclude that the whole expression is $\gg (\log q_n)^{-4}$, so there is no polynomial rate of convergence to zero along any subsequence of $\{q_n\}_{n\geq 1}$. In fact, [64, Lemma 3.2] shows that a decay of the form $\exp(-(\log\log q_n)^{1+\delta})$ for any $\delta > 0$ would be enough for Möbius disjointness, but that too is false by our counterexample.

4.6 Extension of general rigidity results to ϕ of non-zero mean

Recall that a topological dynamical system (X, T) is called *rigid* if for each $\nu \in M(X, T)$ there exists a sequence $\{r_n\}_{n\geq 1}$ of positive integers such that $g \circ T^{r_n} \to g$ in $L^2(\nu)$ for all $g \in L^2(\nu)$.

By theorems of Herman [50, page XIII.4.8] and Gabriel, Lemańczyk, and Liardet [40, Théorème 1.1], if $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and ϕ is absolutely continuous, has topological degree zero, and satisfies $\widehat{\phi}(0) = 0$, then the skew product $T_{\alpha,\phi}$ is rigid, and in fact they show that $T_{\alpha,\phi}^{q_n} \to \mathrm{Id}$ uniformly by obtaining

$$\lim_{n\to\infty} \sup_{x\in\mathbb{T}} |S_{q_n}(\phi)(x)| = 0.$$

Lemańczyk and Mauduit [74, Theorem 1] (see also [1, Corollary 2.8]) generalized[1] these theorems to show rigidity (though not uniformly) of $T_{\alpha,\phi}$ for all $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and $\phi \in L^2(\mathbb{T})$ (of topological degree zero) satisfying $\widehat{\phi}(0) = 0$ and $\widehat{\phi}(m) = o(1/|m|)$.

The techniques of this chapter may be employed to extend both results to cover the case $\widehat{\phi}(0) \neq 0$. Furthermore, in the case $\phi \in C^{1+\varepsilon}$ we can modify Lemma 26 to recover a uniform polynomial rate of rigidity instead of just the result in $L^2(\nu)$ presented previously.

[1] If $\phi : \mathbb{T} \to \mathbb{R}$ is absolutely continuous then $\phi' \in L^1(\mathbb{T})$, so the Riemann-Lebesgue lemma gives $\widehat{\phi}(m) = o(1/|m|)$.

Uniform rigidity for ϕ absolutely continuous

Proposition 8. *If $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and ϕ is absolutely continuous of topological degree zero, then the skew product $T_{\alpha,\phi}$ is uniformly rigid.*

Proof. We can simply use the original result for the zero mean case to conclude that there is $\lambda(n) \to \infty$ as $n \to \infty$ such that

$$\sup_{x\in\mathbb{T}} |S_{q_n}(\phi - \widehat{\phi}(0))(x)| \leq \lambda(n)^{-1},$$

so choose $\ell_n \in \mathbb{Z}$ with

$$0 < \ell_n \leq \lambda(n)^{1/2} \quad \text{and} \quad \|\ell_n q_n \widehat{\phi}(0)\| < \lambda(n)^{-1/2}$$

using Dirichlet's approximation theorem to get

$$\begin{aligned}\|S_{\ell_n q_n}(\phi)(x)\| &\leq \|\ell_n q_n \widehat{\phi}(0)\| + |S_{\ell_n q_n}(\phi - \widehat{\phi}(0))(x)| \\ &< \lambda(n)^{-1/2} + \sum_{k=0}^{\ell_n - 1} |S_{q_n}(\phi - \widehat{\phi}(0))(x + kq_n\alpha)| \ll \lambda(n)^{-1/2} \to 0\end{aligned}$$

uniformly in $x \in \mathbb{T}$. Therefore, $T_{\alpha,\phi}$ is uniformly rigid along the sequence $\{\ell_n q_n\}_{n\geq 1}$.

□

Rigidity for ϕ with tamely decaying Fourier coefficients

Proposition 9. *If $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and $\phi \in L^1(\mathbb{T})$ (of topological degree zero) satisfies $\widehat{\phi}(m) = o(1/|m|)$, then the skew product $T_{\alpha,\phi}$ is rigid.*

Proof (Sketch). We have the Fourier expansion[2] (in $L^2(\mathbb{T})$)

$$\phi(x) = \sum_{q\in\mathbb{Z}} c_q e(qx) \quad \text{with} \quad |c_q| \leq \frac{1}{|q| \cdot \psi(|q|)} \text{ for } q \neq 0,$$

where $\psi : (0,\infty) \to (0,\infty)$ satisfies $\psi(z) \to \infty$ as $z \to \infty$ and for technical reasons we can of course also assume that it is non-decreasing and does not grow too fast, say $\psi(z) \ll_\phi z^{1/100}$.

With the conditions above, we can show that there is a sequence of positive integers $\{r_n\}_{n\geq 1}$ such that

$$\int_{\mathbb{T}\times\mathbb{T}} d(T^{r_n}_{\alpha,\phi}(x,y),(x,y))^2 \, d\nu(x,y) \ll_\phi \psi(q_n^{1/4})^{-1/100} \to 0 \quad \text{as} \quad n \to \infty$$

[2]It follows from the conditions that $\phi \in L^2(\mathbb{T})$.

for any $\nu \in M(\mathbb{T}^2, T_{\alpha,\phi})$.

The proof is a simple modification of the proof of Lemma 26, substituting q_n^ε with $\psi(q_n)$, so for instance $\ell_n \in \mathbb{Z}$ is chosen so that

$$0 < \ell_n \leq \psi(q_n)^{1/10} \quad \text{and} \quad \|\ell_n q_n c_0\| < \psi(q_n)^{-1/10}.$$

Observe that we do not have multiplicativity of ψ, which is why the bound is not of the form $\psi(r_n)^{-1/100}$, but it is enough to prove that $T_{\alpha,\phi}$ is rigid[3] (the latter bound could be obtained if we imposed extra attainable conditions on ψ).

□

Uniform polynomial rate of rigidity for $\phi \in C^{1+\varepsilon}$

Finally, we point out that the conclusion of Lemma 26 can actually be strengthened to a uniform polynomial rate of rigidity:

Proposition 10. *Let $0 < \varepsilon < \frac{1}{100}$ and $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. If $\phi : \mathbb{T} \to \mathbb{R}$ is of class $C^{1+\varepsilon}$, then there exists an unbounded sequence of positive integers $\{r_n\}_{n\geq 1}$ such that*

$$\sup_{(x,y)\in\mathbb{T}\times\mathbb{T}} d(T_{\alpha,\phi}^{r_n}(x,y),(x,y)) \ll_{\phi,\varepsilon} r_n^{-\varepsilon/200}.$$

Proof (Sketch). We start by substantially modifying the results of Lemma 27 and Lemma 28. Namely one can show, using the same techniques as in the corresponding results of Section 4.3 but this time for the functions

$$f_1(z) = \begin{cases} 2q_k, & \text{if } \|z\| \leq \frac{1}{2q_k} \\ \frac{1}{\|z\|}, & \text{if } \|z\| > \frac{1}{2q_k} \end{cases} \quad \text{and} \quad f_2(z) = \begin{cases} c, & \text{if } \|z\| \leq \frac{1}{c} \\ \frac{1}{\|z\|}, & \text{if } \|z\| > \frac{1}{c}, \end{cases}$$

respectively, that if $\varepsilon > 0$, $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, $k \geq 1$, and $1 \leq c \leq q_k$ then

$$\sum_{0<|q|<q_k} \frac{1}{\|q\alpha\|} \ll q_k \log(q_k + 1) \tag{4.13}$$

and

$$\sum_{q_k\leq|q|<q_{k+1}} \frac{1}{|q|^{1+\varepsilon}} \min\left\{\frac{1}{\|q\alpha\|}, c\right\} \ll_\varepsilon \frac{\log(c+1)}{q_k^\varepsilon}. \tag{4.14}$$

[3]The bound implies rigidity for $T_{\alpha,\phi}$ since the Lipschitz continuous functions on $\mathbb{T}^2$ are dense in $L^2(\nu)$, for any $\nu \in M(\mathbb{T}^2, T_{\alpha,\phi})$. This follows by the Stone-Weierstrass theorem and the fact that $C(\mathbb{T}^2)$ is dense in $L^2(\nu)$, since ν is a Radon measure — see for instance [36, Proposition 7.9].

Then expanding $S_{r_n}(\phi)(x)$ into a Fourier series and trivially bounding it we get

$$d(T^{r_n}_{\alpha,\phi}(x,y),(x,y)) \leq \|r_n\alpha\| + \|c_0 r_n\| + \sum_{q\neq 0} |c_q| \left| \frac{1-e(qr_n\alpha)}{1-e(q\alpha)} \right|,$$

so we can proceed as in the proof of Lemma 26 with the expression above corresponding to (4.6) and the bounds of (4.13) and (4.14) corresponding to Lemma 27 and Lemma 28, respectively, to get the desired uniform polynomial decay.

□

Remark 10. *The proof actually shows that for every $\phi : \mathbb{T} \to \mathbb{R}$ of class $C^{1+\varepsilon}$ and of mean zero,*

$$\sup_{x\in\mathbb{T}} |S_{q_n}(\phi)(x)| \ll_{\phi,\varepsilon} q_n^{-\varepsilon/200}, \tag{4.15}$$

since in that case we can take $\ell_n = 1$ throughout the argument.

Remark 11. *Even though Proposition 10 gives a stronger result than Lemma 26, we chose to emphasize the latter in our presentation because the L^2 methods employed there seem more suitable for generalization (and the proof is slightly more complicated). For instance, an approach to Proposition 9 using L^∞ methods would already be frustrated by the presence of the extra logarithmic factor in* (4.13)*, if the decay of the Fourier coefficients is sufficiently slow. Therefore, the use of L^2 methods seems to allow us to go a bit further.*

4.7 Smooth flows on $\mathbb{T}^2$ and Rokhlin extensions

We can adapt the result of this chapter, following [64], to give Möbius disjointness for new cases of smooth flows on the torus and Rokhlin extensions.

Smooth flows on $\mathbb{T}^2$

For $\alpha \in \mathbb{R}\setminus\mathbb{Q}$, let $f : \mathbb{T} \to \mathbb{R}$ be a strictly positive continuous function. Let

$$\mathbb{T}^f := \{(x,s) \in \mathbb{T}\times\mathbb{R} : 0 \leq s \leq f(x)\}/\sim,$$

where $\sim$ denotes the equivalence relation $(x, s+f(x)) \sim (R_\alpha(x), s)$ in $\mathbb{T}\times\mathbb{R}$ and $R_\alpha : \mathbb{T} \to \mathbb{T}$ is the irrational rotation by α. We can define a *special flow* $T^f = \{T^f_t\}_{t\in\mathbb{R}}$ over R_α with roof function f, which acts on $\mathbb{T}^f$ by

$$T^f_t(x,s) := (x, s+t)$$

for all $(x,s) \in \mathbb{T}^f$. More explicitly, if we extend a previous definition to

$$S_N(f)(x) := \begin{cases} \sum_{0 \le j < N} f(R_\alpha^j(x)), & \text{if } N > 0 \\ 0, & \text{if } N = 0 \\ \sum_{N \le j < 0} f(R_\alpha^j(x)), & \text{if } N < 0 \end{cases}$$

then

$$T_t^f(x,s) = (R_\alpha^N(x), s + t - S_N(f)(x))$$

for all $(x,s) \in \mathbb{T}^f$, where $N = N(x,s,t) \in \mathbb{Z}$ is such that

$$S_N(f)(x) \le s + t < S_{N+1}(f)(x),$$

which exists and is unique as f is continuous and strictly positive.

Every sufficiently smooth area-preserving flow on $\mathbb{T}^2$ with no fixed points or closed orbits can be represented by such a special flow for f with corresponding smoothness properties (see [25]).

We have the following consequence of our work:

Corollary 4. *Let $\varepsilon > 0$ and $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. If $f \in C^{1+\varepsilon}$ then all the maps of the special flow $T^f = \{T_t^f\}_{t \in \mathbb{R}}$ over the irrational rotation R_α are Möbius disjoint.*

Proof. There is a natural quotient metric D making $\mathbb{T}^f$ a compact metric space (see [24, Appendix 9.1]), and it satisfies

$$D(T_t^f(x,s),(x,s)) \le |t| \quad \text{for all} \quad t \in \mathbb{R} \quad \text{and} \quad (x,s) \in \mathbb{T}^f. \tag{4.16}$$

Denote $\beta := \widehat{f}(0) > 0$ and let q_n be the denominators of convergents of the continued fraction of α, as before. For a fixed $t \in \mathbb{R}$, let $v_n \in \mathbb{Z}$ be such that

$$0 < v_n \le q_n^{1+\gamma} \quad \text{and} \quad \left\| v_n \frac{t}{q_n \beta} \right\| < q_n^{-1-\gamma},$$

where $\gamma > 0$ will be chosen later to be sufficiently small (v_n exists by Dirichlet's approximation theorem). Then there is $j_n \in \mathbb{Z}$ such that

$$|t v_n - j_n q_n \beta| < \frac{q_n \beta}{q_n^{1+\gamma}} \ll_f q_n^{-\gamma} \tag{4.17}$$

and

$$|j_n| < \left| \frac{v_n t}{q_n \beta} \right| + 1 \ll_{f,t} q_n^{\gamma}. \tag{4.18}$$

For every $(x,s) \in \mathbb{T}^f$ we have

$$\begin{aligned}
&D(T^f_{tv_n}(x,s),(x,s)) = D(T^f_{tv_n - S_{j_n q_n}(f)(x)} \circ T^f_{S_{j_n q_n}(f)(x)}(x,s),(x,s)) \\
&\leq D(T^f_{tv_n - S_{j_n q_n}(f)(x)}(R_\alpha^{j_n q_n}(x),s),(R_\alpha^{j_n q_n}(x),s)) + D((R_\alpha^{j_n q_n}(x),s),(x,s)) \\
&\leq |S_{j_n q_n}(f-\beta)(x)| + |tv_n - j_n q_n \beta| + \|j_n q_n \alpha\| \\
&\ll_{f,t} q_n^\gamma \cdot \sup_{z \in \mathbb{T}} |S_{q_n}(f-\beta)(z)| + q_n^{-\gamma} + q_n^\gamma \cdot \|q_n \alpha\|,
\end{aligned}$$

where we have used (4.16), (4.17), and (4.18). Choosing $\gamma := \varepsilon/1000$ and using (4.15) (we could also take the L^2 norm and use the proof of Lemma 26) we get the bound $\ll_{f,t,\varepsilon} q_n^{-\varepsilon/1000} < v_n^{-\varepsilon/2000}$, which gives a polynomial rate of rigidity for $T_t^f : \mathbb{T}^f \to \mathbb{T}^f$ along the (unbounded, unless $t=0$) sequence $\{v_n\}_{n\geq 1}$, and this implies Möbius disjointness for T_t^f.

□

Rokhlin extensions

As before, let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and let $R_\alpha : \mathbb{T} \to \mathbb{T}$ denote the irrational rotation by α. Given a continuous function $f : \mathbb{T} \to \mathbb{R}$, a compact metric space (Y,ρ) and a continuous flow $L = \{L_t\}_{t\in\mathbb{R}}$ acting on Y, we can define a *Rokhlin extension* $E_{f,L}$ of R_α, acting on $\mathbb{T} \times Y$ by

$$E_{f,L}(x,y) := (R_\alpha(x), L_{f(x)}(y))$$

for all $(x,y) \in \mathbb{T} \times Y$ (observe that if $Y = \mathbb{T}$ and L is the linear flow we recover the Anzai skew product $T_{\alpha,f}$). We have the following disjointness result in this case:

Corollary 5. *Let $\varepsilon > 0$. If $f \in C^{1+\varepsilon}$ has mean zero and the flow $L = \{L_t\}_{t\in\mathbb{R}}$ is uniformly Lipschitz continuous in t, then $E_{f,L}$ is Möbius disjoint.*

Proof. If D denotes the product metric in $\mathbb{T} \times Y$ then

$$D(E^{q_n}_{f,L}(x,y),(x,y)) = \|q_n\alpha\| + \rho(L_{S_{q_n}(f)(x)}(y),y) \ll_L \|q_n\alpha\| + |S_{q_n}(f)(x)|,$$

where the implied constant does not depend on (x,y). Using (4.15) we get a polynomial rate of rigidity for $E_{f,L}$ along $\{q_n\}_{n\geq 1}$ (we could also take the L^2 norm and use the proof of Lemma 26), so the corollary follows.

□

Appendix A

ZERO-DENSITY FOR TWISTS OF PRIMITIVE FORMS

The purpose of this appendix is to obtain a zero-density estimate for character twists of a fixed form f that holds in the generality required for our application in Chapter 2 and is efficient in the Q-aspect. We use the notation of (2.32) for the number of zeros in a rectangle.

Proposition 11 (Zero-density for twists in degree two)**.** *Let $f \in S_k(\Gamma_0(N), \xi)$ be a primitive holomorphic modular form of arbitrary weight k, level N, and nebentypus ξ. Then for any $Q \geq 2$, $T \geq 2$, $\varepsilon > 0$, and $\frac{1}{2} + \varepsilon \leq \alpha \leq 1$, there exists some A depending only on ε such that*

$$\sum_{\substack{q \leq Q \\ (q,N)=1}} \sideset{}{^*}\sum_{\chi \pmod q} N_{f\otimes\chi}(\alpha, T) \ll_{f,\varepsilon} \left((QT)^{4+\varepsilon} + \left(Q^2T\right)^{c(\alpha)} \right)^{1-\alpha} \log^A(QT),$$

where

$$c(\alpha) := \min\left\{ \frac{3}{2-\alpha}, \frac{3}{3\alpha - 1} \right\}$$

and $\sum^*$ denotes summation over primitive characters.

The proof uses standard methods for large values of Dirichlet polynomials, and we closely follow the argument of Iwaniec-Kowalski [59] for the case of Dirichlet L-functions, with the necessary technical modifications to deal with our case of degree two (mostly complications coming from larger conductor). The advantage of this approach is that we do not have to deal with moments of L_f, where only limited information is available (as we do not have access to the fourth moment). Proposition 11 is not particularly efficient in the T-aspect, however this is irrelevant as T is fixed in our desired application. For T small in terms of Q (in particular for T fixed), Proposition 11 improves (in all ranges of α) results of Zhang [107] valid for f of level $N = 1$.

Proof of Proposition 11. Let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be given by

$$g(x) := \kappa \int_x^\infty \exp\left(-y - \frac{1}{y}\right) \frac{dy}{y},$$

where $\kappa := (2K_0(2))^{-1}$ is a normalizing constant so that $g(0) = 1$. Then one may check that the Mellin transform

$$\hat{g}(z) := \int_0^\infty g(x)x^{z-1}\,dx$$

is odd and has a pole at $z = 0$, and that $z\hat{g}(z)$ is analytic. In addition, we have the bounds

$$0 < g(x) < \kappa e^{-x}, \tag{A.1}$$

$$0 < 1 - g(x) < \kappa e^{-1/x}, \tag{A.2}$$

and

$$\hat{g}(z) \ll |z|^{|\Re(z)|-1}e^{-\frac{\pi}{2}|\Im(z)|} \tag{A.3}$$

uniformly for $z \in \mathbb{C}$. We refer to [59, p. 257-258] for details, where one may combine Euler's reflection formula $\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}$ with Stirling's formula to obtain (A.3) from [59, (10.55)].

Our preliminary goal is to obtain a convenient approximate functional equation for $L_{f\otimes\chi}(s)$, where from now on we assume that $s = \sigma + it$ with $\frac{1}{2} + \varepsilon \leq \sigma \leq 1$ and χ is a primitive character modulo q, where $(q, N) = 1$. We evaluate the sum

$$B_f(s,\chi) := \sum_{n=1}^\infty \lambda_f(n)\chi(n)n^{-s}g\left(\frac{n}{X}\right), \tag{A.4}$$

where $X > 0$ will be chosen later. By contour integration,

$$\begin{aligned} B_f(s,\chi) &= \frac{1}{2\pi i}\int_{(1)} L_f(s+u,\chi)X^u\hat{g}(u)\,du \\ &= L_f(s,\chi) + \frac{1}{2\pi i}\int_{(-1)} L_f(s+u,\chi)X^u\hat{g}(u)\,du. \end{aligned} \tag{A.5}$$

Since $(q, N) = 1$ and $\chi \pmod q$ is primitive, $L_f(z,\chi) = L_{f\otimes\chi}(z)$ and $f \otimes \chi$ is a primitive form in $S_k(\Gamma_0(Nq^2), \xi\chi^2)$, so we have the functional equation

$$L_f(z,\chi) = \epsilon_{f\otimes\chi}(Nq^2)^{\frac{1}{2}-z}\gamma_k(z)L_{\overline{f}}(1-z,\overline{\chi}),$$

where $|\epsilon_{f\otimes\chi}| = 1$ and

$$\gamma_k(z) := (2\pi)^{2z-1}\frac{\Gamma\left(1 - z + \frac{k-1}{2}\right)}{\Gamma\left(z + \frac{k-1}{2}\right)}.$$

Using this functional equation, the integral over $\Re(u) = -1$ in (A.5) is equal to $-\epsilon_{f\otimes\chi}B_f^*(s,\chi)$, where

$$B_f^*(s,\chi) := \frac{1}{2\pi i}\int_{(1)} (Nq^2)^{\frac{1}{2}-s+u}\gamma_k(s-u)L_{\overline{f}}(1-s+u,\overline{\chi})X^{-u}\hat{g}(u)\,du.$$

Expanding $L_{\overline{f}}$ into a Dirichlet series we get

$$B_f^*(s,\chi) = \sum_{m=1}^{\infty} \lambda_{\overline{f}}(m)\overline{\chi}(m)m^{s-1}h(Xm), \tag{A.6}$$

with

$$h(y) := \frac{1}{2\pi i}\int_{(1)} (Nq^2)^{\frac{1}{2}-s+u}\gamma_k(s-u)y^{-u}\hat{g}(u)\,du. \tag{A.7}$$

In conclusion, collecting the expressions above we obtain

$$L_{f\otimes\chi}(s) = L_f(s,\chi) = B_f(s,\chi) + \epsilon_{f\otimes\chi}B_f^*(s,\chi), \tag{A.8}$$

where $B_f(s,\chi)$ and $B_f^*(s,\chi)$ are given by (A.4) and (A.6), respectively, and $X > 0$ is arbitrary.

By Euler's reflection formula and Stirling's formula, we have

$$\gamma_k(z) \ll_k |z|^{1-2\Re(z)} \tag{A.9}$$

uniformly in the half-plane $\Re(z) \leq \frac{1}{2}$. Using the bounds (A.3) and (A.9) and moving the integral (A.7) sufficiently to the right, say to the line

$$\Re(u) = \max\left(1, \frac{1}{3}\left(\frac{y}{Nq^2|s|^2}\right)^{\frac{1}{3}}\right),$$

one obtains the rough uniform bound

$$h(y) \ll_k \frac{Nq^2|s|^2}{y}\exp\left(-\frac{1}{3}\left(\frac{y}{Nq^2|s|^2}\right)^{\frac{1}{3}}\right).$$

Therefore, $h(mX)$ is quite small as long as m is a bit larger than $Nq^2|s|^2X^{-1}$. More precisely, by (A.6) and Deligne's bound we have

$$B_f^*(s,\chi) = \sum_{m\leq Y} \lambda_{\overline{f}}(m)\overline{\chi}(m)m^{s-1}h(mX) + O_f\left(\frac{1}{XY}\right) \tag{A.10}$$

provided

$$XY \geq Nq^2|s|^2\log^4(Nq^2|s|^2). \tag{A.11}$$

Now we write (A.7) as

$$h(y) := \frac{1}{2\pi i}\int_{(\eta)} (Nq^2)^{\frac{1}{2}-s+2\sigma-u}\gamma_k(s-2\sigma+u)y^{u-2\sigma}\hat{g}(2\sigma-u)\,du$$

by changing u into $2\sigma - u$ and then moving the line of integration to $\Re(u) = \eta$, where $1 < \eta < 2\sigma$. Inserting this into (A.10) we get

$$B_f^*(s,\chi) = \frac{1}{2\pi i}\int_{(\eta)} \left(\sum_{m\le Y} \lambda_{\overline{f}}(m)\overline{\chi}(m)m^{-\overline{s}+u-1}\right) W(u)\,du + O_f\left(\frac{1}{XY}\right), \tag{A.12}$$

where

$$W(u) := (Nq^2)^{\frac{1}{2}-s+2\sigma-u}\gamma_k(s-2\sigma+u)X^{u-2\sigma}\hat{g}(2\sigma-u).$$

Choose $\eta = 1+\varepsilon$, which satisfies $1 < \eta < 2\sigma$ and $-\sigma+\eta \le \frac{1}{2}$. By (A.3) and (A.9), for $u = \eta + iv$ we have

$$\begin{aligned} W(u) &\ll \left(Nq^2\left(|s|+|v|\right)^2\right)^{\frac{1}{2}+\sigma-\eta} X^{\eta-2\sigma}(2\sigma-\eta)^{-1}e^{-\frac{\pi}{2}|v|} \\ &\ll (2\sigma-\eta)^{-1}\left(\frac{Nq^2|s|^2}{X^2}\right)^{\frac{1}{2}+\sigma-\eta} X^{1-\eta}e^{-|v|}. \end{aligned}$$

Assuming that

$$X^2 \ge Nq^2|s|^2, \tag{A.13}$$

since $\sigma \ge \frac{1}{2}+\varepsilon$ we get

$$W(u) \ll \varepsilon^{-1}X^{-\varepsilon}e^{-|v|}.$$

Therefore, (A.12) becomes

$$B_f^*(s,\chi) \ll_f \varepsilon^{-1}X^{-\varepsilon}\int_{-\infty}^{\infty}\left|\sum_{m\le Y}\lambda_f(m)\chi(m)m^{-s+\varepsilon+iv}\right| e^{-|v|}\,dv + \frac{1}{XY}. \tag{A.14}$$

Denote $D := 2\sqrt{N}QT$ and $\mathcal{L} := 2\log D$. As a reminder, we have $s = \sigma + it$ with $\frac{1}{2}+\varepsilon \le \sigma \le 1$, $\chi \pmod q$ primitive with $(q,N) = 1$, and from now on we also assume $|t| \le T$ and $q \le Q$. Choose

$$X = D\mathcal{L} \qquad \text{and} \qquad Y = D\mathcal{L}^3,$$

so that conditions (A.11) and (A.13) are satisfied. Then by (A.1) the sum in (A.4) can be reduced to $n \le Y$ up to an error of $O(D^{-2})$, so that combining it with (A.8) and (A.14) we get

$$\begin{aligned} L_f(s,\chi) = &\sum_{n\le Y}\lambda_f(n)\chi(n)n^{-s}g\left(\frac{n}{X}\right) \\ &+ O_{f,\varepsilon}\left(X^{-\varepsilon}\int_{-\infty}^{\infty}\left|\sum_{n\le Y}\lambda_f(n)\chi(n)n^{-s+\varepsilon+iv}\right| e^{-|v|}\,dv + D^{-2}\right). \end{aligned} \tag{A.15}$$

Let $1 \leq M \leq D$ and

$$M_f(s,\chi) := \sum_{m \leq M} b_f(m)\chi(m)m^{-s},$$

where the coefficients b_f are inverses of λ_f under Dirichlet convolution, i.e. are given by

$$\sum_{n=1}^{\infty} b_f(n)n^{-s} := \prod_{p \text{ prime}} \left(1 - \lambda_f(p)p^{-s} + \xi(p)p^{-2s}\right) \qquad \text{for } \Re(s) > 1,$$

so that Deligne's bound implies $|b_f(n)| \leq d(n)$. From (A.15) we obtain

$$\begin{aligned} L_f(s,\chi)M_f(s,\chi) = &\sum_{n \leq MY} a_f(n)\chi(n)n^{-s} \\ &+ O_{f,\varepsilon}\left(\mathcal{L}\int_{-\infty}^{\infty} \left|\sum_{n \leq MY} a_f(n,v)\chi(n)n^{-s}\right| e^{-|v|}\,dv + D^{-2}M^{\frac{1}{2}}\right), \end{aligned} \tag{A.16}$$

where

$$a_f(n) := \sum_{\substack{cm=n \\ c \leq Y, m \leq M}} \lambda_f(c) g\left(\frac{c}{X}\right) b_f(m) \ll d_4(n)$$

by (A.1) and similarly

$$a_f(n,v) := \sum_{\substack{cm=n \\ c \leq Y, m \leq M}} \lambda_f(c) \left(\frac{c}{Y}\right)^{\varepsilon+iv} b_f(m) \ll d_4(n).$$

For $n \leq M$, by (A.2) we have the more precise estimates

$$a_f(n) = \sum_{cm=n} \lambda_f(c)b_f(m)\left(1 + O\left(e^{-X/c}\right)\right) = \mathbb{1}_{n=1} + O\left(d_4(n)D^{-2}\right)$$

and

$$a_f(n,v) \ll \left(\frac{n}{Y}\right)^{\varepsilon} \sum_{cm=n} |\lambda_f(c)|\,|b_f(m)| \leq \left(\frac{n}{Y}\right)^{\varepsilon} d_4(n).$$

As a consequence,

$$\left|\sum_{n \leq M} a_f(n,v)\chi(n)n^{-s}\right| \leq Y^{-\varepsilon} \sum_{n \leq M} d_4(n)n^{-\frac{1}{2}} \ll Y^{-\varepsilon}M^{\frac{1}{2}}\log^3(2M).$$

We want this to be $O_\varepsilon(\mathcal{L}^{-2})$, which holds assuming for instance

$$M \leq D^{\varepsilon}. \tag{A.17}$$

In that case, using the bounds above in (A.16) gives

$$L_f(s,\chi)M_f(s,\chi) = 1 + \sum_{M<n\le MY} a_f(n)\chi(n)n^{-s} + O_{f,\varepsilon}\left(\mathcal{L}\int_{-\infty}^{\infty}\left|\sum_{M<n\le MY} a_f(n,v)\chi(n)n^{-s}\right| e^{-|v|}\,dv + \mathcal{L}^{-1}\right). \tag{A.18}$$

To unify the treatment of the sum and integral, we consider the measure

$$d\mu := \frac{1}{3}e^{-|v|}dv + \frac{1}{3}\delta(v),$$

where dv denotes the Lebesgue measure on $\mathbb{R}$, $\delta(v)$ is the point measure at $v = 0$, and the factor $\frac{1}{3}$ is a normalization that makes $\int_{-\infty}^{\infty} d\mu = 1$. Then (A.18) can be written as

$$L_f(s,\chi)M_f(s,\chi) - 1 \ll_{f,\varepsilon} \mathcal{L}\int_{-\infty}^{\infty}\left|\sum_{M<n\le MY} a_f(n,v)\chi(n)n^{-s}\right| d\mu(v) + \mathcal{L}^{-1} \tag{A.19}$$

after redefining $a_f(n,0) := a_f(n)$. For convenience, we also redefine $a_f(n,v) := 0$ for $n \le M$ or $n > MY$. From now on the only properties about the coefficients we will use are that they do not depend on s or χ and satisfy $a_f(n,v) \ll d_4(n)$.

Now, assume that ρ is a zero of $L_{f\otimes\chi}(s) = L_f(s,\chi)$ for some primitive $\chi \pmod q$ with $(q,N) = 1$, $q \le Q$, $\frac{1}{2} + \varepsilon \le \alpha \le \Re(\rho) \le 1$, and $|\Im(\rho)| \le T$. If D is sufficiently large in terms of f and ε (which we can assume, otherwise Proposition 11 follows trivially), then (A.19) implies

$$\int_{-\infty}^{\infty}\left|\sum_{M<n\le MY} a_f(n,v)\chi(n)n^{-\rho}\right| d\mu(v) \gg_{f,\varepsilon} \mathcal{L}^{-1}.$$

We break the summation into dyadic segments $J < n \le 2J$ for $J := 2^\ell M$, $0 \le \ell \le L := \lfloor \log Y/\log 2\rfloor \ll \mathcal{L}$. Denote

$$D_\ell(s,\chi) := \int_{-\infty}^{\infty}\left|\sum_{J<n\le 2J} a_f(n,v)\chi(n)n^{-s}\right| d\mu(v).$$

Then for each such zero ρ being counted there exists some ℓ such that

$$D_\ell(\rho,\chi) \gg_{f,\varepsilon} \mathcal{L}^{-2}. \tag{A.20}$$

If $\mathcal{S}_\ell$ denotes the set of relevant pairs (ρ, χ) satisfying (A.20) and $R_\ell := |\mathcal{S}_\ell|$, then the total number R of zeros being counted in Proposition 11 satisfies

$$R \leq \sum_{\ell=0}^{L} R_\ell \ll \mathcal{L} \max_{0\leq \ell \leq L} R_\ell.$$

Raising $D_\ell(s,\chi)$ to a suitable power $2r$, for $r \geq 2$ (depending on J), we get

$$\begin{aligned} D_\ell(s,\chi)^{2r} &\leq \int_{-\infty}^{\infty} \left| \sum_{J<n\leq 2J} a_f(n,v)\chi(n)n^{-s} \right|^{2r} d\mu(v) \\ &=: \int_{-\infty}^{\infty} \left| \sum_{P<n\leq 2^r P} c_f(n,v)\chi(n)n^{-s} \right|^{2} d\mu(v), \end{aligned}$$

where $P := J^r$ falls in the segment

$$Z \leq P \leq (MY)^2 + Z^{\frac{3}{2}} \tag{A.21}$$

for Z that will be chosen later satisfying

$$MY \leq Z \ll D^{100}. \tag{A.22}$$

Observe that an integer $r \geq 2$ such that (A.21) holds exists. From now on we choose

$$M = D^{\frac{\varepsilon}{4}},$$

so that r is bounded in terms of ε, by (A.22). Observe that condition (A.17) is automatically satisfied.

By (A.20), we conclude that

$$R_\ell \ll_{f,\varepsilon} \mathcal{L}^{4r} \int_{-\infty}^{\infty} \sum_{(\rho,\chi)\in\mathcal{S}_\ell} \left| \sum_{P<n\leq 2^r P} c_f(n,v)\chi(n)n^{-\rho} \right|^{2} d\mu(v). \tag{A.23}$$

The coefficients satisfy $c_f(n,v) \ll_r d_{4r}(n)$, as $a_f(n,v) \ll d_4(n)$, so

$$\sum_{P<n\leq 2^r P} |c_f(n,v)|^2 n^{-2\alpha} \leq P^{1-2\alpha}\mathcal{L}^B \tag{A.24}$$

for some B depending only on r (and therefore ε). We can now apply results about large values of Dirichlet polynomials to the integrand of (A.23), after separating the zeros ρ for each given χ into $O(\mathcal{L})$ families of 1-spaced points. Let $H := Q^2T$.

Suppose that $\frac{1}{2} + \varepsilon \leq \alpha \leq \frac{3}{4}$. By (A.24) and the large sieve inequality [59, Theorem 9.13], we have

$$\begin{aligned} R_\ell &\ll_{f,\varepsilon} (P+H)P^{1-2\alpha}\mathcal{L}^C \\ &\ll \left((MY)^{4(1-\alpha)} + Z^{3(1-\alpha)} + HZ^{1-2\alpha}\right)\mathcal{L}^C \end{aligned}$$

for some C depending only on ε, where we have used (A.21). If $H \leq (MY)^{3-2\alpha}$, choose $Z = MY$, which trivially satisfies (A.22), so

$$R_\ell \ll_{f,\varepsilon} (MY)^{4(1-\alpha)}\mathcal{L}^C \leq D^{(4+\varepsilon)(1-\alpha)}\mathcal{L}^{C+6}$$

and we are done. If instead $H \geq (MY)^{3-2\alpha}$, then we can make the optimal choice $Z = H^{\frac{1}{2-\alpha}}$ and (A.22) is satisfied, so we get

$$R_\ell \ll_{f,\varepsilon} \left((MY)^{4(1-\alpha)} + H^{\frac{3(1-\alpha)}{2-\alpha}}\right)\mathcal{L}^C \leq \left(D^{(4+\varepsilon)(1-\alpha)} + H^{\frac{3(1-\alpha)}{2-\alpha}}\right)\mathcal{L}^{C+6}$$

as desired.

Finally, suppose that $\frac{3}{4} \leq \alpha \leq 1$. By the Halász-Montgomery-Huxley method [59, Theorem 9.15], we have

$$R_\ell \ll_{f,\varepsilon} \left(P + R_\ell^{\frac{2}{3}} H^{\frac{1}{3}} P^{\frac{1}{3}}\right) P^{1-2\alpha}\mathcal{L}^C$$

for some C depending only on ε, which implies

$$\begin{aligned} R_\ell &\ll_{f,\varepsilon} \left(P^{2-2\alpha} + HP^{4-6\alpha}\right)\mathcal{L}^{3C} \\ &\ll \left((MY)^{4(1-\alpha)} + Z^{3(1-\alpha)} + HZ^{4-6\alpha}\right)\mathcal{L}^{3C}, \end{aligned}$$

where again we have used (A.21). If $H \leq (MY)^{2\alpha}$, we choose $Z = MY$, which trivially satisfies (A.22), and get

$$R_\ell \ll_{f,\varepsilon} (MY)^{4(1-\alpha)}\mathcal{L}^{3C} \leq D^{(4+\varepsilon)(1-\alpha)}\mathcal{L}^{3C+3},$$

so we are done. If instead $H \geq (MY)^{2\alpha}$, then we make the optimal choice $Z = H^{\frac{1}{3\alpha-1}}$, which in this case satisfies (A.22). Therefore,

$$R_\ell \ll_{f,\varepsilon} \left((MY)^{4(1-\alpha)} + H^{\frac{3(1-\alpha)}{3\alpha-1}}\right)\mathcal{L}^{3C} \leq \left(D^{(4+\varepsilon)(1-\alpha)} + H^{\frac{3(1-\alpha)}{3\alpha-1}}\right)\mathcal{L}^{3C+3}$$

as desired.

□

BIBLIOGRAPHY

[1] J. Aaronson, M. Lemańczyk, C. Mauduit, and H. Nakada. Koksma's inequality and group extensions of Kronecker transformations. In *Algorithms, fractals, and dynamics (Okayama/Kyoto, 1992)*, pages 27–50. Plenum, New York, 1995.

[2] N. Andersen and J. Thorner. Zeros of GL_2 L-functions on the critical line. *Forum Math.*, 33(2):477–491, 2021.

[3] A. O. L. Atkin and W. C. W. Li. Twists of newforms and pseudo-eigenvalues of W-operators. *Invent. Math.*, 48(3):221–243, 1978.

[4] P. J. Bauer. Zeros of Dirichlet L-series on the critical line. *Acta Arith.*, 93(1):37–52, 2000.

[5] V. Blomer, X. Li, and S. D. Miller. A spectral reciprocity formula and non-vanishing for L-functions on $\mathrm{GL}(4) \times \mathrm{GL}(2)$. *J. Number Theory*, 205:1–43, 2019.

[6] A. R. Booker. Poles of Artin L-functions and the strong Artin conjecture. *Ann. of Math. (2)*, 158(3):1089–1098, 2003.

[7] A. R. Booker. Simple zeros of degree 2 L-functions. *J. Eur. Math. Soc. (JEMS)*, 18(4):813–823, 2016.

[8] A. R. Booker, P. J. Cho, and M. Kim. Simple zeros of automorphic L-functions. *Compos. Math.*, 155(6):1224–1243, 2019.

[9] A. R. Booker and M. Krishnamurthy. A strengthening of the GL(2) converse theorem. *Compos. Math.*, 147(3):669–715, 2011.

[10] A. R. Booker and M. Krishnamurthy. Further refinements of the GL(2) converse theorem. *Bull. Lond. Math. Soc.*, 45(5):987–1003, 2013.

[11] A. R. Booker and M. Krishnamurthy. Weil's converse theorem with poles. *Int. Math. Res. Not. IMRN*, (19):5328–5339, 2014.

[12] A. R. Booker, M. B. Milinovich, and N. Ng. Quantitative estimates for simple zeros of L-functions. *Mathematika*, 65(2):375–399, 2019.

[13] A. R. Booker, M. B. Milinovich, and N. Ng. Subconvexity for modular form L-functions in the t aspect. *Adv. Math.*, 341:299–335, 2019.

[14] J. Bourgain. Möbius-Walsh correlation bounds and an estimate of Mauduit and Rivat. *J. Anal. Math.*, 119:147–163, 2013.

[15] J. Bourgain. On the correlation of the Moebius function with rank-one systems. *J. Anal. Math.*, 120:105–130, 2013.

[16] J. Bourgain, Z. Rudnick, and P. Sarnak. Spatial statistics for lattice points on the sphere I: Individual results. *Bull. Iranian Math. Soc.*, 43(4):361–386, 2017.

[17] J. Bourgain, P. Sarnak, and T. Ziegler. Disjointness of Moebius from horocycle flows. In *From Fourier analysis and number theory to Radon transforms and geometry*. Volume 28, Dev. Math. Pages 67–83. Springer, New York, 2013.

[18] H. M. Bui, B. Conrey, and M. P. Young. More than 41% of the zeros of the zeta function are on the critical line. *Acta Arith.*, 150(1):35–64, 2011.

[19] J. B. Conrey. More than two fifths of the zeros of the Riemann zeta function are on the critical line. *J. Reine Angew. Math.*, 399:1–26, 1989.

[20] J. B. Conrey, D. W. Farmer, J. P. Keating, M. O. Rubinstein, and N. C. Snaith. Integral moments of L-functions. *Proc. London Math. Soc. (3)*, 91(1):33–104, 2005.

[21] J. B. Conrey and A. Ghosh. Simple zeros of the Ramanujan τ-Dirichlet series. *Invent. Math.*, 94(2):403–419, 1988.

[22] J. B. Conrey and H. Iwaniec. The cubic moment of central values of automorphic L-functions. *Ann. of Math. (2)*, 151(3):1175–1216, 2000.

[23] J. B. Conrey, H. Iwaniec, and K. Soundararajan. Critical zeros of Dirichlet L-functions. *J. Reine Angew. Math.*, 681:175–198, 2013.

[24] J.-P. Conze and M. Lemańczyk. Centralizer and liftable centralizer of special flows over rotations. *Nonlinearity*, 31(8):3939–3972, 2018.

[25] I. Cornfeld, S. Fomin, and Y. Sinaĭ. *Ergodic theory*, volume 245 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, 1982, pages x+486. Translated from the Russian by A. B. Sosinskiĭ.

[26] M. Czarnecki. On the curvature of circles and curves in $\mathbb{H}^n$. *Demonstratio Math.*, 34(1):181–186, 2001.

[27] H. Davenport. On some infinite series involving arithmetical functions (II). *Q. J. Math., Oxf. Ser.*, 8(1):313–320, 1937. ISSN: 0033-5606.

[28] W. Duke. Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.*, 92(1):73–90, 1988.

[29] W. Duke, Ö. Imamoḡlu, and Á. Tóth. Geometric invariants for real quadratic fields. *Ann. of Math. (2)*, 184(3):949–990, 2016.

[30] W. Duke and R. Schulze-Pillot. Representation of integers by positive ternary quadratic forms and equidistribution of lattice points on ellipsoids. *Invent. Math.*, 99(1):49–57, 1990.

[31] M. Einsiedler, E. Lindenstrauss, P. Michel, and A. Venkatesh. The distribution of closed geodesics on the modular surface, and Duke's theorem. *Enseign. Math. (2)*, 58(3-4):249–313, 2012.

[32] E. H. El Abdalaoui, M. Lemańczyk, and T. de la Rue. On spectral disjointness of powers for rank-one transformations and Möbius orthogonality. *J. Funct. Anal.*, 266(1):284–317, 2014.

[33] A. Eskin, M. Mirzakhani, and K. Rafi. Counting closed geodesics in strata. *Invent. Math.*, 215(2):535–607, 2019.

[34] D. W. Farmer. Mean value of Dirichlet series associated with holomorphic cusp forms. *J. Number Theory*, 49(2):209–245, 1994.

[35] S. Feng. Zeros of the Riemann zeta function on the critical line. *J. Number Theory*, 132(4):511–542, 2012.

[36] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics. John Wiley & Sons, New York, second edition, 1999, pages xvi+386.

[37] N. Frantzikinakis and B. Host. The logarithmic Sarnak conjecture for ergodic weights. *Ann. of Math. (2)*, 187(3):869–931, 2018.

[38] H. Furstenberg. Strict ergodicity and transformation of the torus. *Amer. J. Math.*, 83:573–601, 1961.

[39] H. Furstenberg. The structure of distal flows. *Amer. J. Math.*, 85:477–515, 1963.

[40] P. Gabriel, M. Lemańczyk, and P. Liardet. Ensemble d'invariants pour les produits croisés de Anzai. *Mém. Soc. Math. France (N.S.)*, 47, 1991.

[41] E. P. Golubeva and O. M. Fomenko. Asymptotic distribution of lattice points on the three-dimensional sphere. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 160(Anal. Teor. Chisel i Teor. Funktsiĭ. 8):54–71, 297, 1987.

[42] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Elsevier/Academic Press, Amsterdam, eighth edition, 2015, pages xlvi+1133. Edited by D. Zwillinger and V. Moll.

[43] B. Green. On (not) computing the Möbius function using bounded depth circuits. *Combin. Probab. Comput.*, 21(6):942–951, 2012.

[44] B. Green and T. Tao. The Möbius function is strongly orthogonal to nilsequences. *Ann. of Math. (2)*, 175(2):541–566, 2012.

[45] J. L. Hafner. Explicit estimates in the arithmetic theory of cusp forms and Poincaré series. *Math. Ann.*, 264(1):9–20, 1983.

[46] J. L. Hafner. Zeros on the critical line for Dirichlet series attached to certain cusp forms. *Math. Ann.*, 264(1):21–37, 1983.

[47] J. L. Hafner. Zeros on the critical line for Maass wave form L-functions. *J. Reine Angew. Math.*, 377:127–158, 1987.

[48] G. Harcos and P. Michel. The subconvexity problem for Rankin-Selberg L-functions and equidistribution of Heegner points. II. *Invent. Math.*, 163(3):581–655, 2006.

[49] D. R. Heath-Brown. Simple zeros of the Riemann zeta function on the critical line. *Bull. London Math. Soc.*, 11(1):17–18, 1979.

[50] M. Herman. Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. *Inst. Hautes Études Sci. Publ. Math.*, 49:5–233, 1979.

[51] J. Hoffstein and P. Lockhart. Coefficients of Maass forms and the Siegel zero. *Ann. of Math. (2)*, 140(1):161–181, 1994. With an appendix by Dorian Goldfeld, Hoffstein, and Daniel Lieman.

[52] W. Huang, Z. Wang, and X. Ye. Measure complexity and Möbius disjointness. *Adv. Math.*, 347:827–858, 2019.

[53] P. Humphries. Equidistribution in shrinking sets and L^4-norm bounds for automorphic forms. *Math. Ann.*, 371(3-4):1497–1543, 2018.

[54] P. Humphries and M. Radziwiłł. Optimal small scale equidistribution of lattice points on the sphere, Heegner points, and closed geodesics. *Comm. Pure Appl. Math.*, to appear (arXiv:1910.01360), 2019.

[55] M. N. Huxley. Large values of Dirichlet polynomials. *Acta Arith.*, 24:329–346, 1973.

[56] A. Ivić. *The Riemann zeta-function*. Dover Publications, Inc., Mineola, NY, 2003, pages xxii+517. Reprint of the 1985 original.

[57] H. Iwaniec. Fourier coefficients of modular forms of half-integral weight. *Invent. Math.*, 87(2):385–401, 1987.

[58] H. Iwaniec. *Spectral methods of automorphic forms*, volume 53 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI; Revista Matemática Iberoamericana, Madrid, second edition, 2002, pages xii+220.

[59] H. Iwaniec and E. Kowalski. *Analytic number theory*, volume 53 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2004, pages xii+615.

[60] H. Iwaniec and P. Sarnak. The non-vanishing of central values of automorphic L-functions and Landau-Siegel zeros. *Israel J. Math.*, 120(part A):155–177, 2000.

[61] H. Jacquet and J. A. Shalika. A non-vanishing theorem for zeta functions of GL_n. *Invent. Math.*, 38(1):1–16, 1976/77.

[62] J. Jung and N. T. Sardari. Intersecting geodesics on the modular surface. *arXiv:2101.08768*, 2021.

[63] M. Jutila. *Lectures on a method in the theory of exponential sums*, volume 80 of *Tata Institute of Fundamental Research Lectures on Mathematics and Physics*. Published for the Tata Institute of Fundamental Research, Bombay; by Springer-Verlag, Berlin, 1987, pages viii+134.

[64] A. Kanigowski, M. Lemańczyk, and M. Radziwiłł. Rigidity in dynamics and Möbius disjointness. *Fund. Math.*, 255(3):309–336, 2021.

[65] S. Katok and P. Sarnak. Heegner points, cycles and Maass forms. *Israel J. Math.*, 84(1-2):193–227, 1993.

[66] N. M. Katz and P. Sarnak. Zeroes of zeta functions and symmetry. *Bull. Amer. Math. Soc. (N.S.)*, 36(1):1–26, 1999.

[67] J. P. Keating and N. C. Snaith. Random matrix theory and $\zeta(1/2+it)$. *Comm. Math. Phys.*, 214(1):57–89, 2000.

[68] O. Klurman, A. P. Mangerel, and J. Teräväinen. Multiplicative functions in short arithmetic progressions. *arXiv:1909.12280*, 2019.

[69] E. Kowalski and P. Michel. Zeros of families of automorphic L-functions close to 1. *Pacific J. Math.*, 207(2):411–431, 2002.

[70] E. Kowalski, P. Michel, and J. VanderKam. Mollification of the fourth moment of automorphic L-functions and arithmetic applications. *Invent. Math.*, 142(1):95–151, 2000.

[71] J. Kułaga-Przymus and M. Lemańczyk. The Möbius function and continuous extensions of rotations. *Monatsh. Math.*, 178(4):553–582, 2015.

[72] S. Lalley. Statistical regularities of self-intersection counts for geodesics on negatively curved surfaces. *Duke Math. J.*, 163(6):1191–1261, 2014.

[73] A. Laurinčikas and K. Matsumoto. The joint universality of twisted automorphic L-functions. *J. Math. Soc. Japan*, 56(3):923–939, 2004.

[74] M. Lemańczyk and C. Mauduit. Ergodicity of a class of cocycles over irrational rotations. *J. London Math. Soc. (2)*, 49(1):124–132, 1994.

[75] R. J. Lemke Oliver and J. Thorner. Effective log-free zero density estimates for automorphic L-functions and the Sato-Tate conjecture. *Int. Math. Res. Not. IMRN*, 2019(22):6988–7036, 2019.

[76] N. Levinson. More than one third of zeros of Riemann's zeta-function are on $\sigma = 1/2$. *Advances in Math.*, 13:383–436, 1974.

[77] Y. V. Linnik. *Ergodic properties of algebraic fields*. Translated from the Russian by M. S. Keane. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 45. Springer-Verlag New York Inc., New York, 1968, pages ix+192.

[78] J. Liu and P. Sarnak. The Möbius function and distal flows. *Duke Math. J.*, 164(7):1353–1399, 2015.

[79] W. Luo and P. Sarnak. Quantum variance for Hecke eigenforms. *Ann. Sci. École Norm. Sup. (4)*, 37(5):769–799, 2004.

[80] B. Martin, C. Mauduit, and J. Rivat. Théoréme des nombres premiers pour les fonctions digitales. *Acta Arith.*, 165(1):11–45, 2014.

[81] C. Matheus. Some quantitative versions of Ratner's mixing estimates. *Bull. Braz. Math. Soc. (N.S.)*, 44(3):469–488, 2013.

[82] K. Matomäki and M. Radziwiłł. Multiplicative functions in short intervals. *Ann. of Math. (2)*, 183(3):1015–1056, 2016.

[83] C. Mauduit and J. Rivat. Sur un problème de Gelfond: la somme des chiffres des nombres premiers. *Ann. of Math. (2)*, 171(3):1591–1646, 2010.

[84] M. B. Milinovich and N. Ng. Simple zeros of modular L-functions. *Proc. Lond. Math. Soc. (3)*, 109(6):1465–1506, 2014.

[85] M. Mirzakhani. Growth of the number of simple closed geodesics on hyperbolic surfaces. *Ann. of Math. (2)*, 168(1):97–125, 2008.

[86] M. Mirzakhani. Simple geodesics and Weil-Petersson volumes of moduli spaces of bordered Riemann surfaces. *Invent. Math.*, 167(1):179–222, 2007.

[87] H. L. Montgomery. Mean and large values of Dirichlet polynomials. *Invent. Math.*, 8:334–345, 1969.

[88] H. L. Montgomery. The pair correlation of zeros of the zeta function. In *Analytic number theory (Proc. Sympos. Pure Math., Vol. XXIV, St. Louis Univ., St. Louis, Mo., 1972)*, pages 181–193, 1973.

[89] H. L. Montgomery. Zeros of L-functions. *Invent. Math.*, 8:346–354, 1969.

[90] A. M. Odlyzko. On the distribution of spacings between zeros of the zeta function. *Math. Comp.*, 48(177):273–308, 1987.

[91] R. Peckner. Möbius disjointness for homogeneous dynamics. *Duke Math. J.*, 167(14):2745–2792, 2018.

[92] K. Pratt, N. Robles, A. Zaharescu, and D. Zeindler. More than five-twelfths of the zeros of ζ are on the critical line. *Res. Math. Sci.*, 7(2):Paper No. 2, 74, 2020.

[93] M. Ratner. The rate of mixing for geodesic and horocycle flows. *Ergodic Theory Dynam. Systems*, 7(2):267–288, 1987.

[94] Z. Rudnick and P. Sarnak. The n-level correlations of zeros of the zeta function. *C. R. Acad. Sci. Paris Sér. I Math.*, 319(10):1027–1032, 1994.

[95] Z. Rudnick and P. Sarnak. Zeros of principal L-functions and random matrix theory. *Duke Math. J.*, 81(2):269–322, 1996.

[96] P. Sarnak. Mobius randomness and dynamics. *Not. S. Afr. Math. Soc.*, 43(2):89–97, 2012.

[97] P. Sarnak. Möbius randomness and dynamics. `https://publications.ias.edu/sarnak/paper/546`, 2011.

[98] P. Sarnak. Reciprocal geodesics. In *Analytic number theory*. Volume 7, Clay Math. Proc. Pages 217–237. Amer. Math. Soc., Providence, RI, 2007.

[99] A. Selberg. On the zeros of Riemann's zeta-function. *Skr. Norske Vid.-Akad. Oslo I*, 1942(10), 1942.

[100] B. F. Skubenko. The asymptotic distribution of integers on a hyperboloid of one sheet and ergodic theorems. *Izv. Akad. Nauk SSSR Ser. Mat.*, 26:721–752, 1962.

[101] K. Soundararajan. Nonvanishing of quadratic Dirichlet L-functions at $s = \frac{1}{2}$. *Ann. of Math. (2)*, 152(2):447–488, 2000.

[102] T. Tao. The Erdös discrepancy problem. *Discrete Anal.*, 2016.

[103] T. Tao. The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi*, 4(e8), 2016.

[104] Z. Wang. Möbius disjointness for analytic skew products. *Invent. Math.*, 209(1):175–196, 2017.

[105] M. P. Young. A short proof of Levinson's theorem. *Arch. Math. (Basel)*, 95(6):539–548, 2010.

[106] M. P. Young. Weyl-type hybrid subconvexity bounds for twisted L-functions and Heegner points on shrinking sets. *J. Eur. Math. Soc. (JEMS)*, 19(5):1545–1576, 2017.

[107] D. Y. Zhang. Zero-density estimates for automorphic L-functions. *Acta Math. Sin. (Engl. Ser.)*, 25(6):945–960, 2009.