THE PETERSSON/KUZNETSOV TRACE FORMULA WITH PRESCRIBED LOCAL RAMIFICATIONS

YUEKE HU

Abstract. In this paper we derive refined Petersson/Kuznetsov trace formulae with prescribed local ramifications. The spectral side of these formulae are much shorter than the standard versions. We use them to study the first moment and the subconvexity bound of certain Rankin-Selberg L-function in a hybrid setting.

Contents

1. Introduction

The Petersson and the Kuznetsov trace formulae are very close in nature, and they can be both derived from a relative trace formula as in [14][13], by integrating pretrace formula against characters over unipotent subgroups, with the difference coming only from the Archimedean component. They have been important tools in analytic number theory to study various types of problems like the moments of L-functions and their subconvexity bound. See for example [1] for a survey.

In this paper we derive refined Petersson/Kuznetsov trace formulae with prescribed local ramifications. More precisely, the spectral side of these formulae consists of newforms which are associated to automorphic representations whose local component at a given place p belongs to a small family of supercuspidal representations or principal series representations.

We shall use them to study the first moment of the Rankin-Selberg $L-$function. In the special case where we know the positivity of the $L-$functions, we further obtain hybrid subconvexity bounds, which is as strong as the Weyl bound in a relatively wide range.

1.1. the classical trace formulae.

Consider for simplicity the classical Petersson trace formula, which is slightly easier to describe:

$$\frac{\Gamma(\kappa-1)}{(4\pi)^{\kappa-1}}\sum_{\varphi}\frac{\lambda_{m_1}(\varphi)\overline{\lambda_{m_2}(\varphi)}}{||\varphi||^2}=N[\delta_{m_1=m_2}+2\pi i^{-\kappa}\sum_{N|c}\frac{\mathrm{KL}(m_1,m_2,c)}{c}J_{\kappa-1}\left(\frac{4\pi\sqrt{m_1m_2}}{c}\right)]. \tag{1.1}$$

Here the sum of φ is over an orthonormal basis (with respect to (2.1)) of holomorphic cuspidal automorphic forms of weight κ, level N and trivial nebentypus. $\lambda_m(\varphi)$ is the $m-$th normalized Fourier coefficient. $\mathrm{KL}(m_1,m_2,c)$ is the classical Kloosterman sum with conductor c:

$$\mathrm{KL}(m_1,m_2,c)=\sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times}e\left(\frac{m_1x+m_2\overline{x}}{c}\right), \tag{1.2}$$

where $\overline{x}$ is the inverse of x in $(\mathbb{Z}/c\mathbb{Z})^\times$. $J_{\kappa-1}$ is a J-Bessel function. The Kloosterman sum can be written as a product of local Kloosterman sums. The formula (1.1) can be obtained from the relative trace formula where the test function f_p at $p|N$ is chosen to be essentially the characteristic function of a congruence subgroup. The $\delta_{m_1=m_2}$ term comes from the first-cell terms in the Bruhat decomposition, and the Kloosterman sum parts come from second-cell terms. See Section 4 or [14][13] for general settings.

Remark 1.1. In applications to depth aspect problems, there are however two major issues with (1.1):

(1) There is an asymmetry between the Archimedean aspect and the level aspect. More precisely, in the Archimedean aspect, the analytic conductor of φ is roughly k^2, whereas the length of the sum in φ is roughly k. On the other hand in the level aspect, the finite conductor of φ is N, whereas the length of the spectral sum is also roughly N. Thus the spectral sum is much longer in the level aspect in terms of the relation with the conductor.

(2) (1.1) picks out newforms as well as old forms on the spectral side. So it is not convenient to use when aiming only for newforms. Contributions from old forms have to be subtracted, which usually make computations more complicated, and also more tricky for depth aspect problems. For this reason, many results using classical approach deal with square-free or even prime levels only.

1.2. **Main results.** For simplicity, we shall be interested in automorphic representation π over $\mathbb{Q}$ with trivial central character and $N = C(\pi) = p^{\mathfrak{c}}$ for some integer $\mathfrak{c} \to \infty$ and $p \neq 2$. A local irreducible smooth representation π_θ at p will be either a supercuspidal representation or a principal series representation, associated to a character θ over some étale quadratic algebra $\mathbb{L}/\mathbb{Q}_p$ by compact induction or parabolic induction as in Section 3.

1.2.1. *the refined Petersson trace formula.* Let $\mathcal{F}_\theta[n]$ be the set of holomorphic newforms of weight κ, level $N = p^{\mathfrak{c}}$ with $\mathfrak{c} \geq 3$, and trivial nebentypus, whose associated local representation π_p belongs to a 'neighborhood' $\pi_\theta[n]$. Equivalently, π_p is associated to $\theta' \in \theta[n]$. See Definition 3.1, 3.2 for precise meanings for $\theta[n]$ and $\pi_\theta[n]$. For the test function given in Section 4.1, we have the following:

Theorem 1.2 (Theorem 4.16)**.**

$$\sum_{\varphi\in\mathcal{F}_\theta[l_0]} \frac{1}{||\varphi||^2}\lambda_{m_1}(\varphi)\,\overline{\lambda}_{m_2}(\varphi) = C_{\mathcal{F}}[l_0]\frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}\left(\delta_{m_1=m_2} + 2\pi i^\kappa \sum_{c_0|c}\frac{G\left(m_1, m_2, \theta, c^{-2}\right)}{c}J_{\kappa-1}\left(\frac{4\pi\sqrt{m_1m_2}}{c}\right)\right)$$

Here $C_{\mathcal{F}}[l_0] \asymp_p N^{1/2}$ is given in (4.38). $l_0 = 0, 1$ depending on $\mathbb{L}$ as in (4.37). c_0 is given in Definition 4.15, and is roughly $p^{\mathfrak{c}/2}$.

$G\left(m_1, m_2, \theta, c^{-2}\right)$ is the generalized Kloosterman which is a product of local factors as in Definition 4.14, where the local factors at $v \neq p$ are the same as the standard Kloosterman, while the local factor $G_p\left(m_1, m_2, \theta, c^{-2}\right)$ given in Lemma 4.5/Definition 4.11 involves the character θ and an integration inside $\mathbb{L}^\times$.

Remark 1.3. The square-root-cancellation type upper bounds are proven in Lemma 4.5, 4.12. The implied constants can depend on some fixed powers of p. But it should be possible to remove this dependence by a more careful study of character sums over residue fields.

We will also explain in Remark 4.7, 4.13 that $G_p\left(m_1, m_2, \theta, c^{-2}\right)$ becomes the standard Kloosterman sum when $v_p(c) \geq \mathfrak{c}$.

Remark 1.4. The main advantage of Theorem 1.2 is that it addresses both issues mentioned in Remark 1.1: it picks out only newforms; the length of the spectral side and the first-cell term have size $C_{\mathcal{F}}[l_0] \asymp N^{1/2}$ compared to N in (1.1). There are also two trade-offs:

(1) The generalized Kloosterman sums are more complicated than the standard Kloosterman sum to analyze;
(2) The length of the sum of Kloosterman sums is longer, in the sense that in Theorem 1.2 $v_p(c) \geq v_p(c_0)$ which is roughly $\frac{\mathfrak{c}}{2}$, while in (1.1) $v_p(c) \geq \mathfrak{c}$.

We shall develop tools and tricks to mitigate these disadvantages. For example, we already discussed the square-root cancellation for the generalized Kloosterman sum in Remark 1.3; In Theorem 1.5 we shall develop a formula picking out a larger family with shorter sum of Kloosterman sums, helping us to reach a balance between the first-cell term and the second-cell terms; In Section 1.3.2 we shall discuss alternative perspective for the generalized Kloosterman sum, and how to deal with the character sum after applying the Voronoï summation, which is a commonly used combo after the Petersson/Kuznetsov trace formula in dealing with many analytic number theory problems.

1.2.2. *Spectral average.* Let l be an integer such that $l_0 \leq l < i_0$, where i_0 is given in Definition 3.1 and is roughly $\frac{\mathfrak{c}}{2}$. Let $c_l = c_0 p^{l-l_0}$, and $\mathcal{F}_\theta[l]$ be as above. Then we have the following:

Theorem 1.5 (Theorem 4.18)**.**

(1.3)

$$\sum_{\varphi\in\mathcal{F}_\theta[l]} \frac{1}{\|\varphi\|^2}\lambda_{m_1}(\varphi)\overline{\lambda}_{m_2}(\varphi) = C_{\mathcal{F}}[l]\frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}\left(\delta_{m_1=m_2} + 2\pi i^\kappa \sum_{c_l|c} \frac{G\left(m_1,m_2,\theta,c^{-2}\right)}{c} J_{\kappa-1}\left(\frac{4\pi\sqrt{m_1 m_2}}{c}\right)\right)$$

Here $C_{\mathcal{F}}[l] \asymp C_{\mathcal{F}}[l_0]p^{l-l_0}$ *is given in* (4.42).

Remark 1.6. One can obtain Theorem 1.5 from Theorem 1.2 by taking a sum. The nontrivial part is however to show that the length of the sum of Kloosterman sums becomes shorter, which comes from a local cancellation. Theorem 1.5 displays a nice transition from Theorem 1.2 to the classical formula (1.1).

1.2.3. *Refined Kuznetsov trace formula.* In this case, let $\mathcal{F}_\theta[n]$ now be the similar set of cuspidal Maass newforms of level $N = p^{\mathfrak{c}}$, trivial nebentypus, whose local component $\pi_p \in \pi_\theta[n]$.

The residue spectrum will not be picked out by our choice of the test function. The contribution from the continuous spectrum will be nontrivial only when allowed π_p is a principal series representation. For this reason, let

$$\epsilon_{\mathbb{L}} = \begin{cases} 1, & \text{if } \mathbb{L} \simeq \mathbb{Q}_p \times \mathbb{Q}_p; \\ 0, & \text{otherwise.} \end{cases} \tag{1.4}$$

When $\epsilon_{\mathbb{L}} = 1$, for each finite order Hecke character $\theta' = (\chi, \chi^{-1})$ such that θ' is unramified when $v \neq p$, and $\theta'_p \in \theta[n]$, define $\varphi_s \in \pi(\chi|\cdot|^s, \chi^{-1}|\cdot|^{-s})$ to be a flat section associated to a L^2-normalized newform, and define

$$E_{\theta',s}(g) = \sum_{\gamma\in B(\mathbb{Q})\backslash \mathrm{GL}_2(\mathbb{Q})} \varphi_s(\gamma g).$$

Then using the general setup from [13], together with the test function at p and the relevant computations for the refined Petersson trace formula above, one can get the following:

Theorem 1.7 (Kuznetsov for prescribed local component)**.** *For* $l_0 \leq l < i_0$*, we have*

(1.5)

$$\sum_{\varphi\in\mathcal{F}_\theta[l]} \frac{\lambda_{m_1}(\varphi)\overline{\lambda_{m_2}(\varphi)}}{\|\varphi\|^2}\frac{h(t_\varphi)}{\cosh(\pi t_\varphi)} + \frac{\epsilon_{\mathbb{L}}}{\pi}\sum_{\theta',\theta'_p\in\theta[l]}\int_{-\infty}^{\infty} \lambda_{m_1}(E_{\theta',s})\overline{\lambda_{m_2}(E_{\theta',s})}h(t)dt$$

$$= C_{\mathcal{F}}[l][\frac{\delta(m_1=m_2)}{\pi^2}\int_{-\infty}^{\infty} h(t)\tanh(\pi t)tdt + \frac{2i}{\pi}\sum_{c_l|c}\frac{G\left(m_1,m_2,\theta,c^{-2}\right)}{c}\int_{-\infty}^{\infty} J_{2it}\left(\frac{4\pi\sqrt{m_1m_2}}{c}\right)\frac{h(t)t}{\cosh(\pi t)}dt].$$

Here h is an even test function with sufficient decay. t_φ *is the spectral parameter of* φ *such that* $\Delta\varphi = (1/4 + t_\varphi^2)\varphi$ *for the Laplace operator* Δ.

Remark 1.8. Note that it is possible to compute $\lambda_m(E_{\theta',s})$ more explicitly in terms of the twisted divisor function and the $L-$function for Hecke characters. We skip the details here.

1.2.4. *Application to the first moment and the hybrid subconvexity bound for the Rankin–Selberg L-function.* We expect several possible applications for the above theorems. One of them is to the vertical Sato-Tate law. Using Theorem 1.2, 1.5 or 1.7, the bound for the generalized Kloosterman sum discussed in Remark 1.3, and the recipe in [2], one should be able to get some variants of the vertical Sato-Tate law for small families of newforms in the depth aspect.

We also wish to explore other possible applications in future works. In this paper we focus on the first moment of the Rankin–Selberg L-functions.

Theorem 1.9. *Let $\mathcal{F}_\theta[l]$ be the set of holomorphic newforms of weight $\kappa \geq 2$, level $N = p^c$ and trivial nebentypus as above. Let g be a self-dual holomorphic cuspidal newform, with square-free level M which is coprime to N, fixed weight $\kappa_g \geq 2$, and central character χ. Then we have*

$$\sum_{f\in\mathcal{F}_\theta[l]} \frac{L(f\times g,1/2)}{\|f\|^2} \ll_{p,\epsilon} (MN)^\epsilon \left(N^{1/2}p^l + N^{1/4}M^{1/2}p^{-l/2}\right). \tag{1.6}$$

Furthermore suppose that $L(f \times g, 1/2) \geq 0$ for all $f \in \mathcal{F}_\theta[l]$. Suppose that $N = M^\delta$ for $0 < \delta < \infty$, so that the finite conductor $C(f \times g) = M^{2+2\delta}$ for any $f \in \mathcal{F}_\theta[l]$. By picking l to be the closest integer to $\log_p\left(M^{1/3}N^{-1/6}\right)$ while $1 \leq l < i_0$, we get that

$$L(f\times g,1/2) \ll_{\epsilon,p} M^{\max\{\frac{1}{2},\frac{1+\delta}{3},\frac{\delta}{2}\}+\epsilon}. \tag{1.7}$$

In particular we obtain a hybrid subconvexity bound for δ in any compact subset of $(0,\infty)$, which is furthermore a Weyl bound in the range $1/2 \leq \delta \leq 2$.

Remark 1.10. The condition that $L(f \times g, 1/2) \geq 0$ for all f can be guaranteed when, for example, g is dihedral. See the discussion in [6, Section 1.1].

Remark 1.11. Note that from the proof in Section 6, the $N^{1/2}p^l$ part comes from the first-cell terms and $N^{1/4}M^{1/2}p^{-l/2}$ part comes from the second-cell terms. Thus it is actually possible to obtain an asymptotic formula when $N^{1/2}p^{3l}$ is sufficiently large compared to M.

Remark 1.12. Compared to [5] [6], this result has two differences/improvements. First of all, [5] [6] assume N to be square-free. Secondly, they obtain a Weyl-type bound only at $\delta = 1/2$.

We make a more detailed comparison of the method in this paper with the one used in [9] (which extends [5] in some sense). The current method has the following advantages:

(I) It made use of the flexibility of Theorem 1.5, and the resulting subconvexity bound in Theorem 1.9 is stronger than both [9, Theorem 1.8] (which obtains Weyl-type strength at $\delta = 2$) and the analogue of [6, Corollary 1], allowing Weyl-type subconvexity bound in a wide hybrid range.
(II) It covers the case where π_p is a principal series representation. The treatments for the principal series case and the supercuspidal case are relatively uniform.
(III) It does not require any $\epsilon-$value condition for the Archimedean components.
(IV) The refined Petersson/Kuznetsov trace formulae are probably applicable to many other problems.

The method in [9] involves using the relative trace formula associated to Waldspurger's period integral on some quaternion algebra. This quaternion algebra is assumed to be a division algebra at all Archimedean places (which translates into $\epsilon-$value conditions). The method there has the following advantages:

(i) It does not require M to be square-free.

(ii) It is also used to prove a hybrid subconvexity bound [9, Theorem 1.10] in the joint ramification case.
(iii) It works for general number fields, and does not rely on the Ramanujan conjecture.

We do believe that some of the differences are amenable with extra work. For example, (I)-(III) may also be achieved by the method of [9]. On the other hand, (i) (ii) may also be recovered using the method in this paper, by employing a more flexible version of the Voronoï summation formula.

1.3. **Basic strategies.**

1.3.1. *Deriving the refined Petersson/Kuznetsov trace formula.* The classical formula (1.1) can be obtained by setting the local test function for the relative trace formula to be the characteristic function of the related congruence subgroup for the newform, as is done in [14][13]. The first idea to derive Theorem 1.2 is relatively straightforward, that is, to use instead suitable cut-off of the local matrix coefficient for the newform as the test function.

The matrix coefficient itself however is not very convenient to directly make use of. So far we have some understandings about its support, level (from [8, Proposition 2.12]) and size (from [11, Theorem 5.4]).

Our approach in this paper is to make use of the special test vectors, i.e., the minimal vectors for the supercuspidal representations discussed in [10] [9] and the microlocal lifts for the principal series representations discussed in [16]. These test vectors have the property that a large compact open subgroup acts on them by a character $\tilde{\theta}$, which can uniquely identify the test vector only from members in $\pi_\theta[l_0]$ (See Proposition 3.13/Corollary 3.23 for more details). Using the relation between these special test vectors and the newforms in Corollary 3.16/Lemma 3.24, we construct test functions in Definition 3.19, 3.26 from a linear combination of translates of $\tilde{\theta}$, which exactly pick out the newforms from $\pi_\theta[l_0]$. See Proposition3.20, 3.27.

The second-cell terms from the relative trace formula for the constructed local test function can be reduced to the computations for $\tilde{\theta}$ by a change of variables, giving rise to the generalized Kloosterman sums in Lemma 4.5/Definition 4.11. The explicit shape of these character sums allows us to prove the square-root cancellation (up to a bounded power of p), and also detect cancellations when taking averages in Theorem 1.5.

1.3.2. *Alternative description and the character sum after the Voronoï summation.* In Lemma 5.2, we show that the local test function we have constructed and used actually coincides with the matrix coefficient of the newform in the range we are interested in.

This alternative perspective also turns out to be quite useful. To explain this, we remark that in applications the Petersson/Kuznetsov trace formula is often followed by the use of the Voronoï summation formula. In the classical setting, the Kloosterman sum becomes the Ramanujan sum

$$\widetilde{\mathrm{KL}}(m_1, m_2, a, c) = \sum_{x\in(\mathbb{Z}/c\mathbb{Z})^\times} e\left(\frac{m_1 + m_2 a}{c} x\right). \tag{1.8}$$

Here a is an additional parameter, which can be -1 for example. The Ramanujan sum has the property that its average size is roughly 1 when, for example, taking a sum in m_1.

On the other hand for the generalized Kloosterman sum $G(m_1, m_2, \theta, \mu)$, the corresponding character sum, which occurs in the proof of Lemma 6.3 and is denoted by $\tilde{G}(m_1, m_2, a, \theta, \mu)$ in Definition 5.7, becomes more complicated to analyze. Take $\mu = \frac{1}{c^2}$ and $k = v_p(c)$. Recall from Remark 1.3 that when $k \geq \mathfrak{c}$, $G(m_1, m_2, \theta, \mu)$ becomes the classical Kloosterman sum, so $\tilde{G}(m_1, m_2, a, \theta, \mu)$ becomes the Ramanujan sum. We focus on the case $v_p(c_0) \leq k \leq \mathfrak{c}$ now. Using the alternative description above, we can identify $\tilde{G}(m_1, m_2, a, \theta, \mu)$ in the range of interest with the value of the matrix coefficient itself in the proof for Lemma 5.8. Then we apply the known results in [11, Theorem 5.4] on the support and the size of the matrix coefficient for the newform to obtain Lemma 5.8, which says that $\widetilde{G}_p(m_1, m_2, a, \theta, \mu) = 0$ unless

$$v\left(m_2\mu + \frac{am_1}{p^{2k}}\right) \geq -\mathfrak{c}, \tag{1.9}$$

in which case we have

$$\widetilde{G}_p(m_1, m_2, a, \theta, \mu) \ll_p p^{\frac{3k-\mathfrak{c}}{2}}. \tag{1.10}$$

Note that when $k = v_p(c_0)$, the congruence condition (1.9) is (almost) automatic, and the upper bound in (1.10) shows square-root cancellation. Thus $\widetilde{G}_p(m_1, m_2, a, \theta, \mu)$ displays a transition from Ramanujan-sum-type behavior to the square-root-cancellation behavior when k goes from $\mathfrak{c}$ to roughly $\frac{\mathfrak{c}}{2}$.

1.3.3. *Studying moments and hybrid subconvexity bounds.* The strategy to use the approximate functional equation, the Petersson/Kuznetsov trace formula, and then the Voronoï summation formula etc., is relatively standard. We have taken some arguments and results directly from, for example, [15] [6]. The main new ingredients are the refined Petersson trace formula in Theorem 1.5 with a flexible parameter l, and the study of the character sum $\widetilde{G}_p(m_1, m_2, a, \theta, \mu)$.

By choosing l properly, we can to some extent balance the contributions from the first-cell terms and the second-cell terms, obtaining Weyl-type subconvexity bound in a relatively large hybrid range.

1.4. **The Structure of the paper.** In Section 2 we introduce some basic notations and results.

In Section 3 we review some basic properties for the minimal vectors and the microlocal lifts, discuss their relations with the newforms, and construct test functions which pick out small families of newforms.

In Section 4 we use the relative trace formula for period integrals on unipotent subgroups to derive Theorem 1.2, 1.5 and 1.7.

In Section 5 we relate the test functions constructed in Section 3 with the matrix coefficient for the newform. Then we prove Lemma 5.8 for the character sum $\widetilde{G}_p(m_1, m_2, a, \theta, \mu)$.

In Section 6 we review a special version of the Voronoï summation formula, and apply the techniques developed so far to prove Theorem 1.9.

2. Preliminaries

2.1. **Notations.** Globally we shall work with the rational field $\mathbb{Q}$. Many of the discussions also hold for general number fields.

Let $\mathbb{A}$ be the ring of adeles over $\mathbb{Q}$, and $\mathbb{A}_{fin}$ be the finite adeles. We fix an additive character ψ on $\mathbb{Q}\backslash\mathbb{A}$, which is a product of local additive characters ψ_v, where $\psi_\infty(x) = e^{-2\pi ix}$, and $\psi_p(x) = e^{2\pi ix'}$ where $x' \in \mathbb{Q}$ and $x' \equiv x \bmod \mathbb{Z}_p$.

Let $\mathbb{F}$ denote a p-adic local field, $O_\mathbb{F}$ be its ring of integers and ϖ be a uniformizer with order of residue field $p \neq 2$. Let $U_\mathbb{F}(n) = 1 + \varpi^n O_\mathbb{F}$ when $n \geq 1$, and $U_\mathbb{F}(0) = O_\mathbb{F}^\times$.

Let $\mathbb{L}$ be a quadratic étale algebra over $\mathbb{F}$. When $\mathbb{L}$ is a field, let $e_\mathbb{L}$ be the ramification index of $\mathbb{L}$. Let $O_\mathbb{L}$, $\varpi_\mathbb{L}$ and $U_\mathbb{L}(n)$ be defined similar as for $\mathbb{F}$.

If $\mathbb{L} = \mathbb{F} \times \mathbb{F}$ splits, let $e_\mathbb{L} = 1$. Let $U_\mathbb{L}(0) = O_\mathbb{L}^\times = O_\mathbb{F}^\times \times O_\mathbb{F}^\times$, and $U_\mathbb{L}(n) = 1 + \varpi_\mathbb{F}^n(O_\mathbb{F} \times O_\mathbb{F})$.

Let θ be a character over $\mathbb{L}$ with $\theta|_{\mathbb{F}^\times} = 1$. Let $\mathfrak{c}(\theta)$ be the level of θ. If $\mathbb{L}$ splits, then we can write $\theta = \left(\chi, \chi^{-1}\right)$, and $\mathfrak{c}(\theta) = \mathfrak{c}(\chi)$.

For GL_2, let Z be its center, N be the unipotent subgroup. Over $\mathbb{F}$, let K be the standard maximal compact open subgroup $\mathrm{GL}_2(O_\mathbb{F})$. We also denote $G = \mathrm{PGL}_2$. We denote

$$n(x) = \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix},\ a(y) = \begin{pmatrix} y & \\ & 1 \end{pmatrix}.$$

Let π be an irreducible cuspidal automorphic representation of GL_2 with trivial central character. Let π_v denote its local component at v. Let $\mathfrak{c}(\pi_v)$ be the level of π_v.

Haar measures are normalized so that $\mathrm{Vol}\,(\mathbb{Q}\backslash\mathbb{A}) = 1$, $\mathrm{Vol}(K) = \mathrm{Vol}(Z\backslash ZK) = 1$.

For an automorphic cuspidal form φ, define

$$||\varphi||^2 =<\varphi,\varphi>= \int\limits_{Z(\mathbb{A})\mathrm{GL}_2(\mathbb{Q})\backslash \mathrm{GL}_2(\mathbb{A})} |\varphi(g)|^2 dg. \tag{2.1}$$

2.2. A basic result on characters.

Lemma 2.1. *Suppose that either p is large enough and $i = 1$, or i is large enough. The p-adic logarithm* $\log$ *is a group isomorphism from $U_\mathbb{F}(i)$ with multiplication with to $U_\mathbb{F}(i)$ with addition. There exists $\alpha_\nu \in \left(\varpi^{-\mathfrak{c}(\nu)+\mathfrak{c}(\psi_\mathbb{F})}O_\mathbb{F}/\varpi^{-i+\mathfrak{c}(\psi_\mathbb{F})}\right)^\times$ such that*

$$\nu(1+u) = \psi_\mathbb{F}\left(\alpha_\nu \log(1+u)\right), \forall u \in \varpi_\mathbb{F}^i O_\mathbb{F},$$

where $\log(1+u)$ *is defined by the standard Taylor expansion for logarithm*

$$\log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} + \cdots.$$

On the other hand if $p \neq 2$ and $i \geq \mathfrak{c}(\nu)/2$, we have

$$\nu(1+u) = \psi_\mathbb{F}(\alpha_\nu u).$$

Note that we formulate this lemma for general $c(\psi_\mathbb{F})$ because we will also apply it to characters over $\mathbb{L}$ later on.

2.3. **Kirillov model, Whittaker model and unitary pairings.** This subsection is purely local so we skip the subscript v from some of the notations.

For a fixed additive character ψ, the Kirillov model of π is a unique realization of π on a subspace of $C^\infty(\mathbb{F}^\times) \cap \mathcal{S}(\mathbb{F})$ such that

$$\pi\left(\begin{pmatrix} a_1 & m \\ 0 & a_2 \end{pmatrix}\right)\varphi(x) = w_\pi(a_2)\,\psi\left(ma_2^{-1}x\right)\varphi\left(a_1a_2^{-1}x\right), \tag{2.2}$$

where w_π is the central character for π. Let W_φ be the Whittaker function associated to φ. Then it is related to the Kirillov model by

$$\varphi(\alpha) = W_\varphi\left(\begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix}\right),$$

$$W_\varphi(g) = \pi(g)\varphi(1).$$

When π is unitary, one can define the $G-$invariant unitary pairing on the Kirillov model by

$$<\varphi_1,\varphi_2>= \int\limits_{\mathbb{F}^\times} \varphi_1(x)\,\overline{\varphi_2}(x)\,d^*x. \tag{2.3}$$

On the other hand, if $\pi = \pi(\chi_1, \chi_2)$ is a principal series representation with χ_i unitary, the unitary pairing can be alternatively defined by

$$< f_1, f_2 >= \int_K f_1(k)\overline{f}_2(k)dk. \tag{2.4}$$

Here $f_i \in \pi$ are element in the parabolic induction model, and K is a fixed maximal compact open subgroup.

2.4. **Global Whittaker function.** Let W_φ be now the global Whittaker function associated to a holomorphic newform φ and the fixed additive character ψ. It can be computed as

$$W_\varphi(g) = \int_{t\in N(\mathbb{Q})\backslash N(\mathbb{A})} \varphi(n(t)g)\psi(-t)dt. \tag{2.5}$$

W_φ factorizes into a product of local Whittaker functions

$$W_\varphi(g) = \prod_v W_v(g). \tag{2.6}$$

Here W_∞ is the Whittaker function associated to the lowest weight element in a discrete series representation of weight κ over $\mathbb{R}$. We have explicit expression

$$W_\infty\left(\begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix}\right) = \begin{cases} y^{\kappa/2}e^{-2\pi y}e^{2\pi i x}, & \text{if } y > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

On the other hand, W_v is the Whittaker function associated to the local newform at finite place v with $W_v(1) = 1$. They are closely related to the classical Fourier coefficients. More explicitly for a positive integer m,

$$\prod_{v \text{ finite}} W_v\left(\begin{pmatrix} m & \\ & 1 \end{pmatrix}\right) = |m|^{-1/2}\lambda_m(\varphi). \tag{2.8}$$

Here $\lambda_m(\varphi)$ is normalized so that $\lambda_1(\varphi) = 1$ and $\lambda_m(\varphi) \ll m^\epsilon$ by the Ramanujan conjecture.

2.5. **Hecke algebra action.** We shall choose a test function $f = f_\infty \times f_{fin}$ on $G(\mathbb{A})$ (which can be view as a function on $\mathrm{GL}_2(\mathbb{A})$ invariant by $Z(\mathbb{A})$), where f_{fin} is smooth on $G\left(\mathbb{A}_{fin}\right)$ and compactly supported mod center, and $f_\infty \in C(G(\mathbb{R}))$ is sufficiently differentiable and with proper decay (the exact requirements depend on whether we are deriving a Petersson trace formula or a Kuznetsov trace formula). We define the Hecke algebra action both globally and locally as

$$\rho(f)F(h) = \int_{G(\mathbb{A})} f(g)F(hg)dg, \ \pi_v(f_v)\varphi_v = \int_{G(\mathbb{Q}_v)} f_v(g)\pi_v(g)\varphi_v dg.$$

3. Minimal vector, microlocal lifts and newforms

This section is purely local so we skip subscript v from all notations.

3.1. **Small family.**

Definition 3.1. Let $\mathbb{L}$ be an étale quadratic algebra over $\mathbb{F}$. Let $\theta_i, i = 1, 2$ be characters over $\mathbb{L}$ such that $\theta_i|_{\mathbb{F}^\times} = 1$ and $\mathfrak{c}(\theta_1) = \mathfrak{c}(\theta_2)$. Denote

$$i_0 = \mathfrak{c}(\theta)/e_{\mathbb{L}},$$

which is always an integer by $\theta_i|_{\mathbb{F}^\times} = 1$. For $0 \leq n < i_0$, denote $\theta_1 \sim_n \theta_2$ if $\mathfrak{c}\left(\theta_1^{-1}\theta_2\right) \leq e_{\mathbb{L}} n$. For a fixed character θ with $\theta|_{\mathbb{F}^\times} = 1$, denote

$$\theta[n] = \{\theta' \text{ over } \mathbb{L} | \mathfrak{c}(\theta') = \mathfrak{c}(\theta), \theta'|_{\mathbb{F}^\times} = 1, \theta' \sim_n \theta\}.$$

Definition 3.2. Let $\pi_\theta[n] = \{\pi' \simeq \pi_{\theta'} | \theta' \in \theta[n]\}$.

Here $\pi_{\theta'}$ is the representation associated to θ either by the compact induction theory or the parabolic induction theory depending on $\mathbb{L}$ is a field or not. See Section 3.2 3.3 for more details.

Remark 3.3. When $n < i_0$, there is a bijection between $\theta[n]$ and $\pi_\theta[n]$. This is however not true when $n = i_0$, as $\pi_\theta \simeq \pi_{\overline{\theta}}$.

Lemma 3.4. *Let $\pi' = \pi_{\theta'}$ for θ' defined over the same $\mathbb{L}$ as θ, and $\mathfrak{c}(\theta') = \mathfrak{c}(\theta) \geq 2$. Then $\pi' \in \pi_\theta[n]$ for $n < i_0$ iff $C(\pi_{\theta^{-1}} \times \pi_{\theta'}) \leq C(\pi_\theta) p^{2n+e_{\mathbb{L}}-1}$.*

Proof. As θ and θ' are defined over the same $\mathbb{L}$, $C(\pi_{\theta^{-1}} \times \pi_{\theta'}) = C(\pi_{\theta^{-1}\theta'}) C(\pi_{\theta^{-1}\overline{\theta'}})$. Since $p \neq 2$ and $\mathfrak{c}(\theta) \geq 2$, at least one of $\mathfrak{c}\left(\theta^{-1}\theta'\right), \mathfrak{c}\left(\theta^{-1}\overline{\theta'}\right)$ is $\mathfrak{c}(\theta)$. As $\pi_{\theta'} \simeq \pi_{\overline{\theta'}}$, we can assume WLOG that $\mathfrak{c}\left(\theta^{-1}\overline{\theta'}\right) = \mathfrak{c}(\theta)$.

Now $\pi' \in \pi_\theta[n]$ iff $\mathfrak{c}\left(\theta^{-1}\theta'\right) \leq e_{\mathbb{L}} n$. It remains to use that $\mathfrak{c}(\pi_\theta) = \frac{2}{e_{\mathbb{L}}}\mathfrak{c}(\theta) + e_{\mathbb{L}} - 1$ in general. (See the list before Definition 3.9 for the supercuspidal representation cases. It is also true for the parabolic induction case.) □

Lemma 3.5.

$$[\theta[1] : \theta[0]] = pL_{\mathbb{F}}^{-1}(1, \epsilon_{\mathbb{L}/\mathbb{F}}) = \begin{cases} p-1, & \textit{if } \mathbb{L} \textit{ splits,} \\ p+1, & \textit{if } \mathbb{L} \textit{ is an inert field extension,} \\ p, & \textit{if } e_{\mathbb{L}} = 2. \end{cases}$$

For $1 < n < i_0$,

$$[\theta[n] : \theta[n-1]] = p.$$

Proof. For $n \geq 1$, let $\hat{\mathbb{F}}^\times\{n\} = \{\chi \text{ over } \mathbb{F}^\times, \mathfrak{c}(\chi) \leq n\}$. When $\mathbb{L}$ splits, the claims can be easily proven as we have an identification

$$\begin{aligned} \theta[n]/\theta[n-1] &\to \hat{\mathbb{F}}^\times\{n\}/\hat{\mathbb{F}}^\times\{n-1\} \\ \theta' &\mapsto \chi \text{ if } \theta^{-1}\theta' = \left(\chi, \chi^{-1}\right). \end{aligned} \tag{3.1}$$

When $\mathbb{L}$ is a field define $\hat{\mathbb{L}}^\times\{n\}$ similarly. Then we have a short exact sequence

$$1 \to \theta[n]/\theta[n-1] \xrightarrow{\iota} \hat{\mathbb{L}}^\times\{ne_{\mathbb{L}}\}/\hat{\mathbb{L}}^\times\{(n-1)e_{\mathbb{L}}\} \xrightarrow{\sigma} \hat{\mathbb{F}}^\times\{n\}/\hat{\mathbb{F}}^\times\{n-1\} \to 1.$$

Here $\iota(\theta') = \theta^{-1}\theta'$, and $\sigma(\theta) = \theta|_{\mathbb{F}^\times}$. The lemma follows from counting $\hat{\mathbb{L}}^\times\{ne_{\mathbb{L}}\}/\hat{\mathbb{L}}^\times\{(n-1)e_{\mathbb{L}}\}$ and $\hat{\mathbb{F}}^\times\{n\}/\hat{\mathbb{F}}^\times\{n-1\}$, which can be done by using the Pontryagin duality for finite groups. □

Remark 3.6. It is also direct to see that $\sharp\theta[0] = 1$ when $\mathbb{L}$ is inert, and $\sharp\theta[0] = 2$ when $\mathbb{L}$ is ramified. When $\mathbb{L}$ is split, $\theta[0]$ is however not finite.

Lemma 3.7. *Let $\mathbb{L}$ be an étale quadratic algebra over $\mathbb{F}$, $x \in O_{\mathbb{L}}^{\times}$, $j \geq 1$. Then*

$$\frac{1}{[\theta[j] : \theta[0]]} \sum_{\theta' \in \theta[j]/\sim_0} \theta'(x) = \begin{cases} \theta(x), & \text{if } x \in O_{\mathbb{F}}^{\times} U_{\mathbb{L}}(e_{\mathbb{L}} j), \\ 0, & \text{otherwise.} \end{cases}$$

Proof. When $x \in ZU_{\mathbb{L}}(e_{\mathbb{L}} j)$, we have $\theta'(x) = \theta(x)\theta^{-1}\theta'(x) = \theta(x)$ as $\mathfrak{c}\left(\theta^{-1}\theta'\right) \leq e_{\mathbb{L}} j$.

On the other hand, note that $[\theta[j] : \theta[0]] = \sharp\left(O_{\mathbb{L}}^{\times}/O_{\mathbb{F}}^{\times} U_{\mathbb{L}}(e_{\mathbb{L}} j)\right)$ by, for example, Lemma 3.5. So $\theta[j]/\sim_0$ is the Pontryagin dual of $O_{\mathbb{L}}^{\times}/O_{\mathbb{F}}^{\times} U_{\mathbb{L}}(e_{\mathbb{L}} j)$. The sum is thus vanishing because of the orthogonality of the characters.

□

For any $\theta' \in \theta[n]$, there is an element $\alpha_{\theta'} \in \left(\varpi_{\mathbb{L}}^{-\mathfrak{c}(\theta)+\mathfrak{c}(\psi_{\mathbb{L}})} O_{\mathbb{L}}/\varpi^{-i+\mathfrak{c}(\psi_{\mathbb{L}})} O_{\mathbb{L}}\right)^{\times}$ by Lemma 2.1 with

$$\theta'(1+u) = \psi_{\mathbb{L}}(\alpha_{\theta'} \log(1+u)), \forall u \in \varpi_{\mathbb{L}}^{i} O_{\mathbb{L}}. \tag{3.2}$$

$\theta'|_{\mathbb{F}^{\times}} = 1$ implies that α'_{θ} (and also α_{θ}) can be chosen to be imaginary, i.e. $\overline{\alpha}'_{\theta} = -\alpha'_{\theta}$ where $x \mapsto \overline{x}$ is the nontrivial automorphism of $\mathbb{L}/\mathbb{F}$.

Lemma 3.8. *Fix $n < i_0$. Suppose that p is large enough or $1 \leq j < n$ is large enough. For any $\theta' \in \theta[n]$, let $\alpha_{\theta'}$ be an imaginary element associated to θ' by Lemma 2.1. Then we have the following bijection*

$$\begin{aligned} \theta[n]/\sim_j &\to \alpha_{\theta} U_{\mathbb{F}}(i_0 - n)/U_{\mathbb{F}}(i_0 - j) \\ \theta' &\mapsto \alpha_{\theta'} \end{aligned} \tag{3.3}$$

Here j being large enough is similar to i large enough in Lemma 2.1 with $\mathbb{F}$ replaced by $\mathbb{L}$.

Proof. We write $\alpha'_{\theta} = \alpha_{\theta} u$ for $u \in O_{\mathbb{F}}^{\times}$ as $\mathfrak{c}(\theta) = \mathfrak{c}(\theta')$. From $\mathfrak{c}\left(\theta^{-1}\theta'\right) \leq e_{\mathbb{L}} n$, we get that $\theta^{-1}\theta'$ is trivial on $U_{\mathbb{L}}(e_{\mathbb{L}} n)$, whose image under log is $\varpi_{\mathbb{L}}^{e_{\mathbb{L}} n} O_{\mathbb{L}} = \varpi^{n} O_{\mathbb{L}}$. As the associated constant to $\theta^{-1}\theta'$ is $\alpha'_{\theta} - \alpha_{\theta} = \alpha_{\theta}(u-1)$, we get that

$$\psi_{\mathbb{L}}(\alpha_{\theta}(u-1)x) = 1, \ \forall x \in \varpi^{n} O_{\mathbb{L}}.$$

This implies that $u \in U_{\mathbb{F}}(i_0 - n)$.

On the other hand, if $\alpha'_{\theta} \in \alpha_{\theta} U_{\mathbb{F}}(i_0 - j)$, then by (3.2) we get that $\mathfrak{c}\left(\theta^{-1}\theta'\right) \leq e_{\mathbb{L}} j$.

To show that the map is a bijection, it remains to see that the cardinalities of both sides agree using Lemma 3.5. □

3.2. **Supercuspidal case.** We now discuss the representation π associated to θ over $\mathbb{L}$. We consider first the case $\mathbb{L}$ is a quadratic field extension over $\mathbb{F}$, and thus π is supercuspidal. The detailed construction can be found in, for example, [3] with some different conventions.

3.2.1. *Review.* Let $\mathbb{F}$ be a p-adic local field, $\mathbb{L} = \mathbb{F}\left(\sqrt{D}\right)$ be a quadratic field extension with ramification index $e_{\mathbb{L}}$. In [9][12], we assumed that $v_{\mathbb{F}}(D) = 0$ or 1, and used the following embedding of $\mathbb{L}$ as a standard embedding:

$$x + y\sqrt{D} \mapsto \begin{pmatrix} x & y \\ yD & x \end{pmatrix}. \tag{3.4}$$

We fix an additive character ψ such that $\mathfrak{c}(\psi) = 0$. Then $\mathfrak{c}(\psi_{\mathbb{L}}) = -e_{\mathbb{L}} + 1$.

The supercuspidal representations are parameterized via compact induction by characters θ over some quadratic field extension $\mathbb{L}$. More specifically we have the following quick guide.

Case 1. $\mathfrak{c}(\pi) = 2n+1$ corresponds to $e_{\mathbb{L}} = 2$ and $\mathfrak{c}(\theta) = 2n$.
Case 2. $\mathfrak{c}(\pi) = 4n$ corresponds to $e_{\mathbb{L}} = 1$ and $\mathfrak{c}(\theta) = 2n$.
Case 3. $\mathfrak{c}(\pi) = 4n+2$ corresponds to $e_{\mathbb{L}} = 1$ and $\mathfrak{c}(\theta) = 2n+1$.

Definition 3.9. For $e_{\mathbb{L}} = 1, 2$, let

$$\mathfrak{A}_{e_{\mathbb{L}}} = \begin{cases} M_2(O_{\mathbb{F}}), \text{ if } e_{\mathbb{L}} = 1, \\ \begin{pmatrix} O_{\mathbb{F}} & O_{\mathbb{F}} \\ \varpi O_{\mathbb{F}} & O_{\mathbb{F}} \end{pmatrix}, \text{ otherwise.} \end{cases}$$

Its Jacobson radical is given by

$$\mathcal{B}_{e_{\mathbb{L}}} = \begin{cases} \varpi M_2(O_{\mathbb{F}}), \text{ if } e_{\mathbb{L}} = 1, \\ \begin{pmatrix} \varpi O_{\mathbb{F}} & O_{\mathbb{F}} \\ \varpi O_{\mathbb{F}} & \varpi O_{\mathbb{F}} \end{pmatrix}, \text{ otherwise.} \end{cases}$$

Define the filtration of compact open subgroups as follows:

$$K_{\mathfrak{A}_{e_{\mathbb{L}}}}(n) = 1 + \mathcal{B}_{e_{\mathbb{L}}}^n, \tag{3.5}$$

Note that each $K_{\mathfrak{A}_{e_{\mathbb{L}}}}(n)$ is normalised by $\mathbb{L}^\times$ which is embedded as in (3.4).

Denote $J = \mathbb{L}^\times K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lfloor \mathfrak{c}(\theta)/2 \rfloor)$, $J^1 = U_{\mathbb{L}}(1) K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lfloor \mathfrak{c}(\theta)/2 \rfloor)$, $H^1 = U_{\mathbb{L}}(1) K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lceil \mathfrak{c}(\theta)/2 \rceil)$. Then θ on $\mathbb{L}^\times$ can be extended to be a character $\tilde{\theta}$ on H^1 by

$$\tilde{\theta}(l(1+x)) = \theta(l)\psi \circ \mathrm{Tr}(\alpha_\theta x), \tag{3.6}$$

where $l \in \mathbb{L}^\times$, $1 + x \in K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lceil \mathfrak{c}(\theta)/2 \rceil)$ and $\alpha_\theta \in \mathbb{L}^\times \subset M_2(\mathbb{F})$ is associated to θ by Lemma 2.1 under the fixed embedding.

When $\mathfrak{c}(\theta)$ is even, $H^1 = J^1$ and $\tilde{\theta}$ can be further extended to J by the same formula. In this case denote $\Lambda = \tilde{\theta}$ and $\pi_\theta = c - \mathrm{Ind}_J^G \Lambda$ is an irreducible supercuspidal representation. $\pi_\theta \simeq \pi_{\theta'}$ if and only if $\theta = \theta'$ or $\overline{\theta}'$.

When $\mathfrak{c}(\theta)$ is odd, J^1/H^1 is a two dimensional vector space over the residue field. This case only occurs when $\mathfrak{c}(\pi) = 4n + 2$ as listed above. Then there exists a $q-$dimensional (or $q - 1$ dimensional if $\mathfrak{c}(\pi) = 2$, but we will be mainly interested in the case when $\mathfrak{c}(\pi)$ is large enough) irreducible representation Λ of J such that $\Lambda|_{H^1}$ is a multiple of $\tilde{\theta}$, and

$$\Lambda|_{\mathbb{L}^\times} = \bigoplus_{\theta' \in \theta[1], \theta' \neq \theta, \overline{\theta}} \theta' \tag{3.7}$$

More specifically, let B^1 be any intermediate group between J^1 and H^1 such that B^1/H^1 gives a polarisation of J^1/H^1 under the pairing given by

$$(1 + x, 1 + y) \mapsto \psi \circ \mathrm{Tr}(\alpha_\theta [x, y]). \tag{3.8}$$

Then $\tilde{\theta}$ can be extended to B^1 by the same formula (3.6) and $\Lambda|_{J^1} = \mathrm{Ind}_{B^1}^{J^1} \tilde{\theta}$. Again $\pi_\theta = c{-}\mathrm{Ind}_J^G \Lambda$ is irreducible and supercuspidal in this case, and $\pi_\theta \simeq \pi_{\theta'}$ iff $\theta = \theta'$ or $\overline{\theta}'$. We always have $w_\pi = \theta|_{\mathbb{F}^\times}$.

Note that when $J^1 \neq H^1$, any intermediate subgroup B^1 works, as the pairing (3.8) is skew-symmetric. It will however be convenient to fix a choice of B^1 for later purposes.

Definition 3.10. When $\mathbb{L}$ is inert and $\mathfrak{c}(\theta) = 2n + 1$, let

$$B^1 = U_{\mathbb{L}}(1) K_{\mathfrak{A}_2}(2n+1) \tag{3.9}$$

In the case $J^1 = H^1$, $\mathfrak{c}(\theta)$ even, we take $B^1 = U_{\mathbb{L}}(1) K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\mathfrak{c}(\theta)/2)$.

Definition 3.11. There exists a unique element $\varphi_0 \in \pi$ such that B^1 acts on it by $\tilde{\theta}$. We call any single translate $\pi(g)\varphi_0$ a minimal vector (Type 1 minimal vector in the notation of [9]).

Note that the conjugated group gB^1g^{-1} acts on $\pi(g)\varphi_0$ by the conjugated character $\tilde{\theta}^g$.

Corollary 3.12. *Let Φ_{φ_0} be the matrix coefficient associated to a minimal vector φ_0 as above. Then Φ_{φ_0} is supported on J, and*

$$\Phi_{\varphi_0}(bx) = \Phi_{\varphi_0}(xb) = \tilde{\theta}(b)\,\Phi_{\varphi_0}(x) \tag{3.10}$$

for any $b \in B^1$. Furthermore when $\dim \Lambda \neq 1$*, $\Phi_{\varphi_0}|_{J^1}$ is supported only on B^1.*

Due to the central character, it is clear that ZB^1 acts on φ_θ by a character, which we also denote by $\tilde{\theta}$ without confusion. We also need a converse result.

Proposition 3.13. *Let π be an irreducible smooth representation of $GL_2(\mathbb{F})$, with central character $w_\pi = \theta|_{\mathbb{F}^\times}$ and $\mathfrak{c}(\pi) \geq 3$. Suppose that there exists an element $\varphi \in \pi$ on which ZB^1 acts by a given character $\tilde{\theta}$, then φ is unique up to a constant. Furthermore we must have $\pi \simeq \pi_{\theta'}$ where $\theta' \in \theta[l_0]$, for $l_0 = 1$ when $\mathbb{L}$ is inert, and $l_0 = 0$ when $\mathbb{L}$ is ramified.*

Proof. We consider only the case where $\mathbb{L}$ is inert and $\mathfrak{c}(\theta)$ is odd here, as the other cases are very similar and slightly easier.

By the condition, ZB^1 acts on φ by $\tilde{\theta}$. By the Frobenius reciprocity for compact inductions, we have

$$\mathrm{Hom}_{ZB^1}\left(\tilde{\theta}, \pi|_{ZB^1}\right) = \mathrm{Hom}_G\left(c - \mathrm{Ind}_{ZB^1}^G \tilde{\theta}, \pi\right). \tag{3.11}$$

We study $c - \mathrm{Ind}_{ZB^1}^G \tilde{\theta}$ step by step as the induction of representations is transitive. Since $\mathrm{Ind}_{B^1}^{J_1} \tilde{\theta} = \Lambda|_{J^1}$, we have

$$\mathrm{Ind}_{ZB^1}^{ZJ^1} \tilde{\theta} = \Lambda|_{ZJ^1}.$$

For each $\theta' \in \theta[1]$, let $\Lambda_{\theta'}$ be irreducible representations of J constructed similarly as Λ, which are not equivalent to each other by (3.7). From $\theta' \in \theta[1]$, we get that $\Lambda_{\theta'}|_{ZJ^1} = \Lambda|_{ZJ^1}$.

In particular we have

$$\mathrm{Hom}_J\left(\mathrm{Ind}_{ZB^1}^J \tilde{\theta}, \Lambda_{\theta'}\right) = \mathrm{Hom}_{ZJ^1}\left(\mathrm{Ind}_{ZB^1}^{ZJ_1} \tilde{\theta}, \Lambda|_{ZJ^1}\right) \neq 0.$$

Then we must have

$$\mathrm{Ind}_{ZB^1}^J \tilde{\theta} = \bigoplus_{\theta' \in \theta[1]} \Lambda_{\theta'}$$

by a dimension counting.

Then (3.11) becomes

$$\mathrm{Hom}_{ZB^1}\left(\tilde{\theta}, \pi|_{ZB^1}\right) = \bigoplus_{\theta' \in \theta[1]} \mathrm{Hom}_G\left(c - \mathrm{Ind}_J^G \Lambda_{\theta'}, \pi\right). \tag{3.12}$$

From this we see that the RHS is trivial unless $\pi \simeq \pi_{\theta'}$ for some $\theta' \in \theta[1]$, as $\pi_{\theta'}$'s are irreducible and not mutually equivalent. The claims in the proposition are clear now. ◻

3.2.2. *Kirillov model and recovering the newform.* We also need to describe the minimal vectors explicitly in the Kirillov model.

As we are going to vary θ, we fix a choice of D (unlike [9][12]), and assume

$$\alpha_\theta = \frac{\alpha_0}{\varpi_{\mathbb{L}}^{\mathfrak{c}(\theta)}\sqrt{D}} \mapsto \frac{\alpha_0}{\varpi^{\mathfrak{c}(\theta)/e_{\mathbb{L}}}}\begin{pmatrix}0 & \frac{1}{D}\\ 1 & 0\end{pmatrix}. \tag{3.13}$$

Here we can pick $\alpha_0 \in O_{\mathbb{F}}^\times$ by our assumption $\theta|_{\mathbb{F}^\times} = 1$. We define the intertwining operator from π to its Whittaker model by

$$\varphi \mapsto W_\varphi(g) = \int_{\mathbb{F}} \Phi_{\varphi,\varphi_0}\left(\begin{pmatrix}\frac{\varpi^{\lfloor \mathfrak{c}(\pi)/2\rfloor}}{\alpha_0} & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & n\\ 0 & 1\end{pmatrix}g\right)\psi(-n)\,dn. \tag{3.14}$$

Lemma 3.14. *Up to a constant multiple, a minimal vector φ_0 is given in the Kirillov model by the following:*

1. *When $\mathfrak{c}(\pi) = 4n$, $\varphi_0 = char\left(\varpi^{-2n}\alpha_0 U_{\mathbb{F}}(n)\right)$.*
2. *When $\mathfrak{c}(\pi) = 2n+1$, $\varphi_0 = char(\varpi^{-n}\alpha_0 U_{\mathbb{F}}(\lceil n/2\rceil))$.*
3. *When $\mathfrak{c}(\pi) = 4n+2$, $\varphi_0 = char\left(\varpi^{-2n-1}\alpha_0 U_{\mathbb{F}}(n+1)\right)$.*

The computations are essentially same as in [9, Lemma A.7]. Using the notation $i_0 = \frac{\mathfrak{c}(\theta)}{e_{\mathbb{L}}}$, one can uniformly write

$$\varphi_0 = \sqrt{(p-1)\,p^{\lceil i_0/2\rceil - 1}}\,\mathrm{char}\left(\varpi^{-i_0}\alpha_0 U_{\mathbb{F}}(\lceil i_0/2\rceil)\right). \tag{3.15}$$

Note here we have L^2−normalized φ_0. i_0 is roughly $\frac{\mathfrak{c}(\pi)}{2}$.

Remark 3.15. From the explicit Kirillov model, and the local unitary pairing given by

$$<\varphi_1,\varphi_2> = \int_{x\in\mathbb{F}^\times} \varphi_1(x)\,\overline{\varphi_2(x)}\,d^\times x,$$

one can see that the set

$$B_\pi = \left\{\pi\left(\begin{pmatrix}a & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & n\\ 0 & 1\end{pmatrix}\right)\varphi_0 \mid a\in\mathbb{F}^\times/U_{\mathbb{F}}(\lceil i_0/2\rceil),\, n\in\mathbb{F}/\varpi^{\lfloor i_0/2\rfloor}O_{\mathbb{F}}\right\}$$

forms an orthogonal basis for π, and is invariant by any diagonal translation.

Corollary 3.16. *For $a \in \left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^\times$, let $\varphi_a = \pi\left(\begin{pmatrix}\varpi^{-i_0}a^{-1} & 0\\ 0 & 1\end{pmatrix}\right)\varphi_0$. Then we have for $\varphi_{new} = char\left(O_{\mathbb{F}}^\times\right)$*

$$\varphi_{new} = \frac{1}{\sqrt{(p-1)\,p^{\lceil i_0/2\rceil - 1}}}\sum_{a\in\left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^\times}\varphi_a. \tag{3.16}$$

Note that φ_a can be viewed as the minimal vector associated to the embedding

$$x + y\sqrt{D} \mapsto \begin{pmatrix}x & \frac{y}{a\varpi^{i_0}}\\ yDa\varpi^{i_0} & x\end{pmatrix}. \tag{3.17}$$

Definition 3.17. Define $\Phi_{0,0}(g) =< \pi(g)\varphi_0, \varphi_0 >$ with normalisation $\Phi_{0,0}(1) = 1$,

Define $\tilde{\Phi}_{0,0} = \Phi_{0,0}|_{ZB^1}$. Define in general for $a, a' \in \left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^{\times}$

$$\Phi_{a,a'}(g) = \Phi_{0,0}\left(\begin{pmatrix}\varpi^{i_0}a' & 0\\ 0 & 1\end{pmatrix} g \begin{pmatrix}\varpi^{-i_0}a^{-1} & 0\\ 0 & 1\end{pmatrix}\right), \tilde{\Phi}_{a,a'}(g) = \tilde{\Phi}_{0,0}\left(\begin{pmatrix}\varpi^{i_0}a' & 0\\ 0 & 1\end{pmatrix} g \begin{pmatrix}\varpi^{-i_0}a^{-1} & 0\\ 0 & 1\end{pmatrix}\right).$$

Corollary 3.18. *$\tilde{\Phi}_{a,a}(g) = 0$ unless $g = e\begin{pmatrix}1+x & m\\ 0 & 1\end{pmatrix}$ or $\begin{pmatrix}1+x & m\\ 0 & 1\end{pmatrix}e$ for some $e \in ZU_{\mathbb{L}}(1)$, with embedding as in* (3.17)*, $x \in \varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}$ and $m \in \varpi^{-\lceil i_0/2\rceil}O_{\mathbb{F}}$. In that case, we have*

$$\tilde{\Phi}_{a,a}(g) = \theta(e)\psi(\alpha_0 am). \tag{3.18}$$

Proof. It follows from the explicit conjugation in the definition of $\tilde{\Phi}_{a,a}$, Corollary 3.12, the explicit shape of B^1 in Definition 3.10 and the explicit shape of α_θ as in (3.13). □

Definition 3.19. For a quadratic field extension $\mathbb{L}$ and a character θ on it, choose the local test function to be

$$f(g) = \frac{1}{(p-1)p^{\lceil i_0/2\rceil-1}\mathrm{Vol}(Z\backslash ZB^1)}\sum_{a,a'\in\left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^{\times}}\overline{\tilde{\Phi}}_{a,a'}(g). \tag{3.19}$$

Proposition 3.20. *For f defined in* (3.19)*, and let π be an irreducible smooth representation of $GL_2(\mathbb{F})$ with trivial central character. Then $\pi(f)$ is zero unless $\pi \simeq \pi_{\theta'}$ where $\theta' \in \theta[l_0]$, in which case $\pi(f)$ is the projection to the line generated by the newform.*

Proof. We first discuss $\pi\left(\overline{\tilde{\Phi}}_{0,0}\right)$. If $\pi\left(\overline{\tilde{\Phi}}_{0,0}\right)\varphi \neq 0$, then by a change of variable there exists $\varphi' = \pi\left(\overline{\tilde{\Phi}}_{0,0}\right)\varphi$ such that B^1 acts by $\tilde{\Phi}_{0,0} = \tilde{\theta}$. According to Proposition 3.13, $\pi \simeq \pi_{\theta'}$ for $\theta' \in \theta[l_0]$.

In that case, we also know that φ' must be a multiple of φ_0. We choose the orthonormal basis as in Remark 3.15. Then we have

$$< \pi\left(\overline{\tilde{\Phi}}_{0,0}\right)\varphi, \varphi_0 >=< \varphi, \pi\left(\tilde{\Phi}_{0,0}^{-1}\right)\varphi_0 >= \frac{1}{\mathrm{Vol}(Z\backslash ZB^1)} < \varphi, \varphi_0 >,$$

which implies that if $\varphi \in B_\pi$, then $\pi\left(\overline{\tilde{\Phi}}_{0,0}\right)\varphi = 0$ unless $\varphi = \varphi_0$. Thus $\pi\left(\frac{1}{\mathrm{Vol}(Z\backslash ZB^1)}\overline{\tilde{\Phi}}_{0,0}\right)$ is the projection onto the line spanned by φ_0.

Now for any $a, a' \in \left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^{\times}$, $\varphi \in B_\pi$, we have by definition

$$\begin{aligned}\pi\left(\overline{\tilde{\Phi}}_{a,a'}\right)\varphi &= \int_{g\in Z\backslash ZB^1}\tilde{\theta}^{-1}(g)\pi\left(\begin{pmatrix}\varpi^{-i_0}a'^{-1} & \\ & 1\end{pmatrix} g \begin{pmatrix}\varpi^{i_0}a & \\ & 1\end{pmatrix}\right)\varphi\\ &= \pi\left(\begin{pmatrix}\varpi^{-i_0}a'^{-1} & \\ & 1\end{pmatrix}\right)\pi\left(\overline{\tilde{\Phi}}_{0,0}\right)\pi\left(\begin{pmatrix}\varpi^{i_0}a & \\ & 1\end{pmatrix}\right)\varphi.\end{aligned} \tag{3.20}$$

As B_π is invariant by diagonal translates (up to constants), we see from the previous discussion that

$$\pi\left(\frac{1}{\mathrm{Vol}(Z\backslash ZB^1)}\overline{\tilde{\Phi}}_{a,a'}\right)\varphi = 0$$

unless $\varphi = \pi\left(\begin{pmatrix}\varpi^{-i_0}a^{-1} & \\ & 1\end{pmatrix}\right)\varphi_0 = \varphi_a$, in which case it becomes $\varphi_{a'}$. By Corollary 3.16 and Definition 3.19, we get that

$$\begin{aligned}\pi(f)\varphi_{new} &= \frac{1}{(p-1)p^{\lceil i_0/2\rceil-1}}\sum_{a,a'}\pi\left(\frac{1}{\mathrm{Vol}(Z\backslash ZB^1)}\overline{\tilde{\Phi}}_{a,a'}\right)\frac{1}{\sqrt{(p-1)p^{\lceil i_0/2\rceil-1}}}\sum_b\varphi_b\\ &= \frac{1}{(p-1)p^{\lceil i_0/2\rceil-1}}\frac{1}{\sqrt{(p-1)p^{\lceil i_0/2\rceil-1}}}\sum_{a',b}\varphi_{a'} = \varphi_{new}\end{aligned}$$

as $\sharp\left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^\times = (p-1)p^{\lceil i_0/2\rceil-1}$. □

Remark 3.21. It may seem possible and desirable to devise f so that one can take $l_0 = 0$ also for the $e_{\mathbb{L}} = 1$ case. We start with $l_0 = 1$ in this case because of the following two reasons:

(1) When $\mathfrak{c}(\pi_\theta) = 4n+2$, it is still complicated to write down and make use of the matrix coefficients on the whole group J, compared to its restriction to ZB^1.
(2) When $\mathfrak{c}(\pi_\theta) = 4n$, one can easily start from $l_0 = 0$ and $k \geq i_0$. One small benefit to start with $l_0 = 1$ is that the formulations in Theorem 1.2 1.5 are relatively more uniform for the supercuspidal representation cases. The proof of Lemma4.17 in Section 4.6.1 also becomes slightly easier when $k > i_0$ holds.

3.3. **Principal series representation case.** We remark that when the central character w_π is trivial, $p \neq 2$ and $\mathfrak{c}(\pi) \geq 4$, π can not be a Steinberg representation. It also can not be a twisted complementary series representation.

3.3.1. *Microlocal lift and twisting.* Here we recall the microlocal lifts of [16], which is essentially the twisted newforms. For convenience, we mainly restrict ourselves to the case where the central character is trivial, but the approach can be easily extended to more general cases.

We start in slightly more general situations. Let $\pi = \pi(\chi_1, \chi_2)$ be a principal series representation, whose elements $\varphi \in \pi$ satisfies

$$\varphi\left(\begin{pmatrix}a & n\\ 0 & b\end{pmatrix}g\right) = \chi_1(a)\chi_2(b)|\frac{a}{b}|^{1/2}\varphi(g).$$

Let $\pi_1 = \pi\left(1, \chi_1^{-1}\chi_2\right) = \pi\otimes\chi_1^{-1}$, so that $\pi = \pi_1\otimes\chi_1$. Assume that $i_0 = \mathfrak{c}\left(\chi_1^{-1}\chi_2\right)$. Let

$$K_0\left(\varpi^{i_0}\right) = \left\{\begin{pmatrix}a & b\\ c & d\end{pmatrix} \equiv \begin{pmatrix}* & *\\ 0 & *\end{pmatrix} \bmod \varpi^{i_0}\right\}$$

be the usual congruence subgroup.

Lemma 3.22. *The exists a unique (up to constant) element, i.e. the newform, $\varphi_1 \in \pi_1$ such that $K_0\left(\varpi^{i_0}\right)$ acts on φ_1 by $\chi_1^{-1}\chi_2(d)$. The normalised Whittaker function associated to φ_1 is given by*

$$W_{\varphi_1}\left(\begin{pmatrix}\alpha & 0\\ 0 & 1\end{pmatrix}\right) = \sqrt{1-p^{-1}}\begin{cases}p^{-v(\alpha)/2}, & \text{if } v(\alpha)\geq 0\\ 0, & \text{otherwise}\end{cases}.$$

Furthermore if there exists an element φ' from an irreducible smooth admissible representation π' such that $K_0\left(\varpi^{i_0}\right)$ acts on φ' by $\chi_1^{-1}\chi_2(d)$, then $\pi' \simeq \pi\left(\nu_1, \nu_2\chi_1^{-1}\chi_2\right)$ for some unramified characters ν_1, ν_2.

Proof. The existence of φ_1 is simply the newform theory in [4]. In the parabolic induction model, φ_1 is supported only on $BK_0\left(\varpi^{i_0}\right)$. Furthermore, for any $\varphi' \in \pi'$ with the same equivalent property, φ' is in particular invariant by

$$K_1\left(\varpi^{i_0}\right) = \left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} * & * \\ 0 & 1 \end{pmatrix} \bmod \varpi^{i_0}\right\},$$

so $\mathfrak{c}(\pi') \leq i_0$. On the other hand the equivalent property implies that $w_{\pi'}|_{O^\times} = \chi_1^{-1}\chi_2$. Then $\mathfrak{c}(\pi') \geq \mathfrak{c}(w_{\pi'}) = i_0$. This forces π' to be in the specified shape.

The expression for W_{φ_1} follows immediately from, for example, [7, Lemma 2.13] □

For uniformity, let $\mathbb{L}$ denote the diagonal torus, and let θ be the character (χ_1, χ_2). We associate the pair $(\mathbb{L}, \theta)$ to the principal series representation $\pi = \pi(\chi_1, \chi_2)$, and simply write $\pi = \pi_\theta$.

Let $\tilde{\theta}$ be the character on $ZK_0\left(\varpi^{i_0}\right)$ defined by

$$\tilde{\theta}(zg) = \chi_1\chi_2(z)\chi_1(\det g)\chi_1^{-1}\chi_2(d), \tag{3.21}$$

where $z \in Z$, $g \in K_0\left(\varpi^{i_0}\right)$.

Corollary 3.23. *There exists a unique element $\varphi_\theta \in \pi = \pi(\chi_1, \chi_2)$ such that $ZK_0\left(\varpi^{i_0}\right)$ acts on φ_θ by $\tilde{\theta}$. The associated Whittaker function for φ_θ is given by*

$$W_{\varphi_\theta}\left(\begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix}\right) = \sqrt{1 - p^{-1}} \begin{cases} p^{-v(\alpha)/2}\chi_1(\alpha), & \text{if } v(\alpha) \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.22}$$

Conversely, if there is an element $\varphi \in \pi'$ such that $ZK_0\left(\varpi^{i_0}\right)$ acts on it by $\tilde{\theta}$, then $\pi' \simeq \pi(\nu\chi_1, \nu^{-1}\chi_2)$ for some unramified character ν.

Proof. Follows directly from Lemma 3.22 by a twist, and the requirement for the central character to be $\chi_1\chi_2$. □

In particular if we assume the central character to be trivial, we get $\pi' = \pi_{\theta'}$ for some $\theta' \in \theta[0]$ as in Definition 3.1.

3.3.2. *Recovering the newform.*

Lemma 3.24. *Denote $c_1 = \mathfrak{c}(\chi_1)$, $\varphi_a = \pi\left(\begin{pmatrix} 1 & \frac{a}{\varpi^{c_1}} \\ 0 & 1 \end{pmatrix}\right)\varphi_\theta$, and*

$$C_0 = \left(1 - p^{-1}\right)^{3/2} p^{c_1} \int_{x \in O^\times} \chi_1(x)\psi(\varpi^{-c_1}x)\, d^\times x.$$

Then the newform can be written as

$$\varphi_{new} = char(O_{\mathbb{F}}^\times) = \frac{1}{C_0} \sum_{a \in (O/\varpi^{c_1}O)^\times} \chi_1(a)\varphi_a$$

Proof. Note that $C_0 = \sqrt{1 - p^{-1}} \sum_{x \in (O/\varpi^{c_1}O)^\times} \chi_1(x)\psi(\varpi^{-c_1}x)$. In the Kirillov/Whittaker model, we have for $v(x) \geq 0$,

$$\sum_{a\in(O/\varpi^{c_1}O)^\times}\chi_1(a)W_{\varphi_a}\left(\begin{pmatrix}x&0\\0&1\end{pmatrix}\right)=\sum_{a\in(O/\varpi^{c_1}O)^\times}\chi_1(a)\pi\left(\begin{pmatrix}1&\frac{a}{\varpi^{c_1}}\\0&1\end{pmatrix}\right)W_{\varphi_\theta}\left(\begin{pmatrix}x&0\\0&1\end{pmatrix}\right)$$
$$=\sum_{a\in(O/\varpi^{c_1}O)^\times}\chi_1(a)\psi\left(\frac{ax}{\varpi^{c_1}}\right)W_{\varphi_\theta}\left(\begin{pmatrix}x&0\\0&1\end{pmatrix}\right)$$
$$=\sqrt{1-p^{-1}}p^{-v(x)/2}\sum_{a\in(O/\varpi^{c_1}O)^\times}\psi\left(\frac{ax}{\varpi^{c_1}}\right)\chi_1(ax)$$

Here we used Corollary 3.23 for the third line. The sum is automatically 0 when $v(x)<0$. Note that when $v(x)>0$, the sum in a in the last line will be vanishing as the levels do not match. Thus by a change of variable, we have

$$\frac{1}{C_0}\sum_{a\in(O/\varpi^{c_1}O)^\times}\chi_1(a)W_{\varphi_a}\left(\begin{pmatrix}x&0\\0&1\end{pmatrix}\right)=\mathrm{char}(O^\times)=W_{\varphi_{new}}.$$

□

From now on we assume that $\pi=\pi\left(\chi_1,\chi_1^{-1}\right)$, $p\neq 2$ and $\mathfrak{c}(\chi_1)\geq 2$, so that

$$i_0=\mathfrak{c}(\chi_1). \tag{3.23}$$

Then the character $\tilde{\theta}$ can be rewritten as

$$\tilde{\theta}\left(z\begin{pmatrix}a&b\\c&d\end{pmatrix}\right)=\chi_1(a)\chi_1^{-1}(d),\ \text{for}\ \forall\begin{pmatrix}a&b\\c&d\end{pmatrix}\in K_0\left(\varpi^{i_0}\right). \tag{3.24}$$

Definition 3.25. Define $\Phi_{0,0}(g)=<\pi(g)\varphi_\theta,\varphi_\theta>$ with normalisation $\Phi_{0,0}(1)=1$, $\tilde{\Phi}_{0,0}=\Phi_{0,0}|_{ZK_0\left(\varpi^{i_0}\right)}$, and define for $a,a'\in\left(O/\varpi^{i_0}O\right)^\times$

$$\Phi_{a,a'}(g)=\chi_1(a)\chi_1^{-1}(a')\Phi_{0,0}\left(\begin{pmatrix}1&-a'\varpi^{-i_0}\\0&1\end{pmatrix}g\begin{pmatrix}1&a\varpi^{-i_0}\\0&1\end{pmatrix}\right) \tag{3.25}$$

$$\tilde{\Phi}_{a,a'}(g)=\chi_1(a)\chi_1^{-1}(a')\tilde{\Phi}_{0,0}\left(\begin{pmatrix}1&-a'\varpi^{-i_0}\\0&1\end{pmatrix}g\begin{pmatrix}1&a\varpi^{-i_0}\\0&1\end{pmatrix}\right) \tag{3.26}$$

Definition 3.26. Define the following test function

$$f(g)=\frac{1}{(p-1)p^{i_0-1}\mathrm{Vol}(Z\backslash ZK_0(\varpi^{i_0}))}\sum_{a,a'\in\left(O_{\mathbb{F}}/\varpi^{i_0}O_{\mathbb{F}}\right)^\times}\overline{\tilde{\Phi}_{a,a'}}(g). \tag{3.27}$$

Proposition 3.27. *For $\mathbb{L}$ split, f defined in* (3.27) *$w_\pi=1$ and $l_0=0$, Proposition 3.20 is true.*

Proof. The proof is parallel to that of Proposition 3.20. We first specify the orthonormal basis we are going to work with. First of all, the elements in the set

$$\left\{\pi\left(\begin{pmatrix}1&n\\0&1\end{pmatrix}\right)\varphi_\theta\ \middle|\ n\in\mathbb{F}/O_{\mathbb{F}}\right\} \tag{3.28}$$

are orthogonal to each other. The proof for this is exactly the same as (3.30) in the proof of Lemma 3.28. Then we complete an orthonormal basis B_π from (3.28).

As in the proof of Proposition 3.20, we get that $\pi\left(\frac{1}{\mathrm{Vol}\left(Z\backslash ZK_0\left(\varpi^{i_0}\right)\right)}\overline{\tilde{\Phi}}_{0,0}\right)$ is the projection onto the line spanned by φ_θ by Corollary 3.23. Then as $\overline{\chi_1}=\chi_1^{-1}$,

$$\pi\left(\frac{1}{\mathrm{Vol}\left(Z\backslash ZK_0\left(\varpi^{i_0}\right)\right)}\overline{\tilde{\Phi}}_{a,a'}\right)\varphi=\begin{cases}0, & \text{if } \varphi\in B_\pi, \varphi\neq\varphi_a,\\ \chi_1^{-1}(a)\chi_1(a')\varphi_{a'}, & \text{if } \varphi=\varphi_a\end{cases}$$

Using Lemma 3.24, we get that

$$\begin{aligned}\pi(f)\varphi_{new}&=\frac{1}{C_0(p-1)p^{i_0-1}}\sum_{a,a'}\pi\left(\frac{1}{\mathrm{Vol}\left(Z\backslash ZK_0\left(\varpi^{i_0}\right)\right)}\overline{\tilde{\Phi}}_{a,a'}\right)\sum_b\chi_1(b)\varphi_b\\&=\frac{1}{C_0(p-1)p^{i_0-1}}\sum_{a',b}\chi_1(a')\varphi_{a'}=\varphi_{new}.\end{aligned}$$

□

3.3.3. *K−type generated by φ_θ.* Let $K'=\left\{g\in\mathrm{GL}_2(\mathbb{F})\cap\begin{pmatrix}O & \varpi^{-i_0}O\\ \varpi^{i_0}O & O\end{pmatrix}\right\}$. We shall discuss the representation σ of K' generated by φ_θ here, which might have independent interest. It will also be used in Lemma 5.2.

Lemma 3.28. *Let $\pi=\pi\left(\chi_1,\chi_1^{-1}\right)$ be a unitary principal series representation, and $\varphi_\theta\in\pi$ be as in Corollary 3.23. Let σ be the representation of K' generated by φ_θ. The set $\{\pi(g)\varphi_\theta\}_{g\in K'/K_0\left(\varpi^{i_0}\right)}$ provides an orthonormal basis for the representation σ, which is dimension $[K':K_0\left(\varpi^{i_0}\right)]=(p+1)p^{i_0-1}$.*

Note that χ_1 is automatically a unitary character by the setting.

Proof. It is straightforward to verify that we can choose the coset representatives as follows:

$$K'/K_0\left(\varpi^{i_0}\right)=\coprod_{x\in\varpi^{-i_0}O/O}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}\cup\coprod_{x\in\varpi^{-i_0+1}O/O}\begin{pmatrix}0 & \varpi^{-i_0}\\ \varpi^{i_0} & 0\end{pmatrix}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}\tag{3.29}$$

where the RHS has exactly $(p+1)p^{i_0-1}$ elements.

Let g,g' be any different elements from the RHS of (3.29). By the invariance of the unitary pairing, we have $<\pi(g)\varphi_\theta,\pi(g')\varphi_\theta>=<\pi\left(g'^{-1}g\right)\varphi_\theta,\varphi_\theta>$, where $g'^{-1}g\in K'-K_0\left(\varpi^{i_0}\right)$.

Thus for the orthorgonality, it suffices to show that for any coset representative $g\neq 1$,

$$<\pi(g)\varphi_\theta,\varphi_\theta>=0.$$

Let $g=\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}$ for $x\notin O$ first. Then using Corollary 3.23,

$$\begin{aligned}<\pi(g)\varphi_\theta,\varphi_\theta>&=\int_{\alpha\in\mathbb{F}^\times}W_{\varphi_\theta}\left(\begin{pmatrix}\alpha & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}\right)\overline{W_{\varphi_\theta}\left(\begin{pmatrix}\alpha & 0\\ 0 & 1\end{pmatrix}\right)}d^\times\alpha=\int_{v(\alpha)\geq 0}p^{-v(\alpha)}\psi(\alpha x)\,d^\times\alpha\\&=\frac{1}{1-p^{-1}}\int_{v(\alpha)\geq 0}\psi(\alpha x)\,d\alpha=0\end{aligned}\tag{3.30}$$

Now let $g=\begin{pmatrix}0 & \varpi^{-i_0}\\ \varpi^{i_0} & 0\end{pmatrix}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}$ with $v(x)\geq -i_0+1$, let K be the standard maximal compact open subgroup. Then up to a constant multiple, we have by (2.4)

(3.31)

$$<\pi(g)\varphi_\theta,\varphi_\theta>=\int_{k\in K}\varphi_\theta\left(k\begin{pmatrix}0 & \varpi^{-i_0}\\ \varpi^{i_0} & 0\end{pmatrix}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}\right)\overline{\varphi_\theta}(k)\,dk\sim\int_{k\in K_0(\varpi^{i_0})}\varphi_\theta\left(k\begin{pmatrix}0 & \varpi^{-i_0}\\ \varpi^{i_0} & 0\end{pmatrix}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}\right)\tilde{\theta}^{-1}(k)\,dk$$

Here we have used that φ_θ in the parabolic induction model is only supported on $BK_0\left(\varpi^{i_0}\right)$. Writing $k=\begin{pmatrix}k_1 & k_2\\ \varpi^{i_0}k_3 & k_4\end{pmatrix}$, for $k_1,k_4\in O_{\mathbb{F}}^\times$ and $k_2,k_3\in O_{\mathbb{F}}$, we have

$$k\begin{pmatrix}0 & \varpi^{-i_0}\\ \varpi^{i_0} & 0\end{pmatrix}\begin{pmatrix}1 & x\\ 0 & 1\end{pmatrix}=\begin{pmatrix}k_2\varpi^{i_0} & k_1\varpi^{-i_0}+k_2x\varpi^{i_0}\\ k_4\varpi^{i_0} & k_3+k_4x\varpi^{i_0}\end{pmatrix}$$

As $v\left(k_4x\varpi^{i_0}\right)>0$, we need $v(k_3)=0$ for the matrix above to land in the support of φ_θ, which is $BK_0\left(\varpi^{i_0}\right)$. In that case we can write the matrix above as

$$\begin{pmatrix}-\frac{\det k}{k_3+k_4x\varpi^{i_0}} & k_1\varpi^{-i_0}+k_2x\varpi^{i_0}\\ 0 & k_3+k_4x\varpi^{i_0}\end{pmatrix}\begin{pmatrix}1 & 0\\ \frac{k_4\varpi^{i_0}}{k_3+k_4x\varpi^{i_0}} & 1\end{pmatrix},$$

thus

$$<\pi(g)\varphi_\theta,\varphi_\theta>\sim\int_{k\in K_0(\varpi^{i_0})}\chi_1\left(\frac{\det k}{k_3+k_4x\varpi^{i_0}}\right)\chi_1^{-1}\left(k_3+k_4x\varpi^{i_0}\right)\chi_1^{-1}(k_1)\chi_1(k_4)\,dk \tag{3.32}$$

$$=\int_{k\in K_0(\varpi^{i_0})}\chi_1^{-2}\left(\frac{k_3}{k_4}+x\varpi^{i_0}\right)dk=0$$

Here we have used (3.24) for $\tilde{\theta}(k)$, and that $\chi_1(\det k)=\chi_1(k_1k_4)$ as $\mathfrak{c}(\chi_1)=i_0$. □

4. A refined/specialized Petersson trace formula

Fix an étale quadratic algebra $\mathbb{L}$ over $\mathbb{F}=\mathbb{Q}_p$ at a fixed place p, and a character θ on $\mathbb{L}^\times$. Let $\mathfrak{c}(\pi_\theta)$ be the level of π_θ. Fix a weight $\kappa\geq 2$, $\kappa\equiv 0 \bmod 2$. Let n,i_0 be as in Definition 3.1. Define

(4.1) $\mathcal{F}_\theta[n]=\{$holomorphic newforms F of weight κ, level $N=p^c$, and trivial nebentypus s.t. $\pi_p\in\pi_\theta[n]$ where π_p is the local representation associated to $F\}$

We shall develop refined Petersson trace formula where only the members of $\mathcal{F}_\theta[n]$ appear on the spectral side. We shall start with smaller family and get the larger family by summation.

4.1. Test function.

We shall make the standard choice for the local test functions when $v\neq p$. In particular $f_v=\mathrm{char}(Z\mathrm{GL}_2(O_v^\times))$ for any non-archimedean place $v\neq p$. f_∞ is the conjugate of the matrix coefficient for the lowest weight element of π_∞, normalized to be an idempotent element under convolution. Explicitly one can take

$$f_\infty(g)=\begin{cases}\frac{\kappa-1}{4\pi}\frac{\det(g)^{\kappa/2}(2i)^\kappa}{(-b+c+(a+d)i)^\kappa}, & \text{if } \det(g)>0,\\ 0, & \text{otherwise.}\end{cases} \tag{4.2}$$

At the place p, f_p is chosen to be (3.19) or (3.27), depending on $\mathbb{L}$/the local representations we are interested in. By Proposition 3.20, 3.27, $\rho\left(f_p\right)$ is the projection onto the newforms from $\pi_p \simeq \pi_{\theta'}$ where $\theta' \in \theta[l_0]$, and

$$l_0 = \begin{cases} 1, & \text{if } \mathbb{L}/\mathbb{F} \text{ is an inert quadratic field extension,} \\ 0, & \text{otherwise.} \end{cases}$$

4.2. **Relative trace formula for integrals along unipotent orbits.** Let ψ be a fixed additive character of $\mathbb{Q}\backslash\mathbb{A}$. Recall from Definition 3.1 that

$$i_0 = \frac{\mathfrak{c}\,(\theta)}{e_{\mathbb{L}}}. \tag{4.3}$$

Here when $\mathbb{L} \simeq \mathbb{F}\times\mathbb{F}$, we use the convention that $\mathfrak{c}(\theta) = \mathfrak{c}(\chi)$ if $\theta = (\chi, \chi^{-1})$, and $e_{\mathbb{L}} = 1$.

Alternatively one can define

$$i_0 = \lfloor \frac{\mathfrak{c}(\pi_\theta)}{2} \rfloor. \tag{4.4}$$

To get the relative trace formula associated to unipotent period integrals, we start with a pretrace formula for proper f

$$\sum_{\varphi} \frac{1}{||\varphi||^2}\rho\,(f)\,\varphi\,(x)\,\overline{\varphi}\,(y) = \sum_{\gamma\in G(\mathbb{Q})} f(x^{-1}\gamma y) \tag{4.5}$$

The sum in φ is over some orthogonal basis for Automorphic forms, and $||\cdot||$ denotes the L^2-norm.

We choose the orthogonal basis to be extended from $\mathcal{F}_\theta[l_0]$. Then by the choice of f specified in Section 4.1, and Proposition 3.20 3.27, the sum for φ is actually over $\varphi \in \mathcal{F}_\theta[l_0]$ as in (4.1).

Integrating x, y in (4.5) along unipotent subgroups against additive characters, we obtain that that

$$\sum_{\varphi\in\mathcal{F}_\theta[l_0]} \frac{1}{||\varphi||^2} \iint\limits_{t_1,t_2\in N(\mathbb{Q})\backslash N(\mathbb{A})} \rho\,(f)\,\varphi\,(n(t_1))\,\overline{\varphi}\,(n(t_2))\,\psi\,(-m_1t_1 + m_2t_2)\,dt_1dt_2 = \sum_{\gamma\in N(\mathbb{Q})\backslash G(\mathbb{Q})/N(\mathbb{Q})} I(\gamma, f, m_1, m_2), \tag{4.6}$$

where

$$I(\gamma, f, m_1, m_2) = \int\limits_{h\in H_\gamma\backslash H(\mathbb{A})} f\left(\begin{pmatrix}1 & t_1\\ 0 & 1\end{pmatrix}^{-1}\gamma\begin{pmatrix}1 & t_2\\ 0 & 1\end{pmatrix}\right)\psi\,(-m_1t_1 + m_2t_2)\,d\,(t_1,t_2)\,,$$

$H = N\times N$, H_γ is the stabiliser of γ in $H(\mathbb{Q})$.

The period integrals on the left-hand side of (4.6) is directly related to the Whittaker function:

$$\int\limits_{t\in N(\mathbb{Q})\backslash N(\mathbb{A})} \varphi(n(t))\psi(-mt)dt = W_\varphi\left(\begin{pmatrix}m & \\ & 1\end{pmatrix}\right).$$

Using the discussions in Section 2.4 we can rewrite the spectral side of (4.6) as

$$(m_1m_2)^{\kappa/2-1/2}\,e^{-2\pi(m_1+m_2)} \sum_{\varphi\in\mathcal{F}_\theta[l_0]} \frac{1}{||\varphi||^2}\lambda_{m_1}(\varphi)\,\overline{\lambda}_{m_2}(\varphi)\,. \tag{4.7}$$

The main task is, of course, to analyze the geometric side of (4.6). For convenience, denote $f_{a,a'}$ to be the test function which agrees with f at all other places, and at p equals $\overline{\tilde{\Phi}}_{a,a'}$.

Note that using the same computations in [9, Corollary A.6], together with that

$$[\mathbb{L}:\mathbb{F}U_{\mathbb{L}}(1)]=\begin{cases}p+1, & \text{if } e_{\mathbb{L}}=1;\\ 2, & \text{if } e_{\mathbb{L}}=2.\end{cases}$$

we have for supercuspidal representation case

$$\mathrm{Vol}\left(Z\backslash ZB^1\right)=\frac{1}{(p^2-1)\,p^{i_0-1}}. \tag{4.8}$$

On the other hand for principal series representation case, it is also straightforward to check that

$$\mathrm{Vol}\left(Z\backslash ZK_0\left(\varpi^{i_0}\right)\right)=(p+1)\,p^{i_0-1}. \tag{4.9}$$

Denote by $D_{\mathcal{F}}$ the constant appearing in f_p, i.e.,

$$D_{\mathcal{F}}=\frac{1}{(p-1)\,p^{\lceil i_0/2\rceil-1}\mathrm{Vol}\left(Z\backslash ZB^1\right)}\asymp_p p^{c(\pi)/4} \tag{4.10}$$

when π_θ is a supercuspidal representation, and

$$D_{\mathcal{F}}=\frac{1}{(p-1)\,p^{i_0-1}\mathrm{Vol}\left(Z\backslash ZK_0\left(\varpi^{i_0}\right)\right)}\asymp 1 \tag{4.11}$$

when π_θ is a principal series representation.

Define

$$I\left(\gamma,a,a',m_1,m_2\right)=\iint\limits_{H_\gamma\backslash H_{\mathbb{A}}} f_{a,a'}\left(n\left(t_1\right)^{-1}\gamma n\left(t_2\right)\right)\psi\left(-m_1t_1+m_2t_2\right)d\left(t_1,t_2\right) \tag{4.12}$$

Now the geometric side of (4.6) becomes

$$D_{\mathcal{F}}\sum_{a,a'\in\left(O_{\mathbb{F}}/\varpi^{\lceil i_0/d_{\mathbb{L}}\rceil}O_{\mathbb{F}}\right)^\times}\ \sum_{\gamma\in N\backslash G(\mathbb{Q})/N} I\left(\gamma,a,a',m_1,m_2\right). \tag{4.13}$$

Here $d_{\mathbb{L}}=2$ when $\mathbb{L}$ is a field, and $d_{\mathbb{L}}=1$ when $\mathbb{L}$ is split.

Also recall that by the Bruhat decomposition, $N\backslash G(\mathbb{Q})/N$ consists of first-cell terms $\begin{pmatrix}\mu & \\ & 1\end{pmatrix}$ for $\mu\in\mathbb{Q}^\times$, as well as second-cell terms $\begin{pmatrix} & -\mu\\ 1 & \end{pmatrix}$, $\mu\in\mathbb{Q}^\times$. We shall discuss the corresponding orbit integrals $I\left(\gamma,a,a',m_1,m_2\right)$ in the next two subsections.

4.3. **Geometric side: First cell terms.** The manipulations and the local factors at $v\neq p$ for first-cell terms and second-cell terms are the same as in [14, Section 3][13, Section 7]. When $\gamma=\begin{pmatrix}\mu & \\ & 1\end{pmatrix}$, $H_\gamma=\{(n(\mu t),n(t))\in N(\mathbb{Q})^2\}$. We get that

$$\begin{aligned}I\left(\gamma,a,a',m_1,m_2\right)&=\iint\limits_{\{(\mu t,t)\in\mathbb{Q}^2\}\backslash\mathbb{A}^2} f_{a,a'}\left(\begin{pmatrix}\mu & \mu t_2-t_1\\ 0 & 1\end{pmatrix}\right)\psi\left(-m_1t_1+m_2t_2\right)dt_1dt_2\\ &=\int\limits_{x\in\mathbb{A}}\int\limits_{t_2\in\mathbb{Q}\backslash\mathbb{A}} f_{a,a'}\left(\begin{pmatrix}\mu & x\\ 0 & 1\end{pmatrix}\right)\psi\left(m_1x\right)\psi\left(\left(m_2-\mu m_1\right)t_2\right)dx\,dt_2.\end{aligned}$$

Here we made a change of variable $x = \mu t_2 - t_1$. As ψ is nontrivial, the integral in t_2 is nontrivial only when $\mu = \frac{m_2}{m_1}$. In that case, we write $m_1 x = t$ and get that

$$I(\gamma, a, a', m_1, m_2) = \int_{t\in\mathbb{A}} f_{a,a'}\left(\begin{pmatrix} m_2 & t \\ 0 & m_1 \end{pmatrix}\right)\psi(t)\,dt \tag{4.14}$$

which is factorisable. At all finite places, we need $v(m_1) = v(m_2) \geq 0$ for the local factor to be nonvanishing. At ∞, we get $m_1 m_2 > 0$. So $I(\gamma, a, a', m_1, m_2)$ is non-vanishing only when $m_1 = m_2$.

For finite $v \neq p$, we have

$$\int_{t\in\mathbb{Q}_v} f_v\left(\begin{pmatrix} m_1 & t \\ 0 & m_1 \end{pmatrix}\right)\psi_v(t)\,dt = \begin{cases} \|m_1\|_v, & \text{if } v(m_i) \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

For $v = \infty$, $m_1, m_2 > 0$, we have according to [14, Proposition 3.4]

$$\int_{t\in\mathbb{Q}_v} f_v\left(\begin{pmatrix} m_2 & t \\ 0 & m_1 \end{pmatrix}\right)\psi_v(t)\,dt = \frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}(m_1 m_2)^{\kappa/2} e^{-2\pi(m_1+m_2)}.$$

For $v = p$, and $\mathbb{L}$ is a field, we have by Corollary 3.18,

$$\begin{aligned} \int_{t\in\mathbb{Q}_p} f_{a,a',p}\left(\begin{pmatrix} m_1 & t \\ 0 & m_1 \end{pmatrix}\right)\psi_p(t)\,dt &= \int_{t\in\mathbb{Q}_p} \overline{\tilde{\Phi}}_{a,a'}\left(\begin{pmatrix} m_1 & t \\ 0 & m_1 \end{pmatrix}\right)\psi_p(t)\,dt \\ &= \int_{t\in\mathbb{Q}_p} \overline{\tilde{\Phi}}_{0,0}\left(\begin{pmatrix} \frac{a'}{a}m_1 & \varpi^{i_0}a't \\ 0 & m_1 \end{pmatrix}\right)\psi_p(t)\,dt \end{aligned}$$

Note that $\begin{pmatrix} \frac{a'}{a}m_1 & \varpi^{i_0}a't \\ 0 & m_1 \end{pmatrix} \in ZB^1$ iff $a' \equiv a \bmod \varpi_p^{\lceil i_0/2\rceil}$ and $v(t) - v(m_1) \geq -\lceil i_0/2\rceil$, in which case

$$\int_{t\in\mathbb{Q}_p} f_{a,a',p}\left(\begin{pmatrix} m_1 & t \\ 0 & m_1 \end{pmatrix}\right)\psi_p(t)\,dt = \int_{v(t)-v(m_1)\geq-\lceil i_0/2\rceil} \psi_p\left(-\alpha_0 a\frac{t}{m_1}\right)\psi_p(t)\,dt$$

which is nonzero iff $v(m_1) = 0$ and $a \equiv \frac{m_1}{\alpha_0} \bmod \varpi_p^{\lceil i_0/2\rceil}$, in which case the value is $p^{\lceil i_0/2\rceil}$.

In this case we obtain that when $m_1, m_2 \in \mathbb{Z}_{>0}$, $(m_i, p) = 1$,

$$D_{\mathcal{F}} \sum_{a,a'\in\left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^\times} I(\gamma, a, a', m_1, m_2) = \delta_{m_1=m_2}\frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}m_1^{\kappa-1}e^{-4\pi m_1}D_{\mathcal{F}}p^{\lceil i_0/2\rceil} \asymp_p \delta_{m_1=m_2}p^{\mathfrak{c}(\pi)/2}. \tag{4.15}$$

When $v = p$ and $\mathbb{L}$ is split,

$$\begin{aligned}\int\limits_{t\in\mathbb{Q}_p} f_{a,a',p}\left(\begin{pmatrix} m_1 & t \\ 0 & m_1\end{pmatrix}\right)\psi_p(t)\,dt &= \int\limits_{t\in\mathbb{Q}_p} \overline{\tilde{\Phi}}_{a,a'}\left(\begin{pmatrix} m_1 & t \\ 0 & m_1\end{pmatrix}\right)\psi_p(t)\,dt \\ &= \chi_1^{-1}(a)\chi_1(a')\int\limits_{t\in\mathbb{Q}_p}\overline{\tilde{\Phi}}_{0,0}\left(\begin{pmatrix} m_1 & \varpi^{-i_0}m_1(a-a')+t \\ 0 & m_1\end{pmatrix}\right)\psi_p(t)\,dt \\ &= \chi_1^{-1}(a)\chi_1(a')\int\limits_{t\in-\varpi^{-i_0}m_1(a-a')+m_1O_{\mathbb{F}}}\psi_p(t)\,dt \\ &= \delta_{v(m_1)\geq 0}\chi_1^{-1}(a)\chi_1(a')\|m_1\|_v\psi_p\left(-\varpi^{-i_0}m_1(a-a')\right)\end{aligned}$$

The sum over a, a' would now be vanishing unless $v(m_1) = 0$. In that case we obtain that

(4.16)

$$\begin{aligned}D_{\mathcal{F}}\sum_{a,a'\in\left(O_{\mathbb{F}}/\varpi^{i_0}O_{\mathbb{F}}\right)^{\times}} I(\gamma,a,a',m_1,m_2) &= \delta_{m_1=m_2}\frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}m_1^{\kappa-1}e^{-4\pi m_1}D_{\mathcal{F}}|\sum_{a'}\chi_1(a')\psi_p\left(\varpi^{-i_0}m_1a'\right)|^2 \\ &= \delta_{m_1=m_2}\frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}m_1^{\kappa-1}e^{-4\pi m_1}D_{\mathcal{F}}(p-1)p^{i_0-1}\asymp p^{c(\pi)/2}.\end{aligned}$$

4.4. **Geometric side: Second cell term.** This is probably the most technical part of the paper, requiring more careful computations for the test function f_p.

For $v(\mu) \leq 0$ even, denote the classical Kloosterman sum

$$\mathrm{KL}_v(a,b,\mu) = \sum_{t_1,t_2\in\left(\varpi_v^{v(\mu)/2}O_v/O_v\right), t_1t_2\equiv\mu \bmod O_v}\psi_v(at_1+bt_2) \tag{4.17}$$

where the additive character ψ_v is assumed to be unramified.

First of all, as in the standard situation, we have for $\gamma = \begin{pmatrix} 0 & -\mu \\ 1 & 0\end{pmatrix}$, $H_\gamma = 1$ and

$$I(\gamma,a,a',m_1,m_1) = \int\limits_{\mathbb{A}^2} f_{a,a'}\left(\begin{pmatrix} 1 & t_1 \\ 0 & 1\end{pmatrix}^{-1}\begin{pmatrix} 0 & -\mu \\ 1 & 0\end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1\end{pmatrix}\right)\psi(-m_1t_1+m_2t_2)\,dt_1dt_2, \tag{4.18}$$

which is factorisable. The computation at the archimedean place and unramified places are the same as in [14].

At unramified places, the local factor is nonvanishing iff $v(\mu) \leq 0$ is even. Then

$$\begin{aligned}I_v(\gamma,a,a',m_1,m_2) &= \int\limits_{\mathbb{F}_v^2} f_v\left(\begin{pmatrix} -t_1 & -\mu-t_1t_2 \\ 1 & t_2\end{pmatrix}\right)\psi(-m_1t_1+m_2t_2)\,dt_1dt_2 \\ &= \mathrm{KL}_v(m_1,m_2,\mu).\end{aligned} \tag{4.19}$$

At ∞, the local factor is nonvanishing iff $m_i, \mu > 0$, in which case

$$I_\infty(\gamma,a,a',m_1,m_2) = \frac{e^{-2\pi(m_1+m_2)}(4\pi i)^\kappa\sqrt{m_1m_2}^{\kappa-1}}{2(\kappa-2)!}\mu^{1/2}J_{\kappa-1}(4\pi\sqrt{\mu m_1m_2}). \tag{4.20}$$

At the place p, the computations are more complicated. The basic strategy is to compute first $I\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, a, a', m_1, m_2\right)_p$ for a single pair of (a, a'), and then relate to others by a simple change of variable.

4.4.1. *Supercuspidal case.*

Lemma 4.1. *Suppose* $I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, 1, 1, m_1, m_2\right) \neq 0$. *Then we must have* $m_1 \equiv m_2 \equiv \alpha_0$ *mod* $\varpi^{\lceil i_0/2\rceil}$, *where* α_0 *is as in* (3.13).

Proof. By making change of variable $t_2 \to t_2 + \Delta t_2$ for $\Delta t_2 \in \varpi^{-\lceil i_0/2\rceil}O_{\mathbb{F}}$, and note that $\begin{pmatrix}1 & \Delta t_2\\ 0 & 1\end{pmatrix} \in \operatorname{Supp}\tilde{\Phi}_{1,1}$, we get by Corollary 3.18 that the integral is non-vanishing only if

$$\psi(-\alpha_0\Delta t_2)\,\psi(m_2\Delta t_2) = 1,\ i.e.\ m_2 \equiv \alpha_0 \bmod \varpi^{\lceil i_0/2\rceil}.$$

Similarly by a change of variable for t_1, we get that $m_1 \equiv \alpha_0$ mod $\varpi^{\lceil i_0/2\rceil}$. □

To compute $I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, 1, 1, m_1, m_2\right)$ explicitly when $m_1 \equiv m_2 \equiv \alpha_0$ mod $\varpi^{\lceil i_0/2\rceil}$, we care about when $\begin{pmatrix}-t_1 & -\mu - t_1t_2\\ 1 & t_2\end{pmatrix} \in \operatorname{Supp}\tilde{\Phi}_{1,1}$. By considering the determinant, we see that $v(\mu) = -2k$ must be even (including the $e_{\mathbb{L}} = 2$ case, by the choice of $f_{1,1}$). Then

Lemma 4.2. $\begin{pmatrix}-t_1 & -\mu - t_1t_2\\ 1 & t_2\end{pmatrix} \in$ *Supp* $\tilde{\Phi}_{1,1}$ *if and only if all the followings hold*

(i) $\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right) \in ZU_{\mathbb{L}}(1)$.
(ii) $t_2^2 - \frac{1}{D\varpi^{2i_0}} \equiv \mu$ *mod* $\varpi^{v(\mu)+\lceil i_0/2\rceil}$.
(iii) $t_1 \equiv -\frac{\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2$ *mod* $\varpi^{-\lceil i_0/2\rceil}$.

In that case, we have

$$\begin{pmatrix}-t_1 & -\mu - t_1t_2\\ 1 & t_2\end{pmatrix} = \begin{pmatrix}\frac{\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}} & -t_1 - \frac{\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2\\ 0 & 1\end{pmatrix}\begin{pmatrix}t_2 & \frac{1}{D\varpi^{2i_0}}\\ 1 & t_2\end{pmatrix} \tag{4.21}$$

and

$$f_{1,1,p}\left(\begin{pmatrix}-t_1 & -\mu - t_1t_2\\ 1 & t_2\end{pmatrix}\right) = \theta^{-1}\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\alpha_0\left(t_1 + \frac{\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2\right)\right). \tag{4.22}$$

Proof. The matrix decomposition (4.21) is direct to check, while the remaining statements follow directly from the definition $f_{1,1,p} = \overline{\tilde{\Phi}}_{1,1}$ and Corollary 3.18. □

We make an explicit description of admissible values for $v(\mu)$ and $v(t_2)$.

Corollary 4.3. *When the set satisfying (i)-(iii) is non-empty, we must have* $v(\mu) = -2k < -\mathfrak{c}(\pi_\theta)$, *and* $v(t_2) = -k < -i_0$.

Proof. Consider the case $e_{\mathbb{L}} = 1$ first. From Lemma 4.2(i), we get that $v(t_2) < -i_0$. From (ii), we get $v(\mu) = 2v(t_2) < -2i_0 = -\mathfrak{c}(\pi_\theta)$. When $e_{\mathbb{L}} = 2$, we also get $v(t_2) < -i_0$ from (i), and $v(\mu) = 2v(t_2) < -2i_0 - 1 = -\mathfrak{c}(\pi_\theta)$ from (ii). □

Under the conditions in Lemma 4.1, 4.2, we have

(4.23)

$$
\begin{aligned}
& I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, 1, 1, m_1, m_2\right)\\
&= \int_{t_i \text{ satisfying (i)-(iii)}} \theta^{-1}\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\alpha_0\left(t_1 + \frac{\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2\right)\right)\psi\left(-m_1t_1 + m_2t_2\right)dt_1dt_2\\
&= \int_{t_2 \text{ satisfying (i)-(ii)}} \theta^{-1}\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\frac{\alpha_0\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2 + m_2t_2\right)\int_{t_1 \text{ satisfying (iii)}} \psi\left(\left(\alpha_0 - m_1\right)t_1\right)dt_1dt_2\\
&= p^{\lceil i_0/2\rceil}\int_{t_2 \text{ satisfying (i)-(ii)}} \theta^{-1}\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\frac{\alpha_0\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2 + m_2t_2\right)\psi\left(-\left(\alpha_0 - m_1\right)\frac{\mu t_2}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}\right)dt_2\\
&= p^{\lceil i_0/2\rceil}\int_{t_2 \text{ satisfying (i)-(ii)}} \theta^{-1}\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\frac{m_1\mu}{t_2^2 - \frac{1}{D\varpi^{2i_0}}}t_2 + m_2t_2\right)dt_2.
\end{aligned}
$$

Here in the third line, we have used Lemma 4.1, so that the integrand is constant for the integral in t_1 with the domain given in (iii).

For a general pair (a, a'), we have

Lemma 4.4.

$$
I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, a, a', m_1, m_2\right) = I_p\left(\begin{pmatrix}0 & -\mu aa'\\ 1 & 0\end{pmatrix}, 1, 1, a'^{-1}m_1, a^{-1}m_2\right).
$$

Proof. By definition,

(4.24)

$$
\begin{aligned}
I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, a, a', m_1, m_2\right) &= \int_{\mathbb{F}^2} \overline{\tilde{\Phi}}_{a,a'}\left(\begin{pmatrix}1 & t_1\\ 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & t_2\\ 0 & 1\end{pmatrix}\right)\psi\left(-m_1t_1 + m_2t_2\right)dt_1dt_2\\
&= \int_{\mathbb{F}^2} \overline{\tilde{\Phi}}_{1,1'}\left(\begin{pmatrix}a' & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & t_1\\ 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & t_2\\ 0 & 1\end{pmatrix}\begin{pmatrix}a^{-1} & 0\\ 0 & 1\end{pmatrix}\right)\psi\left(-m_1t_1 + m_2t_2\right)dt_1dt_2\\
&= \int_{\mathbb{F}^2} \overline{\tilde{\Phi}}_{1,1'}\left(\begin{pmatrix}1 & a't_1\\ 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}0 & -\mu aa'\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & at_2\\ 0 & 1\end{pmatrix}\right)\psi\left(-m_1t_1 + m_2t_2\right)dt_1dt_2\\
&= \int_{\mathbb{F}^2} \overline{\tilde{\Phi}}_{1,1'}\left(\begin{pmatrix}1 & t_1\\ 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}0 & -\mu aa'\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & t_2\\ 0 & 1\end{pmatrix}\right)\psi\left(-a'^{-1}m_1t_1 + a^{-1}m_2t_2\right)dt_1dt_2
\end{aligned}
$$

(4.25)
$$
= I_p\left(\begin{pmatrix}0 & -\mu aa'\\ 1 & 0\end{pmatrix}, 1, 1, a'^{-1}m_1, a^{-1}m_2\right).
$$

□

Note that a, a' are defined $\mod \varpi^{\lceil i_0/2\rceil}$, and the local integral should be independent of the choice of representatives. Combining the previous lemmas, we get that $I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, a, a', m_1, m_2\right)$ is non-vanishing iff $a'\alpha_0 \equiv m_1 \mod \varpi^{\lceil i_0/2\rceil}$, $a\alpha_0 \equiv m_2 \mod \varpi^{\lceil i_0/2\rceil}$, in which case we simply choose a', a such that $a'\alpha_0 = m_1$, $a\alpha_0 = m_2$.

As a result, we have for fixed m_1, m_2,

(4.26)

$$\sum_{a,a'} I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, a, a', m_1, m_2\right) = I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, \alpha_0^{-1}m_2, \alpha_0^{-1}m_1, m_1, m_2\right) = I_p\left(\begin{pmatrix}0 & -\alpha_0^{-2}\mu m_1 m_2\\ 1 & 0\end{pmatrix}, 1, 1, \alpha_0, \alpha_0\right)$$

$$= p^{\lceil i_0/2\rceil} \int\limits_{t_2 \text{ satisfying (i)}, t_2^2 - \frac{1}{D\varpi^{2i_0}} \equiv \frac{m_1 m_2 \mu}{\alpha_0^2} \mod \varpi^{v(\mu)+\lceil i_0/2\rceil}} \theta^{-1}\left(t_2 + \frac{1}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\frac{m_1 m_2 \mu}{\alpha_0\left(t_2^2 - \frac{1}{D\varpi^{2i_0}}\right)}t_2 + \alpha_0 t_2\right)dt_2$$

$$= p^{\lceil i_0/2\rceil} \int\limits_{t_2+\alpha_\theta \in ZU_{\mathbb{L}}(1), \mathrm{Nm}(t_2+\alpha_\theta)\equiv m_1 m_2 \mu \mod \varpi^{v(\mu)+\lceil i_0/2\rceil}} \theta^{-1}(t_2+\alpha_\theta)\psi\left(\frac{m_1 m_2\mu}{t_2^2 - \frac{\alpha_0^2}{D\varpi^{2i_0}}}t_2 + t_2\right)dt_2$$

In the last line we have made a change of variable $\alpha_0 t_2 \to t_2$, and used that $\alpha_\theta = \frac{\alpha_0}{\sqrt{D}\varpi^{i_0}}$, $\theta|_{\mathbb{F}^\times} = 1$. Note that we can alternatively write

(4.27)

$$\begin{aligned} G_p(m_1, m_2, \theta, \mu) &= \int\limits_{t_2+\alpha_\theta \in ZU_{\mathbb{L}}(1), \mathrm{Nm}(t_2+\alpha_\theta)\equiv m_1 m_2\mu \mod \varpi^{v(\mu)+\lceil i_0/2\rceil}} \theta^{-1}(t_2+\alpha_\theta)\psi\left(\frac{m_1 m_2\mu}{t_2^2 - \frac{\alpha_0^2}{D\varpi^{2i_0}}}t_2 + t_2\right)dt_2 \\ &= \int\limits_{e=t_2+\alpha_\theta \in ZU_{\mathbb{L}}(1), \mathrm{Nm}(e)\equiv m_1 m_2\mu \mod \varpi^{v(\mu)+\lceil i_0/2\rceil}} \theta^{-1}(e)\psi\circ\mathrm{Tr}\left(\frac{1}{2}\left(\frac{m_1 m_2\mu}{e} + e\right)\right)de \end{aligned}$$

Lemma 4.5. *When $k > i_0$, we can adjust the congruence requirement for t_2, i.e.,*

$$(4.28)\qquad \begin{aligned} G_p(m_1, m_2, \theta, \mu) &= \int\limits_{v(t_2)=-k} \theta^{-1}(t_2+\alpha_\theta)\psi\left(\frac{m_1 m_2\mu}{t_2^2 - \frac{\alpha_0^2}{D\varpi^{2i_0}}}t_2 + t_2\right)dt_2 \\ &= \int\limits_{Nm(t_2+\alpha_\theta)\equiv m_1 m_2\mu \mod \varpi^{v(\mu)+i}} \theta^{-1}\left(t_2 + \frac{\alpha_0}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\frac{m_1 m_2\mu}{t_2^2 - \frac{\alpha_0^2}{D\varpi^{2i_0}}}t_2 + t_2\right)dt_2 \end{aligned}$$

for any $0 < i \le \lfloor k/2\rfloor$. In particular we have the square-root cancellation for the generalized Kloosterman sum:

$$G_p(m_1, m_2, \theta, \mu) \ll_p p^{k/2}.$$

Proof. When $k > i_0$, $t_2 + \alpha_\theta \in ZU_{\mathbb{L}}(1)$ follows directly from $v(t_2) = -k$. We apply the p-adic analogue of the stationary phase analysis. Writing $t_2 = t_0(1 + dt)$, with $v(dt) \ge \lceil k/2\rceil$, we have

$$(4.29)\qquad \theta^{-1}\left(t_2 + \frac{\alpha_0}{\sqrt{D}\varpi^{i_0}}\right) = \theta^{-1}\left(t_0 + \frac{\alpha_0}{\sqrt{D}\varpi^{i_0}}\right)\psi\left(\frac{\frac{2\alpha_0^2 t_0 dt}{D\varpi^{2i_0}}}{t_0^2 - \frac{\alpha_0^2}{D\varpi^{2i_0}}}\right),$$

$$\psi\left(\frac{m_1m_2\mu}{t_2^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}t_2+t_2\right)=\psi\left(\frac{m_1m_2\mu}{t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}t_0+t_0\right)\psi\left(-\frac{m_1m_2\mu\left(t_0^2+\frac{\alpha_0^2}{D\varpi^{2i_0}}\right)}{\left(t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}\right)^2}t_0dt+t_0dt\right). \tag{4.30}$$

The stationary point has to satisfy

$$\frac{\frac{2\alpha_0^2}{D\varpi^{2i_0}}}{t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}-\frac{m_1m_2\mu\left(t_0^2+\frac{\alpha_0^2}{D\varpi^{2i_0}}\right)}{\left(t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}\right)^2}+1\equiv 0 \bmod \varpi^{\lfloor k/2\rfloor}. \tag{4.31}$$

This equation factorizes as

$$\left(1-\frac{m_1m_2\mu}{t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}\right)\frac{t_0^2+\frac{\alpha_0^2}{D\varpi^{2i_0}}}{t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}\equiv 0 \bmod \varpi^{\lfloor k/2\rfloor}. \tag{4.32}$$

When $k > i_0$, we have $\lfloor k/2\rfloor \geq \lceil i_0/2\rceil$, and $\frac{t_0^2+\frac{\alpha_0^2}{D\varpi^{2i_0}}}{t_0^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}\not\equiv 0$. Thus the stationary point must satisfy the congruence condition imposed in (4.27), and the nonzero contribution comes only from $\mathrm{Nm}\left(t_0+\alpha_\theta\right)\equiv m_1m_2\mu \bmod \varpi^{v(\mu)+\lfloor k/2\rfloor}$. The square-root cancellation follows directly from this requirement for the stationary point. □

Remark 4.6. The freedom to adjust the congruence condition for t_2 is later used in the proof of Lemma 4.17 to obtain cancellations among second-cell terms for different θ.

Remark 4.7. As a sanity check, we show that when $k \geq \mathfrak{c}(\pi)$, the local integral $G_p(m_1,m_2,\theta,\mu)$ reduces to the usual Kloosterman sum. Indeed in that case, we have $\theta^{-1}\left(t_2+\alpha_\theta\right)=1$ by the level of θ, and

$$\psi\left(\frac{m_1m_2\mu}{t_2^2-\frac{\alpha_0^2}{D\varpi^{2i_0}}}t_2+t_2\right)=\psi\left(t_2+\frac{m_1m_2\mu}{t_2}\left(1+\frac{\alpha_0^2}{Dt_2^2\varpi^{2i_0}}+\cdots\right)\right)=\psi\left(t_2+\frac{m_1m_2\mu}{t_2}\right).$$

4.4.2. *Principal series representation case.* In this case, it is easier to compute $I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix},0,0,m_1,m_2\right)$ first, i.e., to use $\tilde{\Phi}_{0,0}$ as the test function.

Lemma 4.8. $\begin{pmatrix}-t_1 & -\mu-t_1t_2\\ 1 & t_2\end{pmatrix}\in ZK_0\left(\varpi^{i_0}\right)$ *if and only if all the followings hold*

(1) $v(\mu)=-2k$, $v(t_1)=v(t_2)=-k\leq -i_0$*;*
(2) $t_1t_2\equiv -\mu \bmod \varpi^{-k}$.

In that case, we have

$$I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix},0,0,m_1,m_2\right)=\int_{v(t_2)=-k}\chi_1^{-1}(\mu)\chi_1^2(t_2)\psi\left(\frac{m_1\mu}{t_2}+m_2t_2\right)dt_2 \tag{4.33}$$

Proof. Note that this case is very similar to the classical case where f_p is the characteristic function of a congruence subgroup. By considering the determinant, we get that $v(\mu)=-2k$ for some $k\in\mathbb{Z}$.

Thus $\varpi^k \begin{pmatrix} -t_1 & -\mu - t_1 t_2 \\ 1 & t_2 \end{pmatrix} \in K_0\left(\varpi^{i_0}\right)$, giving rise to all the conditions for t_i and k. Then by (3.24), Definition 3.25 and $t_1 \equiv -\frac{\mu}{t_2} \bmod O_{\mathbb{F}}$,

$$I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, 0, 0, m_1, m_2\right) = \int\limits_{v(t_2)=-k} \chi_1^{-1}\left(\varpi^k \frac{\mu}{t_2}\right)\chi_1\left(\varpi^k t_2\right)\psi\left(\frac{m_1\mu}{t_2} + m_2 t_2\right)dt_2$$
$$= \int\limits_{v(t_2)=-k} \chi_1^{-1}(\mu)\chi_1^2(t_2)\psi\left(\frac{m_1\mu}{t_2} + m_2 t_2\right)dt_2 \tag{4.34}$$

□

For a general pair (a, a'), we have

Lemma 4.9.

$$I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, a, a', m_1, m_2\right) = \chi_1^{-1}(a)\psi\left(-m_2 a\varpi^{-i_0}\right)\chi_1(a')\psi\left(m_1 a'\varpi^{-i_0}\right) I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, 0, 0, m_1, m_2\right).$$

Proof. By Definition 3.25,

(4.35)

$$I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, a, a', m_1, m_2\right)$$
$$= \int\limits_{\mathbb{F}^2} \overline{\tilde{\Phi}}_{a,a'}\left(\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\right)\psi(-m_1 t_1 + m_2 t_2)\,dt_1 dt_2$$
$$= \chi_1^{-1}(a)\chi_1(a')\int\limits_{\mathbb{F}^2} \overline{\tilde{\Phi}}_{0,0}\left(\begin{pmatrix} 1 & -a'\varpi^{-i_0} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & a\varpi^{-i_0} \\ 0 & 1 \end{pmatrix}\right)\psi(-m_1 t_1 + m_2 t_2)\,dt_1 dt_2$$
$$= \chi_1^{-1}(a)\psi\left(-m_2 a\varpi^{-i_0}\right)\chi_1(a')\psi\left(m_1 a'\varpi^{-i_0}\right) I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, 0, 0, m_1, m_2\right)$$

□

Corollary 4.10. $\sum\limits_{a,a'} I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, a, a', m_1, m_2\right)$ *is nonzero only when* $v_p(m_i) = 0$*, in which case*

$$\sum_{a,a'} I_p\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}, a, a', m_1, m_2\right) = (p-1)\,p^{i_0-1}\int\limits_{v(t_2)=-k} \chi_1^{-1}(m_1 m_2 \mu)\chi_1^2(t_2)\psi\left(\frac{m_1 m_2 \mu}{t_2} + t_2\right)dt_2.$$

Proof. By the previous discussions, we see indeed that

$$\sum_{a,a'} \chi_1^{-1}(a)\psi\left(-m_2 a\varpi^{-i_0}\right)\chi_1(a')\psi\left(m_1 a'\varpi^{-i_0}\right) \neq 0$$

iff $v_p(m_i) = 0$. In that case, by a change of variable, we have

$$\sum_{a,a'} I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, a, a', m_1, m_2\right) = \chi_1\left(m_1^{-1}m_2\right)\left|\sum_{a'}\chi_1(a')\psi\left(a'\varpi^{-i_0}\right)\right|^2 I_p\left(\begin{pmatrix}0 & -\mu\\ 1 & 0\end{pmatrix}, 0, 0, m_1, m_2\right)$$

$$= \chi_1\left(m_1^{-1}m_2\right)(p-1)\,p^{i_0-1}\int_{v(t_2)=-k}\chi_1^{-1}(\mu)\chi_1^2(t_2)\,\psi\left(\frac{m_1\mu}{t_2}+m_2t_2\right)dt_2$$

$$= (p-1)\,p^{i_0-1}\int_{v(t_2)=-k}\chi_1^{-1}(m_1m_2\mu)\chi_1^2(t_2)\,\psi\left(\frac{m_1m_2\mu}{t_2}+t_2\right)dt_2.$$

□

Definition 4.11. When $\mathbb{L}$ splits, denote

$$G_p(m_1, m_2, \theta, \mu) = \chi_1^{-1}(m_1m_2\mu)\int_{v(t_2)=-k}\chi_1^2(t_2)\,\psi\left(\frac{m_1m_2\mu}{t_2}+t_2\right)dt_2.$$

Lemma 4.12. *$G_p(m_1, m_2, \theta, \mu)$ is vanishing unless there exists t_2 such that $v_p(t_2) = -k$ and $t_2^2 + 2\alpha_{\chi_1}t_2 \equiv m_1m_2\mu \bmod \varpi^{-\lceil 3k/2\rceil}$. In that case, we have*

$$G_p(m_1, m_2, \theta, \mu) = \chi_1^{-1}(m_1m_2\mu)\int_{v(t_2)=-k}\chi_1^2(t_2)\,\psi\left(\frac{m_1m_2\mu}{t_2}+t_2\right)dt_2 \tag{4.36}$$

$$= \chi_1^{-1}(m_1m_2\mu)\int_{t_2^2+2\alpha_{\chi_1}t_2\equiv m_1m_2\mu \bmod \varpi^{v(\mu)+i}}\chi_1^2(t_2)\,\psi\left(\frac{m_1m_2\mu}{t_2}+t_2\right)dt_2.$$

Here $0 < i < \lfloor k/2\rfloor$. *In particular we have*

$$|G_p(m_1, m_2, \theta, \mu)| \ll_p p^{k/2}.$$

Proof. Let $t_2 = t_0(1+dt)$ for $v_p(dt) \geq \lceil k/2\rceil$. Then

$$G_p(m_1, m_2, \theta, \mu) = \chi_1^{-1}(m_1m_2\mu)\sum_{t_0}\chi_1\left(t_0^2\right)\psi\left(\frac{m_1m_2\mu}{t_0}+t_0\right)\int_{dt\in\varpi^{\lceil k/2\rceil}O_{\mathbb{F}}}\psi\left(2\alpha_{\chi_1}dt\right)\psi\left(-\frac{m_1m_2\mu}{t_0}dt+t_0dt\right).$$

The integral in dt is nonvanishing only if

$$2\alpha_{\chi_1} - \frac{m_1m_2\mu}{t_0} + t_0 \equiv 0 \bmod \varpi^{-\lceil k/2\rceil}$$

for some t_0. The claims follow now easily. □

Remark 4.13. Again when $k \geq 2i_0 = \mathfrak{c}(\pi_\theta)$, we get that the stationary points satisfy

$$t_2^2 \equiv m_1m_2\mu \bmod \varpi^{-k-i_0}, \quad \text{so } \chi_1\left(\frac{t_2^2}{m_1m_2\mu}\right) = \chi_1(1) = 1.$$

Then the generalized Kloosterman sum becomes the classical Kloosterman sum.

4.5. **Petersson trace formula for small families.**

Definition 4.14. Define the generalized Kloosterman sum to be

$$G(m_1, m_2, \theta, \mu) = G_p(m_1, m_2, \theta, \mu) \times \prod_{v \neq p \text{ finite}} \mathrm{KL}_v(m_1, m_2, \mu)$$

where $G_p(m_1, m_2, \theta, \mu)$ is given in Lemma 4.5/Definition 4.11 according to whether π_θ is a supercuspidal/principal series representation, and $\mathrm{KL}_v(m_1, m_2, \mu)$ is as in (4.17).

Recall that

$$l_0 = \begin{cases} 1, & \text{if } \mathbb{L} \text{ is an inert field extension} \\ 0, & \text{otherwise.} \end{cases} \tag{4.37}$$

Recall $D_{\mathcal{F}}$ is given in (4.10)/(4.11). Denote

$$C_{\mathcal{F}}[l_0] = D_{\mathcal{F}} \times \begin{cases} p^{\lceil i_0/2 \rceil}, & \text{if } \pi_\theta \text{ is supercuspidal,} \\ (p-1)\, p^{i_0 - 1}, & \text{otherwise.} \end{cases} \tag{4.38}$$

Then in either case, we have $C_{\mathcal{F}}[l_0] \asymp_p p^{i_0} \asymp_p \sqrt{C(\pi)}$, and

$$I_p(\gamma, f, m_1, m_2) = C_{\mathcal{F}}[l_0] G_p(m_1, m_2, \theta, \mu) \tag{4.39}$$

for second cell terms $\gamma = \begin{pmatrix} & -\mu \\ 1 & \end{pmatrix}$.

Definition 4.15. Let $c_0 = p^{i_0+1}$ when π_θ is a supercuspidal representation by Corollary 4.3, and $c_0 = p^{i_0}$ when π_θ is a principal series representation by Lemma 4.8.

Theorem 4.16.

$$\sum_{\varphi \in \mathcal{F}_\theta[l_0]} \frac{1}{\|\varphi\|^2} \lambda_{m_1}(\varphi) \overline{\lambda}_{m_2}(\varphi) = C_{\mathcal{F}}[l_0] \frac{(4\pi)^{\kappa-1}}{(\kappa-2)!} \left(\delta_{m_1=m_2} + 2\pi i^\kappa \sum_{c_0 | c} \frac{G\left(m_1, m_2, \theta, c^{-2}\right)}{c} J_{\kappa-1}\left(\frac{4\pi \sqrt{m_1 m_2}}{c} \right) \right)$$

Proof. Here we collect all the calculations we have done in the last three subsections. We start with the relative trace formula in (4.6). The spectral side is given in (4.7), while the geometric side is set up in (4.13). The first order terms on the geometric side are given in (4.15)/ (4.16).

The second cell terms are given in (4.26)/Corollary 4.10 at p, and in (4.19)(4.20) at other places. Note that the local requirements for μ implies that $\mu = \frac{1}{c^2}$ for $c_0 | c$.

We have also canceled $(m_1 m_2)^{k/2-1/2}\, e^{-2\pi(m_1+m_2)}$ from both sides for the final formula. □

4.6. **Spectral average.** For applications, it is helpful to be able to sum over a larger family than $\theta[l_0]$ on the spectral side, in order to reach a balance between the main terms and the complicated analysis of the error terms. The main idea is that with longer sum on the spectral side, the sum of the generalized Kloosterman sum should be shorter.

Let $l_0 \leq l < i_0$. For any $\theta' \in \theta[l]$, we apply Theorem 4.16 and get

$$\sum_{\varphi \in \mathcal{F}_{\theta'}[l_0]} \frac{1}{\|\varphi\|^2} \lambda_{m_1}(\varphi) \overline{\lambda}_{m_2}(\varphi) = C_{\mathcal{F}}[l_0] \frac{(4\pi)^{\kappa-1}}{(\kappa-2)!} \left(\delta_{m_1=m_2} + 2\pi i^\kappa \sum_{c_0 | c} \frac{G\left(m_1, m_2, \theta', c^{-2}\right)}{c} J_{\kappa-1}\left(\frac{4\pi \sqrt{m_1 m_2}}{c} \right) \right) \tag{4.40}$$

Note that $C_{\mathcal{F}}[l_0]$ depends only on $\mathbb{L}$ and $\mathfrak{c}(\theta)$.

We now take a sum of (4.40) over $\theta' \in \theta[l]/\sim_{l_0}$. The non-trivial observation is that there are further cancellation for the second order terms on the geometric side as below:

Lemma 4.17. *For $v(\mu) = -2k < -2i_0$, we have*

$$\frac{1}{[\theta[l]:\theta[l_0]]}\sum_{\theta'\in\theta[l]/\sim_{l_0}} G_p(m_1,m_2,\theta',\mu) = \begin{cases} G_p(m_1,m_2,\theta,\mu), & \text{if } k\geq v_p(c_0)+l-l_0 \\ 0, & \text{otherwise.}\end{cases} \tag{4.41}$$

Define

$$C_{\mathcal{F}}[l] = C_{\mathcal{F}}[l_0][\theta[l]:\theta[l_0]]. \tag{4.42}$$

It is clear from Lemma 3.5 that

$$C_{\mathcal{F}}[l] \asymp p^{l-l_0} C_{\mathcal{F}}[l_0]. \tag{4.43}$$

From Lemma 4.17, we immediately obtain the following result:

Theorem 4.18. *Let $c_l = c_0 p^{l-l_0}$. Then*

$$\sum_{\varphi\in\mathcal{F}_\theta[l]} \frac{1}{\|\varphi\|^2}\lambda_{m_1}(\varphi)\overline{\lambda}_{m_2}(\varphi) = C_{\mathcal{F}}[l]\frac{(4\pi)^{\kappa-1}}{(\kappa-2)!}\left(\delta_{m_1=m_2} + 2\pi i^\kappa \sum_{c_l|c}\frac{G\left(m_1,m_2,\theta,c^{-2}\right)}{c}J_{\kappa-1}\left(\frac{4\pi\sqrt{m_1m_2}}{c}\right)\right) \tag{4.44}$$

4.6.1. *Proof of Lemma 4.17: supercuspidal representation case.* Consider first the case where π_θ is a supercuspidal representation. Note that $v(\mathrm{Nm}(\alpha_{\theta'})) = -\mathfrak{c}(\pi_\theta)$, and $v_p(c_0) = i_0+1$ in this case. Suppose $k \geq v_p(c_0)+l-l_0$ first. For any $\theta'\in\theta[l]$, we have $\alpha_{\theta'}\in\alpha_\theta U_{\mathbb{F}}(i_0-l)$ by Lemma 3.8. Then we claim that

$$\begin{aligned} G_p\left(m_1,m_2,\theta',\frac{1}{c^2}\right) &= \int_{v(t_2)=-k}\theta'^{-1}(t_2+\alpha_{\theta'})\,\psi\left(\frac{m_1m_2\mu}{\mathrm{Nm}(t_2+\alpha_{\theta'})}t_2+t_2\right)dt_2 \\ &= \int_{v(t_2)=-k}\theta^{-1}(t_2+\alpha_\theta)\,\psi\left(\frac{m_1m_2\mu}{\mathrm{Nm}(t_2+\alpha_\theta)}t_2+t_2\right)dt_2. \end{aligned} \tag{4.45}$$

Here the first equality is Lemma 4.5. By the condition $\alpha_{\theta'}\in\alpha_\theta U_{\mathbb{F}}(i_0-j)$, we have $t_2+\alpha_{\theta'}\in (t_2+\alpha_\theta)U_{\mathbb{L}}(e_{\mathbb{L}}(k-\mathfrak{c}(\pi_\theta)/2+i_0-l))\subset(t_2+\alpha_\theta)U_{\mathbb{L}}(e_{\mathbb{L}}i_0)$. Here we have used that $\mathfrak{c}(\pi_\theta) = 2i_0+e_{\mathbb{L}}-1$. Thus $\theta'^{-1}(t_2+\alpha_{\theta'}) = \theta'^{-1}(t_2+\alpha_\theta)$ as $\mathfrak{c}(\theta) = i_0e_{\mathbb{L}}$; Similarly we have

$$\mathrm{Nm}(t_2+\alpha_{\theta'}) = t_2^2+\mathrm{Nm}(\alpha_{\theta'}) \in \left(t_2^2+\mathrm{Nm}(\alpha_\theta)\right)U_{\mathbb{F}}(2k-\mathfrak{c}(\pi_\theta)+i_0-l)\subset\left(t_2^2+\mathrm{Nm}(\alpha_\theta)\right)U_{\mathbb{F}}(k).$$

Thus by the Taylor expansion, $v(\mu) = -2k < -2i_0$, $v(t_2) = -k$,

$$\frac{m_1m_2\mu}{\mathrm{Nm}(t_2+\alpha_{\theta'})}t_2 \in \frac{m_1m_2\mu}{\mathrm{Nm}(t_2+\alpha_\theta)}t_2 + O_{\mathbb{F}}, \text{ so } \psi\left(\frac{m_1m_2\mu}{\mathrm{Nm}(t_2+\alpha_{\theta'})}t_2\right) = \psi\left(\frac{m_1m_2\mu}{\mathrm{Nm}(t_2+\alpha_\theta)}t_2\right).$$

Lastly $\theta'^{-1}(t_2+\alpha_\theta) = \theta^{-1}(t_2+\alpha_\theta)$, as $\mathfrak{c}\left(\theta^{-1}\theta'\right)\leq e_{\mathbb{L}}l$ while $t_2+\alpha_\theta\in ZU_{\mathbb{L}}\left(\frac{e_{\mathbb{L}}}{2}(2k-\mathfrak{c}(\pi_\theta))\right)\subset ZU_{\mathbb{L}}(e_{\mathbb{L}}l)$. Thus

$$\frac{1}{[\theta[l]:\theta[l_0]]}\sum_{\theta'\in\theta[l]/\sim_{l_0}} G_p(m_1,m_2,\theta',\mu) = G_p(m_1,m_2,\theta,\mu).$$

Consider now the case $v_p(c_0)\leq k < v_p(c_0)+l-l_0$. By the same argument as above, it is clear that for any $\theta_1\in\theta[l]$, and $\theta'\in\theta_1[k+l_0-v_p(c_0)]$, we have $G_p(m_1,m_2,\theta',\mu) = G_p(m_1,m_2,\theta_1,\mu)$.

We shall average over slightly larger family $\theta' \in \theta_1[j]$ for $j = k + l_0 - v_p(c_0) + 1$, so that we will see the cancellation while only have to deal with the first order terms and first digits for the p-adic stationary phase analysis. Note that $j \leq l$ by the condition on k. Then we claim that for any $\theta_1 \in \theta[l]$,

(4.46)
$$\sum_{\theta' \in \theta_1[j]} G_p(m_1, m_2, \theta', \mu) = \sum_{\theta' \in \theta_1[j]} \int_{t_2^2 \equiv m_1 m_2 \mu \bmod \varpi^{v(\mu)+1}} \theta'^{-1}(t_2 + \alpha_{\theta'}) \psi\left(\frac{m_1 m_2 \mu}{\mathrm{Nm}(t_2 + \alpha_{\theta'})} t_2 + t_2\right) dt_2 = 0.$$

Then a further sum over $\theta_1 \in \theta[l]/\sim_j$ would also be vanishing.

For the first equality in (4.46), we apply Lemma 4.5 for $i = 1$. Note that $v_p(t_2^2) < v_p(\mathrm{Nm}(\alpha_{\theta_1}))$ as $k \geq i_0 + 1$ in the supercuspidal representation case, the congruence requirement $\mathrm{Nm}(t_2 + \alpha_\theta) \equiv m_1 m_2 \mu \bmod \varpi^{v(\mu)+1}$ is the same as $t_2^2 \equiv m_1 m_2 \mu \bmod \varpi^{v(\mu)+1}$, which is independent of θ'.

For the second equality of (4.46) , we write $\alpha_{\theta'} = \alpha_{\theta_1} + \alpha_{\theta_1} u$ for $u \in \varpi^{i_0 - j} O_{\mathbb{F}}$. Then by Lemma 3.8, the sum over $\theta_1[j]/\sim_{j-1}$ is parameterized by the sum over $u \in \varpi^{i_0-j} O_{\mathbb{F}}/\varpi^{i_0-j+1} O_{\mathbb{F}}$. By the same argument as above, we have $t_2 + \alpha_{\theta'} \in (t_2 + \alpha_{\theta_1}) U_{\mathbb{L}}(e_{\mathbb{L}}(k - \mathfrak{c}(\pi_\theta)/2 + i_0 - j)) = (t_2 + \alpha_{\theta_1}) U_{\mathbb{L}}(e_{\mathbb{L}} i_0 - 1)$. Then by Lemma 2.1,

$$\theta'^{-1}(t_2 + \alpha_{\theta'}) = \theta'^{-1}(t_2 + \alpha_{\theta_1} + \alpha_{\theta_1} u) = \theta'^{-1}(t_2 + \alpha_{\theta_1}) \psi_{\mathbb{L}}\left(-\alpha_{\theta'} \frac{\alpha_{\theta_1} u}{t_2 + \alpha_{\theta_1}}\right) \tag{4.47}$$
$$= \theta'^{-1}(t_2 + \alpha_{\theta_1}) \psi\left(-\frac{2\alpha_{\theta_1}^2 t_2 u}{\mathrm{Nm}(t_2 + \alpha_{\theta_1})}\right)$$
$$= \theta'^{-1}(t_2 + \alpha_{\theta_1}) \psi\left(-\frac{2\alpha_{\theta_1}^2 u}{t_2}\right).$$

Here in the last line we have used again that $v_p(t_2^2) < v_p(\mathrm{Nm}(\alpha_{\theta_1}))$, and that $v_p\left(\frac{2\alpha_{\theta_1}^2 u}{t_2}\right) \geq -1$ by our choice of j.

Furthermore as $t_2 + \alpha_\theta \in Z U_{\mathbb{L}}\left(\frac{e_{\mathbb{L}}}{2}(2k - \mathfrak{c}(\pi_\theta))\right)$ with $\frac{e_{\mathbb{L}}}{2}(2k - \mathfrak{c}(\pi_\theta)) \geq \frac{e_{\mathbb{L}} j}{2}$. Then

(4.48)
$$\theta'^{-1}(t_2 + \alpha_{\theta_1}) = \theta_1^{-1}(t_2 + \alpha_{\theta_1})(\theta_1 \theta'^{-1})(t_2 + \alpha_{\theta_1}) = \theta_1^{-1}(t_2 + \alpha_{\theta_1}) \psi_{\mathbb{L}}\left(-\alpha_{\theta_1} u \frac{\alpha_{\theta_1}}{t_2}\right) = \theta_1^{-1}(t_2 + \alpha_{\theta_1}) \psi\left(-\frac{2\alpha_{\theta_1}^2 u}{t_2}\right).$$

Similarly one can compute that

$$\psi\left(\frac{m_1 m_2 \mu}{\mathrm{Nm}(t_2 + \alpha_{\theta'})} t_2\right) = \psi\left(\frac{m_1 m_2 \mu}{\mathrm{Nm}(t_2 + \alpha_{\theta_1})} t_2\right) \psi\left(\frac{2 m_1 m_2 \mu \alpha_{\theta_1}^2 u}{t_2^3}\right). \tag{4.49}$$

Piecing together (4.46)(4.47)(4.48)(4.49), we get that

(4.50)
$$\sum_{\theta' \in \theta_1[j]} G_p(m_1, m_2, \theta', \mu)$$
$$= \int_{t_2^2 \equiv m_1 m_2 \mu \bmod \varpi^{v(\mu)+1}} \theta_1^{-1}(t_2 + \alpha_{\theta_1}) \psi\left(\frac{m_1 m_2 \mu}{\mathrm{Nm}(t_2 + \alpha_{\theta_1})} t_2\right) \sum_{u \in \varpi^{i_0-j} O_{\mathbb{F}}/\varpi^{i_0-j+1} O_{\mathbb{F}}} \psi\left(\frac{2(m_1 m_2 \mu - 2t_2^2)\alpha_{\theta_1}^2 u}{t_2^3}\right) dt_2$$
$$= 0.$$

In the last equality we have used that $v_p(m_1m_2\mu - 2t_2^2) = -2k$ as $t_2^2 \equiv m_1m_2\mu \bmod \varpi^{v(\mu)+1}$, and $v_p\left(\frac{2(m_1m_2\mu-2t_2^2)\alpha_{\theta_1}^2}{t_2^3}\right) = -i_0 + j - 1$, thus the sum in u first gives 0.

4.6.2. *Proof of Lemma 4.17: principal series representation case.* Consider now the case where π_θ is a principal series representation. This case is easier than the supercuspidal representation case. In this case, $\theta' = \left(\chi', \chi'^{-1}\right) \in \theta[j]$ if and only if $\mathfrak{c}\left(\chi_1^{-1}\chi'\right) \leq j$. Recall that by Lemma 4.12,

$$G_p(m_1, m_2, \theta', \mu) = \int_{v(t_2)=-k} \chi'\left(\frac{t_2^2}{m_1m_2\mu}\right)\psi\left(\frac{m_1m_2\mu}{t_2} + t_2\right)dt_2$$

$$= \int_{t_2^2+2\alpha_{\chi'}t_2\equiv m_1m_2\mu \bmod \varpi^{v(\mu)+i}} \chi'\left(\frac{t_2^2}{m_1m_2\mu}\right)\psi\left(\frac{m_1m_2\mu}{t_2} + t_2\right)dt_2. \tag{4.51}$$

Recall that in this case $v_p(c_0) = i_0$ and $l_0 = 0$. $0 < i \leq \lfloor k/2 \rfloor$. Note that $v(m_1m_2\mu) = v\left(t_2^2\right)$. When $k \geq i_0 + l$, choose now $i = \min\{\lfloor k/2 \rfloor, k - i_0\}$. Then the points in the integral domain in (4.51) satisfy

$$t_2^2 + 2\alpha_{\chi'}t_2 - m_1m_2\mu \equiv t_2^2 - m_1m_2\mu \equiv 0 \bmod \varpi^{v(\mu)+i},$$

as $v_p\left(\alpha_{\chi_1}t_2\right) = -i_0 - k$. Equivalently we have $\frac{t_2^2}{m_1m_2\mu} \equiv 1 \bmod \varpi^i$.

For such t_2, it is clear that

$$\chi'\left(\frac{t_2^2}{m_1m_2\mu}\right) = \chi_1\left(\frac{t_2^2}{m_1m_2\mu}\right)\chi_1^{-1}\chi'\left(\frac{t_2^2}{m_1m_2\mu}\right) = \chi_1\left(\frac{t_2^2}{m_1m_2\mu}\right)$$

as $\mathfrak{c}\left(\chi_1^{-1}\chi'\right) \leq l \leq \min\{\lfloor k/2 \rfloor, k - i_0\}$. Here we have used that either $\lfloor k/2 \rfloor \geq k - i_0 \geq l$, or $\lfloor k/2 \rfloor < k - i_0$, in which case we have $k \geq 2i_0 + 1$, and thus $l < i_0 \leq \lfloor k/2 \rfloor$. Thus when $k \geq l + i_0$,

$$\frac{1}{[\theta[l] : \theta[l_0]]}\sum_{\theta' \in \theta[l]/\sim_{l_0}} G_p(m_1, m_2, \theta', \mu) = \int_{t_2^2 \equiv m_1m_2\mu \bmod \varpi^{v(\mu)+i}} \chi_1\left(\frac{t_2^2}{m_1m_2\mu}\right)\psi\left(\frac{m_1m_2\mu}{t_2} + t_2\right)dt_2 \tag{4.52}$$

$$= G_p(m_1, m_2, \theta, \mu).$$

On the other hand when $i_0 \leq k < i_0 + l < 2i_0$, we have $\lfloor k/2 \rfloor > k - i_0$. Choose now $i = k - i_0 + 1$. The domain of the integral in (4.51) becomes

$$t_2^2 - m_1m_2\mu \equiv 2\alpha_{\chi'}t_2 \equiv 2\alpha_{\chi_1}t_2 \not\equiv 0 \bmod \varpi^{v(\mu)+i}.$$

Here we have used that when $\theta' \in \theta[l]$, $\alpha_{\chi'} \in \alpha_{\chi_1}U_{\mathbb{F}}(i_0 - l)$. As $i = k - i_0 + 1 \leq l$, we have

$$\frac{t_2^2}{m_1m_2\mu} \not\equiv 1 \bmod \varpi^l.$$

Then we have

$$\frac{1}{[\theta[l] : \theta[l_0]]}\sum_{\theta' \in \theta[l]/\sim_{l_0}} G_p(m_1, m_2, \theta', \mu) \tag{4.53}$$

$$= \frac{1}{[\theta[l] : \theta[l_0]]}\int_{t_2^2 - m_1m_2\mu \equiv 2\alpha_{\chi_1}t_2 \bmod \varpi^{v(\mu)+i}} \sum_{\mathfrak{c}\left(\chi_1^{-1}\chi'\right) \leq l} \chi'\left(\frac{t_2^2}{m_1m_2\mu}\right)\psi\left(\frac{m_1m_2\mu}{t_2} + t_2\right)dt_2 = 0.$$

4.7. **the Refined Kuznetsov trace formula.** The discussions so far also allow us to derive the refined Kuznetsov trace formula in Theorem 1.7 without additional difficulty. Note that the only difference for this case and the Petersson trace formula case is the Archimedean computations, which has already been done in, for example, [13].

We shall skip the details here, leaving them to interested readers.

5. Alternative description and compatibility with Voronoï formula

Again this section is purely local.

5.1. **The relation between the test function and local matrix coefficient.** The construction of the test function f_p is closely related to the restriction of the matrix coefficient of the newform to proper subgroups. We make the relation more clear here for later discussions.

Definition 5.1. Let K' be the maximal compact open subgroup whose elements lie in

$$\begin{pmatrix} O_{\mathbb{F}} & \varpi^{-i_0}O_{\mathbb{F}} \\ \varpi^{i_0}O_{\mathbb{F}} & O_{\mathbb{F}} \end{pmatrix}.$$

Lemma 5.2. *For $\pi = \pi_\theta$, suppose that $\mathfrak{c}(\pi) \geq 3$. Let $\varphi_{new} \in \pi$ be a L^2-normalized newform, and $\Phi_{\varphi_{new}}$ be the associated matrix coefficient. Let $v(\mu) = -2k < -2i_0$. For $v(t_1) = v(t_2) = -k$, we have for test function f_p as specified in Section 4.1 and some positive constant $a_\pi \asymp_p p^{\mathfrak{c}(\pi)/2} \asymp_p C_{\mathcal{F}}[l_0]$,*

$$f_p\left(\begin{pmatrix} -t_1 & -\mu - t_1t_2 \\ 1 & t_2 \end{pmatrix}\right) = a_\pi \overline{\Phi}_{\varphi_{new}}|_{ZK'}\left(\begin{pmatrix} -t_1 & -\mu - t_1t_2 \\ 1 & t_2 \end{pmatrix}\right).$$

Proof. Denote $g = \begin{pmatrix} -t_1 & -\mu - t_1t_2 \\ 1 & t_2 \end{pmatrix}$. Consider the supercuspidal representation case first. By Corollary 3.16,

$$\Phi_{\varphi_{\text{new}}} = \frac{1}{(p-1)\, p^{\lceil i_0/2\rceil - 1}} \sum_{a,a' \in \left(O_{\mathbb{F}}/\varpi^{\lceil i_0/2\rceil}O_{\mathbb{F}}\right)^\times} \Phi_{a,a'}$$

Comparing with Definition 3.19, we get that

$$a_\pi = \frac{1}{\mathrm{Vol}(Z\backslash ZB^1)} \asymp_p p^{\mathfrak{c}(\pi)/2}$$

by (4.8), and it suffices to check by Definition 3.17 that,

$$\Phi_{0,0}|_{ZB^1}(g_{a,a'}) = \Phi_{0,0}|_{ZK}(g_{a,a'}).$$

Here $g_{a,a'} = \begin{pmatrix} \varpi^{i_0}a' & 0 \\ 0 & 1 \end{pmatrix} g \begin{pmatrix} \varpi^{-i_0}a^{-1} & 0 \\ 0 & 1 \end{pmatrix}$, and we have used that

$$\begin{pmatrix} \varpi^{i_0}a' & 0 \\ 0 & 1 \end{pmatrix} K' \begin{pmatrix} \varpi^{-i_0}a^{-1} & 0 \\ 0 & 1 \end{pmatrix} = K.$$

Note that $ZB^1 \subset ZK$. Thus it suffices to show that $g_{a,a'} \in \mathrm{Supp}\, \Phi_{0,0} \cap ZK$ implies $g \in ZB^1$. Indeed in that case, we have $v_p(\det(g_{a,a'})) = v_p(\mu) = -2k$, so

$$\varpi^k g_{a,a'} = \begin{pmatrix} -\varpi^k a^{-1}a't_1 & -(\mu + t_1t_2)\varpi^{i_0+k}a' \\ \varpi^{k-i_0}a^{-1} & \varpi^k t_2 \end{pmatrix} \in K.$$

Note that the lower left element satisfies $v_p(\varpi^{k-i_0}a^{-1}) \geq 1$. Recall that $\mathrm{Supp}\, \Phi_{0,0} \subset J = \mathbb{L}^\times K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lfloor \mathfrak{c}(\theta)/2 \rfloor)$, and when $\mathfrak{c}(\pi) \geq 3$, the lower left entry of any element in $K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lfloor \mathfrak{c}(\theta)/2 \rfloor)$ also satisfies $v_p \geq 1$. Then for $\varpi^k g_{a,a'} \in \mathrm{Supp}\,\Phi_{0,0} \cap K$ implies $\varpi^k g_{a,a'} \in ZJ^1 = ZU_{\mathbb{L}}(1)K_{\mathfrak{A}_{e_{\mathbb{L}}}}(\lfloor \mathfrak{c}(\theta)/2 \rfloor)$. The claim now follows from Corollary 3.12.

The principal series representation case is mostly parallel. In this case by Lemma 3.24, we have

$$\Phi_{\varphi_{\mathrm{new}}} = \frac{1}{|C_0|^2} \sum_{a,a' \in \left(O_{\mathbb{F}}/\varpi^{i_0} O_{\mathbb{F}}\right)^\times} \tilde{\Phi}_{a,a'}(g)$$

Comparing with (3.27), we get that

$$a_\pi = |C_0|^2 \frac{1}{(p-1)\, p^{i_0-1} \mathrm{Vol}\,(Z\backslash ZK_0\,(\varpi^{i_0}))} \asymp_p p^{\mathfrak{c}(\pi)/2},$$

and the lemma is reduced to check that $\mathrm{Supp}\,\Phi_{0,0} \cap ZK' = ZK_0(\varpi^{i_0})$. This follows immediately from Lemma 3.28. □

Remark 5.3. a_π only depends on $\mathbb{L}$ and $\mathfrak{c}(\pi)$, and actually $a_\pi = (1-p^{-1})C_{\mathcal{F}}[l_0]$ for our choice of f_p using a case by case check. But we do not need this property here. The condition $v(\mu) = -2k < -2i_0$ can be easily achieved by using Petersson trace formula for slightly larger family according to Theorem 4.18.

For later applications, we also prove the following lemma

Lemma 5.4. *Let μ and π be as in Lemma 5.2, and $v(t_1) = -k,\ v(t_2) > -k$. Then both f_p and $\Phi_{\varphi_{new}}$ are vanishing.*

Proof. From the computations in Section 4.4, it is straightforward to check that f_p is vanishing. Consider for example the case where π_θ is a principal series representation. By Lemma 4.8, $g = \begin{pmatrix} -t_1 & -\mu - t_1 t_2 \\ 1 & t_2 \end{pmatrix} \in \mathrm{Supp}\,\tilde{\Phi}_{0,0} = ZK_0(\varpi^{i_0})$ only if $v(t_1) = v(t_2) = -k$. For general $\tilde{\Phi}_{a,a'}$, we do translations by $\begin{pmatrix} 1 & \pm\varpi^{-i_0} a \\ & 1 \end{pmatrix}$ on the left or right, which however does not change condition for the valuation of the upper left or lower right entries as $k > i_0$.

On the other hand, let $-j = v(t_2) > -k$, and we apply the extended Iwasawa decomposition in the sense of [7, Lemma 2.1],

$$g = \begin{cases} \varpi^{-j} \begin{pmatrix} \mu\varpi^{2j} & \varpi^j(-\mu - t_1 t_2) \\ & \varpi^j t_2 \end{pmatrix} \begin{pmatrix} 1 & \\ \varpi^j & 1 \end{pmatrix} \begin{pmatrix} t_2^{-1}\varpi^{-j} & \\ & 1 \end{pmatrix}, & \text{if } j \geq 0, \\ \begin{pmatrix} \mu & -t_1 \\ & 1 \end{pmatrix} \begin{pmatrix} 1 & \\ 1 & 1 \end{pmatrix} \begin{pmatrix} & -1 \\ 1 & 1 + t_2 \end{pmatrix}, & \text{otherwise.} \end{cases}$$

One can now check case by case that g is not in the support using [7, Proposition 2.19]. For example when $j \geq 0$, we have $v(a) = 2j - 2k$ for $a = \mu\varpi^{2j}$, while [7, Proposition 2.19] requires $v(a) \geq \min\{0, 2j - \mathfrak{c}(\pi)\} > 2j - 2k$. □

Remark 5.5. With a little extra work, it is possible to show that $\Phi_{\varphi_{\mathrm{new}}}$ is vanishing on the given g when $k = i_0$.

5.2. **Alternative approach to the second cell terms.**

Corollary 5.6. *Let the test function f be as in Section 4.1. Suppose that $v(\mu) = -2k < -2i_0$. Then the the second-cell terms can be alternatively written as*

$$(5.1)\qquad I_p(\gamma, f, m_1, m_2) = \frac{a_\pi}{1-p^{-1}} \int\limits_{v(t_1)=-k} W_{\varphi_{new}}\left(\begin{pmatrix} m_2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}\right)\psi(-m_1t_1)dt_1$$
$$= \frac{a_\pi p^k}{1-p^{-1}} \int\limits_{v(u)=0} W_{\varphi_{new}}\left(\begin{pmatrix} m_2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{u}{p^k} \\ 0 & 1 \end{pmatrix}\right)\psi\left(-\frac{m_1u}{p^k}\right)du.$$

Proof. By Lemma 5.2, Lemma 5.4 and (2.3), we can rewrite

(5.2)

$$I_p(\gamma, f, m_1, m_2) = \int\limits_{\mathbb{F}^2} f_p\left(\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\right)\psi(-m_1t_1 + m_2t_2)dt_1dt_2$$
$$= a_\pi \int\limits_{v(t_1)=-k, v(t_2)\geq -k} \overline{\Phi}_{\varphi_{\text{new}}}\left(\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\right)\psi(-m_1t_1 + m_2t_2)dt_1dt_2$$
$$= a_\pi \int\limits_{t_1,t_2} \overline{w_\pi(-\mu)} \overline{< \pi\left(\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\right)\varphi_{\text{new}}, \pi\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}\right)\varphi_{\text{new}} >}\psi(-m_1t_1 + m_2t_2)dt_1dt_2$$
$$= a_\pi \int\limits_{t_1,t_2}\int\limits_{x\in\mathbb{F}^\times} \overline{W_{\varphi_{\text{new}}}}\left(\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\right) W_{\varphi_{\text{new}}}\left(\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}\right)d^\times x\psi(-m_1t_1 + m_2t_2)dt_1dt_2$$

Here we have used our assumption that w_π is trivial. Now we swap the order and integrate in $v(t_2) \geq -k$ first. Using that

$$W_\varphi\left(\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & t_2 \\ 0 & 1 \end{pmatrix}\right) = \psi(xt_2)W_\varphi\left(\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}\right),$$

we get that the integral in t_2 is nonvanishing iff $x \equiv m_2 \bmod \varpi^k$. As $W_{\varphi_{\text{new}}}(a(x)) = \text{char}\left(O_{\mathbb{F}}^\times\right)(x)$, we get

$$(5.3)\qquad I_p(\gamma, f, m_1, m_2) = a_\pi p^k \int\limits_{v(t_1)=-k}\int\limits_{x\equiv m_2 \bmod \varpi^k} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}\right)\psi(-m_1t_1)d^\times xdt_1$$

We show now that the integrand is a constant function in x when $x \equiv m_2 \bmod \varpi^k$. Note that in the extended Iwasawa decomposition

$$\begin{pmatrix} x & \\ & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix} = t_1\begin{pmatrix} \frac{x\mu\varpi^k}{t_1} & -\frac{x\mu}{t_1} \\ & 1 \end{pmatrix}\begin{pmatrix} 1 & \\ \varpi^k & 1 \end{pmatrix}\begin{pmatrix} t_1^{-1}\varpi^{-k} & \\ & 1 \end{pmatrix},$$

Thus

$$W_{\varphi_{\text{new}}}\left(\begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & t_1 \\ 0 & 1 \end{pmatrix}\right) = \psi\left(-\frac{x\mu}{t_1}\right)W_{\varphi_{\text{new}}}\left(\begin{pmatrix} \frac{x\mu\varpi^k}{t_1} & \\ & 1 \end{pmatrix}\begin{pmatrix} 1 & \\ \varpi^k & 1 \end{pmatrix}\right)$$

which is of level $\leq k$ in x by [8, Proposition 2.12]. Thus the integrand is constant for $x \equiv m_2 \bmod \varpi^k$. The corollary is now clear. □

5.3. **Compatibility with the Voronoi formula.** The alternative description Corollary 5.6 for the second-cell terms for the Petersson/Kuznetsov trace formula allows us to analyze the character sum after applying the Voronoï formula more easily and to reduce the problem to the existing works.

Definition 5.7. For some integer a with $(a, p) = 1$, define

$$\widetilde{G}_p(m_1, m_2, a, \theta, \mu) = \frac{p^k}{1-p^{-1}} \int\limits_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} m_2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{u}{p^k} \\ 0 & 1 \end{pmatrix}\right)\psi\left(-\frac{am_1}{up^k}\right)du$$

The reason we make this definition will be clear in Section 6.

Lemma 5.8. *$\widetilde{G}_p(m_1, m_2, a, \theta, \mu) = 0$ unless $v\left(m_2\mu + \frac{am_1}{p^{2k}}\right) \geq -\mathfrak{c}(\pi)$, in which case we have*

$$\widetilde{G}_p(m_1, m_2, a, \theta, \mu) \ll_p p^{\frac{3k-\mathfrak{c}(\pi)}{2}}.$$

Proof. Our strategy is to reinterpret the integral as the value of the matrix coefficient. By a change of variable and the invariance of the newform, we get

$$\begin{aligned}
(5.4)\quad \widetilde{G}_p(m_1, m_2, a, \theta, \mu) &= p^k \int\limits_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} m_2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{up^k} \\ 0 & 1 \end{pmatrix}\right)\psi\left(-\frac{am_1 u}{p^k}\right)d^\times u \\
&= p^k \int\limits_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} 0 & -m_2\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{u} & \\ & 1 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{p^k} \\ 0 & 1 \end{pmatrix}\right)\psi\left(-\frac{am_1 u}{p^k}\right)d^\times u \\
&= p^k \int\limits_{v(u)=0} w_\pi^{-1}(u) W_{\varphi_{\text{new}}}\left(\begin{pmatrix} u & \\ & 1 \end{pmatrix}\begin{pmatrix} 0 & -m_2\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{p^k} \\ 0 & 1 \end{pmatrix}\right)\psi\left(-\frac{am_1 u}{p^k}\right)d^\times u \\
&= p^k \int\limits_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} u & \\ & 1 \end{pmatrix}\begin{pmatrix} 0 & -m_2\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{p^k} \\ 0 & 1 \end{pmatrix}\right)\overline{W_{\varphi_{\text{new}}}\left(\begin{pmatrix} u & \\ & 1 \end{pmatrix}\begin{pmatrix} 1 & \frac{am_1}{p^k} \\ & 1 \end{pmatrix}\right)}d^\times u \\
&= p^k \Phi_{\varphi_{\text{new}}}\left(\begin{pmatrix} 1 & -\frac{am_1}{p^k} \\ & 1 \end{pmatrix}\begin{pmatrix} 0 & -m_2\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{p^k} \\ 0 & 1 \end{pmatrix}\right)
\end{aligned}$$

Note that

$$\begin{pmatrix} 1 & -\frac{am_1}{p^k} \\ & 1 \end{pmatrix}\begin{pmatrix} 0 & -m_2\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{1}{p^k} \\ 0 & 1 \end{pmatrix} = p^{-k}\begin{pmatrix} -am_1 & -\left(m_2\mu + \frac{am_1}{p^{2k}}\right)p^k \\ p^k & 1 \end{pmatrix}$$

By [11, Theorem 5.4], this matrix is not in the support of the matrix coefficient of the newform unless $v\left(\left(m_2\mu + \frac{am_1}{p^{2k}}\right)p^k\right) \geq k - \mathfrak{c}(\pi)$, in which case $|\Phi_{\varphi_{\text{new}}}| \ll_p p^{\frac{k-\mathfrak{c}(\pi)}{2}}$. The lemma follows easily now. ∎

6. Application to the first moment of the Rankin–Selberg L-function

6.1. **Preparations.** We take a special version of the Voronoï formula from [6, Lemma 7] or [15, Theorem A.4], though a more flexible version would be helpful to extend our main result to more general situations.

Theorem 6.1. *Let $(a, c) = 1$ and h be a smooth compactly supported function in $(0, \infty)$. Let g be a holomorphic modular form of weight κ_g, square-free level M and nebentypus χ. Let $M = M_1M_2$*

with $M_1 = (M, c)$. Then there exists a newform g^* of the same level M and weight κ_g such that

$$\sum_n \lambda_g(n) e\left(\frac{an}{c}\right) h(n) = \frac{2\pi\eta}{c\sqrt{M_2}} \sum_n \lambda_{g^*}(n) e\left(-\frac{\overline{aM_2}n}{c}\right) \int_0^\infty h(\xi) J_{\kappa_g - 1}\left(\frac{4\pi\sqrt{n\xi}}{c\sqrt{M_2}}\right) d\xi. \tag{6.1}$$

Here $\overline{x}$ denotes the multiplicative inverse of $x \bmod c$, and η is a complex number of modulus 1 depending on a, c, g.

The following lemma is straightforward to check using the Chinese remainder theorem:

Lemma 6.2. *Suppose $(n_1, n_2) = 1$, $a_i\overline{a_i} \equiv 1 \bmod n_i$, $i = 1, 2$, $\overline{n_1}n_1 \equiv 1 \bmod n_2$, $\overline{n_2}n_2 \equiv 1 \bmod n_1$. Then*

$$(a_1n_2 + a_2n_1)\left(\overline{a_1}n_2\overline{n_2}^2 + \overline{a_2}n_1\overline{n_1}^2\right) \equiv 1 \bmod n_1n_2.$$

6.2. **the first moment of the Rankin–Selberg L-function and hybrid subconvexity bound.** Recall that $\mathcal{F}_\theta[l]$ is the set of holomorphic newforms of weight κ, level $N = p^{\mathfrak{c}}$ with $\mathfrak{c} \geq 3$, and trivial nebentypus, whose associated local representation $\pi_p \in \pi_\theta[l]$. Let g be a fixed self-dual holomorphic cusp form of weight κ_g, level M and nebentypus χ. We assume M to be square-free and coprime to N. χ is quadratic by that g is self dual.

The implied constant for the bounds $\ll$ are always allowed to depend on ϵ, which we omit from notations. Denote the harmonic average as in [15]

$$\sum_f{}^h \alpha_f := \frac{\Gamma(\kappa - 1)}{(4\pi)^{\kappa - 1}} \sum_f \frac{\alpha_f}{||f||^2}. \tag{6.2}$$

Let M_g be the first moment of the Rankin–Selberg L-functions

$$M_g = \sum_{f \in \mathcal{F}_\theta[l]}{}^h L(f \times g, 1/2) \tag{6.3}$$

Here f is normalized so that $\lambda_f(1) = 1$. We also assume from now on that $\epsilon(f \times g, 1/2) = 1$, since if it is -1, $L(f \times g, 1/2) = 0$. Note that $\epsilon(f \times g, 1/2)$ is the same for any $f \in \mathcal{F}_\theta[l]$. By the approximate functional equation, we get

$$M_g = \sum_{n \geq 1} \frac{2\lambda_g(n)}{\sqrt{n}} V\left(\frac{n}{NM}\right) \sum_{f \in \mathcal{F}_\theta[l]}{}^h \lambda_f(n). \tag{6.4}$$

Multiplying with $\overline{\lambda_f(1)} = 1$ and applying the refined Petersson trace formula in Theorem 4.18, we get that

$$M_g = M_g^d + M_g^{od},$$

where

$$M_g^d = 2C_{\mathcal{F}}[l] V\left(\frac{1}{NM}\right), \tag{6.5}$$

$$M_g^{od} = 4\pi i^\kappa C_{\mathcal{F}}[l] \sum_{c_l | c} \frac{1}{c} \sum_n \frac{\lambda_g(n)}{\sqrt{n}} V\left(\frac{n}{NM}\right) G\left(n, 1, \theta, \frac{1}{c^2}\right) J_{\kappa - 1}\left(\frac{4\pi\sqrt{n}}{c}\right). \tag{6.6}$$

To analyze the off-diagonal term M_g^{od}, we break the sum into dyadic ranges as usual by multiplying with a bump function η_Z, where the size of the sum in n is $Z \ll (NM)^{1+\epsilon}$. Up to a small error, we may assume that $c \ll (MN)^A$ for some fixed large A. Furthermore, we write $c = d_p p^k$ for $k \geq v_p(c_l)$

and $\left(d_p, p\right) = 1$, and organize the sum in c according to d_p and k. We shall however be mainly interested in the case where $k < \mathfrak{c}(\pi)$, as the complementary case is much easier to deal with by Remark 4.7, 4.13. By Definition 4.14, (4.39) and Corollary 5.6,

$$G\left(n, 1, \theta, \frac{1}{c^2}\right) = \frac{1}{C_{\mathcal{F}}[l_0]} \sum_{y \in \left(\mathbb{Z}/d_p\mathbb{Z}\right)^\times} e\left(\frac{\overline{p}^{2k} y}{d_p} + \frac{n\overline{y}}{d_p}\right) I_p(\gamma, f, n, 1)$$

$$= \frac{a_\pi p^k}{C_{\mathcal{F}}[l_0](1-p^{-1})} \sum_{y \in \left(\mathbb{Z}/d_p\mathbb{Z}\right)^\times} e\left(\frac{\overline{p}^{2k} y}{d_p} + \frac{n\overline{y}}{d_p}\right) \int_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{u}{p^k} \\ 0 & 1 \end{pmatrix}\right) e\left(-\frac{nu}{p^k}\right) du \tag{6.7}$$

Because of this, we write

$$M_g^{od} = 4\pi i^\kappa a_\pi \frac{C_{\mathcal{F}}[l]}{C_{\mathcal{F}}[l_0]} \sum_{c_l | c} \sum_{Z \ll (MN)^{1+\epsilon}} \frac{1}{c} K_{c,Z} \tag{6.8}$$

where

$$\begin{aligned} K_{c,Z} =& \frac{p^k}{1-p^{-1}} \sum_n \frac{\lambda_g(n)\eta_Z(n)}{\sqrt{n}} V\left(\frac{n}{NM}\right) J_{\kappa-1}\left(\frac{4\pi\sqrt{n}}{c}\right) \\ & \times \sum_{y \in \left(\mathbb{Z}/d_p\mathbb{Z}\right)^\times} e\left(\frac{\overline{p}^{2k} y}{d_p} + \frac{n\overline{y}}{d_p}\right) \int_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{u}{p^k} \\ 0 & 1 \end{pmatrix}\right) e\left(-\frac{nu}{p^k}\right) du \\ =& \frac{p^k}{1-p^{-1}} \sum_{y \in \left(\mathbb{Z}/d_p\mathbb{Z}\right)^\times} e\left(\frac{\overline{p}^{2k} y}{d_p}\right) \int_{v(u)=0} W_{\varphi_{\text{new}}}\left(\begin{pmatrix} 0 & -\mu \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & \frac{u}{p^k} \\ 0 & 1 \end{pmatrix}\right) \\ & \times \left[\sum_n \frac{\lambda_g(n)\eta_Z(n)}{\sqrt{n}} V\left(\frac{n}{NM}\right) J_{\kappa-1}\left(\frac{4\pi\sqrt{n}}{c}\right) e\left(\frac{n\overline{y}}{d_p}\right) e\left(-\frac{nu}{p^k}\right)\right] du \end{aligned} \tag{6.9}$$

Here in the second equality we have swapped the order of the sum in n and the sum/integral in y/u, as the integral in u is essentially a finite sum.

Lemma 6.3. *For $L = \frac{\sqrt{Z}}{c}$, we have*

$$K_{c,Z} \ll (cMZ)^\epsilon \begin{cases} \left(1 + \frac{M_2}{d_p p^{2k-\mathfrak{c}(\pi)}}\right) p^{\frac{k-\mathfrak{c}(\pi)}{2}} \sqrt{\frac{Z}{M_2}} \frac{1}{L} & \text{when } L \gg 1; \\ \left(1 + \frac{M_2}{d_p p^{2k-\mathfrak{c}(\pi)}} \frac{1}{L^2}\right) p^{\frac{k-\mathfrak{c}(\pi)}{2}} \sqrt{\frac{Z}{M_2}} L & \text{when } L \ll 1. \end{cases}$$

Proof. Denote

$$K_{c,Z}(y,u) = \sum_n \frac{\lambda_g(n)\eta_Z(n)}{\sqrt{n}} V\left(\frac{n}{NM}\right) J_{\kappa-1}\left(\frac{4\pi\sqrt{n}}{c}\right) e\left(\frac{n\overline{y}}{d_p}\right) e\left(-\frac{nu}{p^k}\right),$$

for which we wish to apply the Voronoï summation formula in Theorem 6.1. In particular Lemma 6.2 implies that

$$e\left(\frac{n\overline{y}}{d_p}\right) e\left(-\frac{nu}{p^k}\right) = e\left(\frac{\overline{y}p^k - ud_p}{c} n\right)$$

on the left-hand side of (6.1) with $a = \overline{y}p^k - ud_p$ becomes

$$e\left(-\frac{\overline{aM_2n}}{c}\right) = e\left(-\frac{\overline{M_2}y\overline{p}^{2k}}{d_p}n\right)e\left(\frac{\overline{M_2ud_p^2}}{p^k}n\right),$$

on the right-hand side of (6.1). Thus

$$K_{c,Z}(y,u) = \frac{2\pi\eta}{c\sqrt{M_2}}\sum_n \lambda_{g^*}(n)e\left(-\frac{\overline{M_2}y\overline{p}^{2k}}{d_p}n\right)e\left(\frac{\overline{M_2ud_p}^2}{p^k}n\right)I(n) \tag{6.10}$$

where

$$I(n) = \int_0^\infty \frac{V\left(\frac{x}{NM}\right)\eta_Z(x)}{\sqrt{x}}J_{\kappa-1}\left(\frac{4\pi\sqrt{x}}{c}\right)J_{\kappa-1}\left(\frac{4\pi\sqrt{nx}}{c\sqrt{M_2}}\right)dx. \tag{6.11}$$

Then

$$K_{c,Z} = \frac{2\pi\eta}{c\sqrt{M_2}}\sum_n \lambda_{g^*}(n)\widetilde{\mathrm{KL}}(1-\overline{M_2}n, d_p)\widetilde{G}_p\left(n, 1, -\overline{M_2d_p}^2, \theta, \frac{1}{c^2}\right)I(n). \tag{6.12}$$

Here

$$\widetilde{\mathrm{KL}}(1-\overline{M_2}n, d_p) = \sum_{y\in(\mathbb{Z}/d_p\mathbb{Z})^\times} e\left(\frac{\overline{p}^{2k}y(1-\overline{M_2}n)}{d_p}\right) = \sum_{y\in(\mathbb{Z}/d_p\mathbb{Z})^\times} e\left(\frac{y(1-\overline{M_2}n)}{d_p}\right)$$

is the Ramanujan sum. If

$$(1-\overline{M_2}n, d_p) = d_{p,n}, \tag{6.13}$$

then

$$|\widetilde{\mathrm{KL}}(1-\overline{M_2}n, d_p)| \ll d_{p,n}c^\epsilon. \tag{6.14}$$

$\widetilde{G}_p\left(n, 1, -\overline{M_2d_p^2}, \theta, \frac{1}{c^2}\right)$ is as in Definition 5.7, which by Lemma 5.8 is nonzero only when

$$v_p\left(\frac{1}{c^2} - \frac{\overline{M_2d_p^2}n}{p^{2k}}\right) = v_p(1-\overline{M_2}n) - 2k \geq -\mathfrak{c}(\pi), \tag{6.15}$$

in which case

$$|\widetilde{G}_p| \ll_p p^{\frac{3k-\mathfrak{c}(\pi)}{2}}. \tag{6.16}$$

On the other hand, let $L = \frac{\sqrt{Z}}{c}$, $Q = \frac{\sqrt{nZ}}{c\sqrt{M_2}}$. The function $I(n)$ restricts the sum to essentially (up to $(cZM)^\epsilon$)

$$|L-Q| \ll 1, \text{ or equivalently } |1-\sqrt{\frac{n}{M_2}}| \ll L^{-1}. \tag{6.17}$$

In this range we have

$$I(n) \ll \sqrt{Z}\frac{L}{(1+L)^{3/2}}\frac{Q}{(1+Q)^{3/2}} \ll \begin{cases} \frac{\sqrt{Z}}{L}, & \text{if } L \gg 1; \\ \sqrt{Z}L, & \text{if } L \ll 1 \end{cases}. \tag{6.18}$$

by [6, Lem2.1].

Now the number of n satisfying the bound in (6.17) with the congruence conditions (6.13), (6.15) can be controlled by

$$\ll \left(1 + \frac{M_2}{d_{p,n}p^{2k-\mathfrak{c}(\pi)}} \frac{(1+L)^2}{L^2}\right)(cMZ)^{\epsilon}.$$

For each of these terms in (6.12) we apply the bound $\lambda_{g^*}(n) \ll n^{\epsilon}$ and (6.14), (6.16) (6.18). The lemma is then clear. □

Lemma 6.4. *For M_g^{od} as in* (6.6)*, we have*

$$M_g^{od} \ll_{p,\epsilon} (MN)^{\epsilon}\left(N^{1/4}p^{l/2} + N^{1/4}M^{1/2}p^{-l/2}\right).$$

Proof. We shall focus on the case when $v_p(c_l) \leq k \leq \mathfrak{c}(\pi)$, as the case when $k \geq \mathfrak{c}(\pi)$ will be easier (and one can use the argument in [6] with slight modifications). For conciseness we drop all ϵ-terms in our computations.

By (6.8), (4.43), $a_\pi \asymp_p p^{\mathfrak{c}(\pi)/2} = N^{1/2}$ from Lemma 5.2, and Lemma 6.3, we get

(6.19)
$$M_g^{od} \ll N^{1/2}p^l \sum_{Z\ll(MN)^{1+\epsilon}} \sum_{M_2|M} \sum_{v_p(c_l)\leq k\leq \mathfrak{c}(\pi)} [\sum_{C\ll\sqrt{Z}} \sum_{c=d_pp^k\asymp C,(M/M_2)|d_p} \frac{1}{c}\left(1+\frac{M_2}{d_pp^{2k-\mathfrak{c}(\pi)}}\right)p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{c}{\sqrt{Z}}$$
$$+ \sum_{C\gg\sqrt{Z}} \sum_{c} \frac{1}{c}\left(1+\frac{M_2}{d_pp^{2k-\mathfrak{c}(\pi)}}\frac{c^2}{Z}\right)p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{\sqrt{Z}}{c}]$$

Here the sum over $c \asymp C$ is over dyadic intervals. It seems easier to discuss the contribution of each term inside the square bracket separately. In particular we have

$$\sum_{Z,M_2,k}\sum_{C\ll\sqrt{Z}}\sum_{c}\frac{1}{c}p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{c}{\sqrt{Z}} \ll \sum_{Z,M_2,k}\sum_{C\ll\sqrt{Z}}\frac{CM_2}{p^kM}\frac{1}{c}p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{c}{\sqrt{Z}}$$
$$\ll \sum_{Z,M_2,k}\sum_{C\ll\sqrt{Z}}\frac{C\sqrt{M_2}}{M}\frac{1}{p^{k/2+\mathfrak{c}(\pi)/2}} \ll \sum_k \frac{1}{p^{k/2}} \ll \frac{1}{N^{1/4}p^{l/2}}. \tag{6.20}$$

Here the range of the summation for Z, M_2, k, c are the same as in (6.19). In the last estimate we have used that $k \geq v_p(c_l)$ which is $\mathfrak{c}(\pi)/2 + l$ up to a bounded constant. Similarly we have

$$\sum_{Z,M_2,k}\sum_{C\ll\sqrt{Z}}\sum_{c}\frac{1}{c}\frac{M_2}{d_pp^{2k-\mathfrak{c}(\pi)}}p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{c}{\sqrt{Z}} \ll \frac{M^{1/2}}{N^{1/4}p^{3l/2}}, \tag{6.21}$$

$$\sum_{Z,M_2,k}\sum_{C\gg\sqrt{Z}}\sum_{c}\frac{1}{c}p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{\sqrt{Z}}{c} \ll \frac{1}{N^{1/4}p^{l/2}}, \tag{6.22}$$

$$\sum_{Z,M_2,k}\sum_{C\gg\sqrt{Z}}\sum_{c}\frac{1}{c}\frac{M_2}{d_pp^{2k-\mathfrak{c}(\pi)}}\frac{c^2}{Z}p^{\frac{k-\mathfrak{c}(\pi)}{2}}\sqrt{\frac{Z}{M_2}}\frac{\sqrt{Z}}{c} \ll \frac{M^{1/2}}{N^{1/4}p^{3l/2}}. \tag{6.23}$$

The lemma follows easily now. □

We can now prove Theorem 1.9. From (6.3), Lemma 6.4 and (6.5),

$$M_g = M_g^{od} + M_g^{d} \ll (MN)^{\epsilon}\,[N^{1/2}p^{l} + N^{1/4}M^{1/2}p^{-l/2}] \tag{6.24}$$

Here we have used that $C_{\mathcal{F}}[l] \asymp N^{1/2}p^l$. Recall that $l_0 \leq l < i_0$. We make different choices according to the relation between N and M as follows:

(1) When $N \leq \sqrt{M}$, we choose $l = i_0 - 1$, so $p^l \asymp N^{1/2}$ and

$$M_g \ll (MN)^{\epsilon}\,\sqrt{M}.$$

(2) When $\sqrt{M} \leq N \leq M^2$, we choose $1 \leq l < i_0$ such that $p^l \asymp \left(\frac{M}{\sqrt{N}}\right)^{1/3}$, and

$$M_g \ll (MN)^{1/3+\epsilon}.$$

(3) When $N > M^2$, we choose $l = 1$ and

$$M_g \ll (MN)^{\epsilon}\,N^{1/2}.$$

Theorem 1.9 now follows easily.

Remark 6.5. If we work with the Maass forms instead of the holomorphic modular forms, the Ramanujan conjecture does seem important for the bound in Lemma 6.3. It is unlikely that a Ramanujan-conjecture-on-average type of result would suffice. After all, the sum in n in (6.12) is over a thin arithmetic progression, especially when N is large compared to M.

On the other hand, a reasonable bound towards the Ramanujan conjecture can still give a slightly weaker hybrid subconvexity bound.

References

[1] Valentin Blomer. The relative trace formula in analytic number theory. https://arxiv.org/abs/1912.08137.

[2] Valentin Blomer, Jack Buttcane, and Nicole Raulf. A Sato–Tate law for GL(3). *Commentarii Mathematici Helvetici*, 89:895–919, 2014.

[3] C. Bushnell and G. Henniart. *The Local Langlands Conjecture for* GL(2). Springer-Verlag, Berlin, 2006.

[4] William Casselman. On some results of Atkin and Lehner. *Mathematische Annalen*, 1973.

[5] Brooke Feigon and David Whitehouse. Averages of central L-values of Hilbert modular forms with an application to subconvexity. *Duke Math. J.*, 149(2):347–410, 2009.

[6] Roman Holowinsky and Nicolas Templier. First moment of Rankin–Selberg central L-values and subconvexity in the level aspect. *The Ramanujan Journal*, 33(1):131–155, January 2014.

[7] Yueke Hu. Triple product formula and the subconvexity bound of triple product L-function in level aspect. *American Journal of Mathematics*, 139(1):215–259, 2017.

[8] Yueke Hu. Triple product formula and mass equidistribution on modular curves of level N. *IMRN*, 2018(9):2899–2943, 2018.

[9] Yueke Hu and Paul D. Nelson. New test vector for waldspurger's period integral, relative trace formula, and hybrid subconvexity bounds. *In preparation. partially available at arXiv:1810.11564.*

[10] Yueke Hu, Paul D. Nelson, and Abhishek Saha. Some analytic aspects of automorphic forms on GL(2) of minimal type. *Comm. Math. Helv.*, 94(4):767–801, 2019.

[11] Yueke Hu and Abhishek Saha. Sup-norms of eigenfunctions in the level aspect for compact arithmetic surfaces, II: newforms and subconvexity. *arXiv:1905.06295 [math]*, December 2019.

[12] Yueke Hu, Hongbo Yin, and Jie Shu. Waldspurger's period integral for newforms. *To appear on Acta Arith. arXiv:1907.11428.*

[13] A. Knightly and C. Li. Kuznetsov's trace formula and the Hecke eigenvalues of Maass forms. *Memoirs of the American Mathematical Society*, 224(1055):vi+132, 2013.

[14] Andrew Knightly and Charles Li. A relative trace formula proof of the Petersson trace formula. *Acta Arith.*, 122(3):297–313., 2006.

[15] E. Kowalski, P. Michel, and J. VanderKam. Rankin-Selberg L-functions in the level aspect. *Duke Mathematical Journal*, 114(1):123–191, 2002.
[16] Paul D. Nelson. Microlocal lifts and quantum unique ergodicity on $\mathrm{GL}(2, \mathbb{Q}_p)$. *arXiv:1601.02528.*

Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China