Invent. math. 139, 1–39 (2000)
Digital Object Identifier (DOI) 10.1007/s002229900017

Inventiones mathematicae

A problem of Linnik for elliptic curves and mean-value estimates for automorphic representations⋆,⋆⋆

W. Duke, **E. Kowalski**⋆⋆⋆

Department of Mathematics, Rutgers University, New-Brunswick, NJ 08903, USA
(e-mail: duke@math.rutgers.edu)

Oblatum 27-XII-1996 & 4-V-1999 / Published online: 18 October 1999

1 Introduction

When Linnik introduced the classical large-sieve in 1941 [Lin], he was motivated by the following problem: given a non-trivial primitive character χ modulo q, how large (compared to q) can be the first n such that $\chi(n) \neq 1$?

From the Riemann Hypothesis one can deduce (see [Mon] chapter 13 for instance)

$$n \ll (\log q)^2$$

and the (weaker) conjecture $n \ll q^{\varepsilon}$ for all $\varepsilon > 0$ is known as Vinogradov's conjecture.

Linnik's technique makes it possible to prove that the number of exceptions to these conjectures is extremely small. For example, let $N(Q, \alpha)$ be the number of primitive characters χ of modulus $q \leqslant Q$ such that

$$\chi(n) = 1$$

for all $n \leqslant (\log Q)^{\alpha}$, $(n, q) = 1$, $\alpha > 1$ being given; then from the large-sieve inequality for Dirichlet characters, we can derive

$$N(Q, \alpha) \ll Q^{2/\alpha+\varepsilon} \tag{1}$$

for all $\varepsilon > 0$, whereas there are about Q^2 primitive characters of modulus at most Q [1].

⋆ With an appendix by D. Ramakrishnan.

⋆⋆ Research supported in part by NSF Grant No. DMS-9507797.

⋆⋆⋆ *Current address*: Department of Mathematics, Fine Hall, Princeton, NJ 08544-1000, USA (e-mail: ekowalsk@math.princeton.edu)

[1] What Linnik actually did was, for χ a real character, to assume $\chi(n) = 1$ in the larger range $n \leqslant Q^{\varepsilon}$ and prove that there are at most $O(\log\log Q)$ possible χ of level less than Q. He was also using his own additive form of the large-sieve.

Moreover, because the exponent is less than 1 for $\alpha > 2$ and there are about Q real characters of modulus less than Q, our statement also proves that there are very few exceptions for real characters, which corresponds to the problem of the least quadratic non-residue. In particular, for Q tending to infinity, the probability that two real characters take the same values for all primes $p \leqslant (\log Q)^{\alpha}$ tends to zero.

In recent years, it has been widely perceived that elliptic curves over **Q** are a natural analogue of real Dirichlet characters. In this context, the corresponding problem would be, given two elliptic curves E and F of conductor less than Q, how large (always compared to Q) can n be if E and F have the same number of points modulo p for all primes less than n and yet are not isogenous?

This problem was considered by Serre, for instance, in [Ser]. Assuming the Riemann-Hypothesis for Artin L-functions, he showed that in this case too $n \ll (\log Q)^2$ follows.

In this paper, we are able to prove some analogues of (1).

Theorem 1 *Let $M(Q, \alpha)$ be the maximal number of isogeny classes of semi-stable elliptic curves over* **Q** *with conductor less than or equal to Q which for every prime $p \leqslant (\log Q)^{\alpha}$ have a fixed number of points modulo p.*

Then we have for any $\varepsilon > 0$

$$M(Q, \alpha) \ll Q^{8/\alpha+\varepsilon}.$$

It follows from this and a lower bound for the number of isogeny classes of semi-stable elliptic curves with conductor less than Q that the probability that two such elliptic curves have this property tends to zero as Q tends to infinity, if α is large enough. We also have other results in more general cases.

As in Linnik's original treatment, we attack the problem by means of an analytic inequality for a larger class of objects encompassing the elliptic curves, namely holomorphic cusp-forms of weight two. Thus we use the Theorem of Wiles [Wil], and its further extensions, which prove the modularity of many elliptic curves over **Q**, to embed the set of isogeny classes of modular elliptic curves over **Q** in the set of primitive cusp forms[2].

However, due to incomplete knowledge of lower bounds for the Fourier coefficients of cusp forms it will appear that this inequality is not sufficient to prove the result we are seeking. We have to supplement its use by that of another similar inequality for the coefficients of the symmetric square L-function of cusp forms, and also appeal to a result of Ramakrishnan about the possible multiplicity of the symmetric square, the proof of which appears in an Appendix to this paper.

[2] Also called "newforms" in the literature; we use the vocabulary of [Miy] to emphasize again the analogy with Dirichlet characters.

This other inequality requires the study of Rankin-Selberg convolutions of $GL(3)$ automorphic forms, and it is actually not much harder to prove a generalization of our mean-value estimate to all $GL(n)$, in the context of automorphic representations satisfying the Ramanujan-Petersson bound. The result is related to the large-sieve, although it is not as powerful as one could expect; roughly it corresponds to the case of sums much longer than the conductor of the forms appearing.

We will first study the Linnik problem for primitive cusp forms, stating the analytic results required for the proof before showing how our main result follows from this.

We then apply the theorem to elliptic curves, with a short preliminary discussion of estimates for the number of isogeny classes of elliptic curves over $\mathbf{Q}$ with conductor at most Q. Both upper and lower bound are used in our theorem. The problem of finding an upper bound for a *fixed* level q was recently considered by Brumer and Silverman [B-S]. Their individual bound can be strengthened on average and we show how this is done.

It is then time to come back to the proof of the mean-value estimate, in its full generality for $GL(n)$. This result may be of independent interest, although it falls short of the hypothetical large-sieve inequalities which can be expected by analogy with the case of Dirichlet characters (namely, the case of $GL(1)$). A variant for Maass forms is used by Luo in [Luo].

Acknowledgement. We wish to thank H. Iwaniec for helpful discussions about this paper, and especially for suggesting to look at the symmetric squares to circumvent the difficulties with lower-bounds for the Fourier coefficients. We also thank A. Brumer and J. Silverman for communicating their result about the number of elliptic curves of a given conductor and allowing us to present here the straightforward application of their ideas which strengthens their bound on average.

Notational remark. When using Vinogradov's $\ll$ notation, it will often occur that we consider inequalities such as "for any $\varepsilon > 0$, it holds $f(x) \ll x^{\varepsilon} g(x)$"; as is customary in this case, the implied constant always depends on ε.

We may also remind that, as is usual in analytic number theory, the ε may be different from line to line in an argument.

2 The Linnik problem for cusp forms

2.1 Notations and statement of the mean-value estimates

Our main result in this section is about (families of) primitive cusp forms having the same Fourier coefficients for the first few primes. For k even, $k \geqslant 2$, we will denote by $S_k(q)^+$ the set of primitive cusp forms of weight k and level q and by $S_k(\leqslant Q)^+$ the set of primitive cusp forms of weight k and level less than or equal to Q. Moreover, for $f \in S_k(\leqslant Q)^+$, we will

write $\lambda_f(n)$ its Hecke eigenvalues, normalized so that the critical line for the L-function

$$L(f,s) = \sum_{n\geqslant 1} \lambda_f(n) n^{-s}$$

is the line $\mathrm{Re}(s) = 1/2$. This means that the Fourier expansion of f is

$$f(z) = \sum_{n\geqslant 1} a_f(n) e(nz)$$

with

$$\lambda_f(p) = p^{-(k-1)/2} a_f(p)$$

for all primes p.

Note that

$$|\lambda_f(n)| \leqslant \tau(n) \tag{2}$$

is then the Ramanujan-Petersson bound (proved by Deligne) for f.

We need some estimates for the cardinality of the various sets appearing. The easiest one is $S_k(\leqslant Q)^+$. Classical results about the genus of the modular curves $X_0(q)$ and the index of $\Gamma_0(q)$ in $SL(2,\mathbf{Z})$ show that $|S_k(\leqslant Q)^+|$ is about kQ^2, more precisely there is a constant $c(k) > 0$ with

$$|S_k(\leqslant Q)^+| \sim c(k) Q^2 \tag{3}$$

(see for instance [Shi] pages 25 and 46). Only the dependence in Q actually matters to us.

We will need to argue, here and in Section 4, in the language of automorphic representations, which is better suited to the various L-functions and to the context of $GL(n)$ automorphic forms. Many useful facts, and precise references, for the analytic properties of general L-functions of automorphic representations of $GL(n)$ which interest us here can be found in the paper [R-S] of Rudnick and Sarnak.

Recall first that there is an injective map $f \mapsto \pi_f$ from $S_k(q)^+$ to a certain subset of the set of cuspidal automorphic representations of $GL(2)$ over $\mathbf{Q}$ (see [Del], or [Gel]). This map is compatible with L-functions in the sense that $L(f) = L^\infty(\pi_f)$, where $L(f)$ is the classical Hecke L-function, defined above, and $L(\pi_f)$ is the Jacquet-Langlands L-function (complete with the Gamma factor at infinity), which is defined in terms of representation theory; here and elsewhere L^∞, for automorphic-representation L-functions, denotes the finite part of such an L-function.

Moreover, Gelbart and Jacquet have described a map $\pi \mapsto \pi^{(2)}$ associating a "symmetric square", a certain automorphic representation of $GL(3)$, to a cuspidal automorphic representation of $GL(2)$ [G-J].

Let $L^\infty(\pi_f^{(2)})$ be the finite part of the L-function of $\pi_f^{(2)}$; it is a Dirichlet series which we write

$$L^\infty(\pi_f^{(2)}, s) = \sum_{n\geqslant 1} \lambda_f^{(2)}(n) n^{-s}.$$

We then claim that for squarefree n, we have

$$\lambda_f^{(2)}(n) = \lambda_f(n^2). \tag{4}$$

This is actually due to Shimura, and it follows from the local computations of [G-J] with the fact that $f \mapsto \pi_f$ preserves L-functions: indeed, writing $\zeta_N(s) = L(s, \epsilon_N)$ for any integer $N \geqslant 1$ where ϵ_N is the trivial Dirichlet character modulo N (so ζ_N is the Riemann zeta function with the Euler factors at $p \mid N$ removed), we have

$$\zeta_{q'}(2s) \sum_{n \geqslant 1} \lambda_f(n^2) n^{-s} = L^\infty(\pi_f^{(2)}, s)$$

(where q' is the conductor of $\pi_f^{(2)}$) whence

$$\lambda_f^{(2)}(n) = \sum_{d^2 \mid n} \epsilon_{q'}(d) \lambda_f\Big(\frac{n^2}{d^4}\Big)$$

which immediately implies (4) for squarefree n.

The automorphic representation $\pi^{(2)}$ is not always cuspidal, however, which means that $L(\pi^{(2)})$ is not always entire. More precisely, Gelbart and Jacquet have established that $\pi^{(2)}$ is non-cuspidal if and only if there exists a non-trivial (primitive) character η such that $\pi = \pi \otimes \eta$. Such representations, and the cusp forms in $S_k(q)^+$ to which they correspond, are called monomial representations. It is known from the work of Hecke, Maass, Langlands and others, that they are the forms obtained from Hecke characters χ of a quadratic extension of $\mathbf{Q}$ by automorphic induction, so that $L(s, \pi) = L(s, \chi)$.

We will write $S_k(\leqslant Q)^\sharp$ for the set of primitive cusp forms of level less than or equal to Q which are not monomial. Then we write $S_k^{(2)}(\leqslant Q)^\sharp$ for the image of $S_k(\leqslant Q)^\sharp$ by the map $f \mapsto \pi_f^{(2)}$. The number of monomial representations can be easily shown to be $\ll Q^{1+\varepsilon}$ for any $\varepsilon > 0$, so the estimate

$$|S_k(\leqslant Q)^\sharp| \sim c(k) Q^2$$

holds again.

The monomial representations are exceptional in many respects, in particular we will see this in the case of the Linnik problem (see the remarks at the end of the next section).

The map $f \mapsto \pi_f \mapsto \pi_f^{(2)}$ is not injective; roughly, twisting by quadratic characters doesn't change the symmetric square, but the corollary to Theorem A of the Appendix shows that this is the only case that can occur. We state it here in the form we will use.

Theorem 2 *(Ramakrishnan) Let f and g be primitive cusp forms of level q_f and q_g. If*

$$\pi_f^{(2)} = \pi_g^{(2)}$$

then there exists a quadratic character χ of conductor d dividing $q_f q_g$ such that

$$\lambda_f(p) = \lambda_g(p)\chi(p)$$

for almost all primes p, or equivalently, by the strong multiplicity one theorem for $GL(2)$ (see [Gel])

$$\pi_f = \pi_g \otimes \chi. \tag{5}$$

Moreover, if q_f and q_g are squarefree, then $f = g$.

(We do not have necessarily $f = g \otimes \chi$, because $g \otimes \chi$ might be non-primitive; but (5) is correct because the tensor product is in the sense of automorphic representations, and $\pi_g \otimes \chi$ is the representation whose L-function coincides with that of $g \otimes \chi$ for all but finitely many places, in other words it corresponds to the "newform" associated to the (possibly) "old-form" $g \otimes \chi$.)

As in Section 4, we suppose that we are given for every $q \geqslant 1$ a subset $S(q) \subset S_k(q)^+$. We then write

$$S(\leqslant Q) = \bigcup_{q \leqslant Q} S(q)$$

and assume that $d \geqslant 0$ is such that

$$|S(\leqslant Q)| = O(Q^d) \tag{6}$$

(note that this holds for any choice of subsets $S(q)$ for $d = 2$; indeed taking $d = 2$ in what follows only results in having a slightly larger constant B_d and is not of great importance, so the reader may prefer to assume $d = 2$ for simplicity).

As above, the superscript $\sharp$ restricts the set to the subset of non-monomial forms, and the superscript (2) to the image of the non-monomial forms by the symmetric square map.

Now we can quote from Section 4 the mean-value estimates that we will require in the proof of the main theorem in the next subsection, namely Corollary 5: if $\beta > 2d + 2$ then for any $\varepsilon > 0$ we have

$$\sum_{f \in S(\leqslant Q)^\sharp} \Big| \sum_{n \leqslant Q^\beta} a_n \lambda_f(n) \Big|^2 \ll Q^{\beta+\varepsilon} \sum_{n \leqslant Q^\beta} |a_n|^2; \tag{7}$$

and Corollary 6: if $\beta > 2(d + 3)$ then for any $\varepsilon > 0$

$$\sum_{\pi_f \in S^{(2)}(\leqslant Q)^\sharp} \Big| \sum_{n \leqslant Q^\beta} a_n \lambda_f^{(2)}(n) \Big|^2 \ll Q^{\beta+\varepsilon} \sum_{n \leqslant Q^\beta} |a_n|^2. \tag{8}$$

From this last equation we deduce by (4), for any $\varepsilon > 0$,

$$\sum_{\pi_f \in S^{(2)}(\leqslant Q)^\sharp} \Big| \sum_{n \leqslant Q^\beta}^\flat a_n \lambda_f(n^2) \Big|^2 \ll Q^{\beta+\varepsilon} \sum_{n \leqslant Q^\beta}^\flat |a_n|^2 \tag{9}$$

for any complex numbers $(a_n)_{1 \leqslant n \leqslant Q^\beta}$, where $\sum^\flat$ denotes a sum restricted to squarefree integers. Remark that this is not a sum over all f since quadratic twists give the same π_f.

2.2 The main result

Now fix a set $\mathcal{P}$ of prime numbers of positive natural density δ (for instance, all primes in an arithmetic progression $an + b$ with $(a, b) = 1$), and a real number $\alpha > 1$. For (non-monomial) primitive forms f and g in $S_k(\leqslant Q)^\sharp$, write $f \sim g$ if $\lambda_f(p) = \lambda_g(p)$ for all primes $p \in \mathcal{P}$, $p \leqslant (\log Q)^\alpha$.[3]

Then clearly $\sim$ is an equivalence relation (depending on $\mathcal{P}$, Q, k, and α) on the finite set $S_k(\leqslant Q)^\sharp$, inducing one on the subset $S(\leqslant Q)^\sharp$, which is thus partitioned into finitely many finite equivalence classes. We will denote by $M_S(\mathcal{P}, Q, \alpha)$ the maximum cardinality of such an equivalence class: in other words, $M_S(\mathcal{P}, Q, \alpha)$ is the largest possible number of non-monomial forms in the set $S(\leqslant Q)$ whose Hecke eigenvalues are all equal for primes $p \leqslant (\log Q)^\alpha$.

The analogue of Linnik's result is the following:

Theorem 3 *There exists a constant $B_d > 0$ such that for all $\alpha > 1$, we have*

$$M_S(\mathcal{P}, Q, \alpha) \ll Q^{\frac{1}{2} + \frac{B_d}{\alpha} + \varepsilon}$$

for all $\varepsilon > 0$, the implied constant depending on ε, $\mathcal{P}$ and the family S.

Furthermore, if T is any fixed set of primes, then the number of elements of any equivalence class whose level is squarefree outside T (that is, $p^2 \mid q$ implies $p \in T$) is $\ll Q^{\frac{B_d}{\alpha} + \varepsilon}$ for any $\varepsilon > 0$, the constant depending further on T.

Moreover, $B_d = 2(d + 3)$ is admissible.

Of course, this result is non-trivial only if $S(\leqslant Q)^\sharp$ contains more elements than the bound given for the exceptions, so the efficiency of our result depends also on a lower bound for the number of forms we are considering. In particular, it is always trivial if $d \leqslant 1/2$ (but the result for forms with almost squarefree conductor is not, for α large enough, as long as $|S(\leqslant Q)|$ is larger than some fixed positive power of Q).

As one immediate corollary, we have for instance, taking $S(q) = S_k(q)$ and using (3):

[3] It is possible to relax this condition by asking that the equality holds only with a "small" number of exceptions (to exclude ramified primes for instance). This complicates the argument slightly so we will not do it here, as our results are not definitive anyway.

Corollary 1 *Fix $\alpha > 2B_2/3$. Then for Q tending to infinity, the probability that two non-monomial primitive forms of level less than Q have the same Hecke eigenvalues for all primes less than $(\log Q)^\alpha$ in a fixed arithmetic progression tends to zero. Here $B_2 = 10$ is admissible.*

Proof of Theorem 3. We will omit $\mathcal{P}$ in the notation and write only $M_S(Q, \alpha)$ in the proof.

Take an equivalence class of cardinality $M_S(Q, \alpha)$ for $\sim$, and an element f in this class, of level q_f.

The idea of the proof is that because of the multiplicativity of the Hecke eigenvalues the hypothesis implies that for any $g \sim f$, we have

$$\lambda_f(n) = \lambda_g(n)$$

for any n such that all its prime factors are in $\mathcal{P}$ and less than $(\log Q)^\alpha$; those n form a rather large set, but on the other hand choosing $a_n = \lambda_f(n)$ in the mean-value estimate (7), we get the same sum over those n with multiplicity $M_S(Q, \alpha)$, and it remains only to find a lower bound for this common sum to get some result by positivity. The quality of the result depends on that of the lower bound, and we will see that this fails to give a good result because of the impossibility to be sure that Fourier coefficients are "large" for enough primes; however the *deus ex machina* is the well-known formula (for p unramified)

$$\lambda_f(p)^2 - \lambda_f(p^2) = 1, \tag{10}$$

which implies that if $\lambda_f(p)$ is "small" then $\lambda_f(p^2)$ can not be, and in this case we use the inequality (9) instead (with $a_n = \lambda_f(n^2)$ this time). This trick has already been used, for instance in [DFI], in other contexts when this problem of the lower bound for Fourier coefficients of cusp forms arose. The great virtue of (10) is its complete uniformity in any parameter involved. Since we are considering very small primes (compared to the conductor), this is absolutely vital.

We now come to the details.

By the assumption, the number of primes in $\mathcal{P}$ less than $(\log Q)^\alpha$ is

$$\pi_{\mathcal{P}}((\log Q)^\alpha) \sim \frac{\delta(\log Q)^\alpha}{\alpha \log\log Q}.$$

Since q_f has only $\ll \log Q$ prime divisors, the set $\mathcal{P}(Q)$ of primes $p \leqslant (\log Q)^\alpha$ not dividing q_f satisfies also

$$|\mathcal{P}(Q)| \sim \frac{\delta(\log Q)^\alpha}{\alpha \log\log Q}.$$

For any $p \in \mathcal{P}(Q)$ we have, as mentioned, $\lambda_f(p)^2 - \lambda_f(p^2) = 1$, so one of the two sets of primes

$$\mathcal{P}_1(Q) = \{p \in \mathcal{P}(Q) \mid |\lambda_f(p)| \geqslant 1/2\}$$

and

$$\mathcal{P}_2(Q) = \{p \in \mathcal{P}(Q) \mid |\lambda_f(p^2)| \geqslant 1/2\}$$

(say $\mathcal{P}_i(Q)$) must satisfy

$$|\mathcal{P}_i(Q)| \geqslant |\mathcal{P}(Q)|/2 \geqslant \frac{\delta(\log Q)^\alpha}{3\alpha \log\log Q}$$

for Q large enough.

If $g \sim f$, we have *a fortiori* $\lambda_g(p) = \lambda_f(p)$ for all primes $p \in \mathcal{P}_i(Q)$ so that, by multiplicativity and the Hecke relations:

$$\lambda_g(n) = \lambda_f(n)$$

if $n = \prod_{p\in\mathcal{P}_i(Q)} p^{v_p(n)}$ has all its prime factors in $\mathcal{P}_i(Q)$.

Among those integers consider the set $\mathcal{N}(Q)$ of squarefree integers n such that n has m (which will be chosen later) prime factors exactly, all in $\mathcal{P}_i(Q)$. From the definition of $\mathcal{P}_i(Q)$, it follows that

$$|\lambda_g(n^i)| \geqslant 2^{-m} \tag{11}$$

for all $n \in \mathcal{N}(Q)$ and all $g \sim f$ (note the i on the left hand side). Let $N = \mathrm{Max}\,(\mathcal{N}(Q))$, so $N \leqslant (\log Q)^{\alpha m} = N'$.

We now assume that m is chosen so that N' is less than, but near, Q^β, with $\beta > 2(d+3)$. Then for $n \leqslant Q^\beta$ take

$$a_n = \begin{cases} \overline{\lambda}_f(n^i) \text{ for } n \in \mathcal{N}(Q) \\ 0 \text{ otherwise} \end{cases}$$

in (7) or (9) if $i = 1$ or $i = 2$, respectively.

In the first case we get by positivity

$$\begin{aligned} M_S(Q,\alpha)\Big|\sum_{n\in\mathcal{N}(Q)} |\lambda_f(n)|^2\Big|^2 &\leqslant \sum_{h\in S(\leqslant Q)^\sharat} \Big|\sum_{n\leqslant Q^\beta}{}^\flat\, a_n\lambda_h(n)\Big|^2 \\ &\ll Q^{\beta+\varepsilon}\sum_{n\leqslant Q^\beta}{}^\flat |a_n|^2 \\ &= Q^{\beta+\varepsilon}\sum_{n\in\mathcal{N}(Q)} |\lambda_f(n)|^2 \end{aligned}$$

for any $\varepsilon > 0$, whence

$$\begin{aligned} M_S(Q,\alpha) &\ll Q^{\beta+\varepsilon}\Big(\sum_{n\in\mathcal{N}(Q)} |\lambda_f(n)|^2\Big)^{-1} \\ &\ll Q^{\beta+\varepsilon}2^m|\mathcal{N}(Q)|^{-1} \end{aligned}$$

by (11).

In the second case, let $M_S^{(2)}(Q,\alpha)$ be the cardinality of the image of the equivalence class of f via $f \mapsto \pi_f^{(2)}$. Then by the second mean-value estimate:

$$\begin{aligned} M_S^{(2)}(Q,\alpha)\Big|\sum_{n\in\mathcal{N}(Q)}|\lambda_f(n^2)|^2\Big|^2 &\leqslant \sum_{\pi_h\in S^{(2)}(\leqslant Q)^\sharp}\Big|\sum_{n\leqslant Q^\beta}^{\flat} a_n\lambda_h(n^2)\Big|^2\\ &\ll Q^{\beta+\varepsilon}\sum_{n\leqslant Q^\beta}^{\flat}|a_n|^2\\ &= Q^{\beta+\varepsilon}\sum_{n\in\mathcal{N}(Q)}|\lambda_f(n^2)|^2 \end{aligned}$$

for any $\varepsilon>0$, and

$$\begin{aligned} M_S^{(2)}(Q,\alpha) &\ll Q^{\beta+\varepsilon}\Big(\sum_{n\in\mathcal{N}(Q)}|\lambda_f(n^2)|^2\Big)^{-1}\\ &\ll Q^{\beta+\varepsilon}2^m|\mathcal{N}(Q)|^{-1} \end{aligned}$$

by (11).

We now choose m and estimate N' and $|\mathcal{N}(Q)|$.

As already mentioned, we select m so that the upper bound N' for N is about the same as Q^β, namely

$$m=\left[\frac{\beta\log Q}{\alpha\log\log Q}\right].$$

Then we have $2^m\ll Q^\varepsilon$ if Q is sufficiently large. Similarly

$$(\delta/3\alpha)^{(\log Q)/(\log\log Q)}\gg Q^{-\varepsilon}.$$

Finally, by unique factorization of integers and Stirling's formula

$$\begin{aligned} |\mathcal{N}(Q)| \geqslant \binom{|\mathcal{P}_i(Q)|}{m} &\gg m^{-1/2}\left(\frac{|\mathcal{P}_i(Q)|}{m}\right)^m\\ &\gg \left(\frac{\delta(\log Q)^\alpha}{3\alpha\log\log Q}\frac{\alpha\log\log Q}{\beta\log Q}\right)^m m^{-1/2}\\ &\gg Q^{-\varepsilon}\left((\log Q)^{\alpha-1}\right)^{\beta(\log Q)(\alpha\log\log Q)^{-1}-1}\\ &\gg Q^{\beta\frac{\alpha-1}{\alpha}-\varepsilon} \end{aligned}$$

(for Q sufficiently large again), so that we get from the above estimate, for any $\varepsilon>0$:

- If $i=1$,
$$M_S(Q,\alpha)\ll Q^{\beta-\beta\frac{\alpha-1}{\alpha}+\varepsilon}=Q^{\frac{\beta}{\alpha}+\varepsilon}$$
which concludes the proof in this case, with a much better exponent actually ($B=2d+2$ is enough then).

- If $i = 2$,

$$M_S^{(2)}(Q,\alpha) \ll Q^{\beta-\beta\frac{\alpha-1}{\alpha}+\varepsilon} = Q^{\frac{\beta}{\alpha}+\varepsilon}$$

and it remains to relate $M_S^{(2)}(Q,\alpha)$ and $M_S(Q,\alpha)$, which may be bigger since $f \mapsto \pi_f^{(2)}$ is not injective.

Take a form g in the equivalence class of f whose symmetric square has maximum multiplicity, say M_g, so

$$M_S(Q,\alpha) \leqslant M_g M_S^{(2)}(Q,\alpha)$$

and choose g furthermore so that its level q_g is the smallest possible.

If h is a form equivalent to g with the same symmetric square, then by Theorem 2 there exists a quadratic character χ, of conductor d, such that $\pi_h = \pi_g \otimes \chi$.
If we write uniquely $d = d_1 d_2$ with $d_1 \mid q_g^\infty$ and $(d_2, q_g) = 1$, then comparing conductors we get an equality

$$q_h = d_2^2 d_1' q_g$$

where $d_1' \mid d_1^2$. For any given d_1, we may have as many as $\sqrt{Q/q_g} \leqslant Q^{1/2}$ possible values of d_2. Since the number of integers less than Q divisible only by primes dividing q_g is $\ll Q^\varepsilon$ for any $\varepsilon > 0$, it follows that

$$M_g \ll Q^{\frac{1}{2}+\varepsilon}$$

so

$$M_S(Q,\alpha) \ll Q^{\frac{1}{2}+\frac{\beta}{\alpha}+\varepsilon}$$

for any $\varepsilon > 0$.

If however we are given a fixed finite set of primes T such that we only consider forms of level squarefree outside T, then clearly from

$$q_h = d_2^2 d_1' q_g$$

we see that the conductor of the character χ must be divisible only by primes in T or dividing q_g. The number of such d less than Q is again $\ll Q^\varepsilon$ for any $\varepsilon > 0$, and the last statement of the theorem follows accordingly.

So the theorem is proved. □

3 The Linnik problem for elliptic curves

3.1 Notations and counting problems

We can now approach the Linnik problem for elliptic curves by means of the L-functions of elliptic curves and their modularity.

Recall that the general modularity conjecture for elliptic curves over $\mathbf{Q}$ says that the map which associates to an elliptic curve E the inverse Mellin transform of its Hasse-Weil zeta function

$$L(E,s)=\prod_p (1-a_E(p)p^{-s}+\epsilon_N(p)p^{1-2s})^{-1}$$

(where N is the conductor of E, and a_E is defined as usual by the equality $|E(\mathbf{F}_p)|=p+1-a_E(p)$, if p doesn't divide N) induces a bijection between the set $Ell(q)$ of isogeny classes of elliptic curves over $\mathbf{Q}$ of conductor q and the set $S_2(q,\mathbf{Z})^+$ of primitive cusp forms of weight two and level q with integer Fourier coefficients. In particular, this would embed $Ell(\leqslant Q)$ (with obvious notation) into the set $S_2(\leqslant Q)^+$.

This modularity conjecture is now, after the breakthrough of Wiles [Wil], known in many cases: according to Diamond's extension of Wiles' result, any elliptic curve $E/\mathbf{Q}$ which doesn't have additive reduction at either 3 or 5 is modular. We will work either with all elliptic curves. assuming the full modularity conjecture, or with classes which are known to be modular. Our results can also be restated as holding for modular elliptic curves.

The cusp form f_E corresponding to a modular elliptic curve E is known to be monomial if and only if E has complex multiplication. In that case, the modularity was already well-known. As we apply our result of the previous section we have to exclude those curves. We will write $Ell(\leqslant Q)^\sharp$ for the set of isogeny classes of non-CM elliptic curves over $\mathbf{Q}$ of conductor less than Q, and $Ell(\leqslant Q)^\flat$ for the subset of $Ell(\leqslant Q)$ given by semi-stable elliptic curves, i.e. those whose conductor is squarefree. We have $Ell(\leqslant Q)^\flat \subset Ell(\leqslant Q)^\sharp$ since CM-curves are not semi-stable. Also, by Wiles's Theorem, the curves in $Ell(\leqslant Q)^\flat$ are modular.

As in the previous subsection, we first need to estimate the cardinality of the sets we will consider. This is a subtler question than the corresponding one for cusp forms.

First we consider the problem of an upper bound for $|Ell(\leqslant Q)|$. We will actually deal with $\mathcal{E}ll(\leqslant Q)$, the set of isomorphism classes of elliptic curves over $\mathbf{Q}$ of conductor less than Q. According to results of Mazur and Kenku (see [Si1], page 265), there are at most 8 isomorphism classes of elliptic curves over $\mathbf{Q}$ isogenous to a given curve $E/\mathbf{Q}$, so all our $O(\ )$ estimates for $\mathcal{E}ll(\leqslant Q)$ will also be true for $Ell(\leqslant Q)$.

Recently, Brumer and Silverman [B-S], proved the estimate

$$|\mathcal{E}ll(q)| \ll q^{1/2+\varepsilon} \tag{12}$$

for all $\varepsilon>0$. This trivially gives $|\mathcal{E}ll(\leqslant Q)| \ll Q^{3/2+\varepsilon}$, but the proof of (12) can actually be extended to give a sharper bound on average.

Proposition 1 *For any $\varepsilon > 0$ it holds*

$$|\mathcal{E}ll(\leqslant Q)| \ll Q^{1+\varepsilon}$$

and

$$|Ell(\leqslant Q)| \ll Q^{1+\varepsilon}.$$

Proof. We have already seen how the second statement follows from the first.

Brumer and Silverman actually count elliptic curves having good reduction outside a given (finite) set of primes S (containing 2 and 3) by writing, for such an elliptic curve $E/\mathbf{Q}$,

$$1728\Delta_E = ad^6 \tag{13}$$

where a is 6-th power free, and observing that $(c_6(E)/d^3, c_4(E)/d^2)$ is then an S-integral point on the elliptic curve $\mathcal{E}_a$ given by

$$\mathcal{E}_a \,:\, Y^3 = X^2 + a$$

so that it only remains to estimate how many a's are possible, how many S-integral points there are on $\mathcal{E}_a$ for a given a and how many different curves E can be associated to the same a.

We begin by writing

$$\begin{aligned} |\mathcal{E}ll(\leqslant Q)| &= \sum_{q\leqslant Q} |\mathcal{E}ll(q)| \\ &\leqslant \sum_{q\leqslant Q} |\mathcal{E}ll(q)'| \end{aligned}$$

where $\mathcal{E}ll(q)'$ is the set of isomorphism classes of elliptic curves over $\mathbf{Q}$ having good reduction outside the set of prime divisors of q, with 2 and 3 added.

Now we rewrite straightforwardly the counting argument of Brumer and Silverman for $\mathcal{E}ll(q)'$, obtaining

$$|\mathcal{E}ll(\leqslant Q)| \leqslant \sum_{q\leqslant Q} \sum_{a\in A(q)} \sum_{P\in\mathcal{E}_a(\mathbf{Z}_q)} |E(P)|$$

where:

- $A(q)$ is the set of possible a's for a given level q.
- $\mathbf{Z}_q$ is $\mathbf{Z}_S$ for S the set of prime factors of q, with 2 and 3 added.
- $E(P)$ is the set of elliptic curves that give the point $P \in \mathcal{E}_a(\mathbf{Z}_q)$ in the way sketched above.

Brumer and Silverman show that the inner sum is $\ll Q^{\varepsilon}$ for any $\varepsilon > 0$, so we get

$$|\mathcal{E}ll(\leqslant Q)| \ll Q^{\varepsilon} \sum_{q\leqslant Q} \sum_{a\in A(q)} |\mathcal{E}_a(\mathbf{Z}_q)|$$

and then, still following their argument, we apply deep bounds of Silverman and Evertse for $|\mathcal{E}_a(\mathbf{Z}_q)|$ to obtain the estimate

$$|\mathcal{E}ll(\leqslant Q)| \ll Q^{\varepsilon} \sum_{q\leqslant Q} \sum_{a\in A(q)} h_3(\mathbf{Q}(\sqrt{-a}))$$

in terms of the 3-part of the class group of the imaginary quadratic field $\mathbf{Q}(\sqrt{-a})$ (here and in the remainder of the argument, ε is different from line to line). This is where the saving on average will come from: whereas no better individual bound for $h_3(\mathbf{Q}(\sqrt{-a}))$ is known than $h_3 \leqslant h \leqslant a^{1/2}(\log 2a)$, Davenport and Heilbronn established a sharp average bound in [D-H]. We now apply it.

For this, write the sum over a as the sum over the squarefree kernels a' of elements of $A(q)$. Using that a is 6-th power free and a q-unit to bound the multiplicity, it follows that the number of a for a given a' is again bounded by Q^{ε}, giving

$$|\mathcal{E}ll(\leqslant Q)| \ll Q^{\varepsilon} \sum_{q\leqslant Q} \sum_{a'} h_3(Q(\sqrt{-a'})).$$

Then we exchange the order of summation; a' being squarefree implies $|a'| \leqslant 1728q \leqslant 1728Q$ (see (13)) and moreover a' divides the discriminant Δ of any curve (of conductor q) for which it may appear, so again because a' is squarefree it must actually divide the conductor q, whence the multiplicity of q for a given a' is less than the number of divisors of q, and thus

$$|\mathcal{E}ll(\leqslant Q)| \ll Q^{\varepsilon} \sum_{|a'|\leqslant 1728Q} h_3(\mathbf{Q}(\sqrt{-a'}))$$

which is $\ll Q^{1+\varepsilon}$ by Theorem 3 of Davenport and Heilbronn, as claimed. □

We need also a lower bound of the form

$$|Ell(\leqslant Q)| \gg Q^{d}$$

for some $d > 1/2$, or for semi-stable curves

$$|Ell(\leqslant Q)^{\flat}| \gg Q^{d}$$

(which of course implies the former inequality). This is proved in [FNT] with $d = 5/6$, namely

$$|Ell(\leqslant Q)^{\flat}| \gg Q^{5/6}. \tag{14}$$

Remark. The case $K = 1$ of the main Theorem of [FNT] is not far from giving also Proposition 1; the difference is that it deals with the discriminant instead of the conductor, but most of the ingredients are present there.

3.2 The Linnik problem for elliptic curves

We will now deduce from Theorem 3 our applications to Linnik's problem for elliptic curves. Assuming the general modularity conjecture, we take for $S(q)$ the set of primitive forms associated to isogeny classes of elliptic curves over $\mathbf{Q}$ with conductor q. From Proposition 1 we can take $d = 1 + \epsilon$ for any $\epsilon > 0$, so the constant B in the theorem may be $B = 8 + 2\epsilon$, and actually ϵ can be absorbed in the other ε from Theorem 3, so $B = 8$ is admissible.

Consider first the case of semi-stable elliptic curves over $\mathbf{Q}$, and take therefore $T = \emptyset$ in the second statement of Theorem 3. In this case we need not assume the modularity conjecture. Using the lower-bound (14), we get Theorem 1 from the introduction, the statement of which we now recall.

Corollary 2 *Let $M(Q, \alpha)$ be the maximal number of isogeny classes of semi-stable elliptic curves over $\mathbf{Q}$ with conductor less than or equal to Q which for every prime $p \leqslant (\log Q)^\alpha$ have a fixed number of points modulo p.*

Then we have for any $\varepsilon > 0$

$$M(Q, \alpha) \ll Q^{B/\alpha+\varepsilon}$$

and this is non-trivial for $\alpha > 6B/5$. Moreover, if we only ask that the curves have a fixed number of points modulo p for p in a fixed set of primes of positive natural density, the bound still holds, with a constant in $\ll$ depending on the set.

Or, applied to a special case in a probabilistic phrasing:

Corollary 3 *Fix $\alpha > 6B/5$. Then for Q tending to infinity, the probability that two semi-stable elliptic curves of conductor $\leqslant Q$ have the same number of points modulo p for all primes less than $(\log Q)^\alpha$ in a fixed arithmetic progression tends to zero.*

We have a somewhat weaker estimate for the general case.

Corollary 4 *Assume the general modularity conjecture.*

Fix $\alpha > 3B$ and a set $\mathcal{P}$ of primes with positive natural density. Then for all $\varepsilon > 0$ the maximal number of isogeny classes of elliptic curves of conductor $\leqslant Q$ without complex multiplication which have the same number of points modulo p for all $p \leqslant (\log Q)^\alpha$ in $\mathcal{P}$ is bounded by $Q^{1/2+B/\alpha+\varepsilon}$ up to a positive constant depending only on ε, α and $\mathcal{P}$.

The probabilistic statement also holds in this general case.

Remarks.

- The case of CM-curves (or monomial forms) is actually different, since an estimate such as the one for general $\mathcal{P}$ in Corollary 2, with exponent tending to zero as α tends to infinity, is false for them. For example, taking all curves

$$E_D \,:\, y^2 = x^3 + D \tag{15}$$

it is known that

$$a_{E_D}(p) = 0$$

for all p congruent to 2 mod 3, p unramified, so if we choose this arithmetic progression as our set $\mathcal{P}$, we have as many as $Q^{1/2}$ elliptic curves of conductor less than Q having the same Fourier coefficients for $p \in \mathcal{P}$.

This shows that our introduction of the symmetric square, because of the lack of lower bound for the Fourier coefficients of cusp-forms, is not purely technical.

However, if we consider all primes, then on the Generalized Riemann Hypothesis two monomial forms are still distinguished by some prime less than $(\log Q)^2$, so the corresponding analogue of Linnik's result should hold.

We can actually prove it: if $E/\mathbf{Q}$ is an elliptic curve with complex multiplication and conductor less than Q, then it follows from our proof of Theorem 3 and the knowledge of the Fourier coefficients of the corresponding primitive forms that the number of isogeny classes of elliptic curves over $\mathbf{Q}$ with complex multiplication and conductor less than Q having the same Fourier coefficients as E for $p \leqslant (\log Q)^\alpha$ is $\ll Q^{B/\alpha+\varepsilon}$ for any $\varepsilon > 0$ and some $B > 0$ (actually, $B = 6$ is enough).

Indeed, there are only a finite number of j-invariants of elliptic curves over $\mathbf{Q}$ with complex multiplication, and each possible j gives rise to a family of twists similar to (15) above (see [Si2], appendix A for instance).

It then suffices to find a lower bound for each family, which is not very difficult (see [DFI] page 224 for the reasoning in the case of $y^2 = x^3 + D$).

Once the lower bound is known, it remains to apply the same proof with the mean-value estimate for $GL(2)$ forms applied to the family of primitive forms associated to complex multiplication curves.

Of course, it is then possible to bring both results together and say that the number of elements in the whole set $Ell(\leqslant Q)$ which have the same Fourier coefficients for all primes $p \leqslant (\log Q)^\alpha$ for α large enough is $\ll Q^{1/2+B/\alpha+\varepsilon}$ for any $\varepsilon > 0$ and some $B > 0$.

4 A mean-value estimate for automorphic representations

The original large-sieve inequality for primitive Dirichlet characters is

$$\sum_{q\leqslant Q}\sideset{}{^*}\sum_{\chi\,(q)}\left|\sum_{n\leqslant N}a_n\chi(n)\right|^2\leqslant(N+Q^2)\sum_{n\leqslant N}|a_n|^2 \tag{16}$$

for any sequence $(a_n)_{n\leqslant N}$ of complex numbers.

This is a kind of quasi-orthogonality statement for the truncated sequences $(\chi(n))_{1\leqslant n\leqslant N}$ considered as elements of a finite dimensional Hilbert space.

After the work of Jacquet and Langlands, it has appeared that Dirichlet characters are only the case $n=1$ of a much more general theory of automorphic representations on the algebraic group $GL(n)$. For such an automorphic representation, a (standard) L-function $L(\pi)$ is also defined; it is the product of a Gamma factor and a Dirichlet series

$$L^\infty(\pi,s)=\sum_{n\geqslant 1}\lambda_\pi(n)n^{-s}.$$

It is expected that the coefficients $\lambda_\pi(n)$ of those L-functions should satisfy inequalities similar to (16) when (large enough) increasing families of automorphic representations, where certain parameters (the conductor, the weight, or others) are bounded, are considered in the outer sum on the left-hand side. The hypothetical bound on the right hand side would be roughly the length N of the inner sum plus the number of representations considered, up to small factors. Some results exist, with the weight varying for instance, for the classical case of $GL(2)$-automorphic forms, see for example [D-I].

We will establish such an estimate for certain families of automorphic representations, but only in the easiest case, when the length N of the sum is much larger than the number of π's.

To define those families, fix first an admissible representation π_∞ of (the Hecke group algebra of) $GL(n,\mathbf{R})$ considered as infinite component of some cuspidal automorphic representation of $GL(n)$ over $\mathbf{Q}$ – for example, if $n=2$, and π_∞ is the discrete series representation $\sigma(\mu_1,\mu_2)$ with $\mu_1\mu_2^{-1}(t)=t^{k-1}\mathrm{sgn}(t)$ (see [Gel] page 91 for the notations), for some integer $k\geqslant 2$, then π_∞ is the infinite component of all automorphic representations corresponding to classical weight k modular forms.

Fix also a character η of the idèle class group of $\mathbf{Q}$.

Then for any integer $q\geqslant 1$ we let $\mathrm{Aut}(q)$ denote the set of cuspidal automorphic representations π of $GL(n)$ over $\mathbf{Q}$ such that:

- π_∞ is the infinite component of π, and η its central character;
- π satisfies the Ramanujan-Petersson conjecture: if

$$L^\infty(\pi,s)=\sum_{n\geqslant 1}\lambda_\pi(n)n^{-s}$$

is the finite part of the standard L-function of π, we have

$$\lambda_\pi(n) \ll n^\varepsilon \tag{17}$$

for any $\varepsilon > 0$;

- The conductor of π is q.

With respect to the Ramanujan-Petersson bound, we recall that because of the Euler product

$$L^\infty(\pi, s) = \prod_p \prod_{1\leqslant j\leqslant n} (1 - \alpha_{\pi,j}(p)p^{-s})^{-1}$$

it is known that (17) implies

$$|\alpha_{\pi,j}(p)| \leqslant 1$$

which shows that the bound (17) is actually uniform with respect to π. This will be important.

It should then be true that $\mathrm{Aut}(q)$ is finite and its cardinality (as a function of q) is bounded by a fixed power of q. We don't actually need this fact.

We further set

$$\mathrm{Aut}(\leqslant Q) = \bigcup_{q\leqslant Q} \mathrm{Aut}(q).$$

Again, $\mathrm{Aut}(\leqslant Q)$ should be finite and its cardinality at most polynomial in Q.

Now we will suppose given for every $q \geqslant 1$ a subset $S(q)$ of $\mathrm{Aut}(q)$ and write

$$S(\leqslant Q) = \bigcup_{q\leqslant Q} S(q).$$

We can now state our result.

Theorem 4 *Fix $n > 0$. Given sets $S(\leqslant Q)$ as above, assume that*

$$|S(\leqslant Q)| = O(Q^d). \tag{18}$$

There exists an absolute constant $B_{n,d} > 0$ such that if $N > Q^\beta$ with $\beta > B_{n,d}$, then for any $\varepsilon > 0$ the inequality

$$\sum_{\pi\in S(\leqslant Q)} \Big|\sum_{n\leqslant N} a_n\lambda_\pi(n)\Big|^2 \ll N^{1+\varepsilon} \sum_{n\leqslant N} |a_n|^2 \tag{19}$$

holds for all complex numbers $(a_n)_{1\leqslant n\leqslant N}$. Moreover, $B_{n,d} = 2d + n$ is admissible.

Proof. The strategy is familiar, being based on the well-known duality principle.

The inequality (19) is equivalent to the estimate

$$||T_{N,Q}||^2 \ll N^{1+\varepsilon}$$

for the norm of the linear operator

$$T_{N,Q} \ : \ (a_n)_{n\leqslant N} \mapsto \Big(\sum_{n\leqslant N} a_n\lambda_\pi(n)\Big)_{\pi\in S(\leqslant Q)}$$

where both the domain and range are finite dimensional Hilbert spaces (with the natural hermitian form). Now by general Hilbert theory, we know that the norm of $T_{N,Q}$ is the same as that of (the conjugate of) its adjoint, which is the operator

$$T^*_{N,Q} \ : \ (\alpha_\pi)_{\pi\in S(\leqslant Q)} \mapsto \Big(\sum_{\pi\in S(\leqslant Q)} \alpha_\pi\lambda_\pi(n)\Big)_{n\leqslant N}.$$

In concrete terms this means that (19) is equivalent to the dual inequality

$$\sum_{n\leqslant N}\Big|\sum_{\pi\in S(\leqslant Q)} \alpha_\pi\lambda_\pi(n)\Big|^2 \ll N^{1+\varepsilon}\sum_\pi |\alpha_\pi|^2. \tag{20}$$

We now choose a smooth, positive, compactly supported test function ψ on $[0,+\infty[$, equal to 1 between 0 and 1, and such that $0\leqslant\psi(x)\leqslant 1$ for all $x\in\mathbf{R}$.

Then by positivity the left-hand side of (20) is less than

$$\sum_{n\geqslant 1}\Big|\sum_{\pi\in S(\leqslant Q)} \alpha_\pi\lambda_\pi(n)\Big|^2\psi(n/N)$$

so it is enough to prove the inequality for this last expression.

This we write, expanding the square and interchanging the order of summation, as

$$\sum_{\pi_1,\pi_2\in S(\leqslant Q)} \alpha_{\pi_1}\overline{\alpha_{\pi_2}}\sum_{n\geqslant 1}\lambda_{\pi_1}(n)\overline{\lambda_{\pi_2}(n)}\psi(n/N).$$

Let us denote by $S_N(\pi_1,\pi_2)$ the inner sum,

$$S_N(\pi_1,\pi_2) = \sum_{n\geqslant 1}\lambda_{\pi_1}(n)\overline{\lambda_{\pi_2}(n)}\psi(n/N).$$

We thus have

$$||T^*_{N,Q}(\alpha)||^2 \leqslant \sum_{\pi_1,\pi_2\in S(\leqslant Q)} \alpha_{\pi_1}\overline{\alpha_{\pi_2}}S_N(\pi_1,\pi_2). \tag{21}$$

We will use the following well-known lemma:

Lemma 1 *Let*

$$Q(\alpha) = \sum_{\pi_1,\pi_2} \alpha_{\pi_1}\overline{\alpha_{\pi_2}}K(\pi_1,\pi_2)$$

be a quadratic form, with $K(\pi_1,\pi_2) \in \mathbf{C}$. *Then we have*

$$||Q|| \leqslant \operatorname*{Max}_{\pi_1} \sum_{\pi_2} |K(\pi_1,\pi_2)|.$$

We are thus reduced to the problem of estimating the sums $S_N(\pi_1,\pi_2)$. This we will achieve by studying the analytic properties of the Dirichlet series

$$L_b(\pi_1 \otimes \tilde{\pi}_2, s) = \sum_{n\geqslant 1} \lambda_{\pi_1}(n)\overline{\lambda_{\pi_2}}(n)n^{-s}$$

(which might be called the "naïve" convolution of the automorphic representations π_1 and π_2) and expressing the sums as Mellin transforms.

The necessary properties of L_b are consequences of a result which compare it to the Rankin-Selberg convolution of π_1 and π_2. In complete generality, Jacquet, Piatetskii-Shapiro and Shalika have developed a theory of Rankin-Selberg convolutions of automorphic representations of $GL(n) \times GL(m)$ ([JPS] and other papers); in particular, they have defined a corresponding L-function and studied its properties (analytic continuation and functional equation). Some points which they didn't treat have been established by various other authors (among whom Shahidi, Moeglin and Waldspurger for instance).

In our case, this allows us to consider the L-function $L(\pi_1 \otimes \tilde{\pi}_2)$ of the representation-theoretic convolution of π_1 and the contragredient representation of π_2. [4]

We will prove below

Proposition 2 *Let* π_1 *and* π_2 *be automorphic representations of* $GL(n)$ *satisfying the Ramanujan-Petersson bound, of conductor* q_1 *and* q_2 *respectively.*

There exists an Euler product

$$H(\pi_1,\pi_2;s) = \prod_p H_p(\pi_1,\pi_2;\, p^{-s})$$

where $H_p(\pi_1,\pi_2)$ *is a rational function for all* p *and a polynomial (of degree bounded by a constant depending only on* n_1 *and* n_2*) for almost all* p, *such that* $H(\pi_1,\pi_2)$ *converges absolutely for* $\mathrm{Re}(s) > 1/2$ *(in particular, has no poles in this region), and*

$$L_b(\pi_1 \otimes \pi_2, s) = H(\pi_1,\pi_2;s)L^{\infty}(\pi_1 \otimes \pi_2, s).$$

[4] In the case of $GL(3)$, the convolution of the symmetric squares of two cusp forms f and g has already been used in other contexts in analytic number theory by Hoffstein and Lockhart [H-L] and by Luo, Rudnick, Sarnak [LRS] to obtain deep results about $GL(2)$ automorphic forms, especially non-holomorphic Maass forms.

Moreover, we have for any $\varepsilon > 0$ *and uniformly for* $\mathrm{Re}(s) = \sigma > 1/2$ *a bound*

$$H(\pi_1, \pi_2; s) \ll [q_1, q_2]^{\varepsilon} H(\sigma)$$

where H *is a fixed Dirichlet series absolutely convergent for* $\mathrm{Re}(s) > 1/2$ *satisfying in this region*

$$H(\sigma) \ll (\sigma - 1/2)^{-A}$$

for some $A > 0$ *depending only on* n_1 *and* n_2.

This reflects the fact that the coefficients of $L^{\infty}(\pi_1 \otimes \pi_2)$ and $L_b(\pi_1 \otimes \pi_2)$ are the same for squarefree integers n (see equation (26) below).

In particular, because $L^{\infty}(\pi_1 \otimes \pi_2)$ has a meromorphic continuation, this gives the analytic continuation of L_b up to the critical line.

If we grant the proposition we can now apply Mellin inversion, namely if we let $\hat{\psi}$ denote the Mellin transform of ψ,

$$\hat{\psi}(s) = \int_0^{+\infty} \psi(x) x^s \frac{dx}{x}$$

then we have

$$\psi(x) = \frac{1}{2\pi i} \int_{(3)} \hat{\psi}(s) x^{-s} ds$$

(the integral being on the line $\mathrm{Re}(s) = 3$ of the complex plane), from which easily follows the basic formula

$$\begin{aligned} S_N(\pi_1, \pi_2) &= \frac{1}{2\pi i} \int_{(3)} N^s \hat{\psi}(s) L_b(\pi_1 \otimes \tilde{\pi}_2, s) ds \\ &= \frac{1}{2\pi i} \int_{(3)} N^s \hat{\psi}(s) H(\pi_1, \tilde{\pi}_2; s) L^{\infty}(\pi_1 \otimes \tilde{\pi}_2, s) ds. \end{aligned}$$

We now move the line of integration to $\mathrm{Re}(s) = 1/2 + c$ where $c < 1/2$ will be chosen later. The Mellin transform $\hat{\psi}$ is easily seen to be holomorphic for $\mathrm{Re}(s) > 0$ and quickly decreasing in any vertical strip $\delta < \mathrm{Re}(s) < b$ ($\delta > 0$); the other terms in the integral being at most of polynomial growth, shifting the contour is possible.

The only singularities we can pick up by doing so are those of $L^{\infty}(\pi_1 \otimes \tilde{\pi}_2)$.

From the Rankin-Selberg theory, those are known. Indeed [M-W] establishes:

Theorem 5 *If there are no* $t \in \mathbf{C}$ *such that* $\pi_1 = \pi_2 \otimes |\cdot|^t$, *then* $L(\pi_1 \otimes \tilde{\pi_2})$ *is entire.*

If $\pi_1 = \pi_2$, *then* $L(\pi_1 \otimes \tilde{\pi_2})$ *has two simple poles at* 0 *and* 1 *and is holomorphic outside those points.*

In our case, π_1 and π_2 having unitary central character η, we can have $\pi_1 = \pi_2 \otimes |\cdot|^t$ only if $t = 0$, so this theorem describes all possible cases where poles may appear in the convolution.

Keeping this in mind we then estimate the integral on the other line, namely

$$\frac{1}{2\pi i}\int_{(1/2+c)} N^s \hat{\psi}(s) H(\pi_1, \tilde{\pi}_2; s) L^\infty(\pi_1 \otimes \tilde{\pi}_2, s) ds.$$

We are only interested in the q-aspect of the matters. By the bounds for H in Proposition 2, for any $\varepsilon > 0$ we have

$$H(\pi_1, \tilde{\pi}_2; 1/2 + c + it) \ll Q^\varepsilon c^{-A}.$$

As for the Rankin-Selberg convolution, after inserting the correct Gamma factors it has a functional equation relating its value at s with that of the contragredient convolution $L(\tilde{\pi}_1 \otimes \pi_2)$ at $1-s$ (see the references to several articles of Shahidi in [M-W]):

$$L(\pi_1 \otimes \tilde{\pi}_2, s) = g(\pi_1 \otimes \tilde{\pi}_2) q(\pi_1 \otimes \tilde{\pi}_2)^{1/2-s} L(\tilde{\pi}_1 \otimes \pi_2, 1-s)$$

where $g(\pi_1 \otimes \tilde{\pi}_2)$ is a complex number of absolute value 1 and $q(\pi_1 \otimes \tilde{\pi}_2)$ is the conductor of $\pi_1 \otimes \tilde{\pi}_2$. By a theorem of Bushnell and Henniart [B-H], it is bounded by the product of the n-th powers of the conductors of π_1 and $\tilde{\pi}_2$, which themselves are at most Q, so that

$$q(\pi_1 \otimes \tilde{\pi}_2) \leqslant (Q^2)^n = Q^{2n}.$$

From the functional equation, Stirling's formula and the convexity principle of Phragmen-Lindelöf, this implies in turn

$$L^\infty(\pi_1 \otimes \tilde{\pi}_2, 1/2 + c + it) \ll Q^{2n(1/4-c/2)}|t|^E = Q^{n/2-nc}|t|^E$$

for some $E > 0$. With the previous bound for H, and using the fact that $\hat{\psi}$ decreases faster than any polynomial on the line, we get the estimate

$$c^{-A} N^{1/2+c} Q^{n/2-nc+\varepsilon}$$

for the integral. Recalling that $N > Q^\beta$ and taking $c = (\log Q)^{-1}$ so that $1 \ll Q^c \ll 1$, $N^c \ll 1$ and $c^{-A} = (\log Q)^A$ we obtain therefore for any $\varepsilon > 0$

$$S_N(\pi_1, \pi_2) = \delta(\pi_1, \pi_2)\hat{\psi}(1) N R_{\pi_1} + O(N^{1/2+n/(2\beta)+\varepsilon}) \tag{22}$$

where $\delta(\pi_1, \pi_2)$ is the Kronecker delta, and R_{π_1} is the residue of the naïve convolution considered when $\pi_1 = \pi_2$, namely

$$R_\pi = H(\pi, \pi; 1)\mathrm{Res}_{s=1}(L^\infty(\pi \otimes \tilde{\pi}, s)).$$

We then claim that for any π (and uniformly in π)

$$R_\pi \ll Q^\varepsilon \tag{23}$$

for all $\varepsilon > 0$.

This is a straightforward consequence of (22) for $\pi_1 = \pi_2 = \pi$ and the Ramanujan-Petersson bound (17) which, we have already mentioned, is uniform in π. [5]

Now according to Lemma 1, the quantity we have to bound is actually

$$\operatorname*{Max}_{\pi_1} \sum_{\pi_2 \in S(\leqslant Q)} |S_N(\pi_1, \pi_2)|.$$

Therefore (22) and Lemma 1 give (see also (18) and (23))

$$\sum_{\pi \in S(\leqslant Q)} \left| \sum_{n \leqslant N} a_n \lambda_\pi(n) \right|^2 \ll (N + N^{1/2+(2d+n)/(2\beta)}) N^\varepsilon \sum_{n \leqslant N} |a_n|^2$$

for any $\varepsilon > 0$, and this implies (20) if $\frac{1}{2} + \frac{2d+n}{2\beta} < 1$, that is $\beta > 2d + n$. □

Proof of Proposition 2. We actually treat a more general case where π_i is an automorphic representation of $GL(n_i)$ for $i = 1, 2$.

Write

$$L^\infty(\pi_i, s) = \sum_{n \geqslant 1} \lambda_i(n) n^{-s}$$

for the finite part of the standard L-functions, and put as above

$$L_b(\pi_1 \otimes \pi_2, s) = \sum_{n \geqslant 1} \lambda_1(n)\lambda_2(n) n^{-s}.$$

We have to compare $L_b(\pi_1 \otimes \pi_2)$ and the Rankin-Selberg convolution $L^\infty(\pi_1 \otimes \pi_2)$.

The Rankin-Selberg convolution has an Euler product by the general theory, and the naïve convolution also has one because it's a Dirichlet series whose coefficients are multiplicative:

$$L_b(\pi_1 \otimes \pi_2, s) = \prod_p \sum_{k \geqslant 0} \lambda_1(p^k)\lambda_2(p^k) p^{-ks}.$$

Therefore, since we claim the existence of an Euler product

$$H(\pi_1, \pi_2) = \prod_p H_p(\pi_1, \pi_2)$$

relating the two, we can proceed locally for each prime p.

For any automorphic L-function, we denote by L_p its p-factor, considered as a polynomial (in p^{-s}) with complex coefficients.

[5] For $n = 2$ or $n = 3$ for the symmetric square, which are the two applications used in the Linnik problem, it is possible to give an elementary proof of (23) - not using Deligne's proof of the Ramanujan-Petersson conjecture - using a trick of Iwaniec, see [Iwa] page 131 for $n = 2$.

Assume first that p is an unramified prime of the Rankin-Selberg convolution. This is true for almost all p, and we will prove now the existence of a polynomial $H_p(\pi_1, \pi_2)$ such that

$$\sum_{k\geqslant 0} \lambda_1(p^k)\lambda_2(p^k)X^k = H_p(\pi_1, \pi_2)L_p(\pi_1 \otimes \pi_2). \tag{24}$$

We know that p is unramified for both π_1 and π_2, so that the p-factor of the standard L-function is

$$L_p(\pi_i)^{-1} = \prod_{1\leqslant j\leqslant n_i} (1 - \alpha_{i,j}X) \tag{25}$$

where $\alpha_{i,j}$ are the Satake parameters of the local representation at p.

Again, the general theory gives the p-factor of the Rankin-Selberg convolution

$$L_p(\pi_1 \otimes \pi_2)^{-1} = \prod_{\substack{1\leqslant j\leqslant n_1\\ 1\leqslant k\leqslant n_2}} (1 - \alpha_{1,j}\alpha_{2,k}X).$$

Assume, to begin with, that the $\alpha_{i,j}$ are all distinct and the $\alpha_{1,j}\alpha_{2,k}$ also. Coming then to the p-factor of the naïve convolution, we deduce from the Dirichlet series for $L^\infty(\pi_i)$ that

$$\begin{aligned}\sum_{k\geqslant 0} \lambda_i(p^k)X^k &= \prod_{1\leqslant j\leqslant n_i} (1 - \alpha_{i,j}X)^{-1}\\ &= \sum_{1\leqslant j\leqslant n_i} \frac{r_{i,j}}{1 - \alpha_{i,j}X}\end{aligned}$$

for some complex numbers $r_{i,j}$ (partial fraction expansion, since the α's are distinct), whence

$$\lambda_i(p^k) = \sum_{1\leqslant j\leqslant n_i} r_{i,j}\alpha_{i,j}^k.$$

This implies

$$\begin{aligned}\sum_{k\geqslant 0} \lambda_1(p^k)\lambda_2(p^k)X^k &= \sum_{k\geqslant 0} \Big(\sum_{\substack{1\leqslant i\leqslant n_1\\ 1\leqslant j\leqslant n_2}} r_{1,i}r_{2,j}\alpha_{1,i}^k\alpha_{2,j}^k \Big) X^k\\ &= \sum_{i,j} \frac{r_{1,i}r_{2,j}}{1 - \alpha_{1,i}\alpha_{2,j}X}.\end{aligned}$$

Reducing to a common denominator, which is exactly $L_p(\pi_1 \otimes \pi_2)$, we get the required formula (24).

Moreover, it is obvious that the coefficients of $H_p(\pi_1, \pi_2)$ are polynomials in the α's and since the Ramanujan bound implies $|\alpha_{i,j}| \leqslant 1$ it follows that those coefficients are bounded by some constants depending only on n_1 and n_2. Hence the absolute convergence (and the absence of poles) in

$\mathrm{Re}(s) > 1/2$ of the product over the unramified primes will follow if we can show that the coefficient of X of $H_p(\pi_1, \pi_2)$ vanishes, since there is no term in p^{-s} then.

But for any rational function

$$r = \frac{f}{g}$$

with polynomials f and g, satisfying $r(0) = 1$, the coefficient of X of the numerator f of r is $f'(0)$, and so equals $g(0)r'(0) + g'(0)$.

If $r = \sum_k b_k X^k$ is the power series development of r, we have therefore

$$f'(0) = g(0)b_1 + g'(0).$$

Assume moreover that $g = \prod_j (1 - \beta_j X)$. Then

$$f'(0) = b_1 - \sum_j \beta_j.$$

Applying this to the local factor of L_b which is of this form, we see that the corresponding coefficient is indeed zero since

$$\lambda_1(p)\lambda_2(p) = \sum_{i,j} \alpha_{1,i}\alpha_{2,j}. \tag{26}$$

We can now use a continuity argument to deduce that the existence of the polynomial H_p satisfying formula (24) and the vanishing of the coefficient of X remain valid when some of the roots of the local L-functions are the same.

It remains to treat the case of the ramified primes. The local factor at p of the L-functions of π_1 and π_2 is still of the form

$$L_p(\pi_i) = \prod_{1 \leqslant j \leqslant n_i'} (1 - \alpha_{i,j} X)^{-1}$$

for some $n_i' \leqslant n_i$. The same proof as the unramified case shows again that the local factor of the bilinear convolution is a rational function which has poles only among the reciprocals of the products $\alpha_{1,j}\alpha_{2,k}$. So we can define $H_p(\pi_1, \pi_2)$ by

$$H_p(\pi_1, \pi_2) = \Big(\sum_{k \geqslant 0} \lambda_1(p^k)\lambda_2(p^k)X^k\Big) L_p(\pi_1 \otimes \pi_2)^{-1} \tag{27}$$

and it's also a rational function.

It remains to establish that the finite product over the ramified primes has no pole for $\mathrm{Re}(s) > 1/2$. But a pole s_0 of $H_p(\pi_1, \pi_2, p^{-s})$ must satisfy

$$\alpha_{1,j}\alpha_{2,k}p^{-s_0} = 1$$

(for some j and k), so by the Ramanujan bound again we get $\mathrm{Re}(s_0) \leqslant 0$.

As for bounding $H(\pi_1, \pi_2; s)$, clearly by the Ramanujan bound the product over the unramified primes is absolutely convergent for $\mathrm{Re}(s) > 1/2$. It is dominated (termwise) by the Euler product H whose factors are obtained by taking the corresponding factor of H_p and replacing each coefficient of the polynomial by its absolute value, which in turn, since the coefficient of X^2 is absolutely bounded (say by A), is dominated by an Euler product which may be written (by factoring by force $\zeta(2s)$) as $\zeta(2s)^A J(s)$ where $J(s)$ is absolutely convergent for $\mathrm{Re}(s) > 1/3$. The estimate

$$H(\sigma) \ll (\sigma - 1/2)^{-A}$$

then follows directly.

We now estimate the product over the ramified primes

$$\prod_{p|[q_1,q_2]} H_p(\pi_1, \pi_2; p^{-s})$$

using (27).

For $L_p(\pi_1 \otimes \pi_2)^{-1}$, which is a polynomial of degree at most $n_1 n_2$ we write, by the Ramanujan bound again:

$$\begin{aligned}\prod_{p|[q_1,q_2]} L_p(\pi_1 \otimes \pi_2) &\leqslant \prod_{p|[q_1,q_2]} (1 + p^{-\sigma})^{n_1 n_2} \\ &\leqslant (\prod_{p|[q_1,q_2]} 2)^{n_1 n_2} \\ &\ll [q_1, q_2]^{\varepsilon}\end{aligned}$$

for any $\varepsilon > 0$, since (see [H-W], chapter 22 for instance) the number of prime divisors of an integer n is $O(\log n / \log\log n)$.

On the other hand, still by Ramanujan, for any $\varepsilon > 0$

$$\begin{aligned}\sum_{k \geqslant 0} \lambda_1(p^k)\lambda_2(p^k) p^{-ks} &\ll \sum_{k \geqslant 0} p^{k(\varepsilon - s)} \\ &= \frac{1}{1 - p^{-s+\varepsilon}}\end{aligned}$$

so that taking the product over $p \mid [q_1, q_2]$ we obtain by the same reasoning the same bound as above for the product of those terms, and in the end

$$\prod_{p|[q_1,q_2]} H_p(\pi_1, \pi_2; p^{-s}) \ll [q_1, q_2]^{\varepsilon}\,.$$

□

It is clear that as n increases the condition $N > Q^{B_{n,d}}$ beyond which the inequality is proved becomes more restricted. It seems that further ideas

are required to establish sharp forms of the large-sieve inequalities in those cases (or even to refute them if they happen to be false). Using variants of a trick due to Viola and Forti for Dirichlet characters, it is likely that a sharp large-sieve inequality would follow if we could take $B_{n,d} = d$ in the mean-value estimate (19).

We now state the corollaries which are used in the Linnik problem, see Subsection 2.1.

First for $n = 2$, and for any weight $k \geqslant 2$, we have mentioned already that for a certain π_∞ and $\eta = 1$, it holds

$$\mathrm{Aut}(q) = S_k(q)^+$$

since Deligne has proved that cusp-forms of weight larger than 2 satisfy the Ramanujan-Petersson bound. Take then as $S(q)$ some subset of the automorphic representations corresponding to $S_k(\leqslant Q)^\sharp$ (recall that $\sharp$ means non-monomial) such that

$$|S(\leqslant Q)| = O(Q^d).$$

Then:

Corollary 5 *If $N = Q^\beta$ with $\beta > 2d + 2$, then for any $\varepsilon > 0$, it holds*

$$\sum_{f \in S(\leqslant Q)^\sharp} \left| \sum_{n \leqslant N} a_n \lambda_f(n) \right|^2 \ll N^{1+\varepsilon} \sum_{n \leqslant N} |a_n|^2$$

for any sequence $(a_n)_{n \leqslant N}$ of complex numbers.

Now for $n = 3$; if $f \in S_k(q)^\sharp$, its infinite component is fixed; let π_∞ be the representation of $GL(3, \mathbf{R})$ which is its local symmetric square (see [G-J]), and again $\eta = 1$.

From the Gelbart-Jacquet theory we see that the image of $S_k(\leqslant Q)^\sharp$ by the map $f \mapsto \pi_f^{(2)}$ already mentioned in Subsection 2.1 is contained in the corresponding $\mathrm{Aut}(q)$; we write again $S^{(2)}(\leqslant Q)^\sharp$ for the image of $S(\leqslant Q)^\sharp$ and apply Theorem 4. Here Q^2 replaces Q and since trivially

$$|S^{(2)}(\leqslant Q)^\sharp| = O(Q^d)$$

we get:

Corollary 6 *If $N = Q^\beta$ with $\beta > 2(d + 3)$, then for any $\varepsilon > 0$, it holds*

$$\sum_{\pi_f \in S^{(2)}(\leqslant Q)^\sharp} \left| \sum_{n \leqslant N} a_n \lambda_f^{(2)}(n) \right|^2 \ll N^{1+\varepsilon} \sum_{n \leqslant N} |a_n|^2$$

for any sequence $(a_n)_{n \leqslant N}$ of complex numbers.

References

[B-H] Bushnell, C.J., Henniart, G.: An upper bound on conductors for pairs, J. Number Theory **65**(2), 183–196 (1997)

[B-S] Brumer, A., Silverman, J.: The number of elliptic curves over **Q** with conductor N, Manuscripta Math. **91**, 95–102 (1996)

[D-H] Davenport, H., Heilbronn, H.: On the density of discriminants of cubic fields. II, Proc. Royal Soc., A **322**, 405–420 (1971) *or* Collected Works of Harold Davenport, vol. 2, 535–550, Academic Press (1977)

[Del] Deligne, P.: Formes modulaires et représentations de $GL(2)$, Modular Forms in One Variable IV, Springer Lecture Notes **749**, 55–105 (1972)

[DFI] Duke, W., Friedlander, J., Iwaniec, H.: Bounds for automorphic L-functions, II, Invent. Math. **115**, 219–239 (1994)

[D-I] Deshouillers, J.M., Iwaniec, H.: Kloosterman sums and Fourier coefficients of cusp forms, Invent. Math. **70**, 219–288 (1983)

[FNT] Fouvry, É., Nair, M., Tenenbaum, G.: L'ensemble exceptionnel dans la conjecture de Szpiro, Bull. Soc. Math. France **120**(4), 483–506 (1992)

[Gel] Gelbart, S.: Automorphic forms on adele groups, Princeton University Press (1975)

[G-J] Gelbart, S., Jacquet, H.: A relation between automorphic representations of $GL(2)$ and $GL(3)$, Ann. Sci. E.N.S 4ème série **11**, 471–552 (1978)

[H-L] Hoffstein, J., Lockhart, P.: Coefficients of Maass forms and the Siegel zero (with an appendix by D. Goldfeld, J. Hoffstein and D. Lieman), Ann. of Math. (2) **140**, 161–181 (1994)

[H-W] Hardy, G.H., Wright, E.M.: An introduction to the theory of numbers, Fifth Edition, Oxford University Press (1979)

[Iwa] Iwaniec, H.: Introduction to the Spectral Theory of Automorphic Forms, Biblioteca de la Revista Matemática Iberoamericana (1995)

[JPS] Jacquet, H., Piatetskii-Shapiro, I. I., Shalika, J. A.: Rankin-Selberg convolutions, Amer. J. Math. **105**, 367–464 (1983)

[Lin] Linnik, Yu. V.: The Large Sieve, Dokl. Akad. Nauk. SSSR **30**, 292–294 (1941), in Russian

[LRS] Luo, W., Rudnick, Z., Sarnak, P.: On Selberg's eigenvalue conjecture, Geom. Funct. Anal. **5**, 387–401 (1995)

[Luo] Luo, W.: Values of symmetric square L-functions at 1, J. Reine Angew. Math. **506**, 215–235 (1999)

[Miy] Miyake, T.: Elliptic Modular Forms, Springer-Verlag (1990)

[Mon] Montgomery, H.L.: Topics in Multiplicative Number Theory, Lect. Notes Math. **227**, Springer-Verlag (1971)

[M-W] Moeglin, C., Waldspurger, J.L.: Pôles des fonctions L de paires pour $GL(N)$, *appendice to* Le spectre résiduel de $GL(n)$, Ann. Sci. ENS (4ème série) **22**, 605–674 (1989)

[R-S] Rudnick, Z., Sarnak, P.: Zeros of principal L-functions and random matrix theory. A celebration of John F. Nash, Jr., Duke Math. J. **81**(2), 269–322 (1996)

[Ser] Serre, J.-P.: Quelques applications du théorème de densité de Chebotarev, Pub. Math. I.H.E.S **54**, 123–201 (1981)

[Shi] Shimura, G.: Introduction to the arithmetic theory of automorphic functions, Iwanami Shoten and Princeton Univ. Press (1971)

[Si1] Silverman, J.: The arithmetic of elliptic curves, Grad. Texts in Math. 106, Springer Verlag (1986)

[Si2] Silverman, J.: Advanced topics in the arithmetic of elliptic curves, Grad. Texts in Math. 151, Springer Verlag (1994)

[Wil] Wiles, A.: Modular elliptic curves and Fermat's last theorem, Ann. of Math. (2) **141**, 443–551 (1995)

Note added in proof: Breuil, Conrad, Diamond and Taylor have announced a proof of the full modularity conjecture for elliptic curves over **Q**. Corollary 4 thus holds unconditionally.

Appendix

Recovering modular forms from squares

Dinakar Ramakrishnan

253-37 Caltech, Pasadena, CA 91125, USA (e-mail: dinakar@cco.caltech.edu)

The purpose of this appendix is to provide a proof of the fact that a holomorphic newform f of weight $2k$, level N and trivial character, with Hecke eigenvalues $\{a_p \mid (p, N) = 1\}$, is determined up to a quadratic twist, in fact *on the nose* if N is square-free, by the knowledge of a_p^2 for all primes p in a set of sufficiently large density. We will in fact prove a more general statement (Theorem A) below, including the case of odd weight and non-trivial character, and also establish a mod ℓ analog, where the twisting character is shown to be unramified at ℓ. We found this result in the summer of 94, and we have since learned that Theorem A has also been known to others, including Don Blasius and J.-P. Serre. Also, Siman Wong has recently come up with a different proof in the weight 2 case (with trivial character). So we do not intend any display of great achievement by this write-up, and we give all the details for ease of use by those working in classical modular forms and number theory. We have also found a non-trivial extension of this result (in characteristic zero) to Maass forms using an array of results on automorphic $L-$functions, and this is the subject matter of another paper [Ra2].
This work was partially supported by an NSF grant. We thank Serre for his helpful comments on an earlier version which led to a finer result.

For every pair of integers $N, k \geqslant 1$, and character $\omega : (\mathbf{Z}/N)^* \to \mathbf{C}^*$, denote by $\mathcal{S}_k^{\text{new}}(N, \omega)$ the set of normalized newforms f of weight k, level N and character ω, with Hecke eigenvalues $a_p(f)$, for all p not dividing N, and corresponding $p-$Euler factors

$$L_p(s, f) = (1 - \alpha_p p^{-s})^{-1}(1 - \beta_p p^{-s})^{-1},$$

where $\alpha_p = \alpha_p(f)$ and $\beta_p = \beta_p(f)$ are non-zero algebraic integers satisfying

$$a_p(f) = \alpha_p + \beta_p, \qquad \text{and} \qquad \omega(p)p^{k-1} = \alpha_p\beta_p.$$

Let us set

$$L_p(s, \mathrm{Ad}(f)) = \left(1 - \frac{\alpha_p}{\beta_p} p^{-s}\right)^{-1} (1 - p^{-s})^{-1} \left(1 - \frac{\beta_p}{\alpha_p} p^{-s}\right)^{-1}.$$

Theorem A. *Let $f \in \mathcal{S}_k^{\mathrm{new}}(N, \omega)$ and $g \in \mathcal{S}_{k'}^{\mathrm{new}}(N', \omega')$, $k \geqslant k'$, be such that, for all primes p outside a set S of Dirichlet density $\delta(S) < \frac{1}{18}$, we have*

$$(*) \qquad L_p(s, \mathrm{Ad}(f)) = L_p(s, \mathrm{Ad}(g)).$$

Then $k = k'$, and there exists a Dirichlet character χ of conductor M dividing NN' such that

$$a_p(f) = a_p(g)\chi(p),$$

all p prime to NN'. In particular, $\omega = \omega'\chi^2$.

If f, g are not of CM type and have weights $k, k' \geqslant 2$, then the same conclusion results if (*) is assumed to hold only for a set of primes of positive density.

When f and g have the **same character**, we can deduce the stronger result below:

Corollary. *Let $f \in \mathcal{S}_k^{\mathrm{new}}(N, \omega)$ and $g \in \mathcal{S}_k^{\mathrm{new}}(N', \omega)$ be such that, for all primes p outside a set S of density $\delta(S) < \frac{1}{18}$, we have*

$$a_p(f)^2 = a_p(g)^2,$$

Then there exists a **quadratic** *character χ of conductor M dividing NN' such that*

$$a_p(f) = a_p(g)\chi(p),$$

for all p not dividing NN'. Moreover, if $\omega = 1$ and N, N' **square-free**, *then $f = g$.*

When f, g are not of CM type and of weight $\geqslant 2$, we get the same conclusion assuming only that $\delta(S)$ is < 1.

Theorem A $\Longrightarrow$ Corollary. The hypotheses imply that $(\alpha_p(f)/\beta_p(f)) + (\beta_p(f)/\alpha_p(f)) + 1$ equals $(\alpha_p(g)/\beta_p(g)) + (\beta_p(g)/\alpha_p(g)) + 1)$, for all p outside S. It is then easy to see that $L_p(s, \mathrm{Ad}(f))$ equals $L_p(s, \mathrm{Ad}(g))$, for all such p. So we may apply the Theorem and deduce the existence of a χ such that $a_p(f) = a_p(g)\chi(p)$, for all p prime to NN'. Comparing squares, we see that χ must be quadratic.

Next let N, N' be square-free, and ω trivial. Suppose χ is non-trivial. Denote by π, π' the cuspidal automorphic representations of $\mathrm{GL}(2, \mathbf{A_Q})$ of trivial central character associated to f, g respectively. Then, up to exchanging f and g if necessary, $N = N(\pi)$ must be $N(\pi' \otimes \chi)$, the conductor of $\pi' \otimes (\chi \circ \det)$. (Here we are identifying χ with the idèle class

character of $\mathbf{Q}$ it defines.) Since $N' = N(\pi')$ is square-free, and since π' has trivial central character, one sees easily from the description of local representations and their conductors in [Ge], p.73, that the $p-$component π'_p must be the unramified special (Steinberg) representation at every prime p dividing N'. One sees then, by using the same theorem (loc. cit.) that $\mathrm{ord}_p(N(\pi' \otimes \chi)) \geqslant 2$, for any p dividing the conductor M of χ. Since $\mathbf{Q}$ has class number 1, there are no unramified characters χ. In other words, $N = N(\pi' \otimes \chi)$ is not square-free, giving the desired contradiction.

□

Proof of Theorem A. We will in fact give **two proofs**. We fix a prime ℓ not dividing NN', and begin with the theorems of Deligne ([De], for $k \geqslant 3$), Eichler-Shimura ([Sh], for $k = 2$), and Deligne-Serre ([DS] for $k = 1$), giving the existence, for $h = f$ or g, of an irreducible, continuous representation

$$\sigma_\ell(h) : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\overline{\mathbf{Q}}_\ell),$$

such that, for any prime p not dividing $N\ell$,

$$\begin{aligned} \mathrm{tr}(\sigma_\ell(h)(Fr_p)) &= a_p(h) = \alpha_p(h) + \beta_p(h), \\ |\alpha_p(h)| &= |\beta_p(h)| = p^{(k(h)-1)/2}, \end{aligned}$$

and

$$\det(\sigma_\ell(h)) = \omega(h)\chi_{\mathrm{cyc}}^{k(h)-1}.$$

Here Fr_p denotes the Frobenius conjugacy class at p, $\overline{\mathbf{Q}}_\ell$ a fixed algebraic closure of $\mathbf{Q}_\ell$, and χ_{cyc} the cyclotomic character given by the Galois action on the inverse system of ℓ^m-th roots of unity. ($\omega(h)$ is ω or ω' depending on whether h is f or g; similarly for $k(h)$.) If we consider the field E generated by the coefficients of f, and a place λ of E above ℓ, then one has in fact a representation of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ into $\mathrm{GL}_2(E_\lambda)$, and our σ_ℓ is its extension to $\overline{\mathbf{Q}}_\ell$. We work over $\overline{\mathbf{Q}}_\ell$ because we will need to appeal to Schur's lemma.

For any two dimensional $\overline{\mathbf{Q}}_\ell-$representation σ_ℓ of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, set

$$\mathrm{Ad}(\sigma_\ell) = \mathrm{sym}^2(\sigma_\ell) \otimes \det(\sigma_\ell)^{-1}.$$

Theorem B. *Let K be a number field, and let σ_ℓ and σ'_ℓ be irreducible two dimensional $\overline{\mathbf{Q}}_\ell$-representations of $Gal(\overline{\mathbf{Q}}/K)$ with Frobenius traces a_P, a'_P (for almost all primes P) and conductors N, N' respectively. Suppose* $\mathrm{Ad}(\sigma_\ell) \simeq \mathrm{Ad}(\sigma'_\ell)$. *Then there exists $\psi_\ell \in Hom_{\mathrm{cont}}(\mathrm{Gal}(\overline{\mathbf{Q}}/K), \overline{\mathbf{Q}}_\ell^*)$ such that*

$$\sigma_\ell \simeq \sigma'_\ell \otimes \psi_\ell.$$

Next let $K = \mathbf{Q}$. Suppose we know either that σ_ℓ and σ'_ℓ are Hodge-Tate (see [Se1]) or that the ratio of their determinants is a finite order character times an even power of χ_{cyc}. Then

$$(**) \qquad \psi_\ell = \chi_{\mathrm{cyc}}^r \nu_\ell,$$

where r is an integer, and ν_ℓ the $\ell-$adic character defined by a Dirichlet character ν.

Theorem B $\Longrightarrow$ Theorem A. Let f, g be as in Theorem A. Since $\sigma_\ell(f)$ and $\sigma_\ell(g)$ are simple, $\mathrm{Ad}(\sigma_\ell(f))$ and $\mathrm{Ad}(\sigma_\ell(g))$ are semisimple, and we claim that they are isomorphic.

Modulo this claim, we proceed as follows. Applying the first part of Theorem B, we get a character ψ_ℓ such that $\sigma_\ell(f) \simeq \sigma_\ell(g)\otimes\psi_\ell$. Comparing determinants, we get for almost all p,

$$\text{(I)} \qquad \psi_\ell(Fr_p)^2 = \chi_{\mathrm{cyc}}(Fr_p)^{k-k'}\omega(p)\omega'(p)^{-1}.$$

At this point, one can use (at least) three different methods to finish the argument. The first uses a theorem of Faltings [Fa], which says that $\sigma_\ell(h)$ is Hodge-Tate for any newform h of conductor prime to ℓ. So, by the second part of Theorem B, k and k' are of the same parity, and we get (**) with $r = (k-k')/2$. Let $H(f)$ (resp. $H(g)$) be the $\mathbf{Q}$-Hodge structure of weight $k-1$ (resp. $k'-1$) associated to (the motive of) f (resp. g). Then we must have $H(f) \simeq H(g)(r)$, where $H(g)(r)$ denotes the Tate twist $H(g)\otimes\mathbf{Q}(r)$. Then r must be zero, since the Hodge type of $H(f)$ (resp. $H(g)$) is $\{(k-1,0),(0,k-1)\}$ (resp. $\{(k'-1,0),(0,k'-1)\}$), while that of $H(g)(r)$ is $\{(k'-1-r,-r),(-r,k'-1-r)\}$. Done.

The second method uses L-functions. Let ν be the finite order character defined as $\psi_\ell\chi_{\mathrm{cyc}}^{(k'-k)/2}$. Then by (I) we have, for every Dirichlet character μ, an identity

$$L_p(s, f\otimes\mu) = L_p(s-(k-k')/2, g\otimes\mu\nu),$$

for all p in the set T of all primes not dividing $\ell NN'$ and the conductor of μ. We may fix a μ, sufficiently ramified at the primes in T, such that the local factors of $f\otimes\mu$ and $g\otimes\nu\mu$ at any prime in T are 1. Interchanging f and g if necessary, we may assume that $k \leqslant k'$. Since the archimedean factor attached to $f\otimes\mu$ is $(2\pi)^{-s}\Gamma(s)$, and since its product with (the global Euler product) $L(s, f\otimes\mu)$ is entire, any pole of the Gamma factor results in a zero of $L(s, f\otimes\mu)$, which is $\prod_{p\notin T} L_p(s, f\otimes\mu)$ by the choice of μ. This happens for example at $s=0$, and consequently, by the identity above, $L(s+(k'-k)/2)$ has a zero at $s=0$, even though its the archimedean factor does not have a pole there (as $k'>k$). Then, by applying the functional equation for $g\otimes\mu\nu$ (which relates s to $k'-s$), we see that $L(s, \overline{g}\otimes\overline{\mu\nu})$ has a zero at $s=(k'+k)/2$. This is absurd (see [JS]) as this point is in the region (resp. on the boundary) of absolute convergence if $k>1$ (resp. $k=1$). So we must have $k=k'$.

The third method is to appeal, for ℓ large enough, to the mod ℓ result proved later in this appendix.

Now we prove the claim. The identity (*) says that the characteristic polynomials of the Frobenius classes Fr_p agree on $\mathrm{Ad}(\sigma_\ell(f))$ and $\mathrm{Ad}(\sigma_\ell(g))$, for all p outside a set S of density $\delta < \frac{1}{18}$. If $\delta(S) = 0$, then by the Tchebotarev density theorem, $\mathrm{Ad}(\sigma_\ell(f))$ and $\mathrm{Ad}(\sigma_\ell(g))$ would be equivalent, and our object is to get the same conclusion under the weaker hypothesis on δ. By [GJ], we know that, for $h = f$ or g, there is an (isobaric) automorphic representation $\mathrm{Ad}(h)$ of $\mathrm{GL}(3, \mathbf{A_Q})$, whose standard L-function identifies, after removing the archimedean factors, with $\prod_p L_p(s-1, \mathrm{Ad}(h))$. It suffices to show that $\mathrm{Ad}(f)$ and $\mathrm{Ad}(g)$ are isomorphic. Suppose not. Then we can find (isobaric) automorphic representations π, π' of $\mathrm{GL}(k, \mathbf{A_Q})$, $k \leqslant 3$, such that $\mathrm{Ad}(f) \simeq \pi \boxplus \eta$ and $\mathrm{Ad}(g) \simeq \pi' \boxplus \eta$, where η is an automorphic representation of $\mathrm{GL}(3-k, \mathbf{A_Q})$, taken to be 0 if $k = 3$. Let $Z_S(s)$ be as in equation (3) of [Ra1]. In the present case, if m (resp. r) denotes the number of cuspidals occurring in the isobaric decomposition [La] of π (resp. π'), necessarily with multiplicity 1, we have $-\mathrm{ord}_{s=1} Z_S(s) = m^2 + r^2$ (compare with (4) of [Ra1]). Since one knows the Ramanujan conjecture for holomorphic forms by Deligne, it is easy to verify that Lemma 2 of [Ra1] holds for π (resp. π') with β less than $k^2 m^2 \delta$ (resp. $k^2 r^2 \delta$). Then the argument of Sect. 2 of [Ra1] shows that we must have $1 \leqslant 2k^2\delta$. Since $\delta < 1/18$ and $k \leqslant 3$, we get the desired contradictiction.

It remains to treat the case when f, g are not of CM type and have weights $\geqslant 2$, with δ assumed to be just < 1. One knows by the works of Serre and Ribet [Ri] that $\sigma_\ell(f)$ is absolutely irreducible under restriction to any open subgroup. We note then that the same must be true for $\mathrm{Ad}(\sigma_\ell(f))$, as otherwise the restriction $\sigma_\ell(f)_K$ will, for some number field K, be induced by a character of $\mathrm{Gal}(\overline{\mathbf{Q}}/F)$, for a quadratic extension F/K (see below), making $\sigma_\ell(f)_F$ reducible. Now, applying Theorem 2 of [Raj] for example, we may conclude that, as $\delta < 1$, $\mathrm{Ad}(\sigma_\ell(f))$ must be isomorphic to $\mathrm{Ad}(\sigma_\ell(g)) \otimes \nu_\ell$, for some one-dimensional ν_ℓ of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ defined by a Dirichlet character. Let K be the cyclic extension of $\mathbf{Q}$ corresponding to ν_ℓ, and let τ be a generator of $\mathrm{Gal}(K/\mathbf{Q})$. Then, since $\mathrm{Ad}(\sigma_\ell(f)_K)$ and $\mathrm{Ad}(\sigma_\ell(g)_K)$ are isomorphic, we may apply Theorem B and conclude that $\sigma_\ell(f)_K \simeq \sigma_\ell(g)_K \otimes \lambda_\ell$, for a character λ_ℓ of $\mathrm{Gal}(\overline{\mathbf{Q}}/K)$. Since $\sigma_\ell(f)_K$ and $\sigma_\ell(g)_K$ are invariant under τ, we get

$$\sigma_\ell(g)_K \otimes (\lambda/\lambda^{[\tau]}) \simeq \sigma_\ell(g)_K.$$

Since $\sigma_\ell(g)$ is irreducible under restriction to any open subgroup, $\sigma_\ell(g)_K$ cannot admit any non-trivial self-twist, and λ must be invariant under τ and hence must extend to a character of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. The rest of the argument goes through as above, and Theorem A follows.

Proof of Theorem B. First we need a simple

Lemma. *Let ρ_ℓ be an irreducible, $n-$dimensional, self-dual $\overline{\mathbf{Q}}_\ell$-representation of Gal$(\overline{\mathbf{Q}}/K)$. Then there exists an invariant non-degenerate bilinear form B on (the space of) ρ_ℓ, which is symmetric or alternating, such that*

(i) B is unique up to a non-zero scalar; and

(ii) If ρ'_ℓ is another irreducible, $n-$dimensional, self-dual $\overline{\mathbf{Q}}_\ell$-representation of Gal$(\overline{\mathbf{Q}}/K)$ with invariant non-degenerate bilinear form B', such that ρ_ℓ and ρ'_ℓ are isomorphic, then they are **isometric** *relative to B and B'.*

Indeed, (i) and the statement above it are immediate consequences of Schur's lemma. Also, since $\overline{\mathbf{Q}}_\ell$ is algebraically closed, cB is isometric to B for any $c \in \overline{\mathbf{Q}}_\ell^*$; hence we get (ii) as well.

Now let σ_ℓ and σ'_ℓ be as in Theorem B. Suppose (the semisimple representation) $\mathrm{Ad}(\sigma_\ell)$ is reducible. Then it must contain a one dimensional summand η_ℓ, say. Then η_ℓ occurs in the (self-dual) $\mathrm{End}(\sigma_\ell) = \sigma_\ell \otimes \sigma_\ell^\vee = \mathrm{Ad}(\sigma_\ell) \oplus 1$. Schur's lemma above forces η_ℓ to be non-trivial. Either η_ℓ is quadratic, or otherwise $\eta_\ell^\vee$ will also occur in $\mathrm{End}(\sigma_\ell)$. In either case, we see that $\mathrm{End}(\sigma_\ell)$ must contain a quadratic character δ_ℓ, say; let F be the corresponding quadratic extension of K with non-trivial automorphism θ. Denote by $\sigma_{F,\ell}$ the restriction of σ_ℓ to Gal$(\overline{\mathbf{Q}}/F)$. We claim (as is well known) that if τ_ℓ is another semisimple representation of Gal$(\overline{\mathbf{Q}}/K)$ whose restriction to Gal$(\overline{\mathbf{Q}}/F)$ is isomorphic to $\sigma_{F,\ell}$, then $\tau_\ell \simeq \sigma_\ell \otimes \delta_\ell^j$, for $j \in \{0, 1\}$. Indeed, by the hypothesis, the restriction of $\eta_\ell := \tau_\ell \otimes \sigma_\ell^\vee$ to Gal$(\overline{\mathbf{Q}}/F)$ contains the trivial representation; so by Frobenius reciprocity, there is a non-trivial homomorphism between η_ℓ and the representation of Gal$(\overline{\mathbf{Q}}/K)$ induced by the trivial representation of Gal$(\overline{\mathbf{Q}}/F)$, which decomposes as $1 \oplus \delta_\ell$. So δ_ℓ^j occurs in η_ℓ, for $j = 0$ or 1. Equivalently, there is an intertwining operator between τ_ℓ and $\sigma_\ell \otimes \delta_\ell^j$, which implies the claim by virtue of the irreducibility of σ_ℓ. Next observe that $\sigma_{F,\ell}$ must be reducible as $\mathrm{End}(\sigma_{F,\ell})$ contains 1 with multiplicity 2 (as the restriction of δ_ℓ to Gal$(\overline{\mathbf{Q}}/F)$ is trivial). Write $\sigma_{F,\ell} = \nu_\ell \oplus \mu_\ell$, with ν_ℓ, μ_ℓ being one-dimensionals of Gal$(\overline{\mathbf{Q}}/F)$. We claim that ν_ℓ is not θ-invariant. Indeed, otherwise μ_ℓ would also be θ-invariant as $\sigma_{F,\ell}$ is, and both ν_ℓ and μ_ℓ would admit extensions to Gal$(\overline{\mathbf{Q}}/K)$ and result in a reducible extension of $\sigma_{\ell,F}$, which is impossible by the claim above. Thus ν_ℓ is not fixed by θ, and so we must have $\sigma_{F,\ell} \simeq \nu_\ell \oplus \nu_\ell^{[\theta]}$. This forces σ_ℓ to be the induced representation $\mathrm{Ind}_F^K(\nu_\ell)$, as this induced representation has the same restriction to Gal$(\overline{\mathbf{Q}}/F)$ as σ_ℓ and is moreover isomorphic to its twist by any character of Gal$(\overline{\mathbf{Q}}/K)$ trivial on Gal$(\overline{\mathbf{Q}}/F)$. Since $\mathrm{End}(\sigma_\ell) = \mathrm{End}(\sigma'_\ell)$, σ'_ℓ must also be of the form $\mathrm{Ind}_F^K(\nu'_\ell)$, for some one-dimensional ν'_ℓ of Gal$(\overline{\mathbf{Q}}/F)$. Since the determinant of $\mathrm{Ind}_F^K(\nu_\ell)$ is the transfer of ν_ℓ to

$\mathrm{Gal}(\overline{\mathbf{Q}}/K)$ times δ_ℓ, we see that

$$\mathrm{Ad}(\sigma_\ell) \simeq \mathrm{Ind}_F^K(\nu_\ell/\nu_\ell^{[\theta]}) \oplus \delta_\ell,$$

and similarly for $\mathrm{Ad}(\sigma'_\ell)$. This implies that, up to replacing ν_ℓ by $\nu_\ell^{[\theta]}$, we have

$$\nu_\ell/\nu_\ell^{[\theta]} = \nu'_\ell/(\nu'_\ell)^{[\theta]}.$$

Then ν_ℓ/ν'_ℓ is $\theta-$invariant, and hence extends to a character ψ_ℓ of $\mathrm{Gal}(\overline{\mathbf{Q}}/K)$. In other words, $\sigma_\ell \simeq \sigma'_\ell \otimes \psi_\ell$, as claimed.

We next consider the case when $\mathrm{Ad}(\sigma_\ell)$ and $\mathrm{Ad}(\sigma'_\ell)$ are irreducible. Let λ_ℓ denote the product of the determinants ω_ℓ, ω'_ℓ of σ_ℓ, σ'_ℓ respectively. Set

$$\eta_\ell := \sigma_\ell \otimes \sigma'_\ell.$$

Then

$$\mathrm{sym}^2(\eta_\ell) \otimes \lambda_\ell^{-1} \simeq \mathrm{Ad}(\sigma_\ell) \otimes \mathrm{Ad}(\sigma'_\ell) \oplus 1.$$

Since $\mathrm{Ad}(\sigma_\ell)$ and $\mathrm{Ad}(\sigma'_\ell)$ are irreducible, self-dual and isomorphic, 1 occurs in their tensor product. Hence the multiplicity of λ_ℓ is greater than 1 in $\mathrm{sym}^2(\eta_\ell)$, showing that η_ℓ is reducible. Now suppose η_ℓ contains a two dimensional summand τ_ℓ, say. Then the one dimensional $\det(\tau_\ell)$ occurs in the exterior square of η_ℓ. But on the other hand, we have

$$\Lambda^2(\eta_\ell) \simeq \mathrm{sym}^2(\sigma_\ell) \otimes \omega'_\ell \oplus \omega'_\ell \otimes \mathrm{sym}^2(\sigma'_\ell),$$

showing that, as the symmetric squares of σ_ℓ and σ'_ℓ are irreducible, there can be no one dimensional summand of $\Lambda^2(\eta_\ell)$. This shows that η_ℓ has no two dimensional summand. Since it is reducible, it must then have a one dimensional summand ν_ℓ, say. Then

$$\sigma_\ell \simeq \sigma_\ell'^{\vee} \otimes \nu_\ell \simeq \sigma'_\ell \otimes \omega_\ell'^{-1}\nu_\ell.$$

So we get the desired ψ_ℓ by taking it to be $\omega_\ell'^{-1}\nu_\ell$.

Now let $K = \mathbf{Q}$. Comparing determinants, we see that $\psi_\ell^2 = \det(\sigma_\ell)\det(\sigma'_\ell)^{-1}$. So we get (**) immediately if the ratio of the determinants is a finite order character times an even power of χ_{cyc}. Finally, suppose σ_ℓ and σ'_ℓ are Hodge-Tate. Then ψ_ℓ will also be Hodge-Tate as it occurs in $\sigma_\ell \otimes (\sigma'_\ell)^{\vee}$. Consequently, it corresponds to an algebraic Hecke character ψ. Since we are working over $\mathbf{Q}$, it must be a finite order character times a power of χ_{cyc}. Done.

For the second proof, we begin by recalling the fact that the adjoint representation Ad: $\mathrm{PGL}(2, \overline{\mathbf{Q}}_\ell) \longrightarrow \mathrm{GL}(3, \overline{\mathbf{Q}}_\ell)$ is isomorphic onto the special orthogonal group $\mathrm{SO}(3, \overline{\mathbf{Q}}_\ell)$. Denote by $\overline{\sigma}_\ell$ (resp. $\overline{\sigma}'_\ell$) the composite of σ_ℓ (resp. σ'_ℓ) with the natural homomorphism of $\mathrm{GL}(2, \overline{\mathbf{Q}}_\ell)$ onto $\mathrm{PGL}(2, \overline{\mathbf{Q}}_\ell)$.

Then it is easy to see that $\mathrm{Ad}(\overline{\sigma}_\ell)$ identifies with the $\mathrm{Ad}(\sigma_\ell)$ defined earlier (above Theorem B). So, by our hypothesis, we get two representations, namely $\mathrm{Ad}(\overline{\sigma}_\ell)$ and $\mathrm{Ad}(\overline{\sigma}'_\ell)$, into $\mathrm{SO}(3, \overline{\mathbf{Q}}_\ell)$, which are equivalent in $\mathrm{GL}(3, \overline{\mathbf{Q}}_\ell)$. Suppose they are irreducible. Then we may apply part (ii) of the Lemma and deduce that they are in fact isometric. By changing the isometry by $-I$ if necessary, we may assume that they are equivalent in $\mathrm{SO}(3, \overline{\mathbf{Q}}_\ell)$. Since Ad is an isomorphism, σ_ℓ and σ'_ℓ define equivalent homomorphisms into $\mathrm{PGL}(2, \overline{\mathbf{Q}}_\ell)$. Hence σ_ℓ must be equivalent to $\sigma'_\ell \otimes \psi_\ell$, for some $\psi_\ell \in \mathrm{Hom}(\mathrm{Gal}(\overline{\mathbf{Q}}/K), \overline{\mathbf{Q}}_\ell^*)$. When $\mathrm{Ad}(\sigma_\ell)$ is reducible, one uses explicit arguments as in the reducible case of the first proof to conclude that $\mathrm{Ad}(\sigma_\ell)$ and $\mathrm{Ad}(\sigma'_\ell)$ are isometric. The rest follows. □

The mod ℓ version. For each newform f, let K_f denote the number field generated by the coefficients of f. If g is another newform, let $\mathfrak{O}_{f,g}$ denote the ring of integers of the compositum $K_f K_g$. For $h = f$ or g, write for p not dividing the level,

$$Q_h(T) = \left(1 - \frac{\alpha_p(h)}{\beta_p(h)}T\right)(1-T)\left(1 - \frac{\beta_p(h)}{\alpha_p(h)}T\right),$$

so that $L_p(s, \mathrm{Ad}(h)) = Q_h(p^{-s})^{-1}$. Note that, since $\alpha_p(h)\beta_p(h) = \omega(h)p^{k(h)-1}$, $\alpha_p(h)$ and $\beta_p(h)$ are invertible modulo any prime ℓ not dividing $pN(h)$.

Theorem C. *Let ℓ be an odd prime number and N, N' positive integers prime to ℓ. Let f (resp. g) be a newform of level N (resp. N'), weight k (resp. k'), and character ω (resp. ω'). Let λ be a prime ideal above ℓ in $\mathfrak{O}_{f,g}$. Suppose we have*

$$(C) \qquad Q_f(T) \equiv Q_g(T) \pmod{\lambda},$$

for all p outside a set S (containing the primes divisors of $\ell NN'$) of density 0. Then $k \equiv k' \pmod{\ell - 1}$, and there exists a character β, unramified at ℓ, such that

$$a_p \equiv b_p\beta(p) \pmod{\lambda},$$

for all p not dividing $\ell NN'$.

Remark. Note that if ω and ω' are the same mod λ, and if $k - k' \equiv 0 \pmod{\ell - 1}$, the hypothesis (C) is equivalent to the congruence

$$a_p^2 \equiv b_p^2 \pmod{\lambda}.$$

In this case β is necessarily quadratic. Moreover, if N and N' are in addition square-free, one can conclude (as in the characteristic zero case) that β is trivial.

Proof. Let $\mathbf{F}_\lambda$ denote the residue field $\mathfrak{O}_{f,g}/\lambda$. Reducing the (integrally defined) ℓ-adic representations associated to f, g modulo λ and extending scalars to $\overline{\mathbf{F}_\lambda}$, we get representations

$$\overline{\sigma}_\lambda : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\overline{\mathbf{F}_\lambda})$$

and

$$\overline{\sigma}'_\lambda : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\overline{\mathbf{F}_\lambda})$$

such that, for all p not dividing $NN'\ell$, $\mathrm{tr}(\overline{\sigma}_\lambda(\mathrm{Fr}_p))$ (resp. $\mathrm{tr}(\overline{\sigma}'_\lambda(\mathrm{Fr}_p))$) is the image of a_p (resp. b_p) in $\overline{\mathbf{F}_\lambda}$. Moreover, by hypothesis, $\det(\overline{\sigma}_\lambda)$ and $\det(\overline{\sigma}'_\lambda)$ both equal $\chi^{k-1}\overline{\omega}$ (resp. $\chi^{k'-1}\overline{\omega}'$), where $\chi : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{F}_\ell^*$ is the mod ℓ cyclotomic character and $\overline{\omega}$ (resp. $\overline{\omega}'$) the reduction (mod λ) of ω (resp. ω'). Clearly, the images of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ under these two representations are finite.

For any $\overline{\mathbf{F}_\lambda}$-representation τ_λ of a finite group G of dimension d, let $\tau_\lambda^{\mathrm{ss}}$ denote its semisimplification. Note that in characteristic ℓ, the semisimplification is determined by the characteristic polynomials of $\tau_\lambda(g)$ for all g in G when $d > \ell$, and also when $d = \ell = 3$ if τ_λ is orthogonal of determinant 1.

By the hypothesis (C), the characteristic polynomials of Fr_p in the adjoint representations of $\overline{\sigma}_\lambda$ and $\overline{\sigma}'_\lambda$ are the same for all p in a set of density 1. Thus, by the Tchebotarev density theorem and the remark above, we see that

$$\mathrm{Ad}(\overline{\sigma}_\lambda^{\mathrm{ss}}) \simeq \mathrm{Ad}(\overline{\sigma}_\lambda'^{\,\mathrm{ss}}).$$

Since $\mathrm{End}(\overline{\sigma}_\lambda^{\mathrm{ss}})$ (resp. $\mathrm{End}(\overline{\sigma}_\lambda'^{\,\mathrm{ss}})$) is $\mathrm{Ad}(\overline{\sigma}_\lambda^{\mathrm{ss}}) \oplus 1$ (resp. $\mathrm{Ad}(\overline{\sigma}_\lambda'^{\,\mathrm{ss}}) \oplus 1$), it follows that $\overline{\sigma}_\lambda$ is irreducible iff $\overline{\sigma}'_\lambda$ is.

First suppose that $\overline{\sigma}_\lambda$ and $\overline{\sigma}'_\lambda$ are irreducible. In this case the detailed ℓ-adic argument given in the proof of (the first part of) Theorem B goes through, with $\overline{\mathbf{Q}}_\ell$ replaced everywhere by $\overline{\mathbf{F}_\lambda}$, once one notes the availability of the relevant form of the Frobenius reciprocity in characteristic ℓ (cf. [A], chap. III, Lemma 6) and the fact that the tensor square of a simple Galois module is semisimple [Se3]. One deduces an isomorphism of $\overline{\sigma}_\lambda$ with $\overline{\sigma}'_\lambda \otimes \nu_\lambda$, for some character ν_λ of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ into $\overline{\mathbf{F}_\lambda}$. Since ω_λ and ω'_λ are the same modulo λ, we see by comparing determinants that ν_λ^2 is $\chi^{k-k'}\overline{\omega}/\overline{\omega}'$. We may write ν_λ as $\chi^j\beta_\lambda$, for some $j \in \{0, \ldots, \ell-2\}$, and a character β_λ unramified at ℓ. Consequently, $k - k' \equiv 2j \pmod{\ell-1}$, $\beta_\lambda^2 = \overline{\omega}/\overline{\omega}'$, and

$$(***) \qquad \overline{\sigma}_\lambda \simeq \overline{\sigma}'_\lambda \otimes \chi^j\beta_\lambda.$$

Let G_ℓ denote the decomposition group at ℓ of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, and let I denote the inertia subgroup. When a_ℓ is not zero modulo λ, one knows by Deligne (cf. [E], Theorem 2.5, for example), that $\overline{\rho}_\lambda|_{G_\ell}$ is reducible, and its semisimplification is of the form $\chi^{k-1}\mu_{1,\lambda} \oplus \mu_{2,\lambda}$, where each $\mu_{j,\lambda}$ is unramified. When a_ℓ is divisible by λ, a result of Fontaine (see [E], Theorem 2.6) asserts that the restriction to G_ℓ is irreducible, while the restriction to I

demposes as $\psi^{k-1} \oplus \psi'^{k-1}$, where ψ, ψ' are the two fundamental characters of level 2 [Se2]. Similarly for the restriction of $\overline{\sigma}'_\lambda$ at ℓ. In either case, we see that the only way (***) can hold is for j to be 0 modulo $\ell - 1$.

It remains to consider when $\overline{\sigma}_\lambda$ (and hence $\overline{\sigma}'_\lambda$) is reducible. Here we may write

$$\overline{\sigma}_\lambda^{\mathrm{ss}} \simeq \eta_\lambda \oplus \chi^{k-1}\overline{\omega}/\eta_\lambda,$$

and

$$\overline{\sigma}_\lambda'^{\,\mathrm{ss}} \simeq \eta'_\lambda \oplus \chi^{k'-1}\overline{\omega}/\eta'_\lambda,$$

for some $\overline{\mathbf{F}_\lambda}^*$-valued characters η_λ, η'_λ of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. Then we have

$$\mathrm{Ad}(\overline{\sigma}_\lambda^{\mathrm{ss}}) \simeq \eta_\lambda^2/\overline{\omega}\chi^{k-1} \oplus 1 \oplus \overline{\omega}\chi^{k-1}/\eta_\lambda^2,$$

and

$$\mathrm{Ad}(\overline{\sigma}_\lambda'^{\,\mathrm{ss}}) \simeq \eta_\lambda'^{\,2}/\overline{\omega}\chi^{k'-1} \oplus 1 \oplus \overline{\omega}\chi^{k'-1}/\eta_\lambda'^{\,2}.$$

Since Ad commutes with semisimplification, it follows, after possibly replacing η_λ with $\chi^{k-1}\overline{\omega}/\eta_\lambda$, that $\eta_\lambda^2/\chi^k = \eta_\lambda'^{\,2}/\chi^{k'}$. Arguing as above, we see that η_λ is of the form $\eta'_\lambda\chi^j\beta_\lambda$, for some $j \in \{0, \dots, \ell-2\}$ with $k - k' \equiv 2j (\mathrm{mod}\ \ell - 1)$, and a character β_λ :$\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \overline{\mathbf{F}_\lambda}^*$, unramified at ℓ, such that $\beta_\lambda^2 = \overline{\omega}/\overline{\omega}'$. We obtain

$$\overline{\sigma}_\lambda'^{\,\mathrm{ss}} \simeq \eta'_\lambda\beta_\lambda\chi^{(k-k')/2} \oplus \beta_\lambda\chi^{(k+k')/2-1}\overline{\omega}'/\eta'_\lambda.$$

The reducibility of $\overline{\sigma}_\lambda$ (resp. $\overline{\sigma}'_\lambda$) forces a_ℓ (resp. b_ℓ) to be non-zero modulo λ, as the restriction of $\overline{\sigma}_\lambda^{ss}$ (resp. $\overline{\sigma}_\lambda'^{\,ss}$) to I must then be given by a direct sum of characters of level 1 [Se2]. Applying Deligne's result on the shape of the restriction to G_ℓ (see above), we see that the only possibility is for k and k' to be congruent modulo $\ell - 1$. Then $\overline{\sigma}_\lambda^{ss}$ is isomorphic to $\overline{\sigma}_\lambda'^{\,ss} \otimes \beta_\lambda$. Done.

References

[A] J.L. Alperin, Local representation theory, Cambridge Studies in Advanced Math. II (1986)

[De] P. Deligne, Formes modulaires et représentations $\ell-$adiques, Sém. Bourbaki 1968/69, no. 355, Springer Lecture Notes **179**

[DS] P. Deligne, J.-P.Serre, Formes modulaires de poids 1, Ann. Scient. ENS (4) **7**, (1974), 507–530

[E] B. Edixhoven, The weight in Serre's conjectures on modular forms, Invent. math. **109** (1992), 563–594

[Fa] G. Faltings, Hodge-Tate structures for modular forms, Math. Ann. **278** (1987), 133–149

[Ge] S. Gelbart, Automorphic Forms on Adele Groups, Ann. Math. Studies **83**, Princeton (1975)

[GJ] S. Gelbart, H. Jacquet, A relation between automorphic representations of GL(2) and GL(3), Ann. Sci. ENS (4) **11** (1979), 471–542

[La] R.P. Langlands, Automorphic representations, Shimura varieties and Motives. Ein Märchen, Proc. Symp. Pure Math. **33**, part 2, AMS (1979), 205–246

[JS] H. Jacquet, J. Shalika, A non-vanishing theorem for zeta functions of GL_n, Invent. math. **38** (1976/77), 1–16
[Ra1] D. Ramakrishnan, A refinement of the strong multiplicity one theorem for GL(2), appendix to ℓ-adic representations associated to modular forms over imaginary quadratic fields. II by R. Taylor, Invent. math. **116** (1994), 645–649
[Ra2] D. Ramakrishnan, Modularity of the Rankin-Selberg L-series, and multiplicity one for SL(2), Ann. of Math., to appear
[Raj] C.S. Rajan, On strong multiplicity one for ℓ-adic representations, IMRN, no. 3 (1998), 161–172
[Ri] K. Ribet, Galois representations attached to eigenforms with nebentypus, in Modular functions of one variable V, Springer Lecture Notes **601**, 17–51
[Se1] J.-P.Serre, Abelian $\ell-$adic Representations, Benjamin Press, NY (1968)
[Se2] J.-P. Serre, Sur les représentations de degré 2 de Gal($\overline{\mathbf{Q}}/\mathbf{Q}$), Duke Math. J. **54** (1987), no. 1, 179–230
[Se3] J.-P. Serre, Sur les semi-simplicité des produits tensoriels de représentations de groupes, Invent. math. **116** (1994), 513–530
[Sh] G. Shimura, Arithmetic Theory of Automorphic Functions, Princeton University Press (1971)

Note added in proof: Blasius has recently brought to my attention that his argument for Theorem A (minus the density assertion) can be found on pages 90–91 of "Higher regulators, Hilbert modular surfaces, and special values of L-functions" by G. Kings, Duke Math. J. **92**, no. 1 (1998), 61–127.