Doubly-Efficient Private Information Retrieval from Modular Forms

Alexandra Henzinger* Stanford

Seyoon Ragavan† MIT

November 9, 2026

Abstract

We present new information-theoretic protocols for s-server, doubly-efficient private information retrieval (PIR) in sublinear server time and *one* extra bit of space on the servers, for every constant $s \geq 2$. On an n-bit database, each of the s PIR servers first builds and stores a preprocessing data structure consisting of $n + 1$ bits (which is unique to each server, and after which the server may discard the original database). Then, each server answers a PIR query in time $T_s(n)$, by reading at most T_s bits of its data structure and sending them to the user.

In the two server setting ($s = 2$), our PIR protocols achieve $T_2 = n^{0.82}$, matching the server time of the state-of-the-art schemes of Henzinger and Ragavan (EUROCRYPT 2026) while shrinking their storage from $O(n\sqrt{\log n})$ to $n + 1$ bits. With $s \geq 3$ servers, our PIR protocols retain the $n + 1$ bits of storage and additionally improve on the server time of all prior information-theoretic s-server schemes with $n^{1+o(1)}$ storage: for example, achieving $T_3 = n^{0.74}$ and $T_{10} = n^{0.61}$ (previously $n^{0.79}$ and $n^{0.75}$). Asymptotically, as the server count s grows, we shrink the number of servers required to achieve a given server-work exponent from $s^{3.9}$ to s.

Our results draw inspiration from matching-vector techniques (Efremenko, STOC 2009), which we generalize to handle degree-two computations in the exponent. Using this generalization, we build doubly-efficient PIR from the new intermediate abstraction of Fourier-sparse cone-delta functions. Then, we construct such Fourier-sparse cone-delta functions via a connection to modular forms and complex analysis: if a Fourier-sparse cone-delta function does not exist, then there exists a modular form which has too many roots, contradicting the valence formula.

*amh@cs.stanford.edu. Supported by the NSF GRFP under Grant No. 2141064, a MIT EECS Great Educators fellowship, gifts from Apple, Google, and Meta, and NSF CNS-2054869.

†sragavan@mit.edu. Supported by NSF CNS-2534400 and in part by Jane Street. This work was done in part while SR was a student researcher at Google Quantum AI. SR's access to ChatGPT Pro was supported by the UK AISI alignment project and the Simons Institute for the Theory of Computing, and his access to Claude Max was supported by Jane Street.

Contents

1 Introduction **1**
- 1.1 Results . . . 1
- 1.2 Guide to the Paper . . . 2

2 Technical Overview **4**
- 2.1 Inspiration from Matching-Vector PIR . . . 4
- 2.2 Doubly-Efficient PIR from Sparse Cone-Delta Functions . . . 5
- 2.3 Sparse Cone-Delta Functions from Modular Forms . . . 7
  - 2.3.1 Connecting to the Weil Representation of $\mathrm{SL}_2(\mathbb{Z})$ . . . 8
  - 2.3.2 Proof Architecture Mirroring Reed–Muller PIR . . . 10
  - 2.3.3 Connecting $\mathrm{SL}_2(\mathbb{Z})$ to Modular Forms . . . 10
- 2.4 Open Questions . . . 14

3 Related Work **15**

4 Preliminaries **18**
- 4.1 Private Information Retrieval . . . 18
- 4.2 Fourier-Sparse Cone-Delta Functions . . . 19

I Doubly-Efficient PIR from Fourier-Sparse Cone-Delta Functions **20**

5 Full Doubly-Efficient PIR Results **21**

6 Doubly-Efficient PIR with Two Servers **22**
- 6.1 Setup . . . 23
- 6.2 Protocol . . . 25
- 6.3 Analysis . . . 25

7 Doubly-Efficient PIR with $s \geq 3$ Servers **26**
- 7.1 Protocol . . . 28
- 7.2 Analysis . . . 28

8 Elementary Results on Sparse Cone-Delta Functions **29**
- 8.1 Simple Constructions for $q = s = 3$ from Hamming Slices . . . 29
- 8.2 Lower Bound . . . 32

9 From Expansions to n-Bit PIR **33**
- 9.1 From Characteristic 0 to Finite Fields . . . 33
- 9.2 From Field Elements to Bits . . . 35

II Fourier-Sparse Cone-Delta Functions from Modular Forms **37**

10 Background: $\mathrm{SL}_2(\mathbb{Z})$, the Weil Representation, and Modular Forms **38**
10.1 $\mathrm{SL}_2(\mathbb{Z})$ and Congruence Subgroups . . . 39
10.2 The Weil Representation . . . 40
10.3 Lattices in $\mathbb{C}$ and the Upper Half-Plane . . . 42
10.4 Modular Forms . . . 42
10.4.1 Weakly Modular Functions . . . 42
10.4.2 The Cusp at Infinity . . . 44
10.4.3 Arbitrary Cusps . . . 46
10.4.4 Composition Laws . . . 47
10.4.5 The Valence Formula . . . 48

11 Warm-Up **49**
11.1 Overview of the Proof . . . 49
11.2 Lattices in $\mathbb{R}^r$ . . . 52
11.3 Main Result . . . 53
11.4 Proof Step 1: Duality . . . 54
11.5 Proof Step 2: Vector and Scalar Theta Modular Forms . . . 55
11.6 Proof Step 3: The Order of the Cusp at 0 . . . 57
11.7 Proof Step 4: Root-Counting Contradiction from the Valence Formula . . . 59

12 Stronger Construction **59**
12.1 Overview of the Proof . . . 60
12.2 Main Result . . . 66
12.3 Proof Step 1: Duality and Symmetrization . . . 67
12.4 Proof Step 2: The Vector-Valued Jacobi Theta Modular Form . . . 68
12.5 Proof Step 3: Two Scalarizations . . . 71
12.5.1 Cusp Data . . . 71
12.5.2 Scalarization via Taylor Coefficients . . . 73
12.5.3 Scalarization via 2-Torsion Evaluations . . . 74
12.6 Proof Step 4: Lower Bounding Cusp Orders . . . 75
12.7 Proof Step 5: Applying the Valence Formula . . . 79

Statement on AI Use **82**

Acknowledgements **82**

Appendices **88**

A Proofs of Formal Theorem Statements **88**

B Background for Part II **89**
B.1 Complex Analysis . . . 89
B.2 Poisson Summation Formula . . . 90

B.3 Gaussian Riemann Sums 91

C Deferred Proofs from Section 10 **92**

D Deferred Proofs from Section 11 **94**

E Deferred Proofs from Section 12 **100**

1 Introduction

A *private information retrieval* (PIR) protocol [CGKS95, KO97] runs between s servers, who each hold a copy of an n-bit database $\mathsf{DB} \in \{0,1\}^n$, and a user, who holds an index $i \in [n]$. The protocol lets the user learn the i^{th} bit of the database, while each server learns nothing about the index i. More formally, each server cannot distinguish (either information-theoretically or computationally) between any two indices i or j that the user might be reading. This privacy property does not come cheap: Beimel, Ishai, and Malkin [BIM04] prove that, in any PIR scheme where the s servers store only the database DB, the servers must touch every bit of DB to answer a PIR query—in effect, running in $\Omega(n)$ time—as otherwise they would learn which locations the user is not reading.

Doubly-efficient PIR escapes this limitation by letting the servers preprocess their database once, ahead of time [BIM04, LMW23]: the s servers first transform the n-bit database DB into a $\mathrm{poly}(n)$-bit data structure of their choosing. Then, the servers use this data structure to answer each PIR query in $o(n)$ time. In this setting, a much weaker version of Beimel, Ishai, and Malkin's lower bound applies: if the servers store b preprocessed bits, the PIR servers must run in $\Omega(n/b)$ time, as they could "logically" touch all bits of the database by probing only a far smaller number of bits in their preprocessed data structure [BIM04].

At the same time, known doubly-efficient PIR schemes are far from this ideal time-space tradeoff, either:

1. saving only polylogarithmic factors in the servers' runtime [Lee26b, Lee26a, Lip09, WY05], or
2. incurring server storage that is concretely enormous (in the computationally-private regime [LMW23, BIPW17, CHR17, OPPW25]) or that grows asymptotically-faster than the database size (in the information-theoretic one [BIM04, LLF+25, GLM+25, HR26]).

A puzzling question remains: to make a private data access, what is the right data structure for the job?

1.1 Results

In this paper, we introduce a new data structure for doubly-efficient PIR. Our approach draws inspiration from matching-vector PIR [Yek08, Efr09, DG16, GKS25] and a long line of work on preprocessing PIR [BIM00, LLF+25, GLM+25, HR26]. While our protocols are simple to describe, their correctness is proven via a connection to modular forms and complex analysis. This yields the following results:

1. We obtain the first doubly-efficient PIR schemes with a constant number of servers, polynomially-sublinear server time $n^{1-\Omega(1)}$, and $n+1$ bits of storage per server. That is, our schemes incur *one bit* of overhead in server storage relative to schemes without preprocessing (although each server must store a different data structure, rather than an identical copy of the database).
2. Among doubly-efficient PIR schemes with information-theoretic security and quasilinear (i.e., $n^{1+o(1)}$) server storage, we improve the communication and server time for all constant server counts $s \geq 3$ over the previous state of the art [HR26]. To be more precise, for a given server work budget, we shrink the number of required servers almost quartically (from $s^{3.9}$ to s) as the number of servers grows—all while achieving $n+1$ bits of storage.

As an added benefit, in our PIR schemes, the preprocessing data structure is simple and efficient to produce: it is a Fourier transform of the database with specific, server-dependent weights applied to each term, computable in $n \cdot \mathsf{polylog}(n)$ time.

We report the precise asymptotic performance of our doubly-efficient PIR, along with that of the prior state-of-the-art doubly-efficient schemes [HR26], in Table 1. Formal theorem statements follow in Section 5.

Table 1: Performance of information-theoretic doubly-efficient PIR with a constant number of servers $s \geq 2$, over a database of n bits: prior state of the art [HR26] and our schemes. Each entry n^ρ indicates that the PIR servers run in time $o(n^\rho)$ per query, and that the PIR communication is $o(n^\rho)$ bits; exponents are rounded up to four decimals. The last column reports the smallest number of servers needed to match the PIR exponent from this work using the prior state-of-the-art scheme [HR26] (restricted to the quasilinear-storage setting, given in their Corollary D.11); its last entry is approximately $s^{3.93}$. The last row gives the number of bits of per-server storage required by each PIR scheme, which stays constant as the server count s grows. (One distinction is that, in our schemes, each server stores a different data structure, rather than an identical copy of the database.) For each s, the best server time/communication/storage is highlighted in green.

	Prior work	**This work**			Num. servers
s	[HR26, D.11]	Track 1 (Thm 5.2)	Track 2 (Thm 5.3)	Track 3 (Thm 5.5)	[HR26] needs to match this work
Server time and communication					
2	$n^{0.8113}$	$n^{0.8113}$	—	$n^{0.8204}$	2
3	$n^{0.7897}$	$n^{0.7445}$	$n^{0.7897}$	$n^{0.7327}$	11
4	$n^{0.7897}$	$n^{0.7445}$	$n^{0.7897}$	$n^{0.7025}$	19
5	$n^{0.7628}$	$n^{0.7445}$	$n^{0.7897}$	$n^{0.6535}$	73
6	$n^{0.7628}$	$n^{0.7445}$	$n^{0.7897}$	$n^{0.6521}$	79
7	$n^{0.7456}$	$n^{0.7445}$	$n^{0.7095}$	$n^{0.6198}$	307
8	$n^{0.7456}$	$n^{0.7445}$	$n^{0.7095}$	$n^{0.6198}$	307
9	$n^{0.7456}$	$n^{0.7445}$	$n^{0.7095}$	$n^{0.6198}$	307
10	$n^{0.7456}$	$n^{0.7445}$	$n^{0.7095}$	$n^{0.6078}$	587
11	$n^{0.7239}$	$n^{0.7445}$	$n^{0.6609}$	$n^{0.5901}$	2161
$s \to \infty$	$n^{\frac{1}{2}+\frac{1+o(1)}{\log_2 s}}$	$n^{0.7445}$	$n^{\frac{1}{2}+\frac{0.467+o(1)}{\log_2 s}}$	$n^{\frac{1}{2}+\frac{0.255+o(1)}{\log_2 s}}$	$s^{2/\log_2(\pi e/6)+o(1)}$
Server storage					
any	$n \cdot \mathrm{polylog}(n)$	$n+1$	$n+1$	$n+1$	$n^{1+o(1)}$

1.2 Guide to the Paper

Our new PIR constructions build on a shared intermediate abstraction which we call *Fourier-sparse cone-delta functions*, a special case of the delta functions defined by Ghasemi and Kopparty in their study of matching-vector PIR [GK26]. This abstraction splits the paper into two parts: Part I builds doubly-efficient PIR from Fourier-sparse cone-delta functions, and gives elementary (but limited) constructions of these functions. Part II constructs even sparser cone-delta functions from modular forms.

Some of our designs and analysis of sparse cone-delta functions rely on tools from complex analysis and number theory. In the interest of readability and modularity, we organize these constructions into three "reading tracks," according to the mathematics they use:

- In Part I, our elementary constructions of Fourier-sparse cone-delta functions, together with our compiler from cone-delta functions to doubly-efficient PIR, yield the "Track 1" results displayed in Table 1. These results include our two-server doubly-efficient PIR protocol, as well as a three-server

	Track 1: Fourier analysis	Track 2: Theta functions	Track 3: Jacobi forms
Main result	Main Theorem 5.2	Main Theorem 5.3	Main Theorem 5.5
Server time with s servers			
$s = 2$	$n^{0.812}$	—	$n^{0.821}$
$s = 3$	$n^{0.745}$	$n^{0.790}$	$n^{0.733}$
$s = 7$	$n^{0.745}$	$n^{0.710}$	$n^{0.620}$
$s \to \infty$	$n^{0.745}$	$n^{1/2+(0.467+o(1))/\log_2 s}$	$n^{1/2+(0.255+o(1))/\log_2 s}$
Mathematics	Fourier analysis over $\mathbb{F}_2$ and $\mathbb{F}_3$; polynomial interpolation	Modular forms for $\Gamma_0(q)$; the Weil representation; theta functions of lattices; the valence formula	All of Track 2, plus: Jacobi forms; the cusps of $\Gamma(q)$; Heisenberg operators; scalarization by Taylor coefficients and by values at 2-torsion points
Path with overviews	§§2.1–2.2	§2.3 and §11.1	§12.1
Path with proofs	§§4–8 and §9.2	§9.1, §10 (except Def 10.12, Lem 10.13, and Cor 10.36), §11; Appendices B–D	the parts of §10 skipped in Track 2, §12, and Appendix E

Table 2: The three tracks of our results on constructing doubly-efficient PIR over n-bit databases. In every scheme, each server stores a server-dependent data structure consisting of $n + 1$ bits. The server-time rows give the exponents of each track's own schemes: an entry n^ρ means that the server time and the communication are $o(n^\rho)$. The best entry for each s is highlighted in green. The "reading path" rows are cumulative: each column lists only what its track adds to the tracks preceding it.

doubly-efficient PIR protocol with $n + 1$ bits of server storage and a server-time exponent better than the prior, quasilinear-storage state of the art [HR26]. (A reader interested in only our cryptographic protocols may focus exclusively on Part I.)

- In Part II, we prove that even sparser cone-delta functions exist. Combining these results with the compiler from Part I immediately gives improved PIR protocols. Part II further splits into two tracks, which draw on the theory of modular forms at various levels of complexity:
  - Track 2 exposes the main conceptual ideas underlying the connection between modular forms and cone-delta functions (using, in particular, the theta functions of lattices).
  - Track 3 gives our strongest results with considerably more complexity (including the use of Jacobi forms).

Table 2 summarizes each track's PIR result, and two paths for reading it: a short one through the informal technical overview ("Path with overviews") and a more detailed, full presentation ("Path with proofs"). We give a self-contained presentation of modular forms and their properties in Section 10, as well as more informal intuitions in Section 2.3.

2 Technical Overview

Next, we give an overview of our techniques. The high-level idea behind state-of-the art, information-theoretic doubly-efficient PIR schemes [BIM04, HR26] is to first build "unbalanced" PIR schemes with extremely short queries of length $\ell = (1 + o(1)) \cdot \log n$ bits. Then, these schemes convert such an unbalanced PIR scheme into a doubly-efficient one by having the s servers each build a data structure that stores a piece of information for every possible PIR query—i.e., consisting of $2^\ell = n^{1+o(1)}$ bits. Each possible PIR query determines a fixed pattern of $n^{1/2+\epsilon(s)}$ locations that the servers look up in this data structure and send back to the client, for constant $\epsilon(s)$ approaching $1/\log s + o(1/\log s)$ as the server count s grows.

We push this technique to its extreme: building PIR with queries of bitlength close to $\log n$ and having each server store a data structure consisting of $n + 1$ bits. As before, each possible query determines a fixed pattern of $n^{1/2+\epsilon(s)}$ locations that the servers look up in this data structure and send back to the client; but now for improved constant $\epsilon(s)$ that approaches $0.255/\log s + o(1/\log s)$ as s grows.

To make this design work, our starting point is to adapt techniques from *matching-vector codes*, which yield the most communication-efficient (but not doubly-efficient) PIR schemes to date [Yek08, Efr09, DG16]. Towards modifying these techniques to produce short enough queries for preprocessing, we study the intermediate abstraction of *delta functions* that are sparse in the Fourier domain, a mathematical object introduced by Ghasemi and Kopparty in their study of matching-vector PIR [GK26]. We give new constructions of Fourier-sparse delta functions on the set $\left\{\mathbf{x} \in \mathbb{Z}_q^d \,:\, \langle \mathbf{x}, \mathbf{x}\rangle \equiv 0 \pmod q\right\}$, which we call *the cone* (it is the zero set of the quadratic form $\langle \mathbf{x}, \mathbf{x}\rangle$, and closed under scaling). We refer to these functions as Fourier-sparse "cone-delta" functions, and we construct them via a connection to *modular forms*. Details on each of these steps follow.

Notation. For a positive integer n, we write $[n] := \{1, \ldots, n\}$ and write $\mathbb{1}[E]$ to denote the indicator of an event E. For an integer $q \geq 2$, we write $\mathbb{Z}_q$ to refer to the integers modulo q; for vectors $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_q^d$, we write their inner product as $\langle \mathbf{u}, \mathbf{v}\rangle := \sum_{k\in[d]} u_k v_k \in \mathbb{Z}_q$. We use the notation $\mathbf{r} \xleftarrow{\text{R}} \mathbb{Z}_q^d$ or $\mathbf{r} \leftarrow \mathbb{Z}_q^d$ to denote a vector sampled uniformly. Logarithms are base 2, unless otherwise specified.

2.1 Inspiration from Matching-Vector PIR

Given an n-bit database, the high-level idea behind s-server matching-vector PIR [Yek08, Efr09, DG16]—which is communication-efficient, but not doubly-efficient—is to associate the indices in $[n]$ with distinct vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{Z}_q^d$, for parameter d that grows with n, and then rely on a simple, seemingly-tautological equation:

$$\forall i \in [n], \qquad \mathsf{DB}_i = \sum_{j\in[n]} \mathsf{DB}_j \cdot \mathbb{1}[i = j] = \sum_{j\in[n]} \mathsf{DB}_j \cdot \mathbb{1}[\mathbf{u}_i = \mathbf{u}_j]. \tag{1}$$

To make this approach work, the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{Z}_q^d$ are selected to have suitable structure: namely, there exists a field $\mathbb{F}$ with a primitive q^{th} root of unity γ and an s-sparse polynomial $P \in \mathbb{F}[X]$ such that

$$\forall i, j \in [n], \qquad P\left(\gamma^{\langle \mathbf{u}_i, \mathbf{u}_j\rangle}\right) = \mathbb{1}[\mathbf{u}_i = \mathbf{u}_j]. \tag{2}$$

Plugging this into equation (1) and writing $P(X) = \sum_{t\in[s]} \alpha_t \cdot X^{c_t}$ gives, for any desired index $i \in [n]$:

$$\begin{aligned} \mathsf{DB}_i &= \sum_{j\in[n]} \mathsf{DB}_j \cdot P\left(\gamma^{\langle \mathbf{u}_i, \mathbf{u}_j\rangle}\right) \\ &= \sum_{j\in[n]} \mathsf{DB}_j \cdot \left(\sum_{t\in[s]} \alpha_t \gamma^{\langle c_t \mathbf{u}_i, \mathbf{u}_j\rangle}\right) \end{aligned}$$

$$= \sum_{t\in[s]} \left(\sum_{j\in[n]} \mathsf{DB}_j \cdot \alpha_t \gamma^{\langle c_t \mathbf{u}_i, \mathbf{u}_j\rangle} \right).$$

This chain of substitutions suggests the following (not yet private) protocol: a user, who wants to read the database record at index $i \in [n]$, sends the vector $c_t\mathbf{u}_i$ to each server $t \in [s]$; server t in turn responds with the sum $\sum_{j\in[n]} \mathsf{DB}_j \cdot \alpha_t \gamma^{\langle c_t\mathbf{u}_i, \mathbf{u}_j\rangle}$. Making this protocol private requires only a small tweak: the user instead samples a random vector $\mathbf{r} \leftarrow \mathbb{Z}_q^d$ and sends the masked vector $\mathbf{r} + c_t\mathbf{u}_i$ to server $t \in [s]$. This masking does little damage to the above computation's correctness: namely,

$$\forall i \in [n], \qquad \underbrace{\mathsf{DB}_i \cdot \gamma^{\langle \mathbf{r}, \mathbf{u}_i\rangle}}_{\text{recovered by the user, who derives } \mathsf{DB}_i} = \sum_{t\in[s]} \underbrace{\left(\sum_{j\in[n]} \mathsf{DB}_j \cdot \alpha_t \gamma^{\langle \mathbf{r} + c_t\mathbf{u}_i, \mathbf{u}_j\rangle} \right)}_{\text{server } t\text{'s PIR answer}}.$$

This recovers exactly the matching-vector PIR protocol of Efremenko [Efr09]. Why is this protocol not doubly-efficient? The reason is that equation (2) forces the inner products $\langle \mathbf{u}_i, \mathbf{u}_j\rangle$ to be different when $i = j$ vs. $i \neq j$. This requirement, combined with the s-sparsity of polynomial P, limits how dense the collection of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ can be as a subset of $\mathbb{Z}_q^d$. In the best-known construction [Efr09], these vectors have dimension $d = \omega(\log n)$, making the queries too large for preprocessing.

Put differently, building a doubly-efficient PIR from the matching-vector template requires finding a replacement for equation (2) that both tests for equality between *any* two vectors in $\mathbb{Z}_q^d$, and can be distributed across s servers while retaining privacy, much as in the above protocol. The former requirement closely mirrors the motivation behind Ghasemi and Kopparty's study of Fourier-sparse delta functions on restricted sets [GK26]; our next insight is to construct such Fourier-sparse delta functions on a new set.

2.2 Doubly-Efficient PIR from Sparse Cone-Delta Functions

To this end, we define the notion of *sparse cone-delta functions*, which is the new machinery at the core of our doubly-efficient PIR protocols. Informally, this is a function that is both sparse in its Fourier representation and that evaluates to the indicator function $\mathbb{1}[\mathbf{x} = 0]$ on the cone (i.e., for all $\mathbf{x} \in \mathbb{Z}_q^d$ with $\langle \mathbf{x}, \mathbf{x}\rangle \equiv 0 \pmod q$). (We defer a more detailed comparison with [GK26] to Section 4.) Looking ahead to our final construction, the reason for the cone is that we generalize matching-vector techniques from the previous subsection to handle a degree-two computation in the exponent of γ—allowing for direct cancellation of any inputs not on the cone.

The rest of this section makes our construction more precise:

Definition 2.1 (Fourier-sparse cone-delta function)**.** *Let q be an odd prime and $\mathbb{F}$ a field containing a primitive q^{th} root of unity γ. In this work, we say that a collection of coefficients $C = \left\{ c_{\mathbf{a}} \in \mathbb{F} : \mathbf{a} \in \mathbb{Z}_q^d \right\}$ is a* k-sparse cone delta function *on $\mathbb{Z}_q^d$ if there are at most k nonzero coefficients $c_{\mathbf{a}}$ in C and it holds that:*

$$\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}} \gamma^{\langle \mathbf{a}, \mathbf{x}\rangle} = \begin{cases} 1, \ \textit{if } \mathbf{x} = 0; \\ 0, \ \textit{if } \mathbf{x} \neq 0 \textit{ lies on the cone (i.e., } \langle \mathbf{x}, \mathbf{x}\rangle \equiv 0 \pmod q\textit{)}; \\ \textit{arbitrary, otherwise.}^1 \end{cases} \tag{3}$$

[1] Placing no restrictions on the sum whenever $\langle \mathbf{x}, \mathbf{x}\rangle \not\equiv 0 \pmod q$ is crucial; if this sum were forced to vanish for all non-zero $\mathbf{x}$, the resulting cone-delta function could not be sparse.

Before presenting our construction of cone-delta functions in Section 2.3, we explain how these functions give rise to doubly-efficient PIR protocols. For any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_q^d$, we have:

$$\mathbb{1}[\mathbf{u} = \mathbf{0}] = \left(\frac{1}{q}\sum_{t\in\mathbb{Z}_q} \gamma^{t\langle \mathbf{u},\mathbf{u}\rangle}\right)\cdot\left(\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}}\gamma^{\langle \mathbf{a},\mathbf{u}\rangle}\right) \quad\Rightarrow\quad \mathbb{1}[\mathbf{u} = \mathbf{v}] = \left(\frac{1}{q}\sum_{t\in\mathbb{Z}_q} \gamma^{t\langle \mathbf{u}-\mathbf{v},\mathbf{u}-\mathbf{v}\rangle}\right)\cdot\left(\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}}\gamma^{\langle \mathbf{a},\mathbf{u}-\mathbf{v}\rangle}\right).$$

To make the cancellation more explicit: in this second equation, the term $\left(\frac{1}{q}\sum_{t\in\mathbb{Z}_q} \gamma^{t\langle \mathbf{u}-\mathbf{v},\mathbf{u}-\mathbf{v}\rangle}\right)$ is the indicator that $\mathbf{u}-\mathbf{v}$ lies on the cone (i.e., $\langle \mathbf{u}-\mathbf{v}, \mathbf{u}-\mathbf{v}\rangle \equiv 0 \pmod q$), since the element γ has order q. Meanwhile, whenever $\mathbf{u}-\mathbf{v}$ lies on the cone, the cone-delta property guarantees that the term $\left(\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}}\gamma^{\langle \mathbf{a},\mathbf{u}-\mathbf{v}\rangle}\right)$ evaluates to $\mathbb{1}[\mathbf{u}-\mathbf{v}=0] = \mathbb{1}[\mathbf{u}=\mathbf{v}]$.

Our q-server, doubly-efficient PIR protocol then follows by plugging the above equations into the template from Section 2.1 and in particular equation (1). More precisely, letting the database consist of $n = q^d$ records $\mathsf{DB}_1, \ldots, \mathsf{DB}_n$, each in the field $\mathbb{F}$, and letting the collection of vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be all of $\mathbb{Z}_q^d$:

$$\begin{aligned}
\mathsf{DB}_i &= \sum_{j\in[n]} \mathsf{DB}_j \cdot \mathbb{1}[\mathbf{u}_i = \mathbf{u}_j] \\
&= \sum_{j\in[n]} \mathsf{DB}_j \cdot \left(\frac{1}{q}\sum_{t\in\mathbb{Z}_q} \gamma^{t\langle \mathbf{u}_i-\mathbf{u}_j,\mathbf{u}_i-\mathbf{u}_j\rangle}\right)\cdot\left(\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}}\gamma^{\langle \mathbf{a},\mathbf{u}_i-\mathbf{u}_j\rangle}\right) \\
&= \frac{1}{q}\sum_{t\in\mathbb{Z}_q}\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}} \sum_{j\in[n]} \mathsf{DB}_j \cdot \gamma^{t\langle \mathbf{u}_i-\mathbf{u}_j,\mathbf{u}_i-\mathbf{u}_j\rangle+\langle \mathbf{a},\mathbf{u}_i-\mathbf{u}_j\rangle} \\
&= \frac{1}{q}\sum_{t\in\mathbb{Z}_q}\sum_{\mathbf{a}\in\mathbb{Z}_q^d} \left(c_{\mathbf{a}}\gamma^{\langle \mathbf{a},\mathbf{u}_i\rangle+t\langle \mathbf{u}_i,\mathbf{u}_i\rangle}\left(\sum_{j\in[n]} \left(\mathsf{DB}_j\gamma^{t\langle \mathbf{u}_j,\mathbf{u}_j\rangle}\right)\cdot\gamma^{\langle -2t\mathbf{u}_i-\mathbf{a},\mathbf{u}_j\rangle}\right)\right).
\end{aligned}$$

Once again, masking the i-dependent terms on the RHS with a random vector is harmless:

$$\mathsf{DB}_i\cdot\gamma^{\langle \mathbf{r},\mathbf{u}_i\rangle} = \frac{1}{q}\sum_{t\in\mathbb{Z}_q}\sum_{\mathbf{a}\in\mathbb{Z}_q^d} \left(c_{\mathbf{a}}\gamma^{\langle \mathbf{a},\mathbf{u}_i\rangle+t\langle \mathbf{u}_i,\mathbf{u}_i\rangle}\left(\sum_{j\in[n]} \left(\mathsf{DB}_j\gamma^{t\langle \mathbf{u}_j,\mathbf{u}_j\rangle}\right)\cdot\gamma^{\langle \mathbf{r}-2t\mathbf{u}_i-\mathbf{a},\mathbf{u}_j\rangle}\right)\right). \tag{4}$$

The above identity then naturally points to the following q-server PIR:

Preprocessing. Each server $t \in \mathbb{Z}_q$ stores the truth table of the following function $f_t : \mathbb{Z}_q^d \to \mathbb{F}$:

$$f_t(\mathbf{x}) := \sum_{j\in[n]} \mathsf{DB}_j\gamma^{t\langle \mathbf{u}_j,\mathbf{u}_j\rangle+\langle \mathbf{x},\mathbf{u}_j\rangle}.$$

This data structure consists of exactly q^d elements in $\mathbb{F}$—matching the dimensions of the original database. A quirk is that each server's data structure depends on its own index $t \in [q]$. To compute it, each server $t \in [q]$ applies a quadratic phase $\gamma^{t\langle \mathbf{u}_j,\mathbf{u}_j\rangle}$ to each entry of the database, then performs a Fourier transform.

Querying. Suppose the user wants to recover the record DB_i for $i \in [n]$. To do this, the user samples a random mask $\mathbf{r} \leftarrow \mathbb{Z}_q^d$ and sends server $t \in [q]$ the query vector

$$\mathsf{qu}_t = \mathbf{r} - 2t\cdot\mathbf{u}_i \in \mathbb{Z}_q^d.$$

Security is immediate: each server's query is masked by the random vector $\mathbf{r}$.

Answering. Given a query $\mathsf{qu} \in \mathbb{Z}_q^d$, the server replies with the set of evaluations $f_t(\mathsf{qu} - \mathbf{a})$ for all vectors $\mathbf{a} \in \mathbb{Z}_q^d$ such that the corresponding coefficient of the cone-delta function $c_{\mathbf{a}} \neq 0$. Rather than computing these values, the server can simply fetch them from its precomputed truth table.

Reconstructing. The user combines the servers' answers according to equation (4), which tells us that:

$$\gamma^{-\langle \mathbf{r}, \mathbf{u}_i \rangle} \cdot \frac{1}{q} \sum_{t \in \mathbb{Z}_q} \sum_{\mathbf{a} \in \mathbb{Z}_q^d} c_{\mathbf{a}} \gamma^{\langle \mathbf{a}, \mathbf{u}_i \rangle + t\langle \mathbf{u}_i, \mathbf{u}_i \rangle} \underbrace{f_t(\mathbf{r} - 2t\mathbf{u}_i - \mathbf{a})}_{\text{included in server } t\text{'s answer}} = \mathsf{DB}_i.$$

Efficiency. Finally, we analyze the scheme's efficiency. The user sends $d \log q$ bits to each each server, who in turn responds with k elements in $\mathbb{F}$ from its preprocessed table, where k is the sparsity of the cone-delta function. Similarly, the server computation and the user's computation both are both $O(k)$ field operations.[2] We arrive at the following theorem (stated for databases whose records are field elements, though a standard transformation to bit-valued databases [DPT10], specified in Section 9, applies):

Informal Theorem 2.2. *Let $q \geq 3$ be a constant prime number of servers, and $\mathbb{F}$ a constant field. If there exists a cone-delta function from $\mathbb{Z}_q^d \to \mathbb{F}$ with Fourier sparsity $k \leq q^{d(1-\Omega(1))}$, then there exists a q-server doubly-efficient PIR scheme operating over databases that consist of q^d records in $\mathbb{F}$. This PIR scheme has information-theoretic security and the following efficiency:*

- *each server stores a data structure that consists of q^d records in $\mathbb{F}$ (and may then discard the original database),*
- *to make a PIR query, the user uploads $d\lceil \log q \rceil$ bits to each PIR server,*
- *to answer a PIR query, each server reads and sends back k elements in $\mathbb{F}$ from its data structure, and*
- *to reconstruct, the user performs $O(k)$ field operations.*

We end this section on two remarks: first, the observation that equation (4) carries out a computation that is degree-two in the vector $\mathbf{u}_i$ in its exponent, which generalizes the approach of classical matching-vector PIR (Section 2.1). The reason this works is that the degree-two term splits into a term $\gamma^{t\langle \mathbf{u}_i, \mathbf{u}_i \rangle}$ that the client applies during reconstruction, a term $\gamma^{t\langle \mathbf{u}_j, \mathbf{u}_j \rangle}$ that the server applies during preprocessing, and remaining linear terms that are handled as in matching-vector PIR. Put more abstractly, this degree-two function in the exponent is computable because its derivative is linear; Liu, Vaikuntanathan, and Wee use similar techniques in their protocols for conditional disclosure of secrets [LVW17, Section 4.1].

Second, looking ahead to Section 7, the protocol for doubly-efficient PIR we just constructed generalizes to arbitrary, non-prime server counts. To this end, we generalize of the notion of sparse, cone-delta functions, as will appear in Definition 4.3 and Section 4.2.

2.3 Sparse Cone-Delta Functions from Modular Forms

With that aside, the remaining piece of the doubly-efficient PIR construction is to show that Fourier-sparse cone-delta functions exist. For simple cases of $q \in \{2, 3\}$, we present elementary constructions of sparse cone-delta functions (or related objects for $q = 2$) that only make use of Fourier analysis, yielding the two-server and three-server PIR results given in Tables 1 and 2 ("Track 1").

Generalizing these constructions to Fourier-sparse cone-delta functions over larger q (which yields doubly-efficient PIR with larger server counts), and even obtaining a stronger sparsity-bound for the case of

[2] Although the user's reconstruction is written as a sum over all $\mathbf{a} \in \mathbb{Z}_q^d$, they only need to handle the k terms for which $c_{\mathbf{a}} \neq 0$.

$q = 3$, turns out to be far more involved and requires the use of modular forms. Before we dive into the proof, we give its high-level intuition: if a Fourier-sparse cone-delta function does *not* exist, then we will show that there exists a modular form which has too many roots, contradicting the valence formula. (In fact, proofs of classic Reed–Muller PIR basically follows this same template, changing "modular form" and "valence formula" to "polynomial" and "degree bound" respectively. We reiterate this parallel in Section 2.3.2.)

We focus on the following warm-up result, which is the core of the "Track 2" results given in Tables 1 and 2 and is proved formally in Section 11:

Informal Theorem 2.3 (Informal version of Corollary 11.10)**.** *Let q be an odd prime, and let $\mathcal{L} \subseteq \mathbb{R}^r$ be an even lattice*[3] *of determinant $\sqrt{q}$, with dual lattice $\mathcal{L}^*$. For each coset c of $\mathcal{L}$ in $\mathcal{L}^*$, let $w_{\mathcal{L}}(c) := \frac{q}{2} \min_{\mathbf{y} \in c} \langle \mathbf{y}, \mathbf{y} \rangle$ be the (scaled) squared length of a shortest vector in c. Then, as $d \to \infty$, there exists a k-sparse cone-delta functions on $\mathbb{Z}_q^d$ with $k = q^{(\eta_{\mathcal{L}} + o(1))d}$, where*

$$\eta_{\mathcal{L}} := \inf_{0 < t \le 1} \log_q \left(t^{-r(q+1)/24} \sum_{c \in \mathcal{L}^*/\mathcal{L}} t^{w_{\mathcal{L}}(c)} \right).$$

Roughly, the cone-delta functions we construct are supported on the frequency vectors $\mathbf{a}$ whose coordinates, identified appropriately[4] with cosets of $\mathcal{L}$ in $\mathcal{L}^*$, have total *cost* $\sum_i w_{\mathcal{L}}(a_i)$ at most $r(q+1)d/24$. The sum over c is a generating function for these costs, and the infimum over t is a Chernoff-type bound on the number of such frequency vectors. For primes $q \equiv 3 \pmod 4$, instantiating Informal Theorem 2.3 with a suitable lattice of rank 2 gives an exponent $\eta_q \le \frac{1}{2} + \frac{0.467 + o(1)}{\log_2 q}$ as $q \to \infty$ (Corollary 11.13). The theorem takes the lattice as input, so a lattice with a smaller $\eta_{\mathcal{L}}$ immediately gives sparser cone-delta functions and, through Main Theorem 5.3, faster PIR schemes; we discuss the search for such lattices in Section 2.4.

The warm-up is the simplest instance of our approach. It already gives q-server PIR schemes whose exponent is $\frac{1}{2} + \frac{0.467 + o(1)}{\log_2 q}$ for large q (Main Theorem 5.3), and it introduces every tool that the stronger construction of Section 12 uses to obtain smaller exponents and arbitrary server counts (Main Theorem 5.5). (See Table 1 for a comparison of the "Track 2" and "Track 3" results.) Our proofs combine two well-known results—specifically, the transformation law of lattice theta functions under the Weil representation [Bor98, Theorem 4.1] and the valence formula [Ran77, Theorem 4.1.4]—in a new way. Specifically, we use them to construct a dual witness to show that sparse cone-delta functions exist.[5] We sketch the main ideas of the stronger construction at the end of Section 2.3.3, deferring its overview to Section 12.1.

From here on, we take $\mathbb{F} = \mathbb{C}$ and $\gamma = \zeta_q := e^{2\pi i/q}$. The coefficients we construct in this subsection lie in $\mathbb{Q}(\zeta_q)$; Section 9.1 transfers them to the finite fields that our PIR schemes use. We describe each step informally, and leave details to the body of the paper.

2.3.1 Connecting to the Weil Representation of $\mathrm{SL}_2(\mathbb{Z})$

Observe that the condition in equation (3) can be compressed in terms of two matrices. The first is, unsurprisingly, the Fourier transform matrix $\rho_S \in \mathrm{GL}_q(\mathbb{C})$ defined by:

$$\rho_S \mathbf{e}_x := \sum_{y \in \mathbb{Z}_q} \zeta_q^{-2xy} \mathbf{e}_y. \tag{5}$$

[3] An even lattice is one where $\langle \mathbf{y}, \mathbf{y} \rangle \in 2\mathbb{Z}$ for all $\mathbf{y} \in \mathcal{L}$.

[4] We may make this identification since $\det \mathcal{L} = \sqrt{q} \Rightarrow |\mathcal{L}^*/\mathcal{L}| = q$, so the quotient group is cyclic of order q.

[5] For $q = 3$ and our particular choice of rank-2 lattice (Definition 11.11), the conclusion was already known: the support is contained in Hamming ball of an elementary construction of Alon and Lubetzky [AL07, proof of Theorem 2.2], though we offer an improved elementary construction in Section 8.1 (which shrinks the exponent to 0.7445 from 0.7897). These simpler results are all subsumed by our strongest 3-server result in Corollary 12.4, which uses heavier modular-form machinery.

Here $\{\mathbf{e}_x\}_{x\in\mathbb{Z}_q}$ is the standard basis of $\mathbb{C}^q$, and for $\mathbf{x}\in\mathbb{Z}_q^d$ we write $\mathbf{e}_{\mathbf{x}} := \mathbf{e}_{x_1}\otimes\cdots\otimes\mathbf{e}_{x_d}$ for the standard basis of $(\mathbb{C}^q)^{\otimes d} = \mathbb{C}^{q^d}$. In these terms, we have:

$$\sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\mathbf{a}}\zeta_q^{\langle \mathbf{a},\mathbf{x}\rangle} = \sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{-2\mathbf{a}}\zeta_q^{\langle -2\mathbf{a},\mathbf{x}\rangle} = \mathbf{c}^\top \rho_S^{\otimes d}\mathbf{e}_{\mathbf{x}}, \tag{6}$$

where we define $\mathbf{c}\in\mathbb{C}^{q^d}$ to collect the coefficients $c_{-2\mathbf{a}}$. Secondly, the condition that $\langle\mathbf{x},\mathbf{x}\rangle \equiv 0 \pmod q$ can be described in terms of the quadratic phase matrix $\rho_T\in \mathrm{GL}_q(\mathbb{C})$ defined by:

$$\rho_T\mathbf{e}_x := \zeta_q^{x^2}\mathbf{e}_x. \tag{7}$$

(The factor -2 in ρ_S comes from the identity $(x+y)^2 - x^2 - y^2 = 2xy$, which ties ρ_S to this quadratic phase.) In these terms, we have:

$$\langle\mathbf{x},\mathbf{x}\rangle \equiv 0 \pmod q \Leftrightarrow \rho_T^{\otimes d}\mathbf{e}_{\mathbf{x}} = \mathbf{e}_{\mathbf{x}}. \tag{8}$$

Putting equations (6) and (8) together gives the following succinct description of a k-sparse cone-delta function: there should exist a $\mathbf{c}\in\mathbb{C}^{q^d}$ with at most k nonzero entries such that:

$$\mathbf{c}^\top\rho_S^{\otimes d}\mathbf{e}_{\mathbf{x}} = \begin{cases} 1, \text{ if } \mathbf{x} = 0; \\ 0, \text{ if } \rho_T^{\otimes d}\mathbf{e}_{\mathbf{x}} = \mathbf{e}_{\mathbf{x}} \text{ and } \mathbf{x}\neq 0; \\ \text{arbitrary, otherwise.}\end{cases} \tag{9}$$

One might hope that a conceptual understanding of the matrices ρ_S and ρ_T will let us reason about sparse cone-delta functions. Miraculously, these same two matrices arise in the representation theory of $\mathrm{SL}_2(\mathbb{Z})$! We unpack this connection next.

The group $\mathrm{SL}_2(\mathbb{Z})$ consists of all 2×2 matrices with integer entries and determinant 1; it is generated by

$$S := \begin{bmatrix} 0 & -1 \\ 1 & 0\end{bmatrix} \text{ and } T := \begin{bmatrix} 1 & 1 \\ 0 & 1\end{bmatrix}.$$

One can intuitively think of this group as (up to sign conventions) encoding the transformations used by the Euclidean algorithm for computing the greatest-common-divisor of two integers: S corresponds to swapping and T corresponds to adding one element to the other. It turns out that there exists a mapping $\rho : \mathrm{SL}_2(\mathbb{Z})\to\mathrm{GL}_q(\mathbb{C})$ such that:

- $\rho(S) = \rho_S$ and $\rho(T) = \rho_T$,
- for all $g,h\in\mathrm{SL}_2(\mathbb{Z})$, we have $\rho(gh)\propto\rho(g)\rho(h)$, by which we mean the two matrices are the same up to a nonzero scalar.[6]

This mapping ρ is known as the *Weil representation* of $\mathrm{SL}_2(\mathbb{Z})$ [Wei64, Gér77, Str13]. Another observation that will be useful for intuition is the following: for any element $g\in\mathrm{SL}_2(\mathbb{Z})$ and vector $\mathbf{v}\in\mathbb{C}^q$, we have:

$$\mathbf{c}^\top\rho(g)^{\otimes d}\mathbf{v}^{\otimes d} = \mathbf{c}^\top\left(\rho(g)\mathbf{v}\right)^{\otimes d} = P_{\mathbf{c}}\left(\rho(g)\mathbf{v}\right), \tag{10}$$

[6] Readers familiar with quantum computing may recognize these matrices: up to a scalar and a relabeling of the basis, ρ_S is the quantum Fourier transform on a single q-dimensional qudit, and ρ_T is a diagonal quadratic phase gate. More generally, each $\rho(g)$ normalizes the group generated by the shift and clock operators X and Z (Definition 10.12 and Lemma 10.13), so it is a qudit Clifford gate; for odd q, the Clifford group modulo phases is isomorphic to $\mathrm{SL}_2(\mathbb{F}_q)\ltimes\mathbb{F}_q^2$ [App05, Theorem 1]. Clifford gates are only defined up to a global phase, which is why ρ is multiplicative only up to scalars.

for a homogeneous degree-d polynomial $P_{\mathbf{c}} : \mathbb{C}^q \to \mathbb{C}$ formed using the entries of $\mathbf{c}$ as its coefficients. The intuitive takeaway of this section is thus the following: to study sparse cone-delta functions over $\mathbb{C}$, a promising route is to study the matrices $\rho(g)^{\otimes d}$ for the Weil representation $\rho : \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{GL}_q(\mathbb{C})$. This in turn boils down to reasoning about compositions $P \circ \rho(g)$ for homogeneous degree-d $P : \mathbb{C}^q \to \mathbb{C}$.

Looking ahead, this motivates modular forms because they can be thought of as "Fourier-analytic" tools for studying representations of $\mathrm{SL}_2(\mathbb{Z})$ (and in particular, the Weil representation).

2.3.2 Proof Architecture Mirroring Reed–Muller PIR

We prove that sparse cone-delta functions exist by a root-counting argument, which splits into four steps (with full definitions in Section 2.3.3):

1. *Assume the contrary and dualize:* suppose that no cone-delta function has Fourier support in a set $S_d \subseteq \mathbb{Z}_q^d$ that we will specify. Then, by duality, there is a vector β, indexed by the points of the cone, with
$$\beta_{\mathbf{0}} = 1, \text{ and } \qquad \sum_{\mathbf{x}} \beta_{\mathbf{x}} \zeta_q^{\langle \mathbf{x},\mathbf{z} \rangle} = 0 \quad \forall \mathbf{z} \in S_d. \tag{11}$$
We build from its entries the coefficients of a homogeneous polynomial P_β of degree d in q variables.

2. *Choose a curve through the high-dimensional space:* the theta function $\Theta_{\mathcal{L}} : \mathcal{H} \to \mathbb{C}^q$ of a lattice $\mathcal{L}$ of rank r, a vector-valued modular form on the upper half-plane $\mathcal{H}$.

3. *Restrict the witness to the curve:* the function $P_\beta \circ \Theta_{\mathcal{L}} : \mathcal{H} \to \mathbb{C}$ is a scalar modular form of weight $dr/2$.

4. *Count roots to get a contradiction:* when S_d is chosen correctly, $P_\beta \circ \Theta_{\mathcal{L}}$ is not identically zero, yet it vanishes at a cusp to higher order than the valence formula, a root-counting bound for modular forms, allows.

In fact, this argument is qualitatively similar to how the reconstruction of classic Reed–Muller PIR [CGKS95, WY05] is shown, except vector-valued and scalar-valued modular forms are replaced with multivariate and univariate polynomials respectively.

2.3.3 Connecting $\mathrm{SL}_2(\mathbb{Z})$ to Modular Forms

The Möbius action of $\mathrm{SL}_2(\mathbb{Z})$ on the upper half-plane. To use all of the matrices $\rho(g)$ at once, we need a space on which $\mathrm{SL}_2(\mathbb{Z})$ acts, for which we turn to lattices. Recall the intuition that $\mathrm{SL}_2(\mathbb{Z})$ encodes the steps of the Euclidean algorithm, which repeatedly swaps two numbers and adds one to the other. Applied to a basis (ω_1, ω_2) of a lattice in $\mathbb{R}^2$, rather than to two integers, these steps produce new bases of the same lattice. More precisely, $g = \left[\begin{smallmatrix} a & b \\ c & d_0 \end{smallmatrix}\right] \in \mathrm{SL}_2(\mathbb{Z})$ sends the basis (ω_1, ω_2) to the basis $(c\omega_2 + d_0\omega_1, a\omega_2 + b\omega_1)$ of the same lattice, and the matrices in $\mathrm{SL}_2(\mathbb{Z})$ are exactly the changes of basis that preserve the lattice and its orientation. (We write d_0 for the lower-right entry, since d is the dimension.) This picture, and the rest of the background in this subsection, is classical; see, e.g., [Ser73, Chapter VII] and [DS05, Shi71].

Now identify $\mathbb{R}^2$ with $\mathbb{C}$ in the natural way, and regard two lattices as equivalent if one is a nonzero complex multiple of the other (i.e., a rotated and rescaled copy of it). Dividing by ω_1 shows that every lattice is equivalent to one with a basis of the form $(1, \tau)$, and by ordering the basis we may take τ in the upper

half-plane $\mathcal{H} := \{\tau \in \mathbb{C} : \Im\tau > 0\}$. Applying g to this basis gives the basis $(c\tau + d_0, a\tau + b)$ of the same lattice, and dividing by $c\tau + d_0$ turns it into the basis $(1, g * \tau)$ of an equivalent lattice (Figure 1), where

$$g * \tau := \frac{a\tau + b}{c\tau + d_0}.$$

That is, $\mathbb{Z} + \mathbb{Z}(g * \tau) = (c\tau + d_0)^{-1}(\mathbb{Z} + \mathbb{Z}\tau)$. Since $\Im(g * \tau) = \Im\tau/|c\tau + d_0|^2$, these Möbius transformations map $\mathcal{H}$ to itself, and this is the space we want: $\mathrm{SL}_2(\mathbb{Z})$ acts on $\mathcal{H}$ by $\tau \mapsto g * \tau$. In this picture, S and T are the two steps of the Euclidean algorithm: S replaces the basis $(1, \tau)$ by $(\tau, -1)$, and T replaces it by $(1, \tau + 1)$. As for the Euclidean algorithm, these two steps generate all of $\mathrm{SL}_2(\mathbb{Z})$ [Ser73, Chapter VII, Theorem 2].

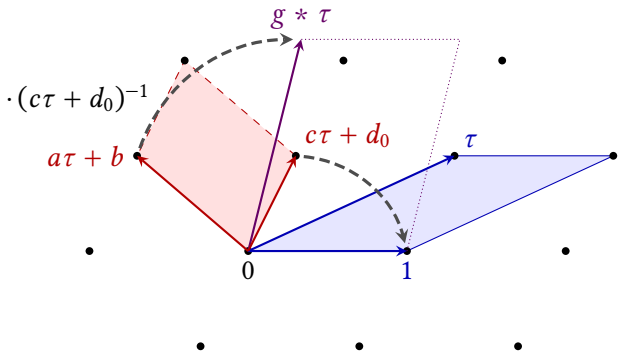


Figure 1: The action of $\mathrm{SL}_2(\mathbb{Z})$ through lattices. The dots form the lattice $\mathbb{Z} + \mathbb{Z}\tau$, with basis $(1, \tau)$ (blue). The matrix g gives another basis $(c\tau + d_0, a\tau + b)$ of the same lattice (red). Rotating and scaling the plane by $(c\tau + d_0)^{-1}$ (dashed arrows) turns this basis into the basis $(1, g * \tau)$ of a similar lattice (purple), where $g * \tau = \frac{a\tau+b}{c\tau+d_0}$.

A crucial ingredient for us will be extending this action to the point at ∞ (which can be thought of as the limit of it as $t \to \infty$). This then also requires us to introduce the boundary points in $\mathbb{Q}$, since $\frac{a\infty+b}{c\infty+d_0} = \frac{a}{c}$, informally speaking. We refer to these points in $\mathbb{Q} \cup \{\infty\}$ as the *cusps* of $\mathcal{H}$ (Figure 2); these will be the evaluation points at which we ultimately count roots to obtain a contradiction.

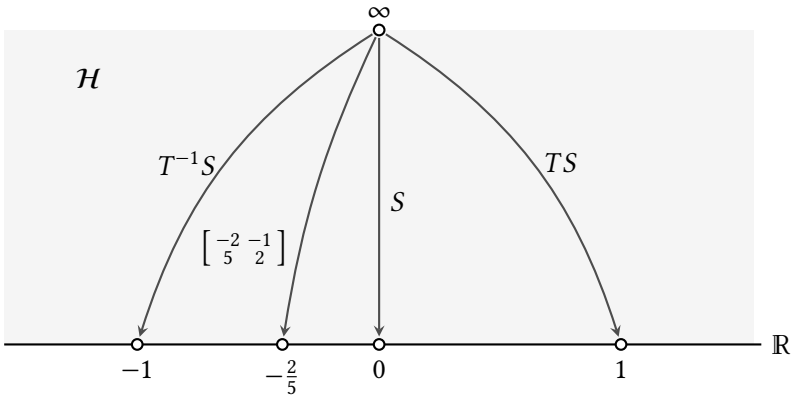


Figure 2: The upper half-plane $\mathcal{H}$, with the cusp ∞ (approached as $\Im\tau \to \infty$) and some of the cusps in $\mathbb{Q}$. Every cusp is the image $g * \infty = \frac{a}{c}$ of ∞ under some $g = \left[\begin{smallmatrix} a & b \\ c & d_0 \end{smallmatrix}\right] \in \mathrm{SL}_2(\mathbb{Z})$; the arrows show one such g for each cusp drawn, where $TS = \left[\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ and $T^{-1}S = \left[\begin{smallmatrix} -1 & -1 \\ 1 & 0 \end{smallmatrix}\right]$.

Modular forms. A *modular form* is a holomorphic (i.e., complex differentiable) function from $\mathcal{H}$ to $\mathbb{C}^q$ that

is "compatible" with the aforementioned action of $\mathrm{SL}_2(\mathbb{Z})$ and its Weil representation in a precise sense (for such vector-valued modular forms, see, e.g., [Bor98, Section 4]). This is analogous to Fourier analysis: a Fourier character is a function from $\mathbb{Z}_q$ to $\mathbb{C}$ that turns translation of its argument into multiplication of its value by a scalar. We depict this analogy in Figure 3, and now turn our attention to defining modular forms more precisely.

Definition 2.4 (Informal)**.** *A modular form of weight k for a subgroup $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ and a representation $\rho : \Gamma \to \mathrm{GL}_\ell(\mathbb{C})$ is a holomorphic function $f : \mathcal{H} \to \mathbb{C}^\ell$ satisfying both of the following:*

- *it is compatible with the Möbius action and ρ, in the sense that:*

$$f\left(\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} * \tau\right) \propto (c\tau + d_0)^k \rho(g) f(\tau) \quad \forall g \in \Gamma, \tau \in \mathcal{H}, \tag{12}$$

  where $\propto$ hides a nonzero constant that may depend on g but not on τ; and

- *very roughly, it does not blow up near any of the cusps in $\mathbb{Q} \cup \{\infty\}$. In this case, we can define an order of vanishing of f at each such cusp according to whether it approaches 0 near the cusp, and if so, how rapidly.*

When $\ell = 1$, we refer to this as a scalar *modular form. The differential terms $d\tau$ in Figure 3 are obtained by noting that $\frac{d}{d\tau}(g * \tau) = \frac{1}{(c\tau+d_0)^2}$.*

There are two useful features of modular forms that we highlight here:

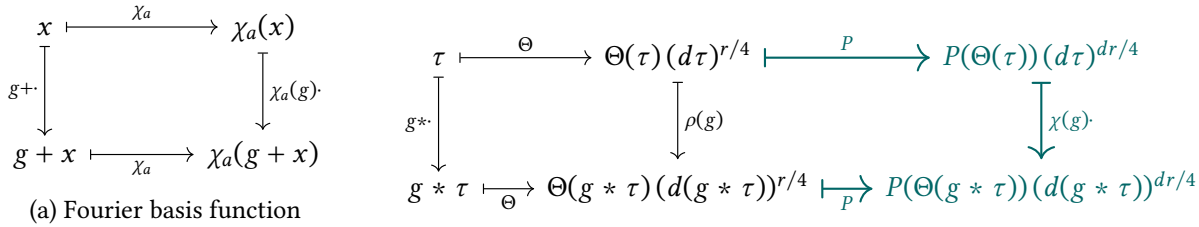
- The constraint of equation (12) together with the requirement that f is holomorphic imposes rigid structure on f which limits how many roots f can have (in analogy to the degree bound for low-degree polynomials). For modular forms, this is formally captured by the *valence formula* (Theorem 10.35), which our proofs make heavy use of.

- Secondly, composition with homogeneous polynomials gives a way to extract a scalar modular form from a vector-valued modular form. Indeed, let P be a homogeneous polynomial of degree d on $\mathbb{C}^\ell$ with $P(\rho(g)\mathbf{v}) \propto P(\mathbf{v})$ for all g in a subgroup Γ. Then, (12) and homogeneity give $P(\Theta(g * \tau)) \propto (c\tau + d_0)^{dk} P(\Theta(\tau))$ for all $g \in \Gamma$, so $P \circ \Theta$ is a scalar modular form of weight dk for Γ. This is the scalarization step of Figure 3(b), and an important ingredient of our proof. It is also how the classical correspondence between weight enumerators of codes and theta series of lattices arises [BE72,NRS06].

Our vector-valued modular form: a lattice theta function. Modular forms are often defined by encoding suitable information about a lattice. In our case, where ρ is the Weil representation, we use *theta functions* of lattices. Let $\mathcal{L} \subseteq \mathbb{R}^r$ be an even lattice (one with $\langle \mathbf{y}, \mathbf{y} \rangle \in 2\mathbb{Z}$ for all $\mathbf{y} \in \mathcal{L}$) of determinant $\sqrt{q}$; its rank r is then even. The dual lattice $\mathcal{L}^* := \{\mathbf{x} \in \mathbb{R}^r : \langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{Z} \text{ for all } \mathbf{y} \in \mathcal{L}\}$ contains $\mathcal{L}$, and $\mathcal{L}^*/\mathcal{L}$ is cyclic of order q, so its elements are the cosets $x\mathbf{a} + \mathcal{L}$, for $x \in \mathbb{Z}_q$ and a generator $\mathbf{a} + \mathcal{L}$. The *theta function* of $\mathcal{L}$ is

$$\Theta_{\mathcal{L}}(\tau) := \sum_{x \in \mathbb{Z}_q} \left(\sum_{\mathbf{y} \in x\mathbf{a} + \mathcal{L}} e^{\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau} \right) \mathbf{e}_x \in \mathbb{C}^q.$$

In the case where τ is purely imaginary, each coordinate of $\Theta_{\mathcal{L}}(\tau)$ is the Gaussian mass of a different coset of $\mathcal{L}$. One can check that $\Theta_{\mathcal{L}}$ is modular for $\mathrm{SL}_2(\mathbb{Z})$ and the Weil representation: for some intuition, equation (12) for $g = T$ follows from reindexing and for $g = S$ it follows from the Poisson summation



(a) Fourier basis function $\chi_a(x) = \zeta_q^{ax}$; $g \in \mathbb{Z}_q$ acts on $\mathbb{Z}_q$ by translation

(b) Vector-valued modular form Θ of weight $\frac{r}{2}$, scalarized by P; $g \in \Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ acts on $\mathcal{H}$ by Möbius maps

Figure 3: Fourier basis functions and modular forms both intertwine the action of a group on their domain with a linear action on their values. The scalarization step (teal) has no counterpart in (a), since every representation of $\mathbb{Z}_q$ splits into characters, while the Weil representation of $\mathrm{SL}_2(\mathbb{Z})$ does not.

formula (for the same reason that Poisson summation arises in lattice-based cryptography; see Section 3). This is well-known [Bor98, Theorem 4.1], but we provide a self-contained proof in Section 11.5. Finally, as $\Im\tau \to \infty$, every term except the one for $\mathbf{y} = \mathbf{0}$ decays, so $\Theta_{\mathcal{L}}(\tau) \to \mathbf{e}_0$.

Root counting for modular forms: the valence formula. The last ingredient for us to sketch is how we bound the root counts of modular forms, analogously to the degree bound for univariate polynomials. The *valence formula* (Theorem 10.35; see, e.g., [Ran77, Theorem 4.1.4]) plays this role: a nonzero modular form of weight k for Γ has total order of vanishing (across equivalence classes of cusps under Γ) at most

$$\frac{k}{12}\,[\mathrm{SL}_2(\mathbb{Z}) : \Gamma]\,.$$

Note that the dependencies on k and $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma]$ above are natural: k can be thought of as a degree-type parameter so larger k should allow for more roots, and larger $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma]$ means Γ is smaller and therefore equation (12) imposes fewer constraints on the modular form, thus allowing for more roots.

Putting the proof together. We can now explain how the conditions on a cone-delta function from Section 2.3.1 reappear in the proof. The witness β of Section 2.3.2 has the same structure, dually to equation (9): it is supported on the cone, so it is fixed by $\rho_T^{\otimes d}$; it satisfies $\beta_{\mathbf{0}} = 1$; and by equation (11) it is orthogonal to $\rho_S^{\otimes d}\mathbf{e}_{\mathbf{z}}$ for every $\mathbf{z} \in S_d$ (the sets S_d we use are closed under scaling by nonzero residues, which absorbs the factor -2 in ρ_S). Each of these becomes a property of the scalar function $P_\beta \circ \Theta_{\mathcal{L}}$:

- *The cone (ρ_T) gives modularity.* Every g in the subgroup

$$\Gamma_0(q) := \left\{ \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \pmod{q} \right\}$$

  acts on $\mathbb{C}^q$, up to a scalar, by permuting the coordinates ($x \mapsto d_0 x$) and multiplying them by quadratic phases. On the cone these phases cancel, just as for ρ_T, and after averaging P_β over the permutations $x \mapsto ux$, the function $P_\beta \circ \Theta_{\mathcal{L}}$ becomes a scalar modular form of weight $dr/2$ for $\Gamma_0(q)$. Since $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma_0(q)] = q + 1$, the valence formula allows it a total order of vanishing of $dr(q+1)/24$.

- *The normalization $\beta_{\mathbf{0}} = 1$ from (11) gives nonvanishing at the cusp at ∞.* As $\Im\tau \to \infty$, we have $\Theta_{\mathcal{L}}(\tau) \to \mathbf{e}_0$, so $P_\beta(\Theta_{\mathcal{L}}(\tau))$ tends to a nonzero multiple of $\beta_{\mathbf{0}}$.

- *The Fourier support (ρ_S) gives vanishing at the cusp at 0.* Because $S * \infty = 0$, it turns out that the order of vanishing of $P_\beta \circ \Theta_{\mathcal{L}}$ at 0 is governed by the lowest-order term of $\tau^{dr/2} P_\beta(\rho_S \Theta_{\mathcal{L}}(\tau))$ that is nonzero.

Expanded as a polynomial in the coordinates of $\Theta_{\mathcal{L}}$, this has no monomial indexed by a frequency vector $\mathbf{t} \in S_d$, by the orthogonality of β given by equation (11). The tth coordinate of $\Theta_{\mathcal{L}}$ is a sum of Gaussians $e^{\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau}$ over a coset of $\mathcal{L}$, so it vanishes at the cusp to order equal to the cost $w_{\mathcal{L}}(t)$ of that coset; a monomial indexed by $\mathbf{t}$ thus vanishes to order at least the total cost $\sum_i w_{\mathcal{L}}(t_i)$.

This tells us how to choose S_d: we work backwards from the contradiction we want. The argument needs $P_\beta \circ \Theta_{\mathcal{L}}$ to vanish at the cusp 0 to order more than $dr(q+1)/24$, the total that the valence formula permits. The orthogonality of the witness removes the contributions of the frequency vectors in S_d, and every other frequency vector contributes at order at least its total cost. So the high vanishing order at 0 is guaranteed as soon as S_d contains every frequency vector of total cost at most $r(q+1)d/24$, and since we want S_d to be small, we take it to consist of exactly these vectors (together with their multiples by nonzero residues). With this choice, $P_\beta \circ \Theta_{\mathcal{L}}$ vanishes at 0 to higher order than the valence formula allows, so no witness exists, and a cone-delta function with Fourier support in S_d does. This is the support in Informal Theorem 2.3: its budget $r(q+1)d/24$ is the total order of vanishing that the valence formula permits, and the lattice enters only through its rank r (the weight of $\Theta_{\mathcal{L}}$ is $r/2$) and the costs of its cosets. We give the details in Section 11.1.

The stronger construction. The stronger construction of Section 12 ("Track 3" in Tables 1 and 2) follows the same high-level architecture, with a number of changes:

- Firstly, instead of constructing Fourier-sparse cone-delta functions, we construct the more general quadratic-phase delta expansions alluded to at the end of Section 2.2. As a result, the dual witness is no longer supported on the cone, so $P_\beta \circ \Theta_{\mathcal{L}}$ is no longer modular for $\Gamma_0(q)$. We use instead the smaller subgroup $\Gamma(q)$ of matrices congruent to the identity modulo q, on which the Weil representation is trivial. It has $\frac{q^2-1}{2}$ classes of cusps rather than two; we will use the analogous condition to equation (11) to establish high-order vanishing at $s(q-1)/2$ of these cusp classes (rather than just at the cusp at 0).

- Next, we give the theta function an additional *elliptic* variable, making it a vector of theta functions with characteristics [Mum83, Chapter I, §3] and, after composing with the witness polynomial, a Jacobi form [EZ85]. We extract two scalar modular forms from it rather than one: one from its Taylor expansion in the elliptic variable around 0 [EZ85, Section 3], and one from its values at the half-periods $\frac{1}{2}$, $\frac{\tau}{2}$ and $\frac{1+\tau}{2}$ [EZ85, Theorem 1.3].

- The Fourier support set S_d is chosen to consist of two components that respectively ensure high-order vanishing of these two scalar modular forms.

  To obtain the final contradiction, we again apply the valence formula to both of these, following prior analyses by [RWR15, Aok22, LW26].

We give a more detailed overview of this proof in Section 12.1.

2.4 Open Questions

We conclude the technical overview with open questions that call for further investigation.

Breaking the $\sqrt{n}$-server-work barrier? Henzinger and Ragavan [HR26] explicitly ask whether there exists doubly-efficient PIR with $O(1)$ servers, $n^{1+o(1)}$ storage, and server work per query $n^{1/2-\Omega(1)}$. None of our new protocols meets this bar. Worse yet, Theorem 8.6 suggests this may be inherent to our techniques: no cone-delta function can have low-enough Fourier sparsity to achieve such a result via our compiler

in Section 7. This state of affairs is particularly puzzling because of a complementary result: server work $n^{o(1)}$ per query is possible if the number of servers is permitted to grow with the database size n, no matter how slowly [GLM+25].

More generally, for any constant server count $s \geq 2$, it remains open what the smallest achievable exponent α_s is that gives PIR with storage $n^{1+o(1)}$ and server work $n^{\alpha_s+o(1)}$ per query. One clean approach to this question would be to find better lattices to plug into Track 2. Main Theorem 5.3 turns every even lattice $\mathcal{L}$ of determinant $\sqrt{q}$ into a q-server PIR scheme with exponent $\eta_{\mathcal{L}}$, so improving Track 2 reduces to a question about lattices alone: which even lattices of determinant $\sqrt{q}$ minimize $\eta_{\mathcal{L}}$? (By Theorem 8.6, no lattice gives an exponent below $\frac{1}{2}$, but it is open whether some lattice matches or beats the stronger construction of Section 12.)

Smooth storage-work tradeoffs? Prior doubly-efficient PIR constructions based on Reed–Muller and multiplicity codes [BIM04, GLM+25, HR26] have the appealing feature of supporting smooth tradeoffs between storage and server work per query. We do not know whether the new protocols in this work can be extended to such a tradeoff.

Other applications of modular forms? Finally, throughout our results, modular forms play a role typically fulfilled by low-degree polynomials (e.g., in PIR). Given the ubiquity of low-degree polynomials in theoretical computer science, as described next in Section 3, we are hopeful that the tools from modular forms developed in this work may have uses elsewhere in coding theory and cryptography.

3 Related Work

Doubly-efficient PIR. Doubly-efficient private information retrieval has been studied in two settings: information-theoretic schemes, which require several non-colluding servers (as in this work), and single-server schemes, which require cryptographic assumptions. In either case, despite decades of study, only a handful of constructions are known, each with severe efficiency limitations shown in Table 3. All prior information-theoretic doubly-efficient PIR schemes [BIM04, LLF+25, GLM+25, HR26] store an $\omega(n)$-bit data structure on each server. All single-server doubly-efficient PIR schemes [BIPW17, LMW23, OPPW24, OPPW25] with polynomially-sublinear server time per query store a data structure of size $\Omega(\mathsf{poly}(\lambda) \cdot \mathsf{polylog}(n) \cdot n)$ bits [CIMR26], which is concretely enormous. In contrast, our construction stores $n + 1$ bits on each server—with no hidden constants.

In response to the elusiveness of doubly-efficient PIR, a weaker notion emerged in the literature: *secret-key* doubly-efficient PIR, which lets a client (using its own secret key) construct the database encoding that is later on stored and served by the server. The first candidates [BIPW17, CHR17] relied on new permuted-code assumptions [BHW19, BHMW21, BW21]. Building on this framework, recent work by Chen, Ishai, Jain, Mour, Rosen, and Xing [CIJ+26] gives candidate doubly-efficient secret-key PIR schemes with $O(n)$ storage and $n^{o(1)}$ communication and online server work. The server's storage can approach the database size, in suitable parameter regimes. Unlike in these schemes, our work does *not* require the PIR data structure to depend on the client's secret key; the servers can preprocess the database once and for all and then use the resulting encoding to answer queries from *all* clients.

Decoding polynomials and Fourier-sparse delta functions. In the matching-vector PIR of Efremenko [Efr09], the number of servers is equal to the number of monomials of an S-decoding polynomial. Two lines of work reduce this number of servers by (A) finding sparser S-decoding polynomials [Efr09, CFL+13, GR27], and (B) designing more server-efficient ways of using them in PIR [DG16, GKS25, ABL25].

Table 3: PIR schemes that use preprocessing to reduce server computation, on a database size $n \in \mathbb{N}$ and security parameter $\lambda \in \mathbb{N}$. We hide doubly-logarithmic factors in n, polylogarithmic factors in λ, and lower-order terms; we use $c \in \mathbb{N}$ to denote an arbitrary constant. Communication costs (in column 3) exclude setup and one-time communication. In all schemes, the communication costs can be made $\mathsf{poly}(\lambda) \cdot \log(n)$, without changing the asymptotic computation or storage costs, by additionally assuming LWE (or RLWE). Concrete storage overheads (in column 6) are measured or estimated on hundreds-of-megabyte to gigabyte-sized databases. *VBB denotes virtual black-box obfuscation, an idealized form of program obfuscation for which no concretely-efficient candidate constructions exist to date.

			Server costs		
Scheme (crypto assumptions)	Num. servers	Communication (asymptotic)	Computation (asymptotic)	Storage (asymptotic)	Storage (concrete overhead)
Linear-time PIR					
HHCG+23, DPC23 (LWE)	1	$\log(n) \cdot \sqrt{n}$	n	$\lambda \cdot \sqrt{n}$	$\approx 1\times$
PIR with polylogarithmic time savings					
Lee26b (LWE)	1	$\mathsf{poly}(\lambda) \cdot \sqrt{n}$	$n/\log^2 n$	$n^{1+o(1)}$	?
Lee26a (QR)	1	$\lambda \cdot \sqrt{n}$	$n/\log^2 n$	$\mathsf{poly}(\lambda) \cdot n^{1+1/c}$	?
WY05	2	$n^{1/3}$	$n/\log^c n$	$\mathsf{poly}(n)$	?
Doubly-efficient PIR with super-constant servers					
LLF+25	$\omega(\log n)$	$n^{o(1)}$	$n^{o(1)}$	$n^{1+o(1)}$	?
GLM+25	$\omega(1)$	$n^{o(1)}$	$n^{o(1)}$	$n^{1+o(1)}$	?
Doubly-efficient PIR with constant servers					
BIPW17, CHR17 (VBB*)	1	$\mathsf{poly}(\lambda) \cdot n^{o(1)}$	$\mathsf{poly}(\lambda) \cdot n^{o(1)}$	$\mathsf{poly}(\lambda) \cdot n^{1+o(1)}$	?
LMW23, OPPW24/25 (RLWE)	1	$\mathsf{poly}(\lambda) \cdot n^{o(1)}$	$\mathsf{poly}(\lambda) \cdot n^{o(1)}$	$\mathsf{poly}(\lambda) \cdot n^{1+o(1)}$	$\approx 10^8 - 10^9\times$
BIM00, GLM+25	2	$n^{0.82}$	$n^{0.82}$	$n^{1.48}$	$\approx 10^5 - 10^8\times$
HR26	2	$n^{0.82}$	$n^{0.82}$	$\sqrt{\log n} \cdot n$	$\approx 10 - 500\times$
This work	2	$n^{0.82}$	$n^{0.82}$	$n + 1$	$\approx 1\times$
HR26	c	$n^{1/2+1/\log_2 c+o(1)}$	$n^{1/2+1/\log_2 c+o(1)}$	$\mathsf{polylog}(n) \cdot n$	?
This work	c	$n^{1/2+0.255/\log_2 c+o(1)}$	$n^{1/2+0.255/\log_2 c+o(1)}$	$n + 1$	$\approx 1\times$

Ghasemi and Kopparty [GK26] recast the former task as a question about delta functions: how sparse in the Fourier basis can a function be, if it equals 1 at the origin and vanishes on the rest of the boolean cube $\{0,1\}^r$? Their lower bounds show that any such function must have Fourier sparsity $\geq r + 1$, as later shown to be tight (under a number-theoretic conjecture) for carefully chosen composite moduli [GR27]. Our lower bound on the sparsity of cone-delta functions (Theorem 8.6) adapts the multiplication argument of [GK26, Lemma 3.7].

Applications of modular forms. Modular forms have found a number of uses throughout theoretical computer science and combinatorics.

Explicit constructions of Ramanujan expander graphs [LPS88, Mar88]. These explicit constructions are Cayley graphs on $\mathrm{SL}_2(\mathbb{F}_p)$ for prime p. Their expansion is proven through bounds on the Fourier coefficients of certain modular forms, whereas our proofs use bounds on their orders of vanishing.

Distance bounds for special binary codes. The weight enumerator of a doubly-even self-dual binary code is invariant under a finite group, which confines it to a small ring of polynomials, and substituting theta

functions into such invariant polynomials gives modular forms [BE72, NRS06]. Mallows and Sloane [MS73] used this rigidity to upper-bound the minimum distance of such codes: beyond a point determined by the length, no such enumerator can have all of its low-weight coefficients vanish, which is the analog of a modular form vanishing to too high an order. Our proofs also substitute theta functions into an invariant polynomial and show that it cannot vanish as much as required, but the polynomial comes from a dual witness rather than from a code.

Sphere packing. The problem of nailing down the densest possible sphere packing in 8-dimensional space was resolved by Viazovska [Via17] using modular forms; soon after, Cohn, Kumar, Miller, Radchenko, and Viazovska [CKM+17] did the same for dimension 24. Very roughly, upper bounding the density can be boiled down to finding a solution to a dual linear program [CE03, Del73]; the aforementioned works used modular forms to construct optimal dual witnesses. We note that this is in contrast to our results, where modular forms are used to rule out — rather than construct — a dual witness to a particular linear system.

Discrete Gaussians in cryptography. The lattice theta function $\Theta_{\mathcal{L}}(\tau)$ we use in Section 11 is closely related to discrete Gaussian masses of lattice cosets; indeed, they coincide when $\tau \in \mathcal{H}$ is purely imaginary. Discrete Gaussian masses are central to lattice-based cryptography; the use of the Poisson summation formula to prove a theta function's transformation under S is analogous to its use for proving the lattice smoothing lemma in cryptography [Ban93, MR07, Reg09]. The smoothing lemma can be interpreted in our language as a statement about the rate of convergence of $\Theta_{\mathcal{L}}(\tau)$ (after appropriate normalization) to the all-ones vector as τ approaches the cusp at 0.

Vanishing bounds for Jacobi forms. Our analysis in Section 12 connects closely to work bounding how much a nonzero Jacobi form can vanish at a cusp [Aok22, LW26]. Useful tools include extracting modular forms from expansions in the elliptic variable and from evaluations at torsion points [EZ85, RWR15, Aok22, LW26], then applying modular-form vanishing bounds. We use this strategy with a Taylor expansion at the origin and evaluations at the three nonzero half-periods. The resulting inequalities rule out Jacobi forms arising from hypothetical dual witnesses, establishing the existence of our quadratic-phase delta expansions.

Elliptic curves. Modular forms are intimately related to elliptic curves, which are solution sets over fields to equations of the form $y^2 = x^3 + ax^2 + bx + c$. One intuition for this connection is that sublattices of $\mathbb{C}$ (which we have already connected to modular forms) correspond in an extremely tight way to elliptic curves over $\mathbb{C}$:

- Each $\tau \in \mathcal{H}$ is associated to an elliptic curve E_τ over $\mathbb{C}$.
- τ, τ' are related by some element of $\mathrm{SL}_2(\mathbb{Z})$ (under the Möbius action) if and only if E_τ and $E_{\tau'}$ are isomorphic.
- The points on E_τ form an abelian group whose structure is identical to that of the torus $\mathbb{C}/(\mathbb{Z} + \mathbb{Z}\tau)$ (see Section 12.1 and Fig. 12).

The arithmetic relationship between elliptic curves and modular forms is central to the proof of Fermat's Last Theorem [Fre86, Ser87, Rib90, Wil95, TW95]. Elliptic-curve groups are also ubiquitous in cryptography [Mil85, Kob87, Men93, KMV00].

4 Preliminaries

4.1 Private Information Retrieval

Definition. Following prior work [BIM04, HR26], for any $s \geq 2$ and ring $\mathcal{R}$, we define an *s-server PIR* protocol over a database in $\mathcal{R}^n$ to be a tuple of four routines (which each may take a security parameter as an implicit input). Crucially, our definition allows each server to store a distinct preprocessed data structure.

$\mathsf{Preprocess}(j, \mathsf{DB}) \to \mathsf{Table}_j$. Given a server index $j \in [s]$ and a database $\mathsf{DB} \in \mathcal{R}^n$, output a preprocessed data structure Table_j to be stored on the j^{th} server.

$\mathsf{Query}(i) \to (\mathsf{st}, \mathsf{qu}_1, \ldots, \mathsf{qu}_s)$. Given an index $i \in [n]$, output a secret client state and s PIR queries to be sent to each of the servers.

$\mathsf{Answer}^{\mathsf{Table}}(\mathsf{qu}) \to \mathsf{ans}$. Given a client query qu, along with oracle access to the preprocessed database structure Table, output a PIR answer ans.

$\mathsf{Reconstruct}(\mathsf{st}, \mathsf{ans}_1, \ldots, \mathsf{ans}_s) \to r$. Given a secret client state and s PIR answers, output the recovered database record $r \in \mathcal{R}$.

We require a PIR scheme to meet two properties, correctness and security, as we define next.

Correctness. At the end of the protocol, the user successfully learns the record of the database that it wants to read: that is, on database size $n \in \mathbb{N}$, for every database $\mathsf{DB} \in \mathcal{R}^n$ and every index $i \in [n]$,

$$\Pr\left[r = \mathsf{DB}_i \;\middle|\; \begin{array}{r} \text{for } j \in [s],\ \mathsf{Table}_j \leftarrow \mathsf{Preprocess}(j, \mathsf{DB}) \\ \mathsf{st}, \mathsf{qu}_1, \ldots, \mathsf{qu}_s \leftarrow \mathsf{Query}(i) \\ \text{for } j \in [s],\ \mathsf{ans}_j \leftarrow \mathsf{Answer}^{\mathsf{Table}_j}(\mathsf{qu}_j) \\ r \leftarrow \mathsf{Reconstruct}(\mathsf{st}, \mathsf{ans}_1, \ldots, \mathsf{ans}_s) \end{array} \right] = 1.$$

Security. At the end of the protocol, each server learns *nothing*about what record the user read. Specifically, we define an s-server PIR protocol to have *information-theoretic privacy* if, on database size $n \in \mathbb{N}$ and for all pairs of indices $i, j \in [n]$, each server's view is identical whether the user is querying for index i or j. Formally, for each $k \in [s]$, it holds that:

$$\left\{ \mathsf{qu}_k \;:\; \mathsf{st}, \mathsf{qu}_1, \ldots, \mathsf{qu}_s \leftarrow \mathsf{Query}(i) \right\} \equiv \left\{ \mathsf{qu}_k \;:\; \mathsf{st}, \mathsf{qu}_1, \ldots, \mathsf{qu}_s \leftarrow \mathsf{Query}(j) \right\}.$$

We define an s-server PIR protocol to have *computational privacy* if, on security parameter $\lambda \in \mathbb{N}$ and on a polynomially-bounded database size $n = N(\lambda) \in \mathbb{N}$, for all pairs of indices $i, j \in [n]$, the above two distributions are computationally indistinguishable.

Efficiency metrics. We measure the efficiency of an s-server PIR scheme over a database in $\mathcal{R}^n$ as in prior work [BIM04, LMW23, HR26]:

- *Communication.* The PIR scheme's communication is the worst-case bit length of all s queries and answers exchanged between the user and the s servers, on any database in $\mathcal{R}^n$.

- *Server storage.* The PIR scheme's server storage is the worst-case bit length of any of the s preprocessed data structures $\mathsf{Table}_1, \ldots, \mathsf{Table}_s$, produced on any database in $\mathcal{R}^n$.

- *Server time.* The PIR scheme's server time is the worst-case runtime of $\mathsf{Answer}^{\mathsf{Table}_j}(\cdot)$, summed across all j from 1 to s and measured in terms of the number of word operations (including but not limited to RAM accesses performed) on any database in $\mathcal{R}^n$.

Finally, following Lin, Mook, and Wichs, we define the following family of PIR schemes with double-efficiency properties [LMW23]: a PIR scheme is *doubly-efficient* if each of the s servers stores a polynomially-large data structure and answers queries in sublinear time. That is, on database size $n \in \mathbb{N}$, the server storage is $\mathsf{poly}(n)$ bits and the server time is $o(n)$.

4.2 Fourier-Sparse Cone-Delta Functions

We next define the mathematical abstractions at the core of our new template for doubly-efficient PIR. The following two definitions specialize those of Ghasemi and Kopparty [GK26] to the abelian group $\mathbb{Z}_q^d$:

Definition 4.1 (Fourier sparsity [GK26, Definition 3.1])**.** *Let q be a prime and $\mathbb{F}$ be a field containing a primitive q^{th} root of unity γ. For any function $f : \mathbb{Z}_q^d \to \mathbb{F}$, there exists a unique function $\hat{f} : \mathbb{Z}_q^d \to \mathbb{F}$ such that: $f(\mathbf{x}) = \sum_{\mathbf{w} \in \mathbb{Z}_q^d} \hat{f}(\mathbf{w}) \cdot \gamma^{\langle \mathbf{w}, \mathbf{x} \rangle}$. The function $\hat{f}$ is the Fourier representation of f. Then, the* Fourier sparsity *of the function f is the size of its Fourier representation's support,*

$$\mathsf{supp}\left(\hat{f}\right) = \left\{ \mathbf{w} \in \mathbb{Z}_q^d \,:\, \hat{f}(\mathbf{w}) \neq 0 \right\}.$$

Definition 4.2 (Delta function [GK26, Definition 2.1])**.** *For any subset $S \subseteq \mathbb{Z}_q^d$, a function $\delta : \mathbb{Z}_q^d \to \mathbb{F}$ is a* delta function on S *if it holds that*

$$\delta(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} = 0 \\ 0, & \text{if } \mathbf{x} \in S \backslash \{0\}. \end{cases}$$

In these terms, a k-sparse cone-delta function, as per Definition 2.1, is exactly a delta function for the cone (i.e., the set $\left\{ \mathbf{x} \in \mathbb{Z}_q^d \,:\, \langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod{q} \right\}$) with Fourier sparsity $\leq k$. However, our protocols work with the following more general definition that allows us to accommodate arbitrary server counts $s \leq q$ (rather than only prime server counts q).

Definition 4.3 (Quadratic-phase delta expansion)**.** *A collection of coefficients $\left\{ c_{\mathbf{w},t} \in \mathbb{F} \,:\, \mathbf{w} \in \mathbb{Z}_q^d, t \in \mathbb{Z}_q \right\}$ is a* quadratic-phase delta expansion *over $\mathbb{F}$, of dimension d, and of order q, if it holds that*

$$\sum_{t \in \mathbb{Z}_q} \sum_{\mathbf{w} \in \mathbb{Z}_q^d} c_{\mathbf{w},t} \cdot \gamma^{\langle \mathbf{w}, \mathbf{x} \rangle + t \cdot \langle \mathbf{x}, \mathbf{x} \rangle} = \begin{cases} 1, & \text{if } \mathbf{x} = 0 \\ 0, & \text{if } \mathbf{x} \in \mathbb{Z}_q^d \backslash \{0\}. \end{cases} \tag{13}$$

We say this expansion has support $\mathcal{W} \times \mathcal{T}$ if there exist subsets $\mathcal{W} \subseteq \mathbb{Z}_q^d$ and $\mathcal{T} \subseteq \mathbb{Z}_q$ such that $\mathsf{supp}\left\{ c_{\mathbf{w},t} \right\} \subseteq \mathcal{W} \times \mathcal{T}$. Finally, we say this expansion is (k, s)-sparse *if $|\mathcal{W}| \leq k$ and $|\mathcal{T}| \leq s$.*

We call a set $\mathcal{W} \subseteq \mathbb{Z}_q^d$ *permutation-invariant* if it is invariant under permuting the d coordinates. All the Fourier supports that we construct are permutation-invariant.

Looking ahead to Part I, in all our PIR schemes that build on a (k, s)-sparse quadratic-phase delta expansion, the Fourier sparsity k governs the PIR scheme's communication, while the parameter s determines the required number of non-colluding servers. Finally, we note that the notions in Definitions 2.1 and 4.3 are existentially equivalent when $s = q$, for an odd prime q:

Lemma 4.4. *For any odd prime q, if f is a cone-delta function on $\mathbb{Z}_q^d$ with Fourier sparsity k, then there exists a (k, q)-sparse quadratic-phase delta expansion of order q, of dimension d, and over the same field, with support $\mathsf{supp}(\hat{f}) \times \mathbb{Z}_q$. Conversely, for any s, if there exists a (k, s)-sparse quadratic-phase delta expansion of order q, dimension d, and over a field $\mathbb{F}$, then there exists a cone-delta function with Fourier sparsity k over the same field.*

Proof. First, consider a cone-delta function f with Fourier sparsity k. Then, consider the function

$$g(\mathbf{x}) := \left(\frac{1}{q} \sum_{t \in \mathbb{Z}_q} \gamma^{t\langle \mathbf{x}, \mathbf{x} \rangle} \right) \cdot f(\mathbf{x}).$$

This is a delta function on $\mathbb{Z}_q^d$, since:

1. at 0, both terms evaluate to 1, and
2. at a nonzero point we either have $\langle \mathbf{x}, \mathbf{x} \rangle \not\equiv 0 \pmod{q}$ in which case the first term vanishes, or $\langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod{q}$ in which case the second term vanishes by the cone-delta guarantee.

Expanding $f(\mathbf{x})$ out in its Fourier basis, we get:

$$g(\mathbf{x}) = \sum_{t \in \mathbb{Z}_q} \sum_{\mathbf{w} \in \mathbb{Z}_q^d} \frac{\hat{f}(\mathbf{w})}{q} \cdot \gamma^{\langle \mathbf{w}, \mathbf{x} \rangle + t\langle \mathbf{x}, \mathbf{x} \rangle},$$

so the coefficients $c_{\mathbf{w},t} := \hat{f}(\mathbf{w})/q$ (for all $t \in \mathbb{Z}_q$) form a (k, q)-sparse quadratic-phase delta expansion, completing the argument.

Conversely, consider a (k, s)-sparse quadratic-phase delta expansion $\{c_{\mathbf{w},t}\}$, and let $g(\mathbf{x})$ denote the left-hand side of its defining identity, so that g is the delta function on $\mathbb{Z}_q^d$. Then, whenever $\langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod{q}$, we have:

$$g(\mathbf{x}) = \sum_{\mathbf{w} \in \mathbb{Z}_q^d} \left(\sum_{t \in \mathbb{Z}_q} c_{\mathbf{w},t} \right) \cdot \gamma^{\langle \mathbf{w}, \mathbf{x} \rangle} =: f(\mathbf{x}),$$

and the function f so defined on all of $\mathbb{Z}_q^d$ agrees with the delta function g on the cone, so it is a cone-delta function; its Fourier coefficients $\sum_t c_{\mathbf{w},t}$ are supported on $\mathcal{W}$, so its Fourier sparsity is at most k. □

Part I
Doubly-Efficient PIR from Fourier-Sparse Cone-Delta Functions

Next, we show that the mathematical objects just defined (Fourier-sparse, cone-delta functions and their generalization to quadratic-phase delta expansions, as in Definitions 2.1 and 4.3) give new constructions of doubly-efficient PIR, which are the first to achieve $n+1$ bits of storage on each server. We state the resulting theorems first (Section 5). Then, we present the doubly-efficient PIR constructions for varying server counts (Sections 6 and 7). Finally, we give an elementary construction of the needed, new mathematical objects (Section 8), while deferring more complex constructions to Part II of the paper.

5 Full Doubly-Efficient PIR Results

We state our results on doubly-efficient PIR as three theorems, one for each track in Table 2. All three construct the following kind of scheme:

Definition 5.1. *For an integer $s \geq 2$ and a real number $\rho \in [0,1]$, an* s-server doubly-efficient PIR scheme with exponent ρ *consists of an s-server PIR scheme with information-theoretic security and the following efficiency, on database of $n \in \mathbb{N}$ bits:*

- *Each server stores a data structure consisting of $n+1$ bits (and may then discard the original database). The data structure is unique to each server and computable in $n \cdot \mathsf{polylog}(n)$ time, given public parameters that are shared by the servers and the user and can be sampled in expected $\mathsf{polylog}(n)$ time.*
- *To make a PIR query, the user uploads $O(\log^2 n)$ bits to each server.*
- *To answer a PIR query, each server reads and sends back at most $n^{\rho+o(1)}$ bits of its data structure, resulting in server time and communication $n^{\rho+o(1)}$.*

We give the proofs associated with each theorem statement, which combine results proved in the next sections, in Appendix A.

Main Theorem 5.2 (Track 1: Fourier analysis)**.** *There is a two-server doubly-efficient PIR scheme with exponent $H_2(1/4) < 0.8113$, where H_2 is the binary entropy function, and a three-server doubly-efficient PIR scheme with exponent $\eta_{\text{ends}} < 0.7445$, where $\eta_{\text{ends}} = 0.7444\ldots$ is the constant of Corollary 8.4.*

Main Theorem 5.3 (Track 2: theta functions of lattices)**.** *Let q be an odd prime, and let $\mathcal{L} \subseteq \mathbb{R}^r$ be an even lattice of determinant $\sqrt{q}$ (Section 11.2). For each coset $c \in \mathcal{L}^*/\mathcal{L}$, let $w_{\mathcal{L}}(c) := \frac{q}{2} \min_{\mathbf{y} \in c} \langle \mathbf{y}, \mathbf{y} \rangle$, and let*

$$\eta_{\mathcal{L}} := \frac{1}{\ln q} \inf_{0<t\leq 1} \left(\ln \sum_{c \in \mathcal{L}^*/\mathcal{L}} t^{w_{\mathcal{L}}(c)} - \frac{r(q+1)}{24} \ln t \right).$$

Then there is a q-server scheme with exponent $\eta_{\mathcal{L}}$.

In particular, for every prime $q \equiv 3 \pmod 4$, there is a q-server scheme with exponent η_q, the constant of (35), where $\eta_3 < 0.7897$, $\eta_7 < 0.7095$, $\eta_{11} < 0.6609$, and $\eta_q \leq \frac{1}{2} + \frac{0.467+o(1)}{\log_2 q}$ as $q \to \infty$. Consequently, as $s \to \infty$, there are s-server schemes with exponent $\frac{1}{2} + \frac{0.467+o(1)}{\log_2 s}$.

The exponent of Track 3 is the value of an explicit two-parameter optimization problem, which we define first.

Definition 5.4 (The Track 3 exponent)**.** *Let q be an odd prime and $s \in [2, q]$ an integer. For $\vartheta \in [0,1]$, define the cost $w_{q,\vartheta} : \mathbb{Z}_q \to \mathbb{R}$ and its generating function by*

$$w_{q,\vartheta}(j) := \min_{\substack{k \in \mathbb{Z} \\ k \equiv j \pmod q}} \big(k(k-1) + 2\vartheta k \big), \qquad M_\vartheta(t) := \sum_{j \in \mathbb{Z}_q} t^{w_{q,\vartheta}(j)} \quad (0 < t \leq 1),$$

and write $w_q := w_{q,1/2}$, so that $w_q(j) = \min_{k \equiv j \pmod q} k^2$. Call a pair (ϑ, v) of real numbers feasible *(for q and s) if*

$$0 \leq \vartheta \leq \min\left\{1, \frac{q+1}{6s}\right\} \qquad \text{and} \qquad 0 \leq v \leq \frac{q^2-1}{8s},$$

and for such a pair let

$$\gamma(\vartheta, v) := \frac{q-1}{2}\,\vartheta + \left(1 - \frac{6s}{q+1}\vartheta\right)\left(\frac{q^2-1}{4s} - 2v\right).$$

Finally, define

$$\begin{aligned} E_q^{\mathrm{sq}}(\nu) &:= \frac{1}{\ln q} \inf_{0<t\le 1} \left(\ln M_{1/2}(t) - \nu \ln t\right), \\ E_{q,s}^{\mathrm{edge}}(\vartheta,\nu) &:= \frac{1}{\ln q} \inf_{0<t\le 1} \left(\ln M_{\vartheta}(t) - \gamma(\vartheta,\nu)\ln t\right), \\ \eta_{q,s}(\vartheta,\nu) &:= \max\left(E_q^{\mathrm{sq}}(\nu),\ E_{q,s}^{\mathrm{edge}}(\vartheta,\nu)\right), \\ \eta_{q,s} &:= \inf\left\{\eta_{q,s}(\vartheta,\nu) \,:\, (\vartheta,\nu)\ \textit{feasible}\right\}. \end{aligned} \tag{14}$$

As in Track 2, these are Chernoff bounds on the number of frequency vectors of small total cost. The expansions behind Main Theorem 5.5 are supported on the vectors whose total cost is small for one of two costs on $\mathbb{Z}_q$, together with a few affine images of these vectors (Definition 12.1): the square cost w_q, with budget ν per coordinate, and the tilted edge cost $w_{q,\vartheta}$, with budget $\gamma(\vartheta,\nu)$. The two costs come from the two scalar modular forms in the proof (Section 12.1). The parameters ϑ and ν trade the two budgets against each other, and the feasibility constraints are the conditions that the proof needs.

Main Theorem 5.5 (Track 3: Jacobi forms)**.** *Let q be an odd prime and $s \in [2, q]$ an integer. For every feasible pair (ϑ,ν) (Definition 5.4), there is an s-server scheme with exponent $\eta_{q,s}(\vartheta,\nu)$; in particular, there is an s-server scheme with exponent $\eta_{q,s}$. More concretely, $\eta_{3,2} < 0.8204$, $\eta_{3,3} < 0.7327$, $\eta_{5,4} < 0.7025$, $\eta_{5,5} < 0.6535$, $\eta_{7,6} < 0.6521$, $\eta_{7,7} < 0.6198$, $\eta_{11,10} < 0.6078$, $\eta_{11,11} < 0.5901$, and $\eta_{q,q} \le \frac{1}{2} + \frac{0.255+o(1)}{\log_2 q}$ as $q \to \infty$. Consequently, as $s \to \infty$, there are s-server schemes with exponent $\frac{1}{2} + \frac{0.255+o(1)}{\log_2 s}$.*

Remark 5.6 (Reducing the communication with fully-homomorphic encryption)*. In each of these PIR schemes, a standard transformation (e.g., in [HR26]) uses fully-homomorphic encryption to preserve their efficiency guarantees while reducing the total communication to* $\mathsf{poly}(\lambda, \log n)$ *and downgrading their security guarantee to a computational (rather than information-theoretic) one. The storage and preprocessing time are unaffected, while the server time and upload communication obtain additional* $\mathsf{poly}(\lambda)$ *factors.*

6 Doubly-Efficient PIR with Two Servers

In this section, we give a stand-alone presentation of two-server doubly-efficient PIR scheme in which each server stores exactly as many field elements as the database has records. Our specific construction works over databases consisting of 2^d record in $\mathbb{F}_p$, for parameters d and p that we specify. The full construction appears in Fig. 4; it works much in the same way as the scheme presented in the technical overview (Section 2.2), but uses other quadratic functions than used in our cone-delta definition (Definition 2.1) that work over the special field $\mathbb{F}_2$ used in the exponent.

Applying a standard transformation of Dodis, Pătraşcu, and Thorup [DPT10], which we review in Section 9.2, then gives a doubly-efficient PIR scheme that works over arbitrary n-bit databases. Each PIR server stores an $(n+1)$-bit data structure and runs in $o(n^{0.82})$ time per query, proving the two-server result in Main Theorem 5.2 ("Track 1"). Compared to the prior, doubly-efficient PIR of Henzinger and Ragavan [HR26], this new two-server PIR scheme improves the server-side storage from $1.5\sqrt{\log n}\cdot n$ bits all the way down to $n+1$ bits, while asymptotically matching its per-query server work and communication.

In the remainder of this section, we show that Fig. 4's PIR scheme proves:

Theorem 6.1. *For every integer $d \ge 2$ with $d \equiv 2 \pmod 4$ and every odd prime $p > d$, there is a two-server PIR scheme with information-theoretic security that operates over any database consisting of 2^d records in $\mathbb{F}_p$, with the following efficiency:*

Two-server doubly-efficient PIR. Parameterized by positive integer $d \equiv 2 \pmod 4$ and odd prime $p > d$. The PIR scheme operates over databases consisting of 2^d records in $\mathbb{F}_p$. Define the set $\mathcal{W} \subseteq \mathbb{F}_2^d$ and the coefficients $\{c_{\mathbf{w}}\}_{\mathbf{w}\in\mathcal{W}}$ as in Section 6. For vectors $\mathbf{x} \in \mathbb{F}_2^d$, define the quadratic function in $\mathbb{F}_2^d \to \mathbb{F}_2$ as:

$$Q(\mathbf{x}) = \sum_{1\le i\le j\le d} x_i x_j.$$

$\mathsf{Preprocess}(\mathsf{DB} \in \mathbb{F}_p^{2^d}) \to \mathsf{Table}_0, \mathsf{Table}_1$

- Split DB into 2^d records $\{\mathsf{DB}_{\mathbf{v}}\}_{\mathbf{v}\in\mathbb{F}_2^d}$ each in $\mathbb{F}_p$.
- For $t \in \{0,1\}$, let the function $f_t : \mathbb{F}_2^d \to \mathbb{F}_p$ be
$$f_t(\mathbf{x}) = \sum_{\mathbf{v}\in\mathbb{F}_2^d} \mathsf{DB}_{\mathbf{v}} \cdot (-1)^{Q(\mathbf{x}+\mathbf{v})+t\cdot Q(\mathbf{v})}.$$
- Set Table_t to be the truth table of the function f_t.

$\mathsf{Query}(i \in [2^d]) \to \mathsf{st}, \mathsf{qu}_0, \mathsf{qu}_1$

- Let $\mathbf{i} \in \mathbb{F}_2^d$ be the bit-vector corresponding to index i.
- Sample $\mathbf{r} \xleftarrow{R} \mathbb{F}_2^d$ and set $\mathsf{qu}_0 = \mathbf{r}$ and $\mathsf{qu}_1 = \mathbf{r} + \mathbf{i}$.
- Output $\mathsf{st} = (\mathbf{i}, \mathbf{r})$, qu_0, and qu_1.

$\mathsf{Answer}^{\mathsf{Table}_t}(\mathsf{qu}_t \in \mathbb{F}_2^d) \to \mathsf{ans}_t \in \mathbb{F}_p^{|\mathcal{W}|}$

- Output $\mathsf{ans}_t = \left\{\mathsf{Table}_t[\mathsf{qu}_t + \mathbf{w}]\right\}_{\mathbf{w}\in\mathcal{W}}$.

$\mathsf{Reconstruct}(\mathsf{st}, \mathsf{ans}_0, \mathsf{ans}_1) \to \mathbb{F}_p$

- Parse $(\mathbf{i}, \mathbf{r})$ from st and $\mathsf{Table}_t[\mathsf{qu}_t + \mathbf{w}]$ from each ans_t.
- Output the $\mathbb{F}_p$-element computed as
$$\frac{1}{2}\sum_{t\in\{0,1\}}\sum_{\mathbf{w}\in\mathcal{W}} c_{\mathbf{w}}\,(-1)^{Q(\mathbf{r}+(t+1)\cdot\mathbf{i}+\mathbf{w})+t\cdot Q(\mathbf{i})}\,\mathsf{Table}_t[\mathsf{qu}_t + \mathbf{w}].$$

Figure 4: Two-server doubly-efficient PIR, over database consisting of 2^d records in $\mathbb{F}_p$.

- *each server stores a server-dependent data structure that consists of 2^d elements of $\mathbb{F}_p$, computed from the database in $O(d2^d)$ operations over $\mathbb{F}_p$ (and may then discard the original database),*
- *to make a PIR query, the user uploads d bits to each PIR server, and*
- *to answer a PIR query, each server reads and sends back at most $2^{H_2(1/4)(d+1)}$ elements of $\mathbb{F}_p$ in its data structure, where H_2 is the binary entropy function.*

6.1 Setup

On a database which consists of 2^d records in $\mathbb{F}_p$, for an integer $d \ge 2$ with $d \equiv 2 \pmod 4$ and an odd prime $p > d$, the PIR scheme is parametrized by:

1. A *sparse decoding set* $\mathcal{W}$ defined, for the integer $\ell = (d-2)/4$, as
$$\mathcal{W} = \{\mathbf{w} \in \mathbb{F}_2^d : |\mathbf{w}| \le \ell\} \cup \{\mathbf{w} \in \mathbb{F}_2^d : |\mathbf{w}| \ge d - \ell + 1\}.$$
In other words, the set $\mathcal{W}$ contains all vectors of weight at most ℓ, together with the complements of all vectors of weight at most $\ell - 1$. These vectors are distinct, giving $|\mathcal{W}| = \sum_{j=0}^{\ell}\binom{d+1}{j}$.

2. A *set of public recovery coefficients* $c_{\mathbf{w}} \in \mathbb{F}_p$ for each vector $\mathbf{w} \in \mathcal{W}$, defined to be the Fourier coefficients of the following polynomial, using the parity function $\pi(\mathbf{x}) = |\mathbf{x}| \pmod 2$:
$$G(\mathbf{x}) = \prod_{j=1}^{\ell}\left(1 - \frac{|\mathbf{x}| + \pi(\mathbf{x})}{4j}\right) \in \mathbb{F}_p, \qquad \text{for all } \mathbf{x} \in \mathbb{F}_2^d$$

$$c_{\mathbf{w}} = \frac{1}{2^d} \cdot \sum_{\mathbf{v} \in \mathbb{F}_2^d} G(\mathbf{v}) \cdot (-1)^{\langle \mathbf{w}, \mathbf{v} \rangle}, \qquad \text{for all } \mathbf{w} \in \mathcal{W}$$

These are well defined, since p is odd and $p > d > 4\ell$, so that 2^d and $4j$ for $j \in [\ell]$ are invertible in $\mathbb{F}_p$.

We additionally define the quadratic function $Q : \mathbb{F}_2^d \to \mathbb{F}_2$ as $Q(\mathbf{x}) = \sum_{1 \le i \le j \le d} x_i x_j$.

Properties. These parameter choices generalize the earlier Fourier-sparse delta-function framework, as they use the quadratic cone given by $Q(\mathbf{x}) = 0$, rather than direct vector inner-products $\langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod q$ as in Definition 2.1. (This is necessary since $\langle \mathbf{x}, \mathbf{x} \rangle$ is linear over $\mathbb{F}_2$.) In particular, mirroring the cone-delta definition, these choices offer the following property, which will lie at the core of our PIR scheme: for all vectors $\mathbf{x} \in \mathbb{F}_2^d$ with $Q(\mathbf{x}) = 0$,

$$\sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w}} \cdot (-1)^{Q(\mathbf{x}+\mathbf{w})+Q(\mathbf{w})} = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{0} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

We give a proof of this property next.

Proof. Consider any nonzero vector $\mathbf{x} \in \mathbb{F}_2^d$ with $Q(\mathbf{x}) = 0$. Then, as $Q(\mathbf{x}) = 0$ implies that $|\mathbf{x}| \equiv 0, 3 \pmod 4$, the integer $|\mathbf{x}| + \pi(\mathbf{x})$ must be a multiple of 4. Its possible values are $4, 8, \ldots, 4\ell$, as $d = 4\ell + 2$. Therefore, the function $G(\mathbf{x})$, which is defined as

$$G(\mathbf{x}) = \prod_{j=1}^{\ell} \left(1 - \frac{|\mathbf{x}| + \pi(\mathbf{x})}{4j}\right) \in \mathbb{F}_p,$$

must evaluate to 1 at input $\mathbf{x} = \mathbf{0}$ and evaluate to 0 at every other point $\mathbf{x} \in \mathbb{F}_2^d$ with $Q(\mathbf{x}) = 0$.

Next, we examine the representation of $G(\mathbf{x})$ in the Fourier domain. Applying the identity that, for any bit $b \in \{0, 1\}$ it holds that $b = (1 - (-1)^b)/2$, we see that

$$|\mathbf{x}| = \sum_{i=1}^{d} x_i = \frac{d - \sum_{i=1}^{d} (-1)^{x_i}}{2} \qquad \text{and}$$

$$\pi(\mathbf{x}) = \frac{1 - (-1)^{\pi(\mathbf{x})}}{2} \qquad \text{giving that}$$

$$|\mathbf{x}| + \pi(\mathbf{x}) = \frac{d + 1 - \sum_{i=1}^{d} (-1)^{x_i} - (-1)^{\pi(\mathbf{x})}}{2}.$$

As a result, the function $G(\mathbf{x})$ can be expressed as a multilinear polynomial of degree at most ℓ in the $d + 1$ variables $X_1 = (-1)^{x_1}, \ldots, X_d = (-1)^{x_d}$, and $Y = (-1)^{\pi(\mathbf{x})}$, using $X_i^2 = Y^2 = 1$. In this polynomial, any monomial without the term $Y = (-1)^{\pi(\mathbf{x})}$ can be expressed as $(-1)^{\langle \mathbf{w}, \mathbf{x} \rangle}$, for some vector $\mathbf{w}$ of norm at most ℓ. Any monomial with the term $Y = (-1)^{\pi(\mathbf{x})}$ can be expressed as $(-1)^{\langle \mathbf{1}+\mathbf{w}, \mathbf{x} \rangle}$ for some vector $\mathbf{w}$ of norm at most $\ell - 1$ and where $\mathbf{1}$ is the all-ones vector. Here, $|\mathbf{1} + \mathbf{w}| = d - |\mathbf{w}| \ge d - \ell + 1$, so the vector $\mathbf{1} + \mathbf{w}$ has norm at least $d - \ell + 1$. Each of these vectors is in the sparse-decoding set $\mathcal{W}$, implying that the Fourier support of function G lies in $\mathcal{W}$. That is, the public recovery coefficients $c_{\mathbf{w}}$ include all the non-zero coefficients of the Fourier representation of G.

Finally, using this observation and using that $Q(\mathbf{x} + \mathbf{w}) + Q(\mathbf{w}) = \langle \mathbf{w}, \mathbf{x} + \pi(\mathbf{x}) \cdot \mathbf{1} \rangle$ whenever $Q(\mathbf{x}) = 0$, it must also be that

$$\sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w}} \cdot (-1)^{Q(\mathbf{x}+\mathbf{w})+Q(\mathbf{w})} = \sum_{\mathbf{w} \in \mathcal{W}} \left(\frac{1}{2^d} \cdot \sum_{\mathbf{v} \in \mathbb{F}_2^d} G(\mathbf{v}) \cdot (-1)^{\langle \mathbf{w}, \mathbf{v} \rangle} \right) \cdot (-1)^{Q(\mathbf{x}+\mathbf{w})+Q(\mathbf{w})}$$

$$= \frac{1}{2^d} \sum_{\mathbf{v} \in \mathbb{F}_2^d} G(\mathbf{v}) \sum_{\mathbf{w} \in \mathbb{F}_2^d} (-1)^{\langle \mathbf{w}, \mathbf{v}+\mathbf{x}+\pi(\mathbf{x})\cdot\mathbf{1} \rangle}$$
$$= G(\mathbf{x} + \pi(\mathbf{x}) \cdot \mathbf{1})$$
$$= \begin{cases} 1, & \mathbf{x} = \mathbf{0}, \\ 0, & \mathbf{x} \neq \mathbf{0}. \end{cases}$$

The last equality follows because, if vector $\mathbf{x}$ has even weight, then $G(\mathbf{x} + \pi(\mathbf{x}) \cdot \mathbf{1}) = G(\mathbf{x})$. On the other hand, if vector $\mathbf{x}$ has odd weight and $Q(\mathbf{x}) = 0$, then $|\mathbf{x}| \equiv 3 \pmod 4$, so its complement has positive weight $d - |\mathbf{x}| \equiv 3 \pmod 4$ and also satisfies that $Q(\mathbf{1} + \mathbf{x}) = 0$. This completes the proof. □

6.2 Protocol

We next sketch the two-server, doubly-efficient PIR protocol, with servers indexed by $t \in \{0, 1\}$.

Preprocessing. The servers split the database into 2^d records in $\mathbb{F}_p$, each indexed by the vectors in $\mathbb{F}_2^d$. Server $t \in \{0, 1\}$ then stores the truth table of the following function $f_t : \mathbb{F}_2^d \to \mathbb{F}_p$:

$$f_t(\mathbf{x}) = \sum_{\mathbf{v} \in \mathbb{F}_2^d} \mathsf{DB}_\mathbf{v} \cdot (-1)^{Q(\mathbf{x}+\mathbf{v})+t\cdot Q(\mathbf{v})}.$$

Here, arithmetic in the exponent happens mod 2 (since -1 has order 2), while arithmetic in the base happens mod p—a setup mirroring matching-vector PIR [DG16].

Querying. Letting $\mathbf{i} \in \mathbb{F}_2^d$ be the address of the requested record, the user samples vector $\mathbf{r} \xleftarrow{R} \mathbb{F}_2^d$, then sends query $\mathsf{qu}_0 = \mathbf{r}$ to server 0 and query $\mathsf{qu}_1 = \mathbf{r} + \mathbf{i}$ to server 1.

Answering. For $t \in \{0, 1\}$ and given query qu_t, server t returns all of the evaluations $f_t(\mathsf{qu}_t + \mathbf{w})$ for each vector $\mathbf{w} \in \mathcal{W}$ to the user, by looking up each value in its precomputed truth table.

Reconstructing. The user recovers the chosen $\mathbf{i}^{\text{th}}$ database record, which is a $\mathbb{F}_p$-element, as

$$\mathsf{record} = \frac{1}{2} \sum_{\mathbf{w} \in \mathcal{W}} c_\mathbf{w} \cdot \left((-1)^{Q(\mathsf{qu}_1+\mathbf{w})} \cdot f_0(\mathsf{qu}_0 + \mathbf{w}) + (-1)^{Q(\mathsf{qu}_0+\mathbf{w})+Q(\mathbf{i})} \cdot f_1(\mathsf{qu}_1 + \mathbf{w}) \right) \pmod p.$$

6.3 Analysis

We give a formal proof of Theorem 6.1.

Correctness. Consider a user who aims to read the $\mathbf{i}^{\text{th}}$ database element, for $\mathbf{i} \in \mathbb{F}_2^d$. Expanding out the recovery equation and using the quadratic identity

$$Q(\mathbf{x} + \mathbf{y} + \mathbf{z}) = Q(\mathbf{x} + \mathbf{y}) + Q(\mathbf{x} + \mathbf{z}) + Q(\mathbf{y} + \mathbf{z}) + Q(\mathbf{x}) + Q(\mathbf{y}) + Q(\mathbf{z}) \pmod 2,$$

we see that:

$$\mathsf{record} = \frac{1}{2} \sum_{\mathbf{w} \in \mathcal{W}} c_\mathbf{w} \cdot \left((-1)^{Q(\mathbf{r}+\mathbf{w}+\mathbf{i})} \cdot f_0(\mathsf{qu}_0 + \mathbf{w}) + (-1)^{Q(\mathbf{r}+\mathbf{w})+Q(\mathbf{i})} \cdot f_1(\mathsf{qu}_1 + \mathbf{w}) \right)$$
$$= \frac{1}{2} \sum_{\mathbf{w} \in \mathcal{W}} c_\mathbf{w} \cdot \sum_{\mathbf{v} \in \mathbb{F}_2^d} \mathsf{DB}_\mathbf{v} \cdot (-1)^{Q(\mathbf{r}+\mathbf{w}+\mathbf{i})+Q(\mathbf{r}+\mathbf{w}+\mathbf{v})} \cdot \left(1 + (-1)^{Q(\mathbf{v}+\mathbf{i})}\right)$$

$$= \frac{1}{2} \sum_{\mathbf{v} \in \mathbb{F}_2^d} \mathsf{DB}_{\mathbf{v}} \cdot \Big(\sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w}} \cdot (-1)^{Q(\mathbf{w}+\mathbf{v}+\mathbf{i})+Q(\mathbf{w})} \Big) \cdot (-1)^{Q(\mathbf{r}+\mathbf{v})+Q(\mathbf{r}+\mathbf{i})} \cdot \Big(1 + (-1)^{Q(\mathbf{v}+\mathbf{i})}\Big).$$

In this last equation, if $Q(\mathbf{v}+\mathbf{i}) \neq 0$, then the term $\left(1 + (-1)^{Q(\mathbf{v}+\mathbf{i})}\right)$ is zero and the corresponding part of the sum cancels. On the other hand, if $Q(\mathbf{v}+\mathbf{i}) = 0$, then the relation on the public recovery coefficients in Equation (15) implies that the term $\left(\sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w}} \cdot (-1)^{Q(\mathbf{w}+\mathbf{v}+\mathbf{i})+Q(\mathbf{w})}\right)$ is 1 if $\mathbf{v} = \mathbf{i}$ and 0 otherwise. As a result, the whole equation simplifies out to

$$\mathsf{record} = \frac{1}{2} \cdot \mathsf{DB}_{\mathbf{i}} \cdot \left(1 + (-1)^{Q(\mathbf{0})}\right) = \mathsf{DB}_{\mathbf{i}}.$$

That is, the user correctly recovers its desired database element.

Security. PIR security is immediate: every query sent to each server is distributed uniformly in $\mathbb{F}_2^d$.

Efficiency. Each preprocessed truth table consists of 2^d entries in $\mathbb{F}_p$, exactly matching the size of the original database. On each PIR query, the user sends each server d bits. Each of the servers reads and returns $|\mathcal{W}|$ elements in $\mathbb{F}_p$ to the user, using $O(d)$ operations per element to compute the point $\mathsf{qu}_t + \mathbf{w}$. Since $\ell \leq (d+1)/4$, the binomial bound below, together with the fact that H_2 is increasing on $[0, \frac{1}{2}]$, gives

$$|\mathcal{W}| = \sum_{j=0}^{\ell} \binom{d+1}{j} \leq 2^{H_2(1/4)(d+1)}.$$

To construct its truth table, each server performs at most 2^d sign multiplications followed by one unnormalized Walsh–Hadamard transform, taking $O(d2^d)$ operations in $\mathbb{F}_p$ in total. The signs are determined by $Q(\mathbf{v}) = |\mathbf{v}| \cdot (|\mathbf{v}|+1)/2 \pmod 2$. The public coefficients can be generated in the same asymptotic time. □

7 Doubly-Efficient PIR with $s \geq 3$ Servers

Next, we construct doubly-efficient PIR with $n+1$ bits of storage for any number of servers $s \geq 3$. As in the technical overview (Section 2.2) and in Section 6, each server stores a Fourier transform of the database, whose entries are weighted linear combinations of the database records. Moreover, much as in Section 6, our specific construction works over databases consisting of q^d records in $\mathbb{F}_p$, for parameters q, d, and p that we specify. The full PIR construction appears in Fig. 5; it makes use of a set of public coefficients that correspond to a quadratic-phase delta expansion, as in Definition 4.3.

Applying the standard transformation [DPT10] (given in Section 9.2) then yields a s-server doubly-efficient PIR scheme that works over arbitrary n-bits databases, with $n+1$ bits of storage per server. This PIR construction is key to the results in Main Theorems 5.2, 5.3 and 5.5, which differ only in the quadratic-phase delta expansion that the protocol uses.

Overview. The main difference from the prime-server PIR scheme (given in the technical overview) is how recovery works. In the prime-server scheme with q servers, cancellation happens in two stages: first, summing across all servers' answers removes any database values labeled with a vector $\mathbf{v} \in \mathbb{F}_q^d$ such that $Q(\mathbf{v}-\mathbf{i}) \neq 0$, where $\mathbf{i}$ is the address of the requested record and Q is a quadratic form (in Section 2.2, we take $Q(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle$). Then, applying the coefficients of a Fourier-sparse, cone-delta function (as per Definition 2.1) removes any database values labeled with a vector $\mathbf{v} \in \mathbb{F}_q^d$ such that $Q(\mathbf{v}-\mathbf{i}) = 0$ *except* the $\mathbf{i}^{\text{th}}$ database record. Via this two-step process, the user recovers exactly the desired database record.

This approach, which sums across all q servers, appears to break when fewer than q servers are available. However, in this section, we show that such a PIR scheme can work using $s < q$ servers by choosing

Doubly-efficient PIR with s servers. Parameterized by a prime $q \geq \max\{3, s\}$, an integer d, and a prime p such that the field $\mathbb{F}_p$ contains an element γ of order q. The PIR scheme operates over databases consisting of q^d records in $\mathbb{F}_p$. Define the sets $\mathcal{W}$ and $\mathcal{T}$ (with $|\mathcal{T}| = s$) and the coefficients $\left\{ c_{\mathbf{w},t} \in \mathbb{F}_p \,:\, \mathbf{w} \in \mathcal{W}, t \in \mathcal{T} \right\}$ as in Section 7. They satisfy, for any $\mathbf{x} \in \mathbb{F}_q^d$,

$$\sum_{t \in \mathcal{T}} \sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w},t} \cdot \gamma^{\langle \mathbf{w}, \mathbf{x} \rangle + t \cdot \langle \mathbf{x}, \mathbf{x} \rangle} = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{0} \\ 0, & \text{else.} \end{cases}$$

$\mathsf{Preprocess}(\mathsf{DB} \in \mathbb{F}_p^{q^d}) \to (\mathsf{Table}_t)_{t \in \mathcal{T}}$

- Split DB into q^d records $\{\mathsf{DB}_{\mathbf{v}}\}_{\mathbf{v} \in \mathbb{F}_q^d}$, each in $\mathbb{F}_p$.
- For $t \in \mathcal{T}$, let the function $f_t : \mathbb{F}_q^d \to \mathbb{F}_p$ be
$$f_t(\mathbf{x}) = \sum_{\mathbf{v} \in \mathbb{F}_q^d} \mathsf{DB}_{\mathbf{v}} \cdot \gamma^{\langle \mathbf{x}, \mathbf{v} \rangle + t \cdot \langle \mathbf{v}, \mathbf{v} \rangle}.$$
- Set Table_t to be the truth table of the function f_t.

$\mathsf{Query}(i \in [q^d]) \to \mathsf{st}, (\mathsf{qu}_t)_{t \in \mathcal{T}}$

- Let $\mathbf{i} \in \mathbb{F}_q^d$ be the vector corresponding to index i.
- Sample $\mathbf{r} \xleftarrow{R} \mathbb{F}_q^d$ and set $\mathsf{qu}_t = \mathbf{r} - 2t \cdot \mathbf{i}$ for $t \in \mathcal{T}$.
- Output $\mathsf{st} = (\mathbf{i}, \mathbf{r})$ and $(\mathsf{qu}_t)_{t \in \mathcal{T}}$.

$\mathsf{Answer}^{\mathsf{Table}_t}(\mathsf{qu}_t \in \mathbb{F}_q^d) \to \mathsf{ans}_t \in \mathbb{F}_p^{|\mathcal{W}|}$

- Output $\mathsf{ans}_t = \left\{ \mathsf{Table}_t[\mathsf{qu}_t + \mathbf{w}] \right\}_{\mathbf{w} \in \mathcal{W}}$.

$\mathsf{Reconstruct}(\mathsf{st}, (\mathsf{ans}_t)_{t \in \mathcal{T}}) \to \mathbb{F}_p$

- Parse $(\mathbf{i}, \mathbf{r})$ from st and $\mathsf{Table}_t[\mathsf{qu}_t + \mathbf{w}]$ from each ans_t.
- Output the $\mathbb{F}_p$-element computed as
$$\sum_{t \in \mathcal{T}} \sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w},t} \cdot \mathsf{Table}_t[\mathsf{qu}_t + \mathbf{w}] \cdot \gamma^{\langle t \cdot \mathbf{i} - \mathbf{r} - \mathbf{w}, \mathbf{i} \rangle}.$$

Figure 5: Doubly-efficient PIR with s servers, over databases consisting of q^d records in $\mathbb{F}_p$.

server-dependent recovery coefficients such that both types of cancellation happen together, in the same step. This technique is inspired by sparse decoding polynomials, which reduce the number of servers in classic (not doubly-efficient) matching-vector PIR [Efr09, GKS25, ABL25, BL26, GR27]. The resulting PIR proves that:

Theorem 7.1. *For any number of servers $s \geq 2$ and prime $q \geq \max\{3, s\}$, if there exists a (k, s)-sparse quadratic-phase delta expansion of order q, dimension d, and over a field $\mathbb{F}_p$ (Definition 4.3), then there is an s-server PIR scheme that operates over any database consisting of q^d entries in $\mathbb{F}_p$, with information-theoretic security and the following efficiency:*

- *each server stores a server-dependent data structure that consists of q^d entries in $\mathbb{F}_p$, computed from the database in $O(dq^d)$ operations over $\mathbb{F}_p$ (and may then discard the original database),*
- *to make a PIR query, the user uploads $\lceil d \log_2 q \rceil$ bits to each PIR server, and*
- *to answer a PIR query, each server reads and sends back at most k $\mathbb{F}_p$-elements in its data structure, using $O(d)$ operations per element.*

Remark 7.2. *A (k, s)-sparse quadratic-phase delta expansion over $\mathbb{F}_p$ on $\mathbb{Z}_q^d$ also gives a (k, s)-sparse quadratic-phase delta expansion over $\mathbb{F}_p$ on $\mathbb{Z}_q^{d'}$ for any $d' < d$, by a similar padding argument to Corollary 12.3. So Theorem 7.1 then gives PIR schemes as above over databases of $q^{d'}$ records in $\mathbb{F}_p$, for every $d' \leq d$.*

7.1 Protocol

The scheme is parameterized by a prime $q \geq \max\{3, s\}$, an integer $d \geq 1$, and a prime p such that

1. $\mathbb{F}_p$ contains an element γ of order q and
2. there exists a (k, s)-sparse quadratic-phase delta expansion of order q, of dimension d, and over field $\mathbb{F}_p$, as per Definition 4.3.

The support of this expansion is contained in $\mathcal{W} \times \mathcal{T}$ for some sets $\mathcal{W} \subseteq \mathbb{F}_q^d$ and $\mathcal{T} \subseteq \mathbb{Z}_q$ with $|\mathcal{W}| \leq k$ and $|\mathcal{T}| \leq s$. Since $s \leq q$, we may enlarge $\mathcal{T}$ so that $|\mathcal{T}| = s$, and we index the s servers by the elements $t \in \mathcal{T}$. That is, we have public coefficients $\{c_{\mathbf{w},t} \in \mathbb{F}_p \ : \ \mathbf{w} \in \mathcal{W},\ t \in \mathcal{T}\}$ such that, for every $\mathbf{x} \in \mathbb{F}_q^d$,

$$\sum_{t \in \mathcal{T}} \sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w},t} \cdot \gamma^{\langle \mathbf{w},\mathbf{x} \rangle + t \cdot \langle \mathbf{x},\mathbf{x} \rangle} = \begin{cases} 1, \text{ if } \mathbf{x} = \mathbf{0} \\ 0, \text{ else.} \end{cases} \tag{16}$$

Then, given any database that consists of q^d elements in $\mathbb{F}_p$, the PIR scheme works as follows:

Preprocessing. The servers split the database into q^d records in $\mathbb{F}_p$, each indexed by the vectors in $\mathbb{F}_q^d$. Each server $t \in \mathcal{T}$ stores the truth-table of the following function $f_t \ : \ \mathbb{F}_q^d \to \mathbb{F}_p$:

$$f_t(\mathbf{x}) = \sum_{\mathbf{v} \in \mathbb{F}_q^d} \mathsf{DB}_{\mathbf{v}} \cdot \gamma^{\langle \mathbf{x},\mathbf{v} \rangle + t \cdot \langle \mathbf{v},\mathbf{v} \rangle}.$$

Mirroring the prime-server construction from Section 2.2, arithmetic in the exponent happens modulo q; arithmetic in the base happens modulo p.

Querying. The user maps its index i into a query point $\mathbf{i} \in \mathbb{F}_q^d$ such that the record $\mathsf{DB}_{\mathbf{i}}$ contains the database value he aims to read. Then, the user samples a random mask $\mathbf{r} \xleftarrow{\text{R}} \mathbb{F}_q^d$ and sends server $t \in \mathcal{T}$ the query vector

$$\mathsf{qu}_t = \mathbf{r} - 2t \cdot \mathbf{i} \in \mathbb{F}_q^d.$$

Answering. Given query point qu_t, server t responds with all evaluations $f_t(\mathsf{qu}_t + \mathbf{w})$ for $\mathbf{w} \in \mathcal{W}$, by looking up each value in its precomputed truth table.

Reconstructing. The user combines the servers' answers, weighting them by the public coefficients of the quadratic-phase delta expansion, to recover and output the following $\mathbb{F}_p$-element:

$$\mathsf{record} = \sum_{t \in \mathcal{T}} \sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w},t} \cdot f_t(\mathsf{qu}_t + \mathbf{w}) \cdot \gamma^{\langle t \cdot \mathbf{i} - \mathbf{r} - \mathbf{w},\mathbf{i} \rangle}. \tag{17}$$

7.2 Analysis

We prove each of the scheme's requirements in sequence.

Correctness. The above reconstruction routine exactly outputs the $\mathbf{i}^{\text{th}}$ record in the database since, by expanding expressions and applying the identity $\langle \mathbf{v} - \mathbf{i}, \mathbf{v} - \mathbf{i} \rangle = \langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{i}, \mathbf{i} \rangle - 2\langle \mathbf{i}, \mathbf{v} \rangle$, we see that

$$\begin{aligned} \mathsf{record} &= \sum_{t \in \mathcal{T}} \sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w},t} \cdot \left(\sum_{\mathbf{v} \in \mathbb{F}_q^d} \mathsf{DB}_{\mathbf{v}} \cdot \gamma^{\langle \mathbf{r} - 2t \cdot \mathbf{i} + \mathbf{w},\mathbf{v} \rangle + t \cdot \langle \mathbf{v},\mathbf{v} \rangle} \right) \cdot \gamma^{\langle t \cdot \mathbf{i} - \mathbf{r} - \mathbf{w},\mathbf{i} \rangle} \\ &= \sum_{\mathbf{v} \in \mathbb{F}_q^d} \mathsf{DB}_{\mathbf{v}} \cdot \gamma^{\langle \mathbf{r},\mathbf{v} - \mathbf{i} \rangle} \left(\sum_{t \in \mathcal{T}} \sum_{\mathbf{w} \in \mathcal{W}} c_{\mathbf{w},t} \cdot \gamma^{\langle \mathbf{w},\mathbf{v} - \mathbf{i} \rangle + t \langle \mathbf{v} - \mathbf{i},\mathbf{v} - \mathbf{i} \rangle} \right) = \mathsf{DB}_{\mathbf{i}}. \end{aligned}$$

The final equality here follows directly from property (16); this proves the scheme's correctness.

Security. Security is immediate: each query is masked by the uniformly random vector $\mathbf{r} \in \mathbb{F}_q^d$.

Efficiency. The scheme's efficiency properties follow by inspection. Each server stores q^d entries in $\mathbb{F}_p$ in its preprocessed table. The user sends each server $\lceil d \log_2 q \rceil$ bits. Each server reads and returns $|\mathcal{W}| \le k$ elements in $\mathbb{F}_p$, using $O(d)$ operations per element to compute the point $\mathsf{qu}_t + \mathbf{w}$.

Each preprocessing table is an entry-wise multiplication by $\gamma^{t \cdot \langle \mathbf{v}, \mathbf{v} \rangle}$, followed by a d-dimensional Fourier transform. Applying the transform one coordinate at a time takes $O(dq^d)$ operations, as q is fixed. □

8 Elementary Results on Sparse Cone-Delta Functions

8.1 Simple Constructions for $q = s = 3$ from Hamming Slices

For $q = 3$, we can construct sparse cone-delta functions without using modular forms (whereas Part II will use modular forms). We improve on an elementary Hamming-ball construction of Alon and Lubetzky [AL07], by using Hamming shells from both ends of the weight range, obtaining Fourier sparsity $3^{(0.7444\ldots + o(1))d}$ (Lemma 8.3 and Corollary 8.4); it yields a 3-server doubly-efficient PIR scheme with server work $n^{0.7445}$ (Main Theorem 5.2). (For comparison, the more sophisticated construction in Section 12 gives the exponent 0.7327 for $q = 3$ shown in Table 1; by a lower bound given next, no cone-delta function achieves an exponent below $1/2$.)

Setup. Throughout this subsection, $\mathbb{F}$ is a field containing a primitive cube root of unity γ, and $\mathrm{wt}(\mathbf{x})$ denotes the number of nonzero coordinates of $\mathbf{x} \in \mathbb{Z}_3^d$. Every nonzero element of $\mathbb{Z}_3$ squares to 1, so $\langle \mathbf{x}, \mathbf{x} \rangle \equiv \mathrm{wt}(\mathbf{x}) \pmod 3$, and the cone $\{\mathbf{x} : \langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod 3\}$ consists exactly of the points of weight $3r$ for $r = 0, 1, \ldots, \lfloor d/3 \rfloor$. We look for cone-delta functions of the form $f(\mathbf{x}) = \sum_{\mathbf{a} \in \mathbb{Z}_3^d} c_{\mathrm{wt}(\mathbf{a})} \gamma^{\langle \mathbf{a}, \mathbf{x} \rangle}$, whose Fourier coefficients depend only on the weight of the frequency vector; such an f is a combination of the *shell sums* $\sum_{\mathrm{wt}(\mathbf{a})=j} \gamma^{\langle \mathbf{a}, \mathbf{x} \rangle}$, and its Fourier sparsity is at most $\sum_{j : c_j \neq 0} \binom{d}{j} 2^j$.

Lemma 8.1 (Shell sums and ternary Krawtchouk polynomials [MS77, Chapter 5])**.** *For $0 \le j \le d$ and $\mathbf{x} \in \mathbb{Z}_3^d$ of weight w, the sum $\sum_{\mathrm{wt}(\mathbf{a})=j} \gamma^{\langle \mathbf{a}, \mathbf{x} \rangle}$ depends only on $w = \mathrm{wt}(\mathbf{x})$. Hence we abuse notation and denote this sum by $K_j(w)$. Then, we have:*

$$\sum_{j=0}^{d} K_j(w)\, z^j = (1+2z)^{d-w}(1-z)^w,$$

$$K_j(w) = \sum_{n=0}^{j} (-3)^n\, 2^{j-n} \binom{d-n}{j-n} \binom{w}{n} \qquad \text{for all } j \in [0, d],$$

$$K_{d-i}(w) = 2^{d-w}(-1)^w \sum_{n=0}^{i} \left(-\frac{3}{2}\right)^n 2^{-(i-n)} \binom{d-n}{i-n} \binom{w}{n} \qquad \text{for all } i \in [0, d].$$

The second identity above holds over $\mathbb{Z}$ and the third identity holds over $\mathbb{Z}[1/2]$.

So f is a cone-delta function exactly when:

$$\sum_j c_j K_j(3r) = \mathbb{1}[r = 0], \quad \text{for } r = 0, 1, \ldots, \lfloor d/3 \rfloor. \tag{18}$$

To understand this condition, associate to each $j \in [0, d]$ the vector

$$\mathbf{v}_j := (K_j(0), K_j(3), \ldots, K_j(3\lfloor d/3 \rfloor)). \tag{19}$$

Then (18) says that there is a cone-delta function supported on the Hamming shells from a set $J \subseteq \{0, 1, \dots, d\}$ if and only if the vectors $\mathbf{v}_j$ for $j \in J$ span $(1, 0, \dots, 0)$. For each Hamming shell j, we pay a cost of $\binom{d}{j}2^j$ in Fourier sparsity, and we would like to select a collection of shells J that has total cost as low as possible, while satisfying the spanning condition. We plot the cost of shell j as a function of j in Figure 6: the cost increases to a peak at $j = 2d/3$ and then decreases again.

Bearing this intuition in mind, we now restate the existing construction of cone-delta functions implicit in prior work by Alon and Lubetzky [AL07]. It proceeds by simply taking $J = \{0, 1, \dots, \lfloor d/3 \rfloor\}$. We omit the proof of [AL07] and also note that the below is a special case of Lemma 8.3 which we will prove.

Lemma 8.2 (Hamming ball, adapted from [AL07, proof of Theorem 2.2])**.** *Suppose that $\mathbb{F}$ has characteristic 0 or greater than $\max(\lfloor d/3 \rfloor, 2)$. Then there is a cone-delta function on $\mathbb{Z}_3^d$ over $\mathbb{F}$ whose Fourier support is contained in the Hamming ball $\{\mathbf{a} : \mathrm{wt}(\mathbf{a}) \le \lfloor d/3 \rfloor\}$. Its Fourier sparsity is at most $\sum_{j \le \lfloor d/3 \rfloor} \binom{d}{j} 2^j = 3^{(\eta_{\mathrm{ball}} + o(1))d}$, where $\eta_{\mathrm{ball}} := 1 - \frac{1}{3}\log_3 2 = 0.7896\ldots$.*

However, the picture in Figure 6 points to a way to improve on this construction. Intuitively, we want to span a vector in dimension $\lfloor d/3 \rfloor + 1$, so hopefully it suffices to greedily take the $\lfloor d/3 \rfloor + 1$ lowest-cost shells i.e., some shells close to the $j = 0$ end and some shells close to the $j = d$ end. Fortunately for us, these shells actually do have the desired spanning, and this gives us the improved $0.744\ldots$ exponent. We formally check this next:

Lemma 8.3 (Both ends)**.** *Let $\ell, \ell' \ge 0$ be integers with $\ell + \ell' = \lfloor d/3 \rfloor$, and suppose that $\mathbb{F}$ has characteristic 0 or a prime characteristic $p > \max\{\lfloor d/3 \rfloor, 2\}$. Then the vectors $\mathbf{v}_j$ (defined in (19)) for $j \in J := \{0, 1, \dots, \ell\} \cup \{d - \ell' + 1, \dots, d\}$ are linearly independent. In particular, they span $(1, 0, \dots, 0)$.*

Consequently, there is a cone-delta function on $\mathbb{Z}_3^d$ over $\mathbb{F}$ whose Fourier coefficients depend only on the weight of the frequency vector, and whose Fourier support is contained in $\{\mathbf{a} : \mathrm{wt}(\mathbf{a}) \le \ell\} \cup \{\mathbf{a} : \mathrm{wt}(\mathbf{a}) \ge d - \ell' + 1\}$.

Proof. For $0 \le n \le \lfloor d/3 \rfloor$, let $B_n(r) := \binom{3r}{n} = \frac{1}{n!}\prod_{i=0}^{n-1}(3r - i)$, a polynomial in r over $\mathbb{F}$ of degree exactly n (its leading coefficient $3^n/n!$ is invertible, as the characteristic of $\mathbb{F}$ is 0 or exceeds $\lfloor d/3 \rfloor \ge n$ and is not 3). Define $\lambda := -\frac{1}{8} \in \mathbb{F}$ (which is defined since the characteristic is not 2). By Lemma 8.1 with $w = 3r$, so that $2^{d-w}(-1)^w = 2^d\lambda^r$:

- for $j \le \ell$, the function $K_j(3r)$ is a combination of $B_0, \dots, B_j$ with coefficient $(-3)^j \ne 0$ on B_j; and
- for $i < \ell'$, the function $K_{d-i}(3r)$ is $2^d\lambda^r$ times a combination of $B_0, \dots, B_i$ with coefficient $(-\frac{3}{2})^i \ne 0$ on B_i.

To parse the above, define:

$$\begin{aligned}
\mathbf{b}_j &:= (B_j(0), \dots, B_j(\lfloor d/3 \rfloor)), \quad j \in \{0, 1, \dots, \ell\};\\
\mathbf{b}'_j &:= (0^j, \dots, (\lfloor d/3 \rfloor)^j), \quad j \in \{0, 1, \dots, \ell\};\\
\mathbf{b}_{d-i} &:= (\lambda^0 B_i(0), \lambda^1 B_i(1), \dots, \lambda^{\lfloor d/3 \rfloor} B_i(\lfloor d/3 \rfloor)), \quad i \in \{0, 1, \dots, \ell' - 1\}\\
\mathbf{b}'_{d-i} &:= (\lambda^0 0^i, \lambda^1 1^i, \dots, \lambda^{\lfloor d/3 \rfloor}(\lfloor d/3 \rfloor)^i), \quad i \in \{0, 1, \dots, \ell' - 1\}.
\end{aligned}$$

Then it follows that there exists an upper triangular matrix $\mathbf{M}$ with nonzero diagonal such that:

$$\begin{bmatrix}\mathbf{v}_0 & \dots & \mathbf{v}_\ell & \mathbf{v}_d & \dots & \mathbf{v}_{d-\ell'+1}\end{bmatrix} = \begin{bmatrix}\mathbf{b}_0 & \dots & \mathbf{b}_\ell & \mathbf{b}_d & \dots & \mathbf{b}_{d-\ell'+1}\end{bmatrix}\mathbf{M}.$$

Since $\mathbf{M}$ is invertible, to show that the $\mathbf{v}_j$'s are linearly independent it suffices to show that the $\mathbf{b}_j$'s are linearly independent. A similar argument, now applied to the $\mathbf{b}_j$'s, reduces this to showing that the $\mathbf{b}'_j$'s are

linearly independent (in words, this replaces each evaluation $B_j(r)$ with the leading-order monomial r^j). In other words, we want to show that the matrix:

$$V := \begin{bmatrix} \mathbf{b}'_0 & \dots & \mathbf{b}'_\ell & \mathbf{b}'_d & \dots & \mathbf{b}'_{d-\ell'+1} \end{bmatrix} = \begin{bmatrix} 1 & \dots & 0 & \lambda^0 & \dots & 0 \\ 1 & \dots & 1 & \lambda^1 & \dots & \lambda^1 1^{\ell'-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & (\lfloor d/3 \rfloor)^\ell & \lambda^{\lfloor d/3 \rfloor} & \dots & \lambda^{\lfloor d/3 \rfloor}(\lfloor d/3 \rfloor)^{\ell'-1} \end{bmatrix},$$

is invertible. This is a *confluent Vandermonde matrix,* and [Kra99, Theorem 21, equation (3.2)] tells us that its determinant is

$$\det V = \lambda^{\binom{\ell'}{2}}(\lambda-1)^{(\ell+1)\ell'} \left(\prod_{j=0}^{\ell} j!\right)\left(\prod_{j=0}^{\ell'-1} j!\right),$$

which is nonzero under our characteristic assumptions: $\lambda = -1/8$ and $\lambda - 1 = -9/8$ are nonzero, and all factorials have indices at most $\lfloor d/3 \rfloor$. The desired invertibility follows, completing our proof. □

Corollary 8.4. *Let $H_2(t) := -t\log_2 t-(1-t)\log_2(1-t)$ be the binary entropy function. For every $\alpha \in [0, \frac{1}{3}]$, taking $\ell = \lfloor \alpha d \rfloor$ in Lemma 8.3 gives Fourier sparsity $3^{(\eta(\alpha)+o(1))d}$, where*

$$\eta(\alpha) := \frac{1}{\log_2 3} \max\left\{ H_2(\alpha) + \alpha,\ H_2\left(\tfrac{1}{3} - \alpha\right) + \tfrac{2}{3} + \alpha \right\}.$$

The two terms are equal at $\alpha^* = 0.2994\ldots$, and $\eta_{\text{ends}} := \eta(\alpha^*) = 0.7444\ldots$; in particular, $\eta(0.2994) < 0.7445$.

Proof. The sparsity is at most $\sum_{j\le\ell}\binom{d}{j}2^j + \sum_{i<\ell'}\binom{d}{i}2^{d-i}$, where $\ell' = N - \ell$. In the first sum, the ratio of consecutive terms is $\frac{2(d-j)}{j+1} \ge 1$ for $j \le \frac{2d-1}{3}$, so its terms increase and it is at most $(d+1)\binom{d}{\ell}2^\ell$, as $\ell \le \frac{d}{3}$. In the second sum, the ratio of consecutive terms is $\frac{d-i}{2(i+1)} \ge 1$ for $i \le \frac{d-2}{3}$, so it is at most $(d+1)\binom{d}{\ell'}2^{d-\ell'}$, as $\ell' - 1 \le \frac{d}{3} - 1$. Finally, $\binom{d}{k} \le 2^{dH_2(k/d)}$ for $0 \le k \le d$ (as $\binom{d}{k}t^k(1-t)^{d-k} \le 1$ for $t = k/d$). With $\ell/d \to \alpha$ and $\ell'/d \to \frac{1}{3} - \alpha$, and H_2 continuous, the two sums are $2^{(H_2(\alpha)+\alpha+o(1))d}$ and $2^{(H_2(1/3-\alpha)+2/3+\alpha+o(1))d}$. The numerical values are a direct computation; see Figure 6. □

Remark 8.5 (Relation to Part II). *The construction of Lemma 8.2 has two features. First, the Fourier support consists of the frequency vectors* $\mathbf{a}$ *of small total cost* $\sum_i w(a_i)$*, where the cost* w *of a residue in* $\mathbb{Z}_3$ *is* 0 *for the zero residue and* 1 *otherwise. Second, the vanishing on the cone comes from interpolation in the single variable* $\mathrm{wt}(\mathbf{x})$*, and the degree it needs is dictated by root counting: a univariate polynomial that equals* 1 *at* 0 *and vanishes at* N *prescribed points must have degree at least* N*, and this degree suffices. The second feature does not extend to* $q \ge 5$*, because* $\langle \mathbf{x}, \mathbf{x} \rangle \bmod q$ *is then no longer a function of* $\mathrm{wt}(\mathbf{x})$*: for* $q = 5$*, the points* $(1, 0, \dots, 0)$ *and* $(2, 0, \dots, 0)$ *both have weight* 1*, but* $\langle \mathbf{x}, \mathbf{x} \rangle$ *equals* 1 *and* 4 *respectively. The constructions of Part II keep the first feature, with the support again consisting of the frequency vectors of small total cost, for a cost function on* $\mathbb{Z}_q$ *that we construct from a lattice (Definition 11.5), and they replace interpolation in* $\mathrm{wt}(\mathbf{x})$ *by a root-counting argument for modular forms. At* $q = 3$*, the resulting cost is* 1 *on both nonzero residues, and the support of the warm-up construction (Section 11) is exactly the Hamming ball of Lemma 8.2.*

The construction of Lemma 8.3 gives up the first feature, as the shells of weight close to d *have high cost. The support of the stronger construction of Section 12 also reaches both ends of the weight range: it contains the shifts* $\mathbf{z} + b\mathbf{1}$ *of low-cost vectors* $\mathbf{z}$ *(Definition 12.1), and for* $q = 3$ *and* $b \ne 0$ *these have weight at least* $d - \mathrm{wt}(\mathbf{z})$*.*

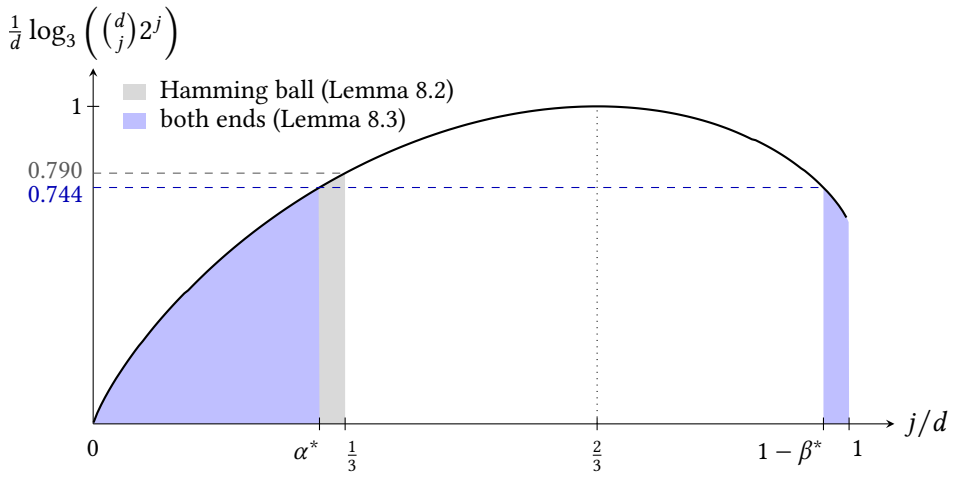


Figure 6: Sizes of the Hamming shells of $\mathbb{Z}_3^d$. The shell of weight $j = td$ has $\binom{d}{j}2^j = 3^{(g(t)+o(1))d}$ points, where $g(t) := (H_2(t) + t)/\log_2 3$; the profile g peaks at $g(\frac{2}{3}) = 1$. Both constructions use $\lfloor d/3 \rfloor + 1$ shells, and their Fourier sparsity is governed by the largest shell they use. The Hamming ball of Lemma 8.2 uses the weights up to $\frac{d}{3}$ (gray), where the shells reach $3^{0.790d}$. The construction of Lemma 8.3 trades the gray shells between $\alpha^* d$ and $\frac{d}{3}$ for the top $\beta^* d$ shells (blue), where the shells are small again; balancing the two ends ($\alpha^* \approx 0.2994$, $\beta^* = \frac{1}{3} - \alpha^* \approx 0.0339$) lowers the largest shell to $3^{0.744d}$ (Corollary 8.4).

8.2 Lower Bound

Every exponent in Table 1 is above $1/2$. We now show that this is unavoidable for our approach: a cone-delta function over any field must have Fourier sparsity at least $q^{d/2-O(1)}$, and by Lemma 4.4, the same bound applies to every quadratic-phase delta expansion. So our doubly-efficient PIR constructions from these objects cannot achieve a server work exponent below $1/2$. The main idea is that a cone-delta function must be a delta function on a large subspace, which forces it to have large Fourier support.

Theorem 8.6. *Let q be an odd prime, and let $\mathbb{F}$ be any field with a primitive q^{th} root of unity γ. Assume there exists a cone delta function with Fourier sparsity k. Then, we have $k \geq q^{2\lfloor d/4 \rfloor}$.*

To set up the proof of this theorem, we use the following intermediate lemmas:

Proposition 8.7. *There exist residues $a, b \in \mathbb{Z}_q$ such that $a^2 + b^2 \equiv -1 \pmod q$.*

Proof. The sets $\{a^2 : a \in \mathbb{Z}_q\}$ and $\{-1 - b^2 : b \in \mathbb{Z}_q\}$ both have size $(q+1)/2$, so they must intersect. The conclusion follows. □

Lemma 8.8. *There exists a subspace V of $\mathbb{Z}_q^4$ of dimension 2 such that $\langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod q$ for all $\mathbf{x} \in V$.*

Proof. Take the a, b given by Proposition 8.7 and consider the subspace of $\mathbb{Z}_q^4$ spanned by $(a, b, 1, 0)$ and $(b, -a, 0, 1)$. For any $x, y \in \mathbb{Z}_q$, we have $x(a, b, 1, 0) + y(b, -a, 0, 1) = (ax + by, bx - ay, x, y)$. Finally, we have $(ax + by)^2 + (bx - ay)^2 + x^2 + y^2 = (a^2 + b^2 + 1)(x^2 + y^2) \equiv 0 \pmod q$. □

Corollary 8.9. *There exists a subspace V of $\mathbb{Z}_q^d$ of dimension $2\lfloor d/4 \rfloor$ such that $\langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod q$ for all $\mathbf{x} \in V$.*

Proof. For the V_0 given in Lemma 8.8, we will take the subspace $V := \bigoplus_{i\in[\lfloor d/4\rfloor]} V_0$, which clearly has dimension $2\lfloor d/4\rfloor$. In other words, we split the d coordinates into $\lfloor d/4\rfloor$ blocks of 4 coordinates each and let V be the set of vectors in $\mathbb{Z}_q^d$ such that the restriction to any of these blocks yields an element of V_0, and any coordinate outside these blocks is zero. The conclusion follows. □

Lemma 8.10. *Suppose $f : \mathbb{Z}_q^d \to \mathbb{F}$ is a delta function on V for a subspace $V \subseteq \mathbb{Z}_q^d$. Then $\hat{f}$ is nonzero in at least $q^{\dim V}$ positions.*

Proof. This is the multiplication argument of [GK26, Lemma 3.7 and Theorem 3.9], applied with the indicator of V. Define $g(\mathbf{x}) := \left(\frac{1}{q^{d-\dim V}} \sum_{\mathbf{y}\in V^\perp} \gamma^{\langle \mathbf{x},\mathbf{y}\rangle}\right) \cdot f(\mathbf{x})$. The first term is 1 if $\mathbf{x} \in V$ and 0 otherwise, so g is the delta function on all of $\mathbb{Z}_q^d$. Consequently, it must have full Fourier support q^d. On the other hand, by definition its Fourier support has size at most $|V^\perp| \cdot |\mathsf{supp}\hat{f}| = q^{d-\dim V} \cdot |\mathsf{supp}\hat{f}|$. Putting these together forces $\hat{f}$ to have support size $\geq q^{\dim V} = |V|$ as claimed. □

Proof of Theorem 8.6. If f is a delta function on the cone, it must in particular be a delta function on the subspace V given by Corollary 8.9 of size $q^{2\lfloor d/4\rfloor}$. Then Lemma 8.10 implies that $\hat{f}$ must have support size at least $q^{2\lfloor d/4\rfloor}$, as desired. □

9 From Expansions to n-Bit PIR

The constructions of Part II give quadratic-phase delta expansions over the field $\mathbb{Q}(\zeta_q)$, while the PIR schemes of Sections 6 and 7 operate over databases of records in a finite field $\mathbb{F}_p$. In this section, we first transfer expansions from $\mathbb{Q}(\zeta_q)$ to $\mathbb{F}_p$, for every sufficiently large prime $p \equiv 1 \pmod q$ (Section 9.1). We then convert PIR schemes over databases of $\mathbb{F}_p$-records into PIR schemes over databases of n bits, with $n+1$ bits of storage per server (Section 9.2). The proofs of the main theorems (Section 5) combine these two steps.

9.1 From Characteristic 0 to Finite Fields

In the following, we provide some building blocks to transition from our cone-delta results over $\mathbb{Q}(\zeta_q)$ to usable cone deltas over finite fields that can be explicitly computed; these are necessary for the PIR protocols of Section 7. To do this, the main observation is that the Fourier support sets used in Theorems 11.9 and 12.2 are symmetric by definition; they are invariant under permutations of the d coordinates. This allows us to impose the same symmetry on the coefficients:

Lemma 9.1. *Assume that there exists a quadratic-phase delta expansion on $\mathbb{Z}_q^d$ for $\mathbb{F} = \mathbb{Q}(\zeta_q)$ with support $\mathcal{W} \times \mathcal{T}$, such that the set $\mathcal{W}$ is invariant under permutations of the d coordinates. Then there also exists a quadratic-phase delta expansion supported on the same $\mathcal{W} \times \mathcal{T}$ with the same symmetry invariance i.e., $c'_{\mathbf{a},t} = c'_{\mathbf{a}',t}$ if $\mathbf{a}, \mathbf{a}'$ are permutations of each other. Going forward, we refer to such a quadratic-phase delta expansion as* symmetric.

Proof. By averaging: define $c'_{\mathbf{a},t} := \frac{1}{d!} \sum_{\pi \text{ permutation on } [d]} c_{\pi(\mathbf{a}),t}$. This automatically has the stated symmetry. To show that this works, we have:

$$\sum_{t\in\mathbb{Z}_q} \sum_{\mathbf{a}\in\mathbb{Z}_q^d} c'_{\mathbf{a},t} \zeta_q^{\langle \mathbf{a},\mathbf{x}\rangle + t\langle \mathbf{x},\mathbf{x}\rangle} = \frac{1}{d!} \sum_{t\in\mathbb{Z}_q} \sum_{\pi} \sum_{\mathbf{a}\in\mathbb{Z}_q^d} c_{\pi(\mathbf{a}),t} \zeta_q^{\langle \mathbf{a},\mathbf{x}\rangle + t\langle \mathbf{x},\mathbf{x}\rangle}$$

$$
\begin{aligned}
&= \frac{1}{d!} \sum_{t \in \mathbb{Z}_q} \sum_{\pi} \sum_{\mathbf{a} \in \mathbb{Z}_q^d} c_{\pi(\mathbf{a}),t} \zeta_q^{\langle \pi(\mathbf{a}), \pi(\mathbf{x}) \rangle + t \langle \pi(\mathbf{x}), \pi(\mathbf{x}) \rangle} \\
&= \frac{1}{d!} \sum_{\pi} \sum_{t \in \mathbb{Z}_q} \sum_{\mathbf{a} \in \mathbb{Z}_q^d} c_{\mathbf{a},t} \zeta_q^{\langle \mathbf{a}, \pi(\mathbf{x}) \rangle + t \langle \pi(\mathbf{x}), \pi(\mathbf{x}) \rangle} \\
&= \frac{1}{d!} \sum_{\pi} \mathbb{1}[\pi(\mathbf{x}) = 0] \\
&= \mathbb{1}[\mathbf{x} = 0],
\end{aligned}
$$

as desired. □

Lemma 9.2. *Assume there exists a symmetric quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ supported on $\mathcal{W} \times \mathcal{T}$. Then there exists a symmetric quadratic-phase delta expansion supported on the same $\mathcal{W} \times \mathcal{T}$ and a positive integer $M \leq \exp(\mathsf{poly}(d))$ such that all coefficients in the delta expansion take the form $\frac{1}{M}(a_0 + a_1\zeta_q + a_2\zeta_q^2 + \ldots + a_{q-2}\zeta_q^{q-2})$, for integers $a_0, \ldots, a_{q-2} \leq \exp(\mathsf{poly}(d))$.*

Proof sketch. For a vector $\mathbf{a} \in \mathbb{Z}_q^d$, define:

$$
\mathsf{Type}(\mathbf{a}) := (|\{i : a_i = j\}|)_{j \in \mathbb{Z}_q}.
$$

It is easy to see that the quadratic-phase delta expansion condition can be written as a system of linear equations over $\mathbb{Q}(\zeta_q)$, where:

- There is one variable $c_{\mathbf{a},t}$ for each $t \in \mathcal{T}$ and each $\mathsf{Type}(\mathbf{a})$ that appears in $\mathcal{W}$; and
- There is one equation for each $\mathsf{Type}(\mathbf{x})$.

The number of variables and equations is thus bounded above by $q(d+1)^q = \mathsf{poly}(d)$. Moreover, each coefficient is a sum of at most q^d powers of ζ_q.

We can further unroll this into a linear system over $\mathbb{Q}$ by writing every element of $\mathbb{Q}(\zeta_q)$ as a $\mathbb{Q}$-linear combination of $1, \zeta_q, \ldots, \zeta_q^{q-2}$, at the expense of a factor of $q-1$ in the number of variables. With this, it is easy to see that all coefficients have absolute value at most q^d.

The resulting linear system is solvable, so it has a solution arising from a full-rank square submatrix. Moreover, Cramer's rule tells us that this solution will be an integer divided by the determinant of this submatrix. Since this is a submatrix of some dimension $d' \leq \mathsf{poly}(d)$ with entries at most q^d in magnitude, it follows that its determinant is $\leq d'! \cdot q^{dd'} \leq \exp(\mathsf{poly}(d))$. The upper bound on the numerators also follows from Cramer's rule. □

Lemma 9.3. *Assume there exists a symmetric quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ with support $\mathcal{W} \times \mathcal{T}$, where $\mathcal{W} \subseteq \mathbb{Z}_q^d$ and $\mathcal{T} \subseteq \mathbb{Z}_q$. Then there is a positive integer $M \leq M_d$, where $M_d := D! \cdot q^{dD}$ for $D := q^2(d+1)^q$ (so $M_d \leq \exp(\mathsf{poly}(d))$), such that, for every prime $p \equiv 1 \pmod{q}$ with $p > M$ and every primitive qth root of unity $\gamma \in \mathbb{F}_p$, there is a symmetric quadratic-phase delta expansion over $\mathbb{F}_p$ supported on $\mathcal{W} \times \mathcal{T}$. Moreover, given p, γ, $\mathcal{T}$ and the list of types of the vectors in $\mathcal{W}$, the coefficients of such an expansion can be computed in $\mathsf{poly}(d, \log p)$ time.*

Proof. Let M be as given by Lemma 9.2; its proof bounds M by $d'! \cdot q^{dd'}$ for some $d' \leq (q-1) \cdot q(d+1)^q \leq D$, so $M \leq M_d$. Let p and γ be as above, so that $p \nmid M$. The symmetric quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ can now be mapped into $\mathbb{F}_p$ by replacing $1/M$ with $M^{-1} \bmod p$, reducing all coefficients, modulo p,

and mapping $\zeta_q \mapsto \gamma$. More formally, we are applying a homomorphism from $\mathbb{Z}[\zeta_q, 1/M] \to \mathbb{F}_p$ given by $\zeta_q \mapsto \gamma$ and $1/M \mapsto M^{-1} \bmod p$.

For the last part, replace ζ_q by γ in the linear system of the proof of Lemma 9.2: it has one unknown $c_{\tau,t}$ for each type τ of the vectors in $\mathcal{W}$ and each $t \in \mathcal{T}$, and one equation for each type σ, so $\mathsf{poly}(d)$ unknowns and equations. A solution over $\mathbb{F}_p$ exists (the expansion above is one), and every solution gives a symmetric quadratic-phase delta expansion over $\mathbb{F}_p$ supported on $\mathcal{W} \times \mathcal{T}$, as both sides of the defining identity depend only on the type of $\mathbf{x}$. For $\mathbf{x}$ of type σ, the coefficient of $c_{\tau,t}$ in the equation for σ is $\gamma^{t\langle \mathbf{x},\mathbf{x}\rangle}$ times the coefficient of $\prod_j Y_j^{\tau_j}$ in the polynomial $\prod_{r\in\mathbb{Z}_q} \left(\sum_{j\in\mathbb{Z}_q} \gamma^{jr} Y_j\right)^{\sigma_r}$, which has $\mathsf{poly}(d)$ monomials and can be expanded with $\mathsf{poly}(d)$ operations over $\mathbb{F}_p$. Gaussian elimination then finds a solution with $\mathsf{poly}(d)$ further operations over $\mathbb{F}_p$, each taking $\mathsf{poly}(\log p)$ time. □

9.2 From Field Elements to Bits

In this section, we give a compiler that transforms a PIR scheme that operates over databases whose entries are field elements, to databases whose entries are bits. The main workhorse is the following theorem by [DPT10]:

Theorem 9.4 ([DPT10])**.** *Let K be an alphabet size and N a number of records. Then, there is an absolute constant $c > 0$ such that, for all integers $K, N \geq 2$, there exists an integer $L \leq \lceil N \log_2 K + cN/\sqrt{K} \rceil$ and deterministic algorithms $\mathsf{Encode}_{K,N} : [K]^N \to \{0,1\}^L$ and $\mathsf{Decode}_{K,N} : [N] \times \{0,1\}^L \to [K]$, such that:*

- *For every $\mathbf{v} \in [K]^N$ and $i \in [N]$, $\mathsf{Decode}_{K,N}(i, \mathsf{Encode}_{K,N}(\mathbf{v})) = v_i$.*
- *Given random access to the encoded string, $\mathsf{Decode}_{K,N}$ uses $\mathsf{poly}(\log N, \log K)$ bit operations.*
- *$\mathsf{Encode}_{K,N}$ uses $N \cdot \mathsf{poly}(\log N, \log K)$ bit operations.*

To encode the servers' tables of $\mathbb{F}_p$-elements compactly with Theorem 9.4, we need a prime p that is very close to a power of 2. The following lemma finds one in a given residue class; we use it both for odd primes q and for $q = 2$, in which case the condition $p \equiv 1 \pmod 2$ just says that p is odd.

Lemma 9.5 (Primes just above a power of two)**.** *Fix a prime q. There is a randomized algorithm that takes as input a positive integer M and a sufficiently large integer n, and constructs an integer B and a prime $p \equiv 1 \pmod q$ such that:*

$$p > M, \qquad p \geq n^4, \qquad 2^B < p \leq 2^B \left(1 + \frac{1}{4n}\right), \qquad B = O(\log M + \log^3 n).$$

This algorithm runs in expected time $\mathsf{poly}(B)$.

Proof. Set

$$B := \max\left(\lceil \log_2(M+1) \rceil, (1 + \lceil \log_2 n \rceil)^3\right), \qquad x := 2^B, \qquad H := \left\lfloor \frac{x}{4n} \right\rfloor.$$

Then $x > M$ and $x \geq n^4$. Recalling that q is fixed, it is straightforward to check that the prime number theorem in arithmetic progressions, with its error term $O(xe^{-c\sqrt{\ln x}})$ for a fixed modulus [Dav80, Chapter 22, p. 133], implies that:

$$\#\{p \in (x, x+H] \,:\, p \text{ prime},\ p \equiv 1 \pmod q\} = (1 + o(1)) \frac{H}{(q-1)\ln x}.$$

Thus, among the integers in this interval congruent to 1 modulo q, a fraction $\Theta_q(1/B)$ are prime. As a result, we can obtain a prime p with $x < p \le x + H$ and $p \equiv 1 \pmod{q}$ by rejection sampling. For this p, we have $p > M$ and:

$$n^{\omega(1)} \le 2^B < p \le 2^B \left(1 + \frac{1}{4n}\right).$$

□

In the following lemma, we show how to handle databases of n-bits by generically reducing to handling databases of $\mathbb{F}_p$-records where p is very close to a power of 2. The lemma takes as input PIR schemes over databases of b^j records, for a constant b: we use $b = q$ for the schemes of Theorem 7.1, and $b = 16$ for the two-server scheme of Theorem 6.1 (see the proof of Main Theorem 5.2).

Lemma 9.6 (PIR for n-bit databases)**.** *Fix integers $s, b \ge 2$. Let n be a sufficiently large database size (in bits), and let d be the largest integer with $b^d \le n$. Suppose there are integers B, p with p a prime, such that:*

$$p \ge n^4, \qquad 2^B \le p \le 2^B \left(1 + \frac{1}{4n}\right), \qquad B, \log p \le \mathsf{polylog}(n).$$

Assume additionally that, for every $j \in [d]$, there is an information-theoretically private s-server PIR scheme over databases of b^j records in $\mathbb{F}_p$, where each server stores b^j elements of $\mathbb{F}_p$, computed in $b^j \cdot \mathsf{polylog}(n)$ time, the user uploads $O(\log n)$ bits, and each server reads and sends back at most k elements of $\mathbb{F}_p$, in $k \cdot \mathsf{polylog}(n)$ time. Then, there is an information-theoretically private s-server PIR scheme for arbitrary n-bit databases with:

- *at most $n + 1$ stored bits per server;*
- *$n \cdot \mathsf{polylog}(n)$ preprocessing time;*
- *$O(\log^2 n)$ upload communication; and*
- *$k \cdot \mathsf{polylog}(n)$ server time and download communication.*

In particular, if $k \le n^{\eta + o(1)}$ for a constant η, this is an s-server scheme with exponent η (Definition 5.1).

Proof. First, it suffices to deal with the case where n is divisible by B: we can use such a protocol for the first $B\lfloor n/B \rfloor$ bits of the database, while the remaining $\le B = \mathsf{polylog}(n)$ bits can just be sent regardless of the client's query.

Bearing this in mind, let $n' = n/B$. Interpret the database as n' records in $\{0, 1, \ldots, 2^B - 1\}$, which can trivially be lifted to n' records in $\mathbb{F}_p$. Now we use the base-b decomposition of n' to chunk the database into mini-databases of size $b^{d_1} \ge b^{d_2} \ge \ldots \ge b^{d_r}$, with $d_1, \ldots, d_r \ge 1$, together with fewer than b leftover records (from the last digit), which can also be sent regardless of the client's query. This is not a new idea e.g., see [HR26, Corollary 3.3]. It is easy to see that we have $r \le O(\log n)$ and $d_1, \ldots, d_r \le d$. We will apply the given PIR schemes to each of these mini-databases. (The client only needs a record from one of the mini-databases, but they will send dummy queries for the others to ensure privacy.) A few efficiency claims become clear already:

- The upload is $O(r \log n) = O(\log^2 n)$.
- The download communication is at most $O(kr \log p) + O(b \log p + B) \le k \cdot \mathsf{polylog}(n)$.

It remains to address storage and server time: if we concatenate the data structures of the given schemes for each mini-database, together with the leftover records, each server needs to store n' records over $\mathbb{F}_p$. To encode this into bits, each server uses the encoding given by Theorem 9.4 to store their data structure. The resulting encoding has size at most:

$$\lceil n' \log_2 p + O(n'/\sqrt{p}) \rceil \le \lceil n'B + n' \log(1 + 1/(4n)) + O(n'/n^2) \rceil$$

$$
\begin{aligned}
&= n + \left\lceil \frac{n}{B} \log(1 + 1/(4n)) + O(1/(nB)) \right\rceil \\
&\le n + \left\lceil \frac{1}{4B \ln 2} + O\left(\frac{1}{nB}\right) \right\rceil \\
&= n + 1,
\end{aligned}
$$

for sufficiently large n. Here, we are using the fact that the conditions force $B \ge \omega(1)$. This implies the claimed storage. To access a stored $\mathbb{F}_p$-element from this encoding, each server needs to make $\mathsf{poly}(\log n, \log p) = \mathsf{polylog}(n)$ accesses to the encoding, giving the claimed $k \cdot \mathsf{polylog}(n)$ server time. Finally, preprocessing computes the data structures of the mini-databases in $n' \cdot \mathsf{polylog}(n)$ time, and then their encoding in $n \cdot \mathsf{polylog}(n)$ time (Theorem 9.4). □

Part II
Fourier-Sparse Cone-Delta Functions from Modular Forms

This part proves the existence results that power the PIR schemes of Part I: cone-delta functions and quadratic-phase delta expansions (Definitions 2.1 and 4.3) over $\mathbb{Z}_q^d$ whose Fourier supports are exponentially smaller than the trivial q^d. Throughout this part, $\mathcal{Q} := \{\mathbf{x} \in \mathbb{Z}_q^d : \langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod q\}$ denotes the cone, the set on which cone-delta functions are constrained (Definition 2.1). We prove two such results: a warm-up (Corollary 11.13), producing the cone-delta functions behind q-server PIR (Main Theorem 5.3), and a stronger theorem (Theorem 12.2, with its support bounds in Corollary 12.4), which trades the number of servers s against the support exponent.

Both proofs follow the architecture of the technical overview (Section 2.3.2), which mirrors the reconstruction argument for Reed–Muller PIR: if the desired function does not exist, duality provides a witness; composing a polynomial built from the witness with a vector-valued theta function produces scalar modular forms; and the witness forces these forms to vanish at the cusps to higher order than the valence formula allows. Recall the dual form of the Reed–Muller argument. There, the database is encoded as a polynomial F of degree at most D in several variables. To read $F(\mathbf{x})$, the user picks a uniformly random direction $\mathbf{h}$, sends the points $\mathbf{x} + j\mathbf{h}$ for $j \in [s]$ to $s \ge D + 1$ servers, and recovers $F(\mathbf{x})$ by interpolating F along the line $j \mapsto \mathbf{x} + j\mathbf{h}$. Interpolation could fail only if some polynomial Δ of degree at most D had $\Delta(\mathbf{x}) = 1$ while vanishing at all s query points; restricted to the line, such a Δ would be a univariate polynomial of degree at most D that is nonzero at $j = 0$ yet has $s > D$ roots, which is impossible. Table 4 lines up the two arguments in the case of the warm-up, and Figures 8 and 13 show how the steps fit together in each of our two proofs. The stronger construction adds an elliptic variable to the theta function, which yields a second scalarization and finer bounds on the orders at the cusps, so the number of servers can drop below q (Table 5).

Relation to the literature. Almost all of the results on modular forms that we use in this part are classical. We restate them, often with proofs, only to keep the presentation self-contained for readers outside the area, and we point to sources along the way. This includes the background of Section 10 (see, e.g., [Ser73, DS05, Shi71]); the Weil representation and its action on the finite Heisenberg group [Wei64, Gér77, Str13]; the transformation laws of lattice theta functions [Bor98, Theorem 4.1] and of theta functions with characteristics [Mum83, Chapter I, §§3 and 7]; and the facts that restricting a Jacobi form to a torsion point, or taking the first nonvanishing coefficient of its Taylor expansion in the elliptic variable,

Step	Reed–Muller PIR	Warm-up (Section 11)
1. Witness	Δ of degree $\leq D$, with $\Delta(\mathbf{x}) = 1$ and $\Delta(\mathbf{x} + j\mathbf{h}) = 0$ for $j \in [s]$	β on the cone, with $\beta_{\mathbf{0}} = 1$ and Fourier sums vanishing on S_d; the polynomial P_β
2. Curve	the line $j \mapsto \mathbf{x} + j\mathbf{h}$	the theta function $\tau \mapsto \Theta_{\mathcal{L}}(\tau) \in \mathbb{C}^q$, for $\tau \in \mathcal{H}$
3. Restriction	$g(j) = \Delta(\mathbf{x} + j\mathbf{h})$, of degree $\leq D$	$P_\beta \circ \Theta_{\mathcal{L}}$, a modular form of weight $dr/2$
4. Nonvanishing	$g(0) = 1$	a nonzero multiple of $\beta_{\mathbf{0}} = 1$ as $\tau \to i\infty$
Vanishing	at $j = 1, \ldots, s$	to order more than $\frac{dr(q+1)}{24}$ at the cusp 0
Root bound	at most D roots	valence formula: total order $\frac{dr(q+1)}{24}$

Table 4: The reconstruction argument for Reed–Muller PIR, in the dual form recalled at the start of Part II, side by side with our proof that sparse cone-delta functions exist, in the case of the warm-up. On the left, Δ is a polynomial of degree at most D, $\mathbf{x}$ is the queried point, $\mathbf{h}$ is the user's random direction, and the $s \geq D + 1$ servers receive the points $\mathbf{x} + j\mathbf{h}$. On the right, $\mathcal{L}$ is a lattice of rank r, and S_d is the set of frequency vectors of small total cost.

produces modular forms [EZ85, Theorem 1.3 and Section 3]. Combining such restrictions and Taylor coefficients with norms and the valence formula in order to bound orders of vanishing is also an established technique [RWR15, Aok22, LW26]. Likewise, substituting a vector of theta functions into a suitably invariant polynomial (Lemma 10.30) is also how the classical correspondence between weight enumerators of codes and theta series of lattices arises [BE72, NRS06].

What we believe is new is the connection itself: the reduction of the existence of Fourier-sparse delta functions to the impossibility of a dual witness, and the encoding of the witness's orthogonality constraints as orders of vanishing of modular forms at cusps — in the stronger construction, via the symmetrized witness, the per-class quantities attached to the selected cusp classes, and the analysis of the edge polynomials in Step 4. To our knowledge, modular forms have not previously been used to construct Fourier-sparse delta functions. The closest arguments we know of are the classical bounds on the minimum distance of self-dual codes [MS73], which also substitute theta functions into an invariant polynomial and bound its order of vanishing; there the polynomial is a weight enumerator, whereas ours comes from a dual witness (Section 3). For $q = 3$, the conclusion of the warm-up also follows from an elementary construction of Alon and Lubetzky (Section 11).

Section 10 collects the background: $\mathrm{SL}_2(\mathbb{Z})$ and its congruence subgroups, the Weil representation, modular forms and their cusp expansions, and the valence formula in the form we use. Section 11 carries out the warm-up via even lattices (for primes $q \equiv 3 \pmod 4$), and Section 12 proves the stronger theorem. Standard complex-analysis facts and routine or lengthy verifications are deferred to the appendix, beginning with Appendix B.1.

10 Background: $\mathrm{SL}_2(\mathbb{Z})$, the Weil Representation, and Modular Forms

Everything in this section is standard; we restate it to fix notation and normalizations, and to keep the presentation self-contained. For modular forms for congruence subgroups, see, e.g., [Ser73, DS05, Shi71]; for vector-valued modular forms for the Weil representation, see, e.g., [Bor98, Section 4]. Standard facts from complex analysis that we use freely are collected in Appendix B.1.

For readers coming from the technical overview, this section makes precise the objects sketched

in Sections 2.3.1 and 2.3.3, and we point back to the figures there along the way. Each piece has a fixed role in the proofs of Sections 11 and 12 (see the flow charts in Figures 8 and 13). The Weil representation (Section 10.2) is the symmetry under which both of our theta functions transform. The upper half-plane and its cusps (Section 10.3) are where these functions live, and the expansions at the cusps (Sections 10.4.2 and 10.4.3) define the orders of vanishing that our proofs bound. The composition laws (Section 10.4.4) contain the scalarization step, and the valence formula (Section 10.4.5) is the root count that ends both proofs.

10.1 $\mathrm{SL}_2(\mathbb{Z})$ and Congruence Subgroups

Definition 10.1. *The group* $\mathrm{SL}_2(\mathbb{Z})$ *consists of matrices* $\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix}$ *with integer entries and determinant* $ad_0 - bc = 1$. *Throughout this part, we write a generic element of* $\mathrm{SL}_2(\mathbb{Z})$ *as* $\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix}$, *reserving the letter* d *for dimension and degree parameters. We single out the two elements*

$$S := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \qquad \text{and} \qquad T := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Fact 10.2. $\mathrm{SL}_2(\mathbb{Z})$ *is generated by* S *and* T. *(Serre [Ser73, Chapter VII, Theorem 2] proves this for* $\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$; *the statement for* $\mathrm{SL}_2(\mathbb{Z})$ *follows since* $-I = S^2 \in \langle S, T\rangle$.*) Moreover, direct computation shows that* $S^2 = (ST)^3 = -I$; *in particular,* S *and* T *obey the relations* $S^4 = I$ *and* $(ST)^3 = S^2$.

Geometrically, S and T are the two steps of the Euclidean algorithm, applied to a basis of a lattice in $\mathbb{C}$ (Section 2.3.3 and Figure 1); this is how $\mathrm{SL}_2(\mathbb{Z})$ will act on the upper half-plane in Section 10.3.

Of particular interest to us will be considering particular subgroups of $\mathrm{SL}_2(\mathbb{Z})$:

Definition 10.3. *For an odd prime* q, *we define the following subgroups of* $\mathrm{SL}_2(\mathbb{Z})$:

$$\Gamma(q) = \left\{ \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \pmod{q} \right\}$$

$$\Gamma_0(q) = \left\{ \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \equiv \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} \pmod{q} \right\}.$$

A congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$ of level q *is any subgroup that contains* $\Gamma(q)$; *in particular,* $\mathrm{SL}_2(\mathbb{Z})$ *itself and* $\Gamma_0(q)$ *are congruence subgroups of level* q. *Equivalently, a subgroup* Γ *is a congruence subgroup of level* q *if and only if a matrix's membership in* Γ *only depends on the values of its entries modulo* q: *if* $\Gamma \supseteq \Gamma(q)$ *and* $g \equiv g' \pmod{q}$, *then* $g'g^{-1} \in \Gamma(q)$ *and so* $g \in \Gamma \Leftrightarrow g' \in \Gamma$; *conversely, a subgroup with the latter property contains every matrix congruent to* I.

The two subgroups play complementary roles for us. The Weil representation of Section 10.2 is trivial on $\Gamma(q)$, and on $\Gamma_0(q)$ it merely permutes the coordinates of $\mathbb{C}^q$ and multiplies them by phases (Fact 10.11). So $\Gamma_0(q)$ is the group of the warm-up (Section 11), where the phases cancel on the cone, and $\Gamma(q)$ is the group of the stronger construction (Section 12), which works with this smaller group at the cost of many more cusps (Figure 9).

Remark 10.4. *We do* not *require a congruence subgroup to contain* $-I$; *indeed,* $-I \notin \Gamma(q)$. *Because* $-I$ *will act trivially on all the geometric objects introduced below, some statements (most importantly the valence*

formula, Theorem 10.35) are stated for subgroups containing $-I$; to apply them to $\Gamma(q)$, we will pass to the congruence subgroup

$$\pm\Gamma(q) := \langle \Gamma(q), -I \rangle = \Gamma(q) \cup (-I)\Gamma(q).$$

Lemma 10.5 (Indices)**.** *Let q be an odd prime. The entrywise-reduction map $\pi : \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}_q)$ is a surjective group homomorphism with kernel $\Gamma(q)$, and $|\mathrm{SL}_2(\mathbb{Z}_q)| = q^3 - q$. Consequently:*

- $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma(q)] = q^3 - q$;
- $[\mathrm{SL}_2(\mathbb{Z}) : \pm\Gamma(q)] = \frac{q^3-q}{2}$; *and*
- $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma_0(q)] = q + 1$.

Proof. The surjectivity of π and the order $|\mathrm{SL}_2(\mathbb{Z}_q)| = q^3 - q$ are standard; see, e.g., [DS05, Section 1.2] (stated there for general level). The indices follow at once: $\Gamma(q) = \pi^{-1}(I)$ is the kernel, so $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma(q)] = |\mathrm{SL}_2(\mathbb{Z}_q)|$; the group $\pm\Gamma(q)$ is the disjoint union of the two cosets $\Gamma(q)$ and $(-I)\Gamma(q)$; and $\Gamma_0(q) = \pi^{-1}(B)$ for the subgroup B of upper-triangular matrices $\left[\begin{smallmatrix} a & b \\ 0 & a^{-1} \end{smallmatrix}\right]$ in $\mathrm{SL}_2(\mathbb{Z}_q)$, which number $q(q-1)$, so that (by surjectivity of π, which puts the cosets of $\Gamma_0(q)$ and of B in bijection) the index is $(q^3 - q)/(q(q-1)) = q + 1$. □

10.2 The Weil Representation

We first fix some notation that will be in force throughout this part: $\zeta_q := e^{2\pi i/q}$, so that ζ_q^t is well-defined for $t \in \mathbb{Z}_q$, and $\{\mathbf{e}_x\}_{x\in\mathbb{Z}_q}$ denotes the standard basis of $\mathbb{C}^q$, indexed by $\mathbb{Z}_q$.

Definition 10.6 (Representation)**.** *A* representation *of a group G on $\mathbb{C}^N$ is a map $\rho : G \to \mathrm{GL}_N(\mathbb{C})$ satisfying $\rho(gh) = \rho(g)\rho(h)$ for all $g, h \in G$. (It follows that $\rho(1) = I$ and $\rho(g^{-1}) = \rho(g)^{-1}$.) When $N = 1$, we identify $\mathrm{GL}_1(\mathbb{C})$ with $\mathbb{C} \setminus \{0\}$ and call ρ* one-dimensional.

Definition 10.7 (Legendre symbol)**.** *For $x \in \mathbb{Z}_q$, define $\left(\frac{x}{q}\right) := 1$ if x is a nonzero square in $\mathbb{Z}_q$, -1 if x is not a square in $\mathbb{Z}_q$, and 0 if $x = 0$. For an integer x, define $\left(\frac{x}{q}\right)$ by first reducing x modulo q.*

Fact 10.8. *The Legendre symbol is multiplicative, i.e., $\left(\frac{xy}{q}\right) = \left(\frac{x}{q}\right)\left(\frac{y}{q}\right)$. (This is a standard consequence of the cyclicity of $\mathbb{Z}_q^*$; see, e.g., [IR90, Chapter 5].)*

Fact 10.9 (Quadratic Gauss sums)**.** *Let q be an odd prime and let μ be an integer with $\mu \not\equiv 0 \pmod q$. Then there exists $\epsilon_{\mu,q} \in \{\pm 1, \pm i\}$ (so in particular $\epsilon_{\mu,q}^4 = 1$ and $|\epsilon_{\mu,q}| = 1$) such that*

$$\sum_{y\in\mathbb{Z}_q} \zeta_q^{\mu y^2} = \epsilon_{\mu,q}^{-1}\sqrt{q}.$$

(This is classical: [IR90, Chapter 6] shows that the sum equals $\left(\frac{\mu}{q}\right) g$ for a constant $g = \sum_{x\in\mathbb{Z}_q}\left(\frac{x}{q}\right)\zeta_q^x$ independent of μ, with $g^2 = \left(\frac{-1}{q}\right) q$; so the square of the sum is $\pm q$, which gives the claim.)

Definition 10.10. *Fix an integer μ with $\mu \not\equiv 0 \pmod q$. Define the matrices $\rho_S, \rho_T \in \mathbb{C}^{q\times q}$ (depending on q and μ) as follows:*

- *ρ_T is diagonal, mapping*

$$\mathbf{e}_x \mapsto \zeta_q^{\mu x^2}\mathbf{e}_x.$$

- *ρ_S is a multiple of the qth order discrete Fourier transform, mapping:*

$$\mathbf{e}_x \mapsto \frac{\epsilon_{\mu,q}}{\sqrt{q}} \sum_{y \in \mathbb{Z}_q} \zeta_q^{-2\mu x y} \mathbf{e}_y.$$

The technical overview (Section 2.3.1) explains why these two matrices are the natural ones for us: ρ_T multiplies by the quadratic phase that defines the cone, and ρ_S is a Fourier transform, so the two conditions defining a cone-delta function are statements about ρ_T and ρ_S. There we used the matrices without the constant $\frac{\epsilon_{\mu,q}}{\sqrt{q}}$ and with $\mu = 1$; with this constant and parameter μ, the following is a representation, not only a representation up to scalars. (Figure 3 compares the Weil representation with the characters of Fourier analysis.)

The matrices ρ_T and ρ_S extend to a representation of $\mathrm{SL}_2(\mathbb{Z})$, the Weil representation, which is trivial on $\Gamma(q)$ and monomial on $\Gamma_0(q)$. This is classical, and we restate it in the form we need. In the language of finite quadratic modules, ρ is the Weil representation attached to the quadratic module $(\mathbb{Z}_q, x \mapsto \mu x^2/q)$, and Definition 10.10 is exactly the normalization of Strömberg [Str13, Definition 5.1] (Scheithauer [Sch09] works with the dual representation); see also [Bor98, Section 4] and, for the finite-field setting, [Wei64, Gér77].

Fact 10.11 (The Weil representation [Str13, Definition 5.1 and Lemmas 5.12 and 5.13])**.** *Let $\mu \not\equiv 0 \pmod q$ be as in Definition 10.10. There is a unique representation ρ of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{C}^q$ with $\rho(T) = \rho_T$ and $\rho(S) = \rho_S$; we call it the* Weil representation *with parameter μ. Moreover:*

- *$\rho(g) = I$ for every $g \in \Gamma(q)$. Consequently, $\rho(g)$ depends only on the reduction of g modulo q, and we write $\rho(\bar{g}) := \rho(g)$ for $\bar{g} \in \mathrm{SL}_2(\mathbb{Z}_q)$ and any lift $g \in \mathrm{SL}_2(\mathbb{Z})$ of $\bar{g}$ (which exists by Lemma 10.5).*
- *For every $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \Gamma_0(q)$ and every $x \in \mathbb{Z}_q$:*

$$\rho(g)\,\mathbf{e}_x = \left(\frac{a}{q}\right) \zeta_q^{\mu b d_0 x^2}\, \mathbf{e}_{d_0 x}. \tag{20}$$

Finally, we record how the Weil representation interacts with the *Heisenberg operators*, the cyclic shift and the diagonal phase matrix on $\mathbb{C}^q$. This is used only in Section 12, where translations of an elliptic variable act through these operators (Lemma 12.15).

Definition 10.12 (Heisenberg operators)**.** *Define the Heisenberg operators $\mathsf{X}, \mathsf{Z} \in \mathbb{C}^{q \times q}$ by $\mathsf{X}\mathbf{e}_r = \mathbf{e}_{(r-1) \bmod q}$ and $\mathsf{Z}\mathbf{e}_r = \zeta_q^r \mathbf{e}_r$. (In coordinates: $(\mathsf{X}\mathbf{y})_j = y_{(j+1) \bmod q}$ and $(\mathsf{Z}\mathbf{y})_j = \zeta_q^j\, y_j$.)*

The Weil representation interacts with these operators in a simple way: conjugating by $\rho(g)$ permutes the operators $\mathsf{X}^m\mathsf{Z}^n$ up to roots of unity.

Lemma 10.13 (Weil conjugation of the Heisenberg operators [Wei64, n° 10, Théorème 1], [Gér77, Theorem 2.4(a)])**.** *Let ρ be the Weil representation with parameter μ (Fact 10.11). For every $g \in \mathrm{SL}_2(\mathbb{Z})$ and all integers m, n, there exist integers m', n', c (depending on g, m, n, and μ) such that:*

$$\rho(g)\,\mathsf{X}^m\mathsf{Z}^n = \zeta_q^c\, \mathsf{X}^{m'}\mathsf{Z}^{n'}\,\rho(g).$$

10.3 Lattices in $\mathbb{C}$ and the Upper Half-Plane

All of our theta functions are sums of Gaussian-type terms $e^{\pi i \tau \langle y, y \rangle}$ over the points of a lattice. The parameter τ controlling the Gaussian lies in the upper half-plane, and the symmetries used in our proofs come from an action of $\mathrm{SL}_2(\mathbb{Z})$ on it. This subsection unpacks that action, the automorphy factor $G(g, \tau)$ through which every transformation law below will be expressed, and the *cusps* (the rational numbers together with ∞) — the points, besides those of $\mathcal{H}$, at which the valence formula will count our forms' zeros.

The technical overview explains this action through lattices (Section 2.3.3): τ stands for the lattice $\mathbb{Z} + \mathbb{Z}\tau$, the matrix g changes its basis to $(c\tau + d_0, a\tau + b)$, and dividing by $c\tau + d_0$ turns this into a basis of the form $(1, g * \tau)$ (Figure 1). In this picture, the automorphy factor $G(g, \tau) = c\tau + d_0$ defined below is exactly the number we divide by, and the chain rule (Proposition 10.17) says that performing two changes of basis in turn multiplies these factors. Figure 2 shows the cusps, and how elements of $\mathrm{SL}_2(\mathbb{Z})$ move ∞ to other cusps.

Definition 10.14. *For a complex number $z \in \mathbb{C}$, let $\Re z$ and $\Im z$ denote its real and imaginary parts. We let $\mathcal{H} \subset \mathbb{C}$ denote the strict upper half-plane, i.e., $\{z \in \mathbb{C} : \Im z > 0\}$. Throughout, we will use τ to denote a complex number in $\mathcal{H}$.*

Fact 10.15. $\mathrm{SL}_2(\mathbb{Z})$ *acts on $\mathcal{H}$ under the following group action:*

$$\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} * \tau := \frac{a\tau + b}{c\tau + d_0}.$$

(Note that under this convention, $-I$ fixes every τ so this is actually an action of $\mathrm{SL}_2(\mathbb{Z})/(\pm I) = \mathrm{PSL}_2(\mathbb{Z})$.)

Definition 10.16. *Define $G(g, \tau) := c\tau + d_0$, where $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z})$. (Note in particular that $G(I, \tau) = 1$ for all τ.)*

Proposition 10.17 (Chain rule [DS05, Lemma 1.2.2(a)])**.** *For any $g, h \in \mathrm{SL}_2(\mathbb{Z})$ and $\tau \in \mathcal{H}$, we have:*

$$G(gh, \tau) = G(g, h * \tau) G(h, \tau).$$

Definition 10.18 (Cusps)**.** *The* cusps *are the elements of $\mathbb{Q} \cup \{\infty\}$. The action of $\mathrm{SL}_2(\mathbb{Z})$ extends to them by the same formula:*

$$\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} * x := \frac{ax + b}{cx + d_0} \quad (x \in \mathbb{Q}), \qquad \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} * \infty := \frac{a}{c},$$

where a fraction with denominator 0 is understood as ∞. (Intuitively, ∞ stands for the point $i\infty$ at the top of the upper half-plane, and the rational cusps are points on its boundary, the real line.) The cusps will be essential to our proofs: the valence formula (Theorem 10.35) counts a modular form's zeros both at points of $\mathcal{H}$ and at cusps, and our arguments work by forcing high-order vanishing precisely at the cusps.

10.4 Modular Forms

10.4.1 Weakly Modular Functions

In the following, let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$ of level q (Definition 10.3), and let $\rho : \Gamma \to \mathrm{GL}_N(\mathbb{C})$ be a representation (Definition 10.6) for a positive integer $N > 0$. We additionally assume throughout that ρ is of *finite order*, meaning that every matrix $\rho(g)$ in its image has finite order. This

assumption is what gives weakly modular functions well-behaved expansions at the cusps (Section 10.4.2), and it holds in all our applications: most importantly, the Weil representation factors through the finite group $\mathrm{SL}_2(\mathbb{Z}_q)$ by Fact 10.11, so it — and every representation obtained from it by restriction or conjugation — has finite order.

Definition 10.19 (Weakly modular functions)**.** *Let k be a non-negative integer. A holomorphic $F : \mathcal{H} \to \mathbb{C}^N$ is weakly modular of weight k for Γ with representation ρ if:*

$$F(g * \tau) = (c\tau + d_0)^k \rho(g) F(\tau), \text{ or more explicitly,} \tag{21}$$
$$F\left(\frac{a\tau + b}{c\tau + d_0}\right) = (c\tau + d_0)^k \rho\left(\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix}\right) F(\tau),$$

for all $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \Gamma$ and $\tau \in \mathcal{H}$. Note that the types here are consistent: $\rho(g) \in \mathbb{C}^{N\times N}$ and it acts on $F(\tau) \in \mathbb{C}^N$ to produce some other vector in $\mathbb{C}^N$.

This is the precise form of the informal definition (12) of the technical overview, and Figure 3(b) draws the transformation law (21) as a commutative diagram.

Proportionality. For two functions F_1, F_2 of the same complex variables (in our applications, some of τ, w, and z, or the coordinates of a vector $\mathbf{v} \in \mathbb{C}^N$), with values in $\mathbb{C}$ or in $\mathbb{C}^N$, we write $F_1 \propto F_2$ if $F_1 = \lambda F_2$ for a nonzero constant scalar $\lambda \in \mathbb{C}$, and $F_1 \propto_\zeta F_2$ if moreover λ can be taken to be a root of unity. Here "constant" means independent of the complex variables; λ may depend on everything else in sight, such as a group element g. We stress that λ is always a *scalar*: these relations never absorb a matrix, so in vector-valued identities such as the transformation law (21), the matrix $\rho(g)$ (like any Heisenberg operator below) is always written out explicitly. Both relations are symmetric and transitive, are preserved when we substitute functions of other variables for the variables (the constant does not change), and are compatible with products, powers, and inverses, as well as with applying a homogeneous polynomial P of degree D to both sides (if $F_1 = \lambda F_2$ then $P(F_1) = \lambda^D P(F_2)$); for $\propto_\zeta$ this uses that the roots of unity form a group. This lets us suppress constants that our arguments never need to know exactly. What they do need is that, for a scalar function, such constants assemble into a representation of finite order. For one-dimensional representations, finite order in the sense above (every value is a root of unity) already implies the uniform statement $\chi^j = 1$:

Lemma 10.20 (Characters of finite order)**.** *Let Γ be a congruence subgroup and $\chi : \Gamma \to \mathbb{C}^\times$ a one-dimensional representation each of whose values $\chi(g)$ is a root of unity. Then $\chi^j = 1$ for some positive integer j. (So for one-dimensional representations of congruence subgroups, the two senses of "finite order" — every value a root of unity, and $\chi^j = 1$ for some $j \geq 1$ — agree, and we use them interchangeably.)*

Proof. Γ has finite index in $\mathrm{SL}_2(\mathbb{Z})$ (it contains $\Gamma(q)$, which has index $q^3 - q$ by Lemma 10.5), and $\mathrm{SL}_2(\mathbb{Z})$ is generated by S and T (Fact 10.2), so Γ is finitely generated [LS77, Chapter II, Proposition 4.2]. Hence $\chi(\Gamma)$ is generated by finitely many roots of unity, so it is a finite cyclic group, and $\chi^j = 1$ for $j := |\chi(\Gamma)|$. □

Lemma 10.21 (Scalar transformation laws)**.** *Let Γ be a congruence subgroup, k a non-negative integer, and $f : \mathcal{H} \to \mathbb{C}$ a holomorphic function that is not identically zero, such that $f(g * \tau) \propto G(g,\tau)^k f(\tau)$ (as functions of τ) for every $g \in \Gamma$; that is, $f(g * \tau) = \chi(g)\, G(g,\tau)^k f(\tau)$ for a nonzero constant $\chi(g)$ (uniquely determined, as $f \not\equiv 0$). Then $\chi : \Gamma \to \mathbb{C}^\times$ is a one-dimensional representation, so f satisfies the transformation law* (21) *with representation χ. If moreover $f(g * \tau) \propto_\zeta G(g,\tau)^k f(\tau)$ for every $g \in \Gamma$, then $\chi^j = 1$ for some*

positive integer j; in particular χ has finite order, so f is weakly modular of weight k for Γ with representation χ.

Proof. For $g_1, g_2 \in \Gamma$, applying the transformation law twice and then Proposition 10.17 gives

$$\chi(g_1g_2)\, G(g_1g_2,\tau)^k f(\tau) = f(g_1g_2 * \tau) = \chi(g_1)\, G(g_1, g_2 * \tau)^k\, \chi(g_2)\, G(g_2,\tau)^k f(\tau) = \chi(g_1)\chi(g_2)\, G(g_1g_2,\tau)^k f(\tau);$$

evaluating at any τ with $f(\tau) \neq 0$ gives $\chi(g_1g_2) = \chi(g_1)\chi(g_2)$. For the second claim, every value $\chi(g)$ is then a root of unity, so Lemma 10.20 applies. □

To informally understand what (21) is saying, note that:

$$\begin{aligned}\frac{d(g*\tau)}{d\tau} &= \frac{d}{d\tau}\frac{a\tau+b}{c\tau+d_0} \\ &= \frac{(c\tau+d_0)a - c(a\tau+b)}{(c\tau+d_0)^2} \\ &= \frac{1}{(c\tau+d_0)^2}.\end{aligned}$$

This allows us to rewrite (21) as:

$$\begin{aligned} F(g*\tau) &= \left(\frac{d(g*\tau)}{d\tau}\right)^{-k/2} \rho(g)F(\tau) \\ \Leftrightarrow F(g*\tau)(d(g*\tau))^{k/2} &= \rho(g)F(\tau)(d\tau)^{k/2}.\end{aligned}$$

In words, F intertwines two actions of g: on the upper half-plane, g moves the point τ to $g * \tau$ (rescaling $(d\tau)^{k/2}$ along the way), while on $\mathbb{C}^N$, $\rho(g)$ acts linearly. This is the left square of the commutative diagram in Figure 3(b), which is drawn there for the theta function Θ, of weight $k = \frac{r}{2}$. From this, we intuitively deduce two things: firstly, that Proposition 10.17 from earlier is indeed the standard chain rule from calculus in disguise; and secondly, to check (21) it suffices to check it for g ranging over a set that generates Γ. Nevertheless, we check this more formally:

Proposition 10.22. *If (21) holds for (g_1,τ) and (g_2,τ) for all $\tau \in \mathcal{H}$, then it also holds for (g_1^{-1},τ) and (g_1g_2,τ).*

Proof. For the product: $F(g_1g_2 * \tau) = G(g_1, g_2*\tau)^k \rho(g_1) F(g_2*\tau) = G(g_1, g_2*\tau)^k G(g_2,\tau)^k \rho(g_1)\rho(g_2)F(\tau)$, which equals $G(g_1g_2,\tau)^k \rho(g_1g_2)F(\tau)$ by Proposition 10.17. For the inverse: substitute $\tau \mapsto g_1^{-1} * \tau$ into the law for g_1 and rearrange, using $G(g_1, g_1^{-1}*\tau)\, G(g_1^{-1},\tau) = G(I,\tau) = 1$ (Proposition 10.17 again). □

10.4.2 The Cusp at Infinity

An important task for us is to understand when a weakly modular function can be extended to include the cusps in $\mathbb{Q} \cup \{\infty\}$ in a reasonable way. This is also where we define the orders of vanishing at cusps, which are the quantities that all of our root counts bound. By "reasonable", we mean that the resulting function will be holomorphic, and this will give us a modular form. This is most naturally done by first addressing the cusp at ∞, so we begin there.

Definition 10.23 (Width of the cusp at ∞)**.** *The* width *of the cusp at infinity is the minimal positive integer h such that $T^h = \begin{bmatrix}1 & h\\ 0 & 1\end{bmatrix} \in \Gamma$. If Γ is a congruence subgroup of level q, then the width is well-defined and divides q (since $T^q \in \Gamma(q) \subseteq \Gamma$); as q is prime, this means $h \in \{1, q\}$, and h need not equal q.*

Letting h be the width of the cusp at ∞, we have:

$$F(\tau + h) = \rho\left(\begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}\right) F(\tau). \tag{22}$$

To analyze this periodicity condition, we first put the matrix $A := \rho(T^h)$ appearing in it — the matrix describing how F transforms over one period — in a canonical form, using the standing finite-order assumption on ρ:

Proposition 10.24 ([HK71, Section 6.4, Theorem 6]). *If $A \in \mathrm{GL}_N(\mathbb{C})$ has finite order, then A is diagonalizable and all of its eigenvalues are roots of unity.*

By Proposition 10.24, we may fix a basis $\mathbf{v}_1, \dots, \mathbf{v}_N$ of $\mathbb{C}^N$ consisting of eigenvectors of A, with dual basis $\mathbf{u}_1, \dots, \mathbf{u}_N$ (characterized by $\mathbf{u}_i^* \mathbf{v}_j = \mathbb{1}[i = j]$), and unique real numbers $\alpha_1, \dots, \alpha_N \in [0, 1)$ such that $A\mathbf{v}_j = e^{2\pi i \alpha_j} \mathbf{v}_j$ for each j. Define:

$$M := \sum_{j \in [N]} \alpha_j \mathbf{v}_j \mathbf{u}_j^*, \tag{23}$$

so that $\exp(2\pi i M) = \sum_{j \in [N]} e^{2\pi i \alpha_j} \mathbf{v}_j \mathbf{u}_j^* = A$. (Here and below we use that the $\mathbf{v}_j \mathbf{u}_j^*$ are idempotents with $\mathbf{v}_i \mathbf{u}_i^* \cdot \mathbf{v}_j \mathbf{u}_j^* = \mathbb{1}[i = j]\, \mathbf{v}_j \mathbf{u}_j^*$ and $\sum_j \mathbf{v}_j \mathbf{u}_j^* = I$, so that $M^m = \sum_j \alpha_j^m \mathbf{v}_j \mathbf{u}_j^*$ for all $m \geq 1$ and the exponential series simplifies as displayed.) This allows us to write Equation (22) as a periodicity condition: since $\exp(-2\pi i M(\tau + h)/h) = \exp(-2\pi i M \tau/h) \exp(-2\pi i M)$ (the two exponents, being multiples of M, commute) and $\exp(-2\pi i M) A = I$, we have

$$\exp(-2\pi i M(\tau + h)/h) F(\tau + h) = \exp(-2\pi i M \tau/h) F(\tau).$$

In other words, the holomorphic function $\exp(-2\pi i M \tau/h) F(\tau)$ is h-periodic. (Geometrically: the map $\tau \mapsto e^{2\pi i \tau/h}$ sends $\mathcal{H}$ onto the punctured unit disk, identifying τ with $\tau + h$; an h-periodic holomorphic function of τ is exactly a holomorphic function of this disk coordinate, and its behavior "at $\tau = i\infty$" becomes its behavior at the puncture; see Figure 7.)

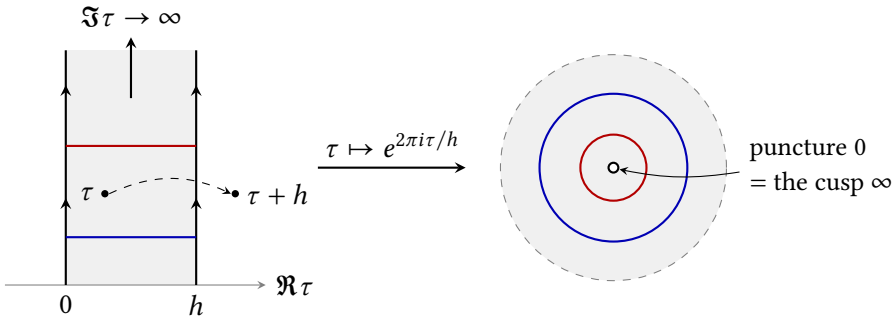


Figure 7: The cusp at ∞, drawn schematically. A function f with $f(\tau + h) = f(\tau)$ is a function on the strip $0 \leq \Re\tau \leq h$ with its two vertical edges glued (left). The map $\tau \mapsto e^{2\pi i \tau/h}$ identifies this glued strip with the punctured unit disk (right): it sends horizontal lines to circles around 0 (higher lines to smaller circles), and $\Im\tau \to \infty$ to the puncture. So the cusp ∞ becomes the point 0: f is holomorphic at ∞ when 0 is a removable singularity, and its order at ∞ is its order of vanishing at 0, the least n with a nonzero coefficient in the expansion $f(\tau) = \sum_n c_n e^{2\pi i n \tau/h}$. For a vector-valued F, the periodicity is twisted by $\rho(T^h)$, and untwisting it shifts the exponents by the numbers α_j (Definition 10.25).

By Fact B.4, it therefore admits a unique infinite series expansion:

$$\exp(-2\pi iM\tau/h)F(\tau) = \sum_{n=-\infty}^{\infty} \mathbf{c}_n e^{2\pi i\tau n/h}, \tag{24}$$

where each $\mathbf{c}_n \in \mathbb{C}^N$, and the series converges absolutely and locally uniformly on $\mathcal{H}$. Expanding this further to recover $F(\tau)$, we have:

$$\begin{aligned} F(\tau) &= \exp(2\pi iM\tau/h) \sum_{n=-\infty}^{\infty} \mathbf{c}_n e^{2\pi i\tau n/h} \\ &= \left(\sum_{j\in[N]} e^{2\pi i\alpha_j\tau/h} \mathbf{v}_j \mathbf{u}_j^* \right) \sum_{n=-\infty}^{\infty} \mathbf{c}_n e^{2\pi i\tau n/h} \text{ (substituting in Equation (23))} \\ &= \sum_{n=-\infty}^{\infty} \sum_{j\in[N]} e^{2\pi i(n+\alpha_j)\tau/h} (\mathbf{u}_j^* \mathbf{c}_n) \mathbf{v}_j \end{aligned}$$

Definition 10.25 (Holomorphic at ∞)**.** *We say that F is holomorphic at the cusp at ∞ if and only if we have $\mathbf{c}_n = 0$ for all $n < 0$. In this event, we define:*

$$\operatorname{ord}_\infty(F) := \min(n + \alpha_j : \mathbf{u}_j^* \mathbf{c}_n \neq 0). \tag{25}$$

We take the convention $\operatorname{ord}_\infty(0) = +\infty$.

Remark 10.26. *Two conventions make Definition 10.25 well-posed. First, because each α_j lies in $[0, 1)$, the exponent $n + \alpha_j$ determines both n (its integer part) and α_j (its fractional part), so the minimum ranges over an unambiguous set of exponents. Second, nothing depends on how the eigenbasis was chosen: for each eigenvalue $e^{2\pi i\alpha}$ of A, the spectral projector $P_\alpha := \sum_{j:\alpha_j=\alpha} \mathbf{v}_j \mathbf{u}_j^*$ depends only on A (it is the projection onto $\ker(A - e^{2\pi i\alpha} I)$ along the sum of the other eigenspaces); hence so does $M = \sum_\alpha \alpha P_\alpha$ and thus the coefficients $\mathbf{c}_n$; and $\mathbf{u}_j^* \mathbf{c}_n \neq 0$ holds for some j with $\alpha_j = \alpha$ if and only if $P_\alpha \mathbf{c}_n \neq 0$.*

10.4.3 Arbitrary Cusps

Next, we characterize what it means for F to be holomorphic at cusps other than ∞. The idea is to act with an element of $\mathrm{SL}_2(\mathbb{Z})$ that sends ∞ to this cusp and then adapt our methods from Section 10.4.2 (Figure 2 shows such elements for a few cusps). To this end, consider a cusp $\mathfrak{a} \in \mathbb{Q}$. Using Definition 10.18, we know there must exist a group element $g_\mathfrak{a} \in \mathrm{SL}_2(\mathbb{Z})$ such that $g_\mathfrak{a} * \infty = \mathfrak{a}$. Accordingly, define:

$$F_\mathfrak{a}(\tau) := G(g_\mathfrak{a}, \tau)^{-k} F(g_\mathfrak{a} * \tau). \tag{26}$$

In addition to sending the cusp at ∞ to $\mathfrak{a}$, this transformation gives us a function that is weakly modular of weight k with respect to the *conjugate* subgroup $\Gamma_\mathfrak{a} := g_\mathfrak{a}^{-1} \Gamma g_\mathfrak{a}$ and the conjugate representation $\rho_\mathfrak{a} : g_\mathfrak{a}^{-1} g g_\mathfrak{a} \mapsto \rho(g)$. ($\Gamma_\mathfrak{a}$ is again a congruence subgroup of the same level: $\Gamma(q)$ is normal in $\mathrm{SL}_2(\mathbb{Z})$, being the kernel of the reduction map of Lemma 10.5, so $\Gamma_\mathfrak{a} \supseteq g_\mathfrak{a}^{-1} \Gamma(q) g_\mathfrak{a} = \Gamma(q)$. Moreover, $\rho_\mathfrak{a}$ is again a representation of finite order, since its image coincides with that of ρ; so the results of Section 10.4.2 apply to $F_\mathfrak{a}$.) To see this, note that:

$$\begin{aligned} F_\mathfrak{a}(g_\mathfrak{a}^{-1} g g_\mathfrak{a} * \tau) &= G(g_\mathfrak{a}, g_\mathfrak{a}^{-1} g g_\mathfrak{a} * \tau)^{-k} F(g g_\mathfrak{a} * \tau) \text{ (by Equation (26))} \\ &= G(g, g_\mathfrak{a} * \tau)^k G(g_\mathfrak{a}, g_\mathfrak{a}^{-1} g g_\mathfrak{a} * \tau)^{-k} \rho(g) F(g_\mathfrak{a} * \tau) \text{ (by Definition 10.19)} \end{aligned}$$

$$= G(g, g_{\mathfrak{a}} * \tau)^k G(g_{\mathfrak{a}}, g_{\mathfrak{a}}^{-1} g g_{\mathfrak{a}} * \tau)^{-k} G(g_{\mathfrak{a}}, \tau)^k \rho(g) F_{\mathfrak{a}}(\tau)$$
$$= G(g_{\mathfrak{a}}^{-1} g g_{\mathfrak{a}}, \tau)^k \rho(g) F_{\mathfrak{a}}(\tau),$$

as claimed. The last step above can be seen from two applications of Proposition 10.17:

$$G(g_{\mathfrak{a}}, g_{\mathfrak{a}}^{-1} g g_{\mathfrak{a}} * \tau) G(g_{\mathfrak{a}}^{-1} g g_{\mathfrak{a}}, \tau) = G(g g_{\mathfrak{a}}, \tau) = G(g, g_{\mathfrak{a}} * \tau) G(g_{\mathfrak{a}}, \tau).$$

With this transformation, the definition of holomorphy at the cusp at $\mathfrak{a}$ is automatic:

Definition 10.27. *The width of the cusp of F at $\mathfrak{a}$ for Γ is defined to be the width of the cusp of $F_{\mathfrak{a}}$ at ∞ for $\Gamma_{\mathfrak{a}}$ (more directly, it is the smallest integer $h > 0$ such that $g_{\mathfrak{a}} \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} g_{\mathfrak{a}}^{-1} \in \Gamma$). F is holomorphic at the cusp $\mathfrak{a}$ if and only if $F_{\mathfrak{a}}$ is holomorphic at ∞. Finally, we define* $\operatorname{ord}_{\mathfrak{a}}(F) = \operatorname{ord}_{\infty}(F_{\mathfrak{a}})$.

Lemma 10.28 (Well-definedness of cusp data)**.** *The width of the cusp at $\mathfrak{a}$, whether F is holomorphic at $\mathfrak{a}$, and $\operatorname{ord}_{\mathfrak{a}}(F)$ do not depend on the choice of $g_{\mathfrak{a}} \in \mathrm{SL}_2(\mathbb{Z})$ with $g_{\mathfrak{a}} * \infty = \mathfrak{a}$. Moreover, if $\mathfrak{b} = \gamma * \mathfrak{a}$ for some $\gamma \in \Gamma$, then the width of the cusp at $\mathfrak{b}$ equals that at $\mathfrak{a}$, and F is holomorphic at $\mathfrak{b}$ if and only if it is holomorphic at $\mathfrak{a}$, in which case* $\operatorname{ord}_{\mathfrak{b}}(F) = \operatorname{ord}_{\mathfrak{a}}(F)$.

Proof. See Appendix C. For scalar forms, the statement about orders is [Ran77, Theorem 4.1.2(i)]. □

Accordingly, we will regard cusps in the same Γ-orbit as equivalent, and speak of the width, holomorphy, and order of F at an equivalence class of cusps.

Finally, we can define a modular form:

Definition 10.29. *Let k be a non-negative integer. A weakly modular $F : \mathcal{H} \to \mathbb{C}^N$ of weight k for Γ with representation ρ is referred to as a* modular form *if it is holomorphic at all cusps.*

10.4.4 Composition Laws

The following lemma turns a vector-valued modular form into a scalar one by substituting it into a polynomial that transforms by a character under the representation. This is also how the classical correspondence between invariant polynomials, such as weight enumerators of self-dual codes, and modular forms arises: by substituting theta functions into them [BE72, NRS06]. In our proofs, it is the step labeled "scalarization via P_β" in Figures 8 and 13, drawn in teal in Figure 3(b). We include a proof for completeness.

Lemma 10.30 (Scalarization)**.** *Consider a modular form $F : \mathcal{H} \to \mathbb{C}^N$ of weight k for Γ with representation $\rho : \Gamma \to \mathrm{GL}_N(\mathbb{C})$. Suppose we also have a homogeneous polynomial $P : \mathbb{C}^N \to \mathbb{C}$ of degree D, not the zero polynomial, such that $P(\rho(g)\mathbf{v}) \propto P(\mathbf{v})$ as functions of $\mathbf{v} \in \mathbb{C}^N$, for every $g \in \Gamma$; that is, for each $g \in \Gamma$ there is a nonzero constant $\chi(g)$ (uniquely determined, as $P \neq 0$) with*

$$P(\rho(g)\mathbf{v}) = \chi(g) P(\mathbf{v}) \qquad \text{for all } \mathbf{v} \in \mathbb{C}^N. \tag{27}$$

Then $\chi : \Gamma \to \mathbb{C}^\times$ is a one-dimensional representation with $\chi^j = 1$ for some positive integer j, and the composed function $f := P \circ F$ is a scalar modular form of weight kD for Γ with representation χ.

Proof. Fix $\mathbf{v}_0 \in \mathbb{C}^N$ with $P(\mathbf{v}_0) \neq 0$. For $g_1, g_2 \in \Gamma$, applying (27) twice gives $\chi(g_1 g_2) P(\mathbf{v}_0) = P(\rho(g_1)\rho(g_2)\mathbf{v}_0) = \chi(g_1)\chi(g_2) P(\mathbf{v}_0)$, so χ is a one-dimensional representation. Moreover, every value of χ is a root of unity: for $g \in \Gamma$, if $\rho(g)^n = I$ (recall that ρ has finite order), then $\chi(g)^n P(\mathbf{v}_0) = P(\rho(g)^n \mathbf{v}_0) = P(\mathbf{v}_0)$, so $\chi(g)^n = 1$. Hence $\chi^j = 1$ for some positive integer j, by Lemma 10.20.

Next, $f = P \circ F$ is holomorphic on $\mathcal{H}$ (a polynomial composed with a holomorphic function), and it is weakly modular: for any $\tau \in \mathcal{H}$ and $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \Gamma$, homogeneity of P gives

$$f(g * \tau) = P(F(g * \tau)) = P\left((c\tau + d_0)^k \rho(g) F(\tau)\right) = (c\tau + d_0)^{kD} P(\rho(g)F(\tau)) = (c\tau + d_0)^{kD} \chi(g) f(\tau).$$

Finally, f is holomorphic at every cusp. The idea is that P turns the cusp expansion of F into one for f, and the covariance (27) keeps the exponents in the right residue class; the verification is routine and deferred to Appendix C. □

Lemma 10.31 (Products [Ran77, Section 4.3])**.** *Let $f_1, f_2 : \mathcal{H} \to \mathbb{C}$ be scalar modular forms for Γ of weights k_1, k_2 with one-dimensional representations χ_1, χ_2. Then $f_1 f_2$ is a scalar modular form of weight $k_1 + k_2$ for Γ with representation $\chi_1\chi_2$, and at every cusp $\mathfrak{a}$ and every $\tau_0 \in \mathcal{H}$:*

$$\operatorname{ord}_{\mathfrak{a}}(f_1 f_2) = \operatorname{ord}_{\mathfrak{a}}(f_1) + \operatorname{ord}_{\mathfrak{a}}(f_2), \qquad \operatorname{ord}_{\tau_0}(f_1 f_2) = \operatorname{ord}_{\tau_0}(f_1) + \operatorname{ord}_{\tau_0}(f_2),$$

where $\operatorname{ord}_{\tau_0}$ denotes the order of vanishing at an interior point (Definition 10.33 below); if either factor is identically zero, both equations hold in $[0, +\infty]$ with the convention $\operatorname{ord}(0) = +\infty$.

Lemma 10.32 (Restriction)**.** *Let $F : \mathcal{H} \to \mathbb{C}^N$ be a modular form of weight k for Γ with representation ρ, and let $\Gamma' \subseteq \Gamma$ be a congruence subgroup of level q. Then F is a modular form of weight k for Γ' with representation $\rho|_{\Gamma'}$.*

Proof. See Appendix C. □

10.4.5 The Valence Formula

The engine of our lower-bound arguments is the valence formula: a nonzero modular form has a fixed total number of zeros — counted over the Γ-equivalence classes of points of $\mathcal{H}$ and of cusps, with appropriate weights — determined only by its weight and the index of Γ. It plays the role that the degree bound plays for univariate polynomials, as in the technical overview (Section 2.3.3): the weight k is a degree-like parameter, and a smaller group Γ imposes fewer constraints and so allows more zeros. We first define the order of vanishing at interior points.

Definition 10.33. *For a holomorphic $f : \mathcal{H} \to \mathbb{C}$ that is not identically zero and a point $\tau_0 \in \mathcal{H}$, let $\operatorname{ord}_{\tau_0}(f)$ denote the order of vanishing of f at τ_0, i.e., the index of the first nonzero coefficient in the Taylor expansion of f at τ_0. We take the convention $\operatorname{ord}_{\tau_0}(0) = +\infty$.*

Remark 10.34. *If $f \not\equiv 0$ is weakly modular of weight k for Γ with a one-dimensional representation χ, then $\operatorname{ord}_{\tau_0}(f)$ depends only on the Γ-orbit of τ_0: for $g \in \Gamma$, the function $\tau \mapsto f(g * \tau) = (c\tau + d_0)^k \chi(g) f(\tau)$ differs from f by a nonvanishing holomorphic factor, and $\tau \mapsto g * \tau$ is a biholomorphism of $\mathcal{H}$, so f vanishes at $g * \tau_0$ to the same order as at τ_0. We also note that the stabilizer $\{g \in \Gamma : g * \tau_0 = \tau_0\}$ of any point of $\mathcal{H}$ is finite (see, e.g., [DS05, Chapter 2]).*

We now state the valence formula, in the form in which we will use it. It is classical, also for nontrivial one-dimensional representations (multiplier systems), and we cite it.

Theorem 10.35 (Valence formula [Ran77, Theorem 4.1.4])**.** *Let Γ be a congruence subgroup of level q containing $-I$, and let $f : \mathcal{H} \to \mathbb{C}$ be a modular form of weight k for Γ with a one-dimensional representation*

χ satisfying $\chi^j = 1$ for some positive integer j, with f not identically zero. Then f has finitely many Γ-equivalence classes of zeros in $\mathcal{H}$, and:

$$\sum_{cusps\ \mathfrak{a}} \frac{\mathrm{ord}_{\mathfrak{a}}(f)}{2} + \sum_{\tau \in \mathcal{H}/\Gamma} \frac{\mathrm{ord}_{\tau}(f)}{|\{g \in \Gamma \,:\, g * \tau = \tau\}|} = \frac{k}{24}\,[\mathrm{SL}_2(\mathbb{Z}) \,:\, \Gamma]\,. \tag{28}$$

Here the first sum ranges over the Γ-equivalence classes of cusps (well-defined by Lemma 10.28), with orders as in Definitions 10.25 and 10.27, and the second sum ranges over the Γ-equivalence classes of zeros of f in $\mathcal{H}$ (well-defined by Remark 10.34). (At each application of this theorem we will verify, or explicitly defer, the hypothesis that some power of χ is trivial. Note also that taking $g = -I$ in the transformation law gives $f(\tau) = (-1)^k \chi(-I) f(\tau)$, so necessarily $\chi(-I) = (-1)^k$.)

Corollary 10.36 (Valence formula for $\Gamma(q)$)**.** *Let $f : \mathcal{H} \to \mathbb{C}$ be a modular form of weight k for $\Gamma(q)$ with a one-dimensional representation χ satisfying $\chi^j = 1$ for some positive integer j, with f not identically zero. Then:*

$$\sum_{cusps\ \mathfrak{c}} \frac{\mathrm{ord}_{\mathfrak{c}}(f)}{2} + \sum_{\tau \in \mathcal{H}/\Gamma(q)} \frac{\mathrm{ord}_{\tau}(f)}{2\,|\{g \in \Gamma(q) \,:\, g * \tau = \tau\}|} = \frac{k}{24} \cdot \frac{q^3 - q}{2}, \tag{29}$$

where the first sum ranges over the $\Gamma(q)$-equivalence classes of cusps, with widths and orders computed with respect to $\Gamma(q)$, and the second over the $\Gamma(q)$-equivalence classes of zeros of f in $\mathcal{H}$.

Proof sketch. By extending f from $\Gamma(q)$ to $\pm\Gamma(q) = \Gamma(q) \cup (-I)\Gamma(q)$ and applying Theorem 10.35 there. □

11 Warm-Up

This section proves our warm-up result, Corollary 11.13: for primes $q \equiv 3 \pmod 4$, a cone-delta function on $\mathbb{Z}_q^d$ with Fourier sparsity $q^{(\eta_q + o(1))d}$. It already produces the objects behind q-server PIR (Main Theorem 5.3), and its proof introduces, in the simplest setting, every tool the stronger construction of Section 12 will reuse. The plan: after an overview of the proof (Section 11.1) and the setup of even lattices (Section 11.2), we state the general result (Theorem 11.9), its consequence for Fourier sparsity, which holds for every lattice the theorem allows (Corollary 11.10), and its concrete rank-2 instantiation (Corollary 11.13); the support-size bounds behind both corollaries are proved in Appendix D.

The ingredients of this argument are classical: the transformation law of lattice theta functions under the Weil representation and the valence formula. What is new is their use to show that sparse cone-delta functions exist (see the start of Part II). Its conclusion is not new for $q = 3$: the instantiation of Appendix D uses the A_2 lattice, the set S_d of Theorem 11.9 is the Hamming ball of radius $d/3$ (as $w_{\mathcal{L}}$ takes the value 1 at both nonzero residues, by Lemma D.2), and a cone-delta function with Fourier support in this ball is also given by an elementary polynomial in the proof of [AL07, Theorem 2.2], specialized to the zero vector.

11.1 Overview of the Proof

This subsection assembles the proof of Theorem 11.9 and its corollaries from the ideas of the technical overview (Sections 2.3.1 to 2.3.3); we give the details in the rest of this section.

The dual object. By duality (Step 1, Section 11.4), it suffices to rule out a *witness*: a vector $(\beta_{\mathbf{x}})_{\mathbf{x} \in Q}$, indexed by the points of the cone Q, with

$$\beta_{\mathbf{0}} = 1 \qquad \text{and} \qquad \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \zeta_q^{\langle \mathbf{a}, \mathbf{x} \rangle} = 0 \quad \text{for all } \mathbf{a} \in S_d \tag{30}$$

(these are (36) and (37) in Step 1). As with the polynomial $P_{\mathbf{c}}$ of the technical overview (Section 2.3.1), we package the witness as the homogeneous polynomial $P_\beta(\mathbf{v}) := \sum_{\mathbf{x}\in Q} \beta_{\mathbf{x}}\, v_{x_1} v_{x_2} \cdots v_{x_d}$ of degree d on $\mathbb{C}^q$. The conditions on β become three statements about P_β:

1. $P_\beta(\mathbf{e}_0) = \beta_{\mathbf{0}} = 1$, since only the term $\mathbf{x} = \mathbf{0}$ survives.
2. $P_\beta(\rho_T \mathbf{v}) = P_\beta(\mathbf{v})$ for every $\mathbf{v}$: substituting $\rho_T \mathbf{v}$ multiplies the term for $\mathbf{x}$ by $\zeta_q^{\mu\langle \mathbf{x},\mathbf{x}\rangle}$, which is 1 on the cone.
3. $P_\beta(\rho_S \mathbf{v})$, as a polynomial in $\mathbf{v}$, has no monomial $v_{z_1} v_{z_2} \cdots v_{z_d}$ with $\mathbf{z} \in S_d$: up to the normalizing constant of ρ_S, the coefficient of this monomial is a sum of the sums $\sum_{\mathbf{x}\in Q} \beta_{\mathbf{x}} \zeta_q^{\langle -2\mu \mathbf{z}', \mathbf{x}\rangle}$ over the rearrangements $\mathbf{z}'$ of $\mathbf{z}$, which vanish by (30) as long as S_d is closed under permuting coordinates and under multiplying by nonzero residues (recall that $\mu \not\equiv 0 \pmod q$, by Proposition 11.7).

   Looking ahead, our choice of S_d is precisely to ensure a large vanishing order of $P_\beta \circ \Theta_{\mathcal{L}}$ at the cusp at 0 via this observation.

The first and third statements are the analogues of $\Delta(\mathbf{x}) = 1$ and of Δ vanishing at the queried points in Reed–Muller PIR (Table 4). We now evaluate P_β along the theta function $\Theta_{\mathcal{L}}$ of an even lattice $\mathcal{L}$ of rank r and determinant $\sqrt{q}$, which is a vector-valued modular form for the Weil representation [Bor98, Theorem 4.1].

Scalarization. By the second statement, $P_\beta \circ \Theta_{\mathcal{L}}$ is invariant under $\tau \mapsto \tau + 1$. Every other $g \in \Gamma_0(q)$ acts on $\mathbb{C}^q$, up to a scalar, by rescaling the indices $x \mapsto d_0 x$ and multiplying by quadratic phases (Fact 10.11). The phases equal 1 on the cone, as in the second statement. To absorb the rescaling, we replace P_β by its average over the rescalings $x \mapsto ux$ for $u \in \mathbb{Z}_q^*$ (Section 11.5); the average still satisfies the three statements, except that its value at $\mathbf{e}_0$ is $q - 1$. With this change, $P_\beta(\rho(g)\mathbf{v}) \propto P_\beta(\mathbf{v})$ for all $g \in \Gamma_0(q)$ (Lemma 11.20), so $f := P_\beta \circ \Theta_{\mathcal{L}}$ is a scalar modular form of weight $dr/2$ for $\Gamma_0(q)$. Substituting theta functions into an invariant polynomial in this way is also how the classical correspondence between weight enumerators of codes and theta series of lattices arises [BE72, NRS06].

The cusp at ∞. As $\Im\tau \to \infty$, we have $\Theta_{\mathcal{L}}(\tau) \to \mathbf{e}_0$, so $f(\tau) \to P_\beta(\mathbf{e}_0) \neq 0$ (Lemma 11.24). In particular, f is not identically zero.

The cusp at 0. To expand f at the cusp 0, we move it to ∞: by the transformation of $\Theta_{\mathcal{L}}$ under S and the homogeneity of P_β,

$$f_0(\tau) := \tau^{-dr/2} f(-1/\tau) \propto P_\beta\big(\rho_S\, \Theta_{\mathcal{L}}(\tau)\big).$$

Write $\Theta_{\mathcal{L}}(\tau) = \sum_{t\in\mathbb{Z}_q} \theta_t(\tau)\, \mathbf{e}_t$, where $\theta_t(\tau) := \sum_{\mathbf{y}\in t\mathbf{a}+\mathcal{L}} e^{\pi i\langle \mathbf{y},\mathbf{y}\rangle\tau}$. By the third statement, when we expand $P_\beta(\rho_S \Theta_{\mathcal{L}}(\tau))$ as a polynomial in the $\theta_t(\tau)$, only the monomials $\theta_{t_1}(\tau)\cdots\theta_{t_d}(\tau)$ with $\mathbf{t} \notin S_d$ survive. Each of them is a power series in $e^{2\pi i\tau/q}$:

$$\theta_{t_1}(\tau)\cdots\theta_{t_d}(\tau) = \sum_{\substack{\mathbf{y}_1,\ldots,\mathbf{y}_d:\\ \mathbf{y}_i\in t_i\mathbf{a}+\mathcal{L}}} e^{2\pi i n\tau/q}, \qquad \text{where } n := \frac{q}{2}\sum_{i=1}^{d}\langle \mathbf{y}_i,\mathbf{y}_i\rangle \geq \sum_{i=1}^{d} w_{\mathcal{L}}(t_i\mathbf{a}),$$

and $w_{\mathcal{L}}(x\mathbf{a}) := \frac{q}{2}\min_{\mathbf{y}\in x\mathbf{a}+\mathcal{L}}\langle \mathbf{y},\mathbf{y}\rangle$ (Definition 11.5). Each exponent n is a nonnegative integer (Proposition 11.4). We call $w_{\mathcal{L}}(x\mathbf{a})$ the *cost* of $x \in \mathbb{Z}_q$. Now take S_d to be the set of $\mathbf{t} \in \mathbb{Z}_q^d$ with $\sum_i w_{\mathcal{L}}(ut_i\mathbf{a}) \leq Bd$ for some $u \in \mathbb{Z}_q^*$; the union over u makes S_d closed under multiplying by nonzero residues, as the third statement requires. Every $\mathbf{t} \notin S_d$ then has $\sum_i w_{\mathcal{L}}(t_i\mathbf{a}) > Bd$, so every surviving exponent exceeds Bd, and f vanishes to order more than Bd at the cusp 0 (Section 11.6).

Counting roots. By the valence formula (Theorem 10.35; see, e.g., [Ran77, Theorem 4.1.4]), the nonzero modular form f, of weight $dr/2$ for $\Gamma_0(q)$, has total order of vanishing $\frac{dr(q+1)}{24}$. All orders are nonnegative, so its order at the cusp 0 is at most $\frac{dr(q+1)}{24}$. Taking $B = \frac{r(q+1)}{24}$ contradicts the previous paragraph. Hence no witness exists, and there is a cone-delta function with Fourier support in S_d (Theorem 11.9). Table 4, in the introduction to this part, summarizes the analogy with the Reed–Muller PIR argument of Section 2.3.2, and Figure 8 shows how the pieces fit together, in the format we reuse for the stronger construction (Figure 13).

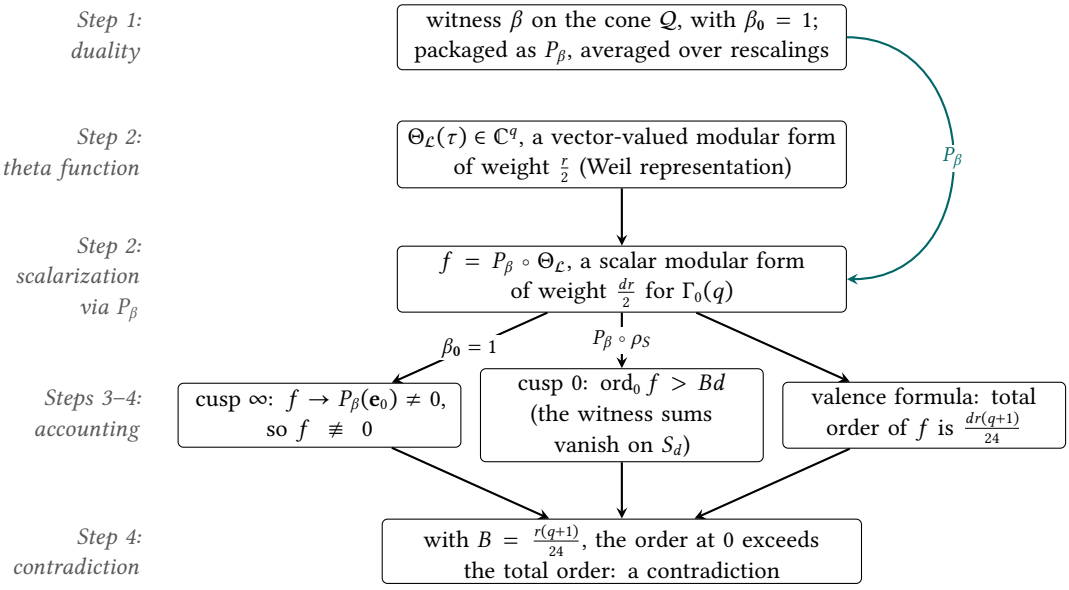


Figure 8: How the proof of Theorem 11.9 fits together, one row per layer of the argument (numbered by the warm-up's steps). The witness and the theta function are built independently, and the witness enters through the polynomial P_β (teal), which turns $\Theta_{\mathcal{L}}$ into a single scalar modular form f. Each property of the witness plays a role: the cone (with the averaging over rescalings) makes f modular for $\Gamma_0(q)$; the normalization $\beta_{\mathbf{0}} = 1$ makes f nonzero at the cusp ∞ (left); and the orthogonality to S_d, which reaches the cusp 0 through $P_\beta \circ \rho_S$, makes f vanish there to order more than Bd (middle), while the valence formula fixes its total order (right). Looking ahead, the stronger construction has a chart of the same shape (Figure 13), in which this chain reappears as the right-hand branch, run at $s \cdot \frac{q-1}{2}$ cusps instead of one.

The size of the support. It remains to bound $|S_d|$, which is a union of $q-1$ sets of the form $\{\mathbf{z} \in \mathbb{Z}_q^d : \sum_i w(z_i) \leq Bd\}$ for a cost function w on $\mathbb{Z}_q$. A moment-generating-function (Markov) bound controls the size of each such set (Lemma D.1), and optimizing over its parameter gives $|S_d| \leq q^{(\eta_{\mathcal{L}} + o(1))d}$ for an explicit exponent $\eta_{\mathcal{L}}$ (Corollary 11.10). For the warm-up, we use a lattice of rank $r = 2$, which exists exactly when $q \equiv 3 \pmod 4$ (Remark 11.12). Its cost function is $w_{\mathcal{L}}(x\mathbf{a}) = x^2$ for $|x| \leq (q+1)/4$ (Lemma D.2), and the budget is $B = (q+1)/12$ per coordinate. Heuristically, S_d is then close to the set of integer vectors in a Euclidean ball of radius about $\sqrt{qd/12}$, which has $q^{d/2 + O(d/\ln q)}$ points. The precise bound is $\eta_q \leq \frac{1}{2} + \frac{0.467 + o(1)}{\log_2 q}$ (Corollary 11.13).

11.2 Lattices in $\mathbb{R}^r$

Definition 11.1 (Lattices, determinants)**.** *Let r be an even positive integer. For a basis matrix $\mathbf{B} \in \mathbb{R}^{r\times r}$ of full rank, we associate to it the lattice $\mathcal{L}(\mathbf{B}) = \mathbf{B}\mathbb{Z}^r \subseteq \mathbb{R}^r$. We define the dual lattice $\mathcal{L}^*$ to be $\mathbf{B}^{-\top}\mathbb{Z}^r$; equivalently, it can be defined by:*

$$\mathcal{L}^* = \{\mathbf{x} \in \mathbb{R}^r \,:\, \langle \mathbf{x}, \mathbf{y}\rangle \in \mathbb{Z} \quad \forall \mathbf{y} \in \mathcal{L}\}.$$

The determinant of $\mathcal{L}$ is defined to be $|\det \mathbf{B}|$, which we will use interchangeably with $\det \mathcal{L}$. We can also associate to a lattice its Gram matrix $\mathbf{G} = \mathbf{B}^\top\mathbf{B}$.

Definition 11.2. *A lattice $\mathcal{L}$ is said to be* even *if for all $\mathbf{x} \in \mathcal{L}$ we have $\langle \mathbf{x}, \mathbf{x}\rangle \in 2\mathbb{Z}$.*

Proposition 11.3. *If $\mathcal{L}$ is even, then we have $\mathcal{L} \subseteq \mathcal{L}^*$. (Note that the converse need not be true e.g., take $\mathcal{L} = \mathbb{Z}^r$.) Moreover, we have $|\mathcal{L}^*/\mathcal{L}| = (\det \mathcal{L})^2$.*

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathcal{L}$, we have:

$$\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{2}\langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle - \frac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle - \frac{1}{2}\langle \mathbf{y}, \mathbf{y}\rangle \in \mathbb{Z}.$$

This implies that $\mathcal{L} \subseteq \mathcal{L}^*$. Secondly, we have $|\mathcal{L}^*/\mathcal{L}| = \det \mathcal{L}/\det \mathcal{L}^* = (\det \mathcal{L})^2$. □

Proposition 11.4. *Let q be an odd prime. Suppose $\mathcal{L}$ is even with $\det \mathcal{L} = \sqrt{q}$. Then:*

- *For any $\mathbf{x} \in \mathcal{L}^*$ and $\mathbf{y} \in \mathcal{L}$, we have $\frac{1}{2}\langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle \equiv \frac{1}{2}\langle \mathbf{x}, \mathbf{x}\rangle \pmod 1$.*
- *For any $\mathbf{x} \in \mathcal{L}^*$, we have $\langle \mathbf{x}, \mathbf{x}\rangle \in \frac{2}{q}\mathbb{Z}$.*

Proof. See Appendix D. □

Definition 11.5. *Let q be an odd prime and let $\mathcal{L} \subseteq \mathbb{R}^r$ be an even lattice of determinant $\sqrt{q}$. For a coset $\mathbf{a} + \mathcal{L} \in \mathcal{L}^*/\mathcal{L}$, we define:*

$$w_{\mathcal{L}}(\mathbf{a}) = w_{\mathcal{L}}(\mathbf{a}+\mathcal{L}) = \frac{q}{2}\min_{\mathbf{x}\in\mathbf{a}+\mathcal{L}} \langle \mathbf{x}, \mathbf{x}\rangle. \tag{31}$$

(The minimum is attained because every ball contains only finitely many points of the coset, as in the proof of Lemma B.7.) By Proposition 11.4, we have $w_{\mathcal{L}}(\mathbf{a}) \in \mathbb{Z}_{\geq 0}$.

Definition 11.6. *Let q be an odd prime and let $\mathcal{L} \subseteq \mathbb{R}^r$ be an even lattice of determinant $\sqrt{q}$. The group $\mathcal{L}^*/\mathcal{L}$ then has order $(\det \mathcal{L})^2 = q$ (Proposition 11.3), so it is cyclic. Let $\mathbf{a} + \mathcal{L}$ be an arbitrary generator of the group (so $\mathbf{a} \in \mathcal{L}^*$).*

For an input $x \in \mathbb{Z}_q$, define $\phi(x) \in \mathbb{R}/\mathbb{Z}$ as follows: take an arbitrary representative $\mathbf{y} \in x\mathbf{a} + \mathcal{L}$, and output $\frac{1}{2}\langle \mathbf{y}, \mathbf{y}\rangle \bmod 1$. By Proposition 11.4, $\phi(x)$ is well-defined (i.e., it does not depend on the chosen representative $\mathbf{y}$), and by taking $\mathbf{y} = x\mathbf{a}$, we can write

$$\phi(x) = \frac{1}{2}\langle \mathbf{a}, \mathbf{a}\rangle x^2 \bmod 1 = \frac{\mu x^2}{q} \bmod 1,$$

where $\mu := \frac{q}{2}\langle \mathbf{a}, \mathbf{a}\rangle$ is an integer by Proposition 11.4. We call μ the phase coefficient *associated to $\mathcal{L}$ and $\mathbf{a}$.*

Similarly, for inputs $x, y \in \mathbb{Z}_q$, we define $\psi(x, y) \in \mathbb{R}/\mathbb{Z}$ as follows: take arbitrary representatives $\mathbf{x} \in x\mathbf{a} + \mathcal{L}, \mathbf{y} \in y\mathbf{a} + \mathcal{L}$, and output $\langle \mathbf{x}, \mathbf{y}\rangle \bmod 1$. This is well-defined by definition of the dual lattice and again we can take:

$$\psi(x, y) = \langle \mathbf{a}, \mathbf{a}\rangle xy \bmod 1 = \frac{2\mu xy}{q} \bmod 1.$$

Proposition 11.7. *In the setting of Definition 11.6, the phase coefficient satisfies* $\mu \not\equiv 0 \pmod q$.

Proof. See Appendix D. □

The phase coefficient is further constrained by the rank of the lattice, via a classical formula of Milgram; we will use this to pin down the Gauss-sum constant in the theta transformation law.

Fact 11.8 (Milgram's formula, specialized [MH73, Appendix 4, p. 127])**.** *In the setting of Definition 11.6, the Gauss-sum constant of the phase coefficient* μ *— defined by Fact 10.9, which applies since* $\mu \not\equiv 0 \pmod q$ *by Proposition 11.7 — satisfies*

$$\epsilon_{\mu,q} = (-i)^{r/2}.$$

(Indeed, since $t \mapsto t\mathbf{a} + \mathcal{L}$ *enumerates* $\mathcal{L}^*/\mathcal{L}$ *and every* $\mathbf{y} \in t\mathbf{a} + \mathcal{L}$ *has* $\frac{1}{2}\langle \mathbf{y}, \mathbf{y}\rangle \equiv \phi(t) \equiv \mu t^2/q \pmod 1$*, we have* $\sum_{t\in\mathbb{Z}_q} \zeta_q^{\mu t^2} = \sum_{\mathbf{y}+\mathcal{L}\in\mathcal{L}^*/\mathcal{L}} e^{\pi i\langle \mathbf{y},\mathbf{y}\rangle}$*. Milgram's formula evaluates the latter sum, for any even lattice, as* $\sqrt{|\mathcal{L}^*/\mathcal{L}|}\, e^{2\pi i\sigma/8}$*, where* σ *is the signature of the inner product; our form is positive definite, so* $\sigma = r$ *and the sum is* $\sqrt{q}\, i^{r/2}$*. Comparing with* $\sum_t \zeta_q^{\mu t^2} = \epsilon_{\mu,q}^{-1}\sqrt{q}$ *from Fact 10.9 gives* $\epsilon_{\mu,q} = i^{-r/2} = (-i)^{r/2}$*, an ordinary integer power since* r *is even.)*

11.3 Main Result

The main theorem of this section is:

Theorem 11.9. *Let* q *be an odd prime, and let* $\mathcal{L} \subseteq \mathbb{R}^r$ *be an even lattice of determinant* $\sqrt{q}$*. The group* $\mathcal{L}^*/\mathcal{L}$ *has order* q*, so it is cyclic with some generator* $\mathbf{a} + \mathcal{L}$ *for a generator* $\mathbf{a} \in \mathcal{L}^*$*. Let* $d > 0$ *be an arbitrary positive integer, and define the set:*

$$S_d = \bigcup_{u\in\mathbb{Z}_q\setminus\{0\}} \left\{ \mathbf{z} \in \mathbb{Z}_q^d \,:\, \sum_{i=1}^{d} w_{\mathcal{L}}(uz_i\mathbf{a}) \le \frac{r(q+1)}{24} d \right\}. \tag{32}$$

Then there exists a cone-delta function $f : \mathbb{Z}_q^d \to \mathbb{Q}(\zeta_q)$ *such that* $\hat{f}$ *is supported on* S_d*.*

(Throughout this part, we use the notions of Section 4.2 over the field $\mathbb{Q}(\zeta_q)$, with $\gamma = \zeta_q$.)

To turn Theorem 11.9 into a bound on Fourier sparsity, it remains to bound $|S_d|$. The only information about the lattice that this bound uses is the distribution of the costs $w_{\mathcal{L}}$ over the q cosets of $\mathcal{L}$ in $\mathcal{L}^*$, which we record in the generating function

$$M_{\mathcal{L}}(t) := \sum_{\mathbf{y}+\mathcal{L}\in\mathcal{L}^*/\mathcal{L}} t^{w_{\mathcal{L}}(\mathbf{y})} \qquad (0 < t \le 1). \tag{33}$$

Corollary 11.10. *Let* q *be an odd prime, and let* $\mathcal{L} \subseteq \mathbb{R}^r$ *be an even lattice of determinant* $\sqrt{q}$*. For* d *a growing parameter, the set* S_d *of Theorem 11.9 has size* $|S_d| \le q^{(\eta_{\mathcal{L}}+o(1))d}$*; hence there exists a cone-delta function* $f : \mathbb{Z}_q^d \to \mathbb{Q}(\zeta_q)$ *whose Fourier support is contained in* S_d*, and so has Fourier sparsity at most* $q^{(\eta_{\mathcal{L}}+o(1))d}$*, where*

$$\eta_{\mathcal{L}} := \frac{1}{\ln q} \inf_{0<t\le 1} \left(\ln M_{\mathcal{L}}(t) - \frac{r(q+1)}{24} \ln t \right). \tag{34}$$

Note that $\eta_{\mathcal{L}}$ depends only on the lattice $\mathcal{L}$, and not on the choice of the generator $\mathbf{a} + \mathcal{L}$ in Theorem 11.9.

We instantiate this with the following lattice, for which $M_{\mathcal{L}}$, and hence $\eta_{\mathcal{L}}$, can be computed explicitly.

Definition 11.11 (Concrete instantiation of Theorem 11.9)**.** *Assume q is a prime $\equiv 3 \pmod 4$. Let $r = 2$ and $\mathbf{B} := \begin{bmatrix} \sqrt{2} & 1/\sqrt{2} \\ 0 & \sqrt{q/2} \end{bmatrix}$. One can then verify $\mathbf{G} = \mathbf{B}^\top \mathbf{B} = \begin{bmatrix} 2 & 1 \\ 1 & (q+1)/2 \end{bmatrix}$. This has determinant q and moreover it is even on the diagonal and integral in all entries so the resulting lattice $\mathcal{L}(\mathbf{B})$ is even and has determinant $\sqrt{q}$. We choose our generator $\mathbf{a} + \mathcal{L}$ of $\mathcal{L}^*/\mathcal{L}$ to have representative $\mathbf{a} = (0, \sqrt{2/q})$.*

Remark 11.12. *The condition $q \equiv 3 \pmod 4$ in Definition 11.11 is necessary: an even lattice $\mathcal{L} \subseteq \mathbb{R}^2$ of determinant $\sqrt{q}$ exists if and only if $q \equiv 3 \pmod 4$. Indeed, as $\mathcal{L}$ is even, it is contained in $\mathcal{L}^*$ (Proposition 11.3), so its Gram matrix has integer entries and an even diagonal: it has the form $\begin{bmatrix} 2a & b \\ b & 2c \end{bmatrix}$ for integers a, b, c. Its determinant is $(\det \mathcal{L})^2 = q$, so $q = 4ac - b^2 \equiv -b^2 \pmod 4$. As q is odd, b is odd, so $q \equiv -1 \pmod 4$. Conversely, for $q \equiv 3 \pmod 4$, Definition 11.11 gives such a lattice.*

This gives the following; here and below, $K := (q+1)/4$ (an integer for the primes $q \equiv 3 \pmod 4$ considered).

Corollary 11.13. *For any fixed prime $q \equiv 3 \pmod 4$ and d a growing parameter, there exists a cone-delta function $f : \mathbb{Z}_q^d \to \mathbb{Q}(\zeta_q)$ with Fourier sparsity at most $q^{(\eta_q + o(1))d}$, where*

$$\eta_q := \frac{1}{\ln q} \inf_{0<t\le 1} \ln\left(t^{-(q+1)/12} \left(1 + 2\sum_{u=1}^{K} t^{u^2} + 2\sum_{u=K+1}^{(q-1)/2} t^{u^2 - qu + qK} \right) \right). \tag{35}$$

In particular, $\eta_3 \le 0.789691$, $\eta_7 \le 0.709410$, and $\eta_{11} \le 0.660807$. As $q \to \infty$ along primes $q \equiv 3 \pmod 4$, we have $\eta_q \le \frac{1}{2} + \frac{C_{\text{warmup}} + o(1)}{\ln q} \le \frac{1}{2} + \frac{0.467 + o(1)}{\log_2 q}$, where $C_{\text{warmup}} \le 0.3234$ is the constant of Lemma D.3.

We prove Corollaries 11.10 and 11.13 in Appendix D: the first by a Chernoff bound on the number of frequency vectors of small total cost, and the second by computing the costs of the cosets of the lattice of Definition 11.11.

Other lattices. Corollary 11.10 is a modular piece of our approach: it accepts any even lattice $\mathcal{L}$ of determinant $\sqrt{q}$, and through Main Theorem 5.3 every such lattice yields a q-server PIR scheme with exponent $\eta_{\mathcal{L}}$. A lattice with a smaller $\eta_{\mathcal{L}}$ therefore immediately gives a better scheme, with no change to the rest of the argument. By (34), $\eta_{\mathcal{L}}$ is small when the shortest vectors in the nonzero cosets of $\mathcal{L}$ in $\mathcal{L}^*$ are long relative to the rank r, which sets the budget $r(q+1)/24$. The lattice of Definition 11.11 is one choice, whose costs we can compute in closed form for every prime $q \equiv 3 \pmod 4$. Other lattices also cover primes that it does not. For instance, for $q = 5$, where no such lattice of rank 2 exists (Remark 11.12), consider the root lattice $A_4 = \{\mathbf{x} \in \mathbb{Z}^5 : \sum_i x_i = 0\}$, of rank 4. It is even with determinant $\sqrt{5}$, and its nonzero cosets in A_4^* consist of the vectors with coordinate sum 0 whose coordinates are all congruent to $\frac{c}{5}$ modulo 1, for $c \in \{1, 2, 3, 4\}$. The shortest such vector takes the value $\frac{c-5}{5}$ in c coordinates and $\frac{c}{5}$ in the other $5 - c$, so it has squared length $\frac{c(5-c)}{5}$. The costs are therefore $w_{A_4} = 2$ on two cosets and 3 on the other two, so $M_{A_4}(t) = 1 + 2t^2 + 2t^3$, and the infimum in (34), attained at $t = \frac{1}{2}$, gives $\eta_{A_4} = \log_5 \frac{7}{2} < 0.7784$. This yields 5-server PIR schemes with a smaller exponent than the $q = 3$ case of Corollary 11.13 (0.7897, which is what the Track 2 column of Table 1 lists for five servers), though not smaller than those of Tracks 1 and 3. Whether some lattice improves on Definition 11.11, or even on the stronger construction of Section 12, is open; we discuss this in Section 2.4.

11.4 Proof Step 1: Duality

Now we begin attacking Theorem 11.9. Unwinding the definitions, the theorem asserts the solvability of a finite linear system. Indeed, a function with $\hat{f}$ supported on S_d is exactly a combination $f(\mathbf{x}) =$

$\sum_{\mathbf{z}\in S_d} \hat{f}(\mathbf{z})\,\zeta_q^{\langle \mathbf{x},\mathbf{z}\rangle}$, and being a cone-delta function constrains its values only on the cone $\mathcal{Q}$: the value must be 1 at $\mathbf{x} = \mathbf{0}$ and 0 at the other points of $\mathcal{Q}$. So, with the following vectors indexed by elements $\mathbf{x} \in \mathcal{Q}$:

- the delta vector, which is 1 when $\mathbf{x} = 0$ and 0 everywhere else; and
- for every $\mathbf{z} \in S_d$, the vector $\mathbf{v_z}$ with entry $\zeta_q^{\langle \mathbf{x},\mathbf{z}\rangle}$ in index $\mathbf{x}$,

Theorem 11.9 asserts precisely that the vectors $\mathbf{v_z}$ span the delta vector, with coefficients (namely, the Fourier coefficients $\hat{f}(\mathbf{z})$) in the field $\mathbb{Q}(\zeta_q)$.

Suppose for the sake of contradiction that they do not. Two pieces of finite-dimensional linear algebra convert this assumption into the dual object we will work with for the rest of the proof. First, *the base field does not matter*: the vectors in question have entries in $\mathbb{Q}(\zeta_q)$, and the solvability of a finite linear system over a field is determined by a comparison of matrix ranks, which is unchanged under field extension; so the delta vector is not in the span of the $\mathbf{v_z}$ even with coefficients in $\mathbb{C}$. Second, *duality*: if a vector of $\mathbb{C}^{\mathcal{Q}}$ lies outside a subspace, then some linear functional vanishes on the subspace but not on that vector (complete a basis of the subspace by the vector, and further to a basis of $\mathbb{C}^{\mathcal{Q}}$, and take the corresponding dual-basis functional). Writing such a functional as $\mathbf{w} \mapsto \sum_{\mathbf{x}\in\mathcal{Q}} \beta_{\mathbf{x}} w_{\mathbf{x}}$ for a coefficient vector $\beta \in \mathbb{C}^{\mathcal{Q}}$, normalized so that its value on the delta vector is 1, we have:

$$\beta_0 = 1 \tag{36}$$

$$\sum_{\mathbf{x}\in\mathcal{Q}} \beta_{\mathbf{x}} \zeta_q^{\langle \mathbf{x},\mathbf{z}\rangle} = 0 \quad \forall \mathbf{z} \in S_d. \tag{37}$$

Going forward, our goal is to show that the above two equations imply a contradiction.

11.5 Proof Step 2: Vector and Scalar Theta Modular Forms

In the following, we build on the notation in Sections 10.2 and 11.2. In particular, let $\mathbf{a} \in \mathcal{L}^*$ be an arbitrary generator of $\mathcal{L}^*/\mathcal{L}$. Throughout the proof, $\mu \in \mathbb{Z}$ denotes the phase coefficient associated to $\mathcal{L}$ and $\mathbf{a}$ in Definition 11.6, and ρ, ρ_S, ρ_T denote the Weil representation of Definition 10.10 and Fact 10.11 with this parameter μ. (Recall that Definition 10.10 requires $\mu \not\equiv 0 \pmod q$, which Proposition 11.7 guarantees.)

Definition 11.14. *Define* $\Theta_{\mathcal{L}} : \mathcal{H} \to \mathbb{C}^q$ *by:*

$$\Theta_{\mathcal{L}}(\tau) = \sum_{x\in\mathbb{Z}_q} \left(\sum_{\mathbf{y}\in x\mathbf{a}+\mathcal{L}} \exp\left(\pi i \langle \mathbf{y},\mathbf{y}\rangle \tau\right) \right) \mathbf{e}_x. \tag{38}$$

We want to show that $\Theta_{\mathcal{L}}$ is a vector-valued modular form of weight $r/2$ for $\mathrm{SL}_2(\mathbb{Z})$ with the Weil representation ρ (note $r/2 \in \mathbb{Z}$, as r is even). The modularity of lattice theta functions with respect to the Weil representation is classical: for a positive definite even lattice of even rank r, it is the special case $p = 1$, $\alpha = \beta = 0$ of [Bor98, Theorem 4.1]. (There, the scalar prefactor of $\rho(S)$ is written as $\sqrt{i}^{\,-r} = (-i)^{r/2}$, which equals our $\epsilon_{\mu,q}$ by Milgram's formula, Fact 11.8.) Our setting nevertheless admits a short self-contained verification, which we give. We proceed in steps: convergence and holomorphy on $\mathcal{H}$ (Lemma 11.15), the transformation laws under the generators T and S (Lemmas 11.16 and 11.17), and holomorphy at the cusps together with the behavior at ∞ (Lemma 11.18).

Lemma 11.15. *The sum in* (38) *is well-defined at every* τ *and* $\Theta_{\mathcal{L}}(\tau)$ *is holomorphic on* $\mathcal{H}$.

Proof. See Appendix D. □

Lemma 11.16 ([Bor98, Theorem 4.1])**.** *For any $\tau \in \mathcal{H}$, we have:*

$$\Theta_{\mathcal{L}}(\tau + 1) = \rho_T \Theta_{\mathcal{L}}(\tau),$$

for ρ_T defined in Definition 10.10. In other words, noting that $G(T, \tau) = 1$, we have:

$$\Theta_{\mathcal{L}}(T * \tau) = G(T, \tau)^{r/2} \rho(T) \Theta_{\mathcal{L}}(\tau).$$

Proof. See Appendix D. □

Lemma 11.17 ([Bor98, Theorem 4.1])**.** *For any $\tau \in \mathcal{H}$, we have:*

$$\Theta_{\mathcal{L}}(-1/\tau) = \tau^{r/2} \rho_S \Theta_{\mathcal{L}}(\tau),$$

for ρ_S defined in Definition 10.10. In other words, noting that $G(S, \tau) = \tau$, we have:

$$\Theta_{\mathcal{L}}(S * \tau) = G(S, \tau)^{r/2} \rho(S) \Theta_{\mathcal{L}}(\tau).$$

Proof. See Appendix D. □

Lemmas 11.16 and 11.17 together with Proposition 10.22 imply that $\Theta_{\mathcal{L}}$ is a weakly modular function of weight $r/2$ for $\mathrm{SL}_2(\mathbb{Z})$ with the Weil representation ρ of Fact 10.11. It remains to address the cusps:

Lemma 11.18. *$\Theta_{\mathcal{L}}$ is holomorphic at every cusp, and is therefore a modular form of weight $r/2$ for $\mathrm{SL}_2(\mathbb{Z})$ with representation ρ. Moreover, $\mathrm{ord}_\infty(\Theta_{\mathcal{L}}) = 0$; more precisely, we have*

$$\lim_{\Im \tau \to \infty} \Theta_{\mathcal{L}}(\tau) = \mathbf{e}_0,$$

uniformly in $\Re \tau$.

Proof. See Appendix D. □

Having constructed the vector-valued modular form, we next scalarize it, by combining it with the vector β from Step 1. The polynomial below is engineered to meet three constraints at once. Its coefficients are the entries of β, arranged so that evaluating it on vectors of qth roots of unity produces sums of exactly the shape that (37) controls; this is what will let us convert the duality conditions into vanishing Fourier coefficients in Step 3. It is homogeneous of degree d, so composing it with the weight-$r/2$ form $\Theta_{\mathcal{L}}$ will produce a scalar form of integer weight $dr/2$. And it is averaged over $u \in \mathbb{Z}_q^*$: under the Weil action of $\Gamma_0(q)$, the coordinates of $\mathbb{C}^q$ are rescaled in index by the (invertible) diagonal entry d_0 and multiplied by quadratic phases (Fact 10.11); the average absorbs the index rescaling, and the restriction of the support to the cone Q makes the quadratic phases equal to 1 — which is why the composition will transform by a mere character.

Definition 11.19. *Define the polynomial $P_\beta : \mathbb{C}^q \to \mathbb{C}$ as follows:*

$$P_\beta(X_0, \ldots, X_{q-1}) := \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \prod_{i=1}^{d} X_{ux_i \bmod q}.$$

By construction, P_β is homogeneous with degree d.

Next, we will show that P_β satisfies (27) by appealing to the explicit formula of Fact 10.11. Recall the definition of the subgroup $\Gamma_0(q)$ of $\mathrm{SL}_2(\mathbb{Z})$ from Definition 10.3.

Lemma 11.20. *For any $g \in \Gamma_0(q)$, we have $P_\beta(\rho(g)\mathbf{v}) \propto P_\beta(\mathbf{v})$ as functions of $\mathbf{v} \in \mathbb{C}^q$.*

Proof. See Appendix D. □

Consequently, by Lemmas 10.30, 10.32 and 11.18 ($\Theta_{\mathcal{L}}$ is a modular form for $\mathrm{SL}_2(\mathbb{Z})$; restrict it to $\Gamma_0(q)$ and compose with P_β, which is not the zero polynomial since $P_\beta(\mathbf{e}_0) = (q-1)\beta_0 \neq 0$, as computed in Lemma 11.24 below), the composed function $P_\beta \circ \Theta_{\mathcal{L}}$ is a scalar modular form of weight $dr/2$ for $\Gamma_0(q)$, with the one-dimensional representation χ defined by (27) for $P = P_\beta$; moreover, $\chi^j = 1$ for some positive integer j. We will also need the classification of the cusp classes of $\Gamma_0(q)$:

Proposition 11.21 ([DS05, Section 3.8])**.** *Under the action of $\Gamma_0(q)$, the cusps $\mathbb{Q} \cup \{\infty\}$ split into two equivalence classes:*

- *The ∞ class contains ∞ and all $a/b \in \mathbb{Q}$ where $\gcd(a, b) = 1$ and b is divisible by q.*
- *The 0 class contains all $a/b \in \mathbb{Q}$ where $\gcd(a, b) = 1$ and b is not divisible by q.*

11.6 Proof Step 3: The Order of the Cusp at 0

This step is where we use the duality of Step 1. We will expand $P_\beta \circ \Theta_{\mathcal{L}}$ at the cusp at 0 and track its coefficients: each term of the expansion comes from a tuple of dual-lattice vectors, and the coefficient multiplying it is an average of sums of exactly the shape of (37), evaluated at frequency vectors built from the tuple's residue classes. The frequencies in S_d are the ones on which the sums in (37) vanish; and a frequency outside S_d has, by the very definition of S_d, a large total cost $\sum_i w_{\mathcal{L}}(t_i \mathbf{a})$ — which lower-bounds the Gaussian decay exponent of the corresponding terms. In this way the exclusion of S_d from the support becomes a strict lower bound on the order of vanishing at the cusp.

Next, we analyze the behavior of $P_\beta \circ \Theta_{\mathcal{L}}$ at the cusp at 0. To do this, we follow the procedure in Section 10.4.3 and start by choosing a group element $g \in \mathrm{SL}_2(\mathbb{Z})$ such that $g * \infty = 0$. We may simply take $g = S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Now, following (26), we define:

$$f_0(\tau) := \tau^{-dr/2} P_\beta(\Theta_{\mathcal{L}}(S * \tau)) = \tau^{-dr/2} P_\beta(\Theta_{\mathcal{L}}(-1/\tau)). \tag{39}$$

This is modular of weight $dr/2$ with respect to the conjugate subgroup $S^{-1}\Gamma_0(q)S$ and the conjugate representation $\overline{\chi}$ defined by $\overline{\chi}(S^{-1}gS) = \chi(g)$. Our goal is to compute $\mathrm{ord}_\infty(f_0)$. To do this, we follow the procedure in Section 10.4.2:

Proposition 11.22. *The width of the cusp at ∞ with respect to $S^{-1}\Gamma_0(q)S$ (equivalently, of the cusp at 0 for $\Gamma_0(q)$) is q.*

Proof. The width is the minimal h such that $T^h \in S^{-1}\Gamma_0(q)S \Leftrightarrow ST^hS^{-1} \in \Gamma_0(q)$. To this end, we compute:

$$\begin{aligned} ST^hS^{-1} &= \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -h & 1 \end{bmatrix}, \end{aligned}$$

which belongs to $\Gamma_0(q)$ if and only if $q \mid h$. Consequently, the width of the cusp is q as claimed. □

Lemma 11.23. *We have* $\operatorname{ord}_0(P_\beta \circ \Theta_{\mathcal{L}}) = \operatorname{ord}_\infty(f_0) > r(q+1)d/24$.

Proof. Now continuing, we have:

$$\overline{\chi}\left(\begin{bmatrix} 1 & q \\ 0 & 1 \end{bmatrix}\right) = \chi\left(\begin{bmatrix} 1 & 0 \\ -q & 1 \end{bmatrix}\right) = 1,$$

because $\begin{bmatrix} 1 & 0 \\ -q & 1 \end{bmatrix} \in \Gamma(q)$, on which ρ is trivial (Fact 10.11): for this element, (27) reads $P_\beta(\mathbf{v}) = \chi\left(\begin{bmatrix} 1 & 0 \\ -q & 1 \end{bmatrix}\right) P_\beta(\mathbf{v})$ for all $\mathbf{v}$, and $P_\beta \neq 0$. Plugging this into (24), we may write:

$$f_0(\tau) = \sum_{n=-\infty}^{\infty} c_n \exp(2\pi i \tau n/q),$$

and our goal is to show that $c_n = 0$ for all $n \leq r(q+1)d/24$. (In particular, this covers all negative n, so the computation below also establishes that f_0 is holomorphic at ∞, i.e., that $P_\beta \circ \Theta_{\mathcal{L}}$ is holomorphic at the cusp at 0.) To do this, let us compute:

$$\begin{aligned}
f_0(\tau) &= \tau^{-dr/2} P_\beta(\Theta_{\mathcal{L}}(-1/\tau)) \\
&= \tau^{-dr/2} P_\beta(\tau^{r/2} \rho_S \Theta_{\mathcal{L}}(\tau)) \quad \text{(Lemma 11.17)} \\
&= P_\beta(\rho_S \Theta_{\mathcal{L}}(\tau)) \quad (P_\beta \text{ is homogeneous of degree } d) \\
&= P_\beta\left(\sum_{t \in \mathbb{Z}_q} \left(\sum_{\mathbf{y} \in t\mathbf{a}+\mathcal{L}} \exp\left(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau\right)\right) \rho_S \mathbf{e}_t\right) \\
&\propto P_\beta\left(\sum_{z \in \mathbb{Z}_q} \left(\sum_{t \in \mathbb{Z}_q} \sum_{\mathbf{y} \in t\mathbf{a}+\mathcal{L}} \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau)\, \zeta_q^{-2\mu t z}\right) \mathbf{e}_z\right) \quad \text{(Definition 10.10)} \\
&= \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \prod_{i=1}^{d} \left(\sum_{t \in \mathbb{Z}_q} \sum_{\mathbf{y} \in t\mathbf{a}+\mathcal{L}} \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau)\, \zeta_q^{-2\mu t u x_i}\right) \\
&= \sum_{\mathbf{t} \in \mathbb{Z}_q^d} \sum_{\substack{\mathbf{y}_1, \ldots, \mathbf{y}_d : \\ \mathbf{y}_i \in t_i \mathbf{a}+\mathcal{L}}} \exp\left(\pi i \sum_{i=1}^{d} \langle \mathbf{y}_i, \mathbf{y}_i \rangle \tau\right) \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}}\, \zeta_q^{-2\mu u \langle \mathbf{t}, \mathbf{x} \rangle}. \qquad (40)
\end{aligned}$$

In the third step from the end, we substituted the definition of ρ_S and pulled its scalar prefactor, which does not depend on t, out of P_β by homogeneity; so f_0 is a fixed nonzero multiple of (40), which does not affect which of its coefficients vanish. In the last step, we expanded the product of the d inner series and moved the (finite) sums over u and $\mathbf{x}$ inside; this rearrangement is justified because each inner series converges absolutely (Lemma B.7, exactly as in the proof of Lemma 11.15), so the product expands into an absolutely convergent sum over tuples. In the display, $\langle \mathbf{t}, \mathbf{x} \rangle := \sum_{i=1}^{d} t_i x_i \in \mathbb{Z}_q$.

We claim that the inner double sum in (40) vanishes whenever $\mathbf{t} \in S_d$. First, observe that S_d is closed under coordinatewise multiplication by any scalar $c \in \mathbb{Z}_q^*$: if $\mathbf{z}$ satisfies the defining condition of S_d for the index u', then $c\mathbf{z}$ satisfies it for the index $u'c^{-1}$, since $w_{\mathcal{L}}\left((u'c^{-1})(cz_i)\mathbf{a}\right) = w_{\mathcal{L}}(u' z_i \mathbf{a})$. Now, for each fixed $u \in \mathbb{Z}_q^*$, the sum over $\mathbf{x}$ in (40) is $\sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}}\, \zeta_q^{\langle \mathbf{x}, -2\mu u \mathbf{t} \rangle}$, which vanishes by (37) whenever $-2\mu u \mathbf{t} \in S_d$; and since $-2\mu u \in \mathbb{Z}_q^*$ (here $\mu \not\equiv 0 \pmod{q}$ by Proposition 11.7, $u \neq 0$, and q is odd), membership $\mathbf{t} \in S_d$ implies $-2\mu u \mathbf{t} \in S_d$ by the closure just observed. This proves the claim.

So the only tuples $(\mathbf{t}, \mathbf{y}_1, \dots, \mathbf{y}_d)$ that contribute nontrivially to (40) have $\mathbf{t} \notin S_d$, which — taking $u = 1$ in the definition of S_d — means:

$$\frac{r(q+1)}{24} d < \sum_{i=1}^{d} w_{\mathcal{L}}(t_i \mathbf{a}) = \frac{q}{2} \sum_{i=1}^{d} \min_{\mathbf{x} \in t_i \mathbf{a} + \mathcal{L}} \langle \mathbf{x}, \mathbf{x} \rangle \le \frac{q}{2} \sum_{i=1}^{d} \langle \mathbf{y}_i, \mathbf{y}_i \rangle,$$

where the final inequality uses $\mathbf{y}_i \in t_i \mathbf{a} + \mathcal{L}$. Each $\langle \mathbf{y}_i, \mathbf{y}_i \rangle$ lies in $\frac{2}{q}\mathbb{Z}$ by Proposition 11.4, and is nonnegative, so each contributing term of (40) equals its coefficient times $\exp(2\pi i n \tau / q)$ for the *integer* $n := \frac{q}{2} \sum_{i=1}^{d} \langle \mathbf{y}_i, \mathbf{y}_i \rangle > r(q+1)d/24$. Collecting the terms of (40) by the value of n (justified by absolute convergence, exactly as in the proof of Lemma 10.30) and invoking the uniqueness in Fact B.4, we conclude that $c_n = 0$ for every $n \le r(q+1)d/24$. That is, $\mathrm{ord}_0(P_\beta \circ \Theta_{\mathcal{L}}) = \mathrm{ord}_\infty(f_0) > r(q+1)d/24$, as desired. □

11.7 Proof Step 4: Root-Counting Contradiction from the Valence Formula

For the coup de grace, we will apply Theorem 10.35 to $P_\beta \circ \Theta_{\mathcal{L}}$ to obtain a contradiction. To do this, we must first verify that $P_\beta \circ \Theta_{\mathcal{L}}$ is not identically 0. This will necessarily make use of (36) (it is the only thing preventing all the $\beta_{\mathbf{x}}$'s from being 0).

Lemma 11.24. *As $\Im\tau \to \infty$, we have $P_\beta(\Theta_{\mathcal{L}}(\tau)) \to (q-1)\beta_0 \neq 0$. Consequently, $P_\beta \circ \Theta_{\mathcal{L}}$ is not identically 0.*

Proof. By Lemma 11.18, $\Theta_{\mathcal{L}}(\tau) \to \mathbf{e}_0$ as $\Im\tau \to \infty$, so by continuity of the polynomial P_β we get $P_\beta(\Theta_{\mathcal{L}}(\tau)) \to P_\beta(\mathbf{e}_0)$. Finally, evaluating P_β at $\mathbf{e}_0$ keeps exactly the terms with $ux_i \equiv 0 \pmod q$ for all i, i.e., with $\mathbf{x} = \mathbf{0}$ (as u is invertible); so $P_\beta(\mathbf{e}_0) = \sum_{u \in \mathbb{Z}_q^*} \beta_0 = (q-1)\beta_0 \neq 0$ by (36). □

Finally, we can finish the proof of Theorem 11.9 by invoking the valence formula:

Proof of Theorem 11.9. We have shown that $P_\beta \circ \Theta_{\mathcal{L}}$ is a nonzero modular form of weight $rd/2$ that has order $> r(q+1)d/24$ at the cusp at 0. On the other hand, Theorem 10.35 applies (we have $-I \in \Gamma_0(q)$, and the representation χ satisfies $\chi^j = 1$ for some positive integer j, by Lemma 10.30); dropping the nonnegative contributions of the interior orders and of the cusp class at ∞ (by Proposition 11.21, the classes of ∞ and 0 are the only two), it implies that:

$$\begin{aligned} \frac{\mathrm{ord}_0(P_\beta \circ \Theta_{\mathcal{L}})}{2} &\le \frac{rd}{48} [\mathrm{SL}_2(\mathbb{Z}) : \Gamma_0(q)] \\ &= \frac{rd(q+1)}{48} \quad \text{(Lemma 10.5)}, \end{aligned}$$

which is a direct contradiction, establishing the theorem. □

12 Stronger Construction

This section proves our main result, Theorem 12.2: sparse quadratic-phase delta expansions, which power s-server PIR for s below q, with the support exponents computed in Appendix E. The proof follows the warm-up, with two new ingredients: an elliptic variable in the theta function, and a second scalarization; Section 12.1 explains both, and Table 5 compares the two proofs step by step.

The new ingredients are again classical in isolation. The components of the Jacobi theta function are, up to normalization, theta functions with rational characteristics [Mum83, Chapter I, §3]; composing it with the witness polynomial yields, up to roots of unity, a Jacobi form of weight $d/2$ (in the variable $z = qw$) in

	Warm-up (Section 11)	Stronger construction (Section 12)
Goal	cone-delta function (q servers)	quadratic-phase delta expansion with $s \le q$ phases (s servers)
Witness	β on the cone $\mathcal{Q} = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{x}\rangle \equiv 0 \pmod q\}$, with $\beta_{\mathbf{0}} = 1$	β on $\{\mathbf{x} : \sum_i x_i \equiv 0\}$, invariant under $\mathbf{x} \mapsto \mathbf{x} + \mathbf{1}$, with $\beta_{\mathbf{0}} = 1$
Theta function	lattice theta function $\Theta_{\mathcal{L}}(\tau) \in \mathbb{C}^q$, of weight $\frac{r}{2}$	Jacobi theta function $\Theta(\tau, w) \in \mathbb{C}^q$, of weight $\frac{1}{2}$
Symmetries absorbed by P_β	$\rho(g)$ for $g \in \Gamma_0(q)$ (after averaging)	the Heisenberg operators X and Z
Group	$\Gamma_0(q)$, with 2 cusp classes	$\Gamma(q)$, with $\frac{q^2-1}{2}$ cusp classes
Cusps where we force high vanishing order	the cusp 0	the $s \cdot \frac{q-1}{2}$ selected classes $\mathfrak{c}_{a,\lambda}$
Scalar modular form(s)	$P_\beta \circ \Theta_{\mathcal{L}}$, of weight $\frac{dr}{2}$	A_m^2 and B^2, of weights $d + 2m$ and $3d$
Costs	shortest squared lengths of the cosets c of $\mathcal{L}$ in $\mathcal{L}^*$, scaled: $w_{\mathcal{L}}(c) = \frac{q}{2} \min_{\mathbf{y} \in c} \langle \mathbf{y}, \mathbf{y}\rangle$ (Definition 11.5)	square cost $w_q(j) = \min_{k \equiv j} k^2$ and tilted edge cost $w_{q,\vartheta}(j) = \min_{k \equiv j}(k(k-1) + 2\vartheta k)$ (Definition 5.4)
Root counting	one valence inequality	two valence inequalities and two per-cusp bounds
Support	vectors $\mathbf{z}$ of total cost $\sum_i w_{\mathcal{L}}(z_i \mathbf{a})$ at most $\frac{r(q+1)}{24} d$, and their scalar multiples (Theorem 11.9)	vectors $\mathbf{z}$ of total square cost $\sum_i w_q(z_i)$ at most $\nu\, d$ or total tilted edge cost $\sum_i w_{q,\vartheta}(z_i)$ at most $\gamma\, d$; their images under $\mathbf{z} \mapsto u\mathbf{z} + b\mathbf{1}$; and these with one coordinate changed arbitrarily (Definition 12.1)

Table 5: The warm-up and the stronger construction side by side.

the sense of [EZ85] (Lemmas 12.19 and 12.20); and the two scalar modular forms of Step 3 are the squares of, respectively, its first nonvanishing Taylor coefficient in the elliptic variable [EZ85, Section 3] and the product of its restrictions to the nonzero 2-torsion points [EZ85, Theorem 1.3]. Bounding orders of vanishing by combining such scalar forms with the valence formula is the strategy of [RWR15, Aok22, LW26] (see the discussion after Lemma 12.42). What is new here is the argument that ties the cusp orders to the witness: the selected cusp classes, the per-class quantities $\ell_{\mathfrak{c}}$ and $H_{\mathfrak{c}}$, and the analysis of the edge polynomials in Step 4.

12.1 Overview of the Proof

This subsection explains how the proof of Theorem 12.2 extends the warm-up (Section 11.1); we give the details in the rest of this section. Table 5 summarizes the changes, and Figure 13 at the end of this subsection shows how the final inequalities fit together.

The target and the witness. A quadratic-phase delta expansion (Definition 4.3) with phases in a set $\mathcal{A} \subseteq \mathbb{Z}_q$ of size s expands the delta function at $\mathbf{0}$ in the characters $\mathbf{x} \mapsto \zeta_q^{\langle \mathbf{z}, \mathbf{x}\rangle + a\langle \mathbf{x}, \mathbf{x}\rangle}$, for $\mathbf{z}$ in the support and $a \in \mathcal{A}$. Using all q phases recovers cone-delta functions: if f is a cone-delta function, then $\frac{1}{q} \sum_{a \in \mathbb{Z}_q} \zeta_q^{a\langle \mathbf{x}, \mathbf{x}\rangle} f(\mathbf{x})$ is

the delta function at $\mathbf{0}$, since the average over a is the indicator function of the cone. With $s < q$ phases, this average is no longer available. Dually (Step 1, Section 12.3), the witness α now satisfies $\alpha_{\mathbf{0}} = 1$ and

$$\sum_{\mathbf{x}\in\mathbb{Z}_q^d} \alpha_{\mathbf{x}} \zeta_q^{a\langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{z}\rangle} = 0 \qquad \text{for all } a \in \mathcal{A} \text{ and } \mathbf{z} \text{ in the support,}$$

and it need not be supported on the cone. We symmetrize it to a witness β that is supported on $\{\mathbf{x} : \sum_i x_i \equiv 0\}$ and invariant under $\mathbf{x} \mapsto \mathbf{x} + \mathbf{1}$ (Proposition 12.7), for a reason that will become clear below, and as in the warm-up we package β as the homogeneous polynomial $P_\beta(\mathbf{v}) := \sum_{\mathbf{x}} \beta_{\mathbf{x}} v_{x_1} v_{x_2} \cdots v_{x_d}$ of degree d.

Where the phases go: the selected cusps. Since β is no longer supported on the cone, P_β is not invariant under ρ_T, and the modularity for $\Gamma_0(q)$ used in the warm-up is lost. We instead work with the smaller subgroup $\Gamma(q) = \{g \in \mathrm{SL}_2(\mathbb{Z}) : g \equiv I \pmod q\}$, on which the Weil representation is trivial (Fact 10.11; [Str13, Lemma 5.12]), so that modularity comes for free. In exchange, $\Gamma(q)$ has $\frac{q^2-1}{2}$ classes of cusps rather than two, and we must keep track of all of them. The class of a cusp is determined by its numerator and denominator in lowest terms (with $\infty = \frac{1}{0}$), taken modulo q and up to a common sign, and every nonzero point of $\mathbb{Z}_q^2$, up to sign, arises in this way (Proposition 12.23; see, e.g., [DS05, Section 3.8]). So the cusp classes lie on the $q+1$ lines through the origin of $\mathbb{Z}_q^2$, with $\frac{q-1}{2}$ classes on each line. We label the line spanned by $(2a, 1)$ by $a \in \mathbb{Z}_q$, and the remaining line, spanned by $(1, 0)$, by ∞. A cusp class is then specified by its label and a multiplier $\lambda \in \{1, \dots, \frac{q-1}{2}\}$, as the class $\pm\lambda(2a, 1)$ or $\pm\lambda(1, 0)$ (Figure 9); for instance, the cusps $\infty = \frac{1}{0}$ and $0 = \frac{0}{1}$ have labels ∞ and 0, both with multiplier 1. As in the warm-up, the behavior at a cusp $g * \infty$ is governed by $P_\beta \circ \rho(g)$. The phase $a \in \mathcal{A}$ becomes visible at the cusp classes $\mathfrak{c}_{a,\lambda} = \pm\lambda(2a, 1)$ with label a, for $\lambda \in \{1, \dots, \frac{q-1}{2}\}$. For a suitable $g_{a,\lambda}$ sending ∞ into this class, $\rho(g_{a,\lambda})$ is a rescaled Fourier transform followed by multiplication by the quadratic phase $\zeta_q^{ay^2}$ (50), so the coefficient of the monomial $v_{x_1} \cdots v_{x_d}$ in $P_\beta \circ \rho(g_{a,\lambda})$ is a constant multiple of

$$\sum_{\mathbf{y}\in\mathbb{Z}_q^d} \beta_{\mathbf{y}}\, \zeta_q^{a\langle \mathbf{y},\mathbf{y}\rangle - \lambda^{-1}\langle \mathbf{x},\mathbf{y}\rangle}.$$

This is a witness sum for the phase a, so it vanishes whenever $-\lambda^{-1}\mathbf{x}$ lies in the support. Each phase $a \in \mathcal{A}$ thus controls the $\frac{q-1}{2}$ cusp classes with label a, and the witness is visible at $s \cdot \frac{q-1}{2}$ *selected* classes in all. In the warm-up, the cone made all phases agree, and a single cusp sufficed: the cusp $0 = S * \infty$, which has label 0.

label	$a=0$	$a=1$	$a=2$	$a=3$	$a=4$	$a=5$	$a=6$	∞
$\lambda=1$	$\boxed{\pm(0,1)}$	$\pm(2,1)$	$\pm(-3,1)$	$\pm(-1,1)$	$\pm(1,1)$	$\pm(3,1)$	$\pm(-2,1)$	$\boxed{\pm(1,0)}$
$\lambda=2$	$\pm(0,2)$	$\pm(-3,2)$	$\pm(1,2)$	$\pm(-2,2)$	$\pm(2,2)$	$\pm(-1,2)$	$\pm(3,2)$	$\pm(2,0)$
$\lambda=3$	$\pm(0,3)$	$\pm(-1,3)$	$\pm(-2,3)$	$\pm(-3,3)$	$\pm(3,3)$	$\pm(2,3)$	$\pm(1,3)$	$\pm(3,0)$

selected classes $\mathfrak{c}_{a,\lambda}$, for $a \in \mathcal{A} = \{0, 1, 2\}$

Figure 9: The $\frac{q^2-1}{2} = 24$ cusp classes of $\Gamma(q)$ for $q = 7$, as nonzero points of $\mathbb{Z}_q^2$ up to sign (with coordinates in $\{-3, \dots, 3\}$), arranged by label and multiplier: the class with label $a \in \mathbb{Z}_q$ and multiplier λ is $\pm\lambda(2a, 1)$, and the class with label ∞ and multiplier λ is $\pm\lambda(1, 0)$. For $\mathcal{A} = \{0, 1, 2\}$, the selected classes are the $s \cdot \frac{q-1}{2} = 9$ entries in the shaded columns. The boxed entries are the warm-up's cusps $0 = \frac{0}{1}$ and $\infty = \frac{1}{0}$.

Tori and torsion points. Recall from the technical overview that $\tau \in \mathcal{H}$ stands for the lattice $\mathbb{Z}\tau + \mathbb{Z}$ in $\mathbb{C}$, up to rotation and scaling (Figure 1). The quotient $\mathbb{C}/(\mathbb{Z}\tau + \mathbb{Z})$ is a *torus*: we draw it as the parallelogram

with vertices 0, 1, τ and $1+\tau$, with opposite sides glued (Figure 12). For an integer $N \geq 1$, its *N-torsion points* are the points w with $Nw \in \mathbb{Z}\tau + \mathbb{Z}$, that is, $w \equiv \frac{a\tau+b}{N}$ for $a, b \in \{0, \dots, N-1\}$; there are N^2 of them, forming a grid in the parallelogram. Two kinds will matter. Translations by the q-torsion points act on our new theta function by the Heisenberg operators (below), and the second scalarization evaluates at the three nonzero 2-torsion points $\frac{1}{2}$, $\frac{\tau}{2}$ and $\frac{1+\tau}{2}$.

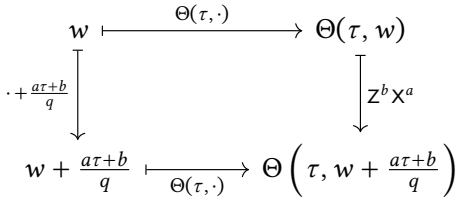
The elliptic variable. Our new theta function has an additional *elliptic* variable $w \in \mathbb{C}$, which we think of as a point of the torus; up to normalization, its coordinates are theta functions with rational characteristics [Mum83, Chapter I, §3]:

$$\Theta(\tau, w) := \sum_{k \in \mathbb{Z}} (-1)^k \exp\left(\frac{\pi i \tau}{q} \left(k - \frac{1}{2}\right)^2 + 2\pi i \left(k - \frac{1}{2}\right) w \right) \mathbf{e}_{k+(q-1)/2 \bmod q} \in \mathbb{C}^q.$$

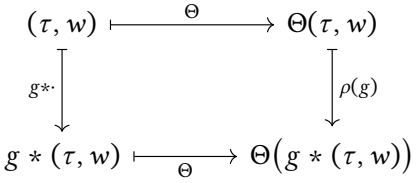
It has two symmetries (Figure 10). First, as in the warm-up, $\mathrm{SL}_2(\mathbb{Z})$ acts, and Θ transforms by the Weil representation (now with $\mu = \frac{q+1}{2}$), up to an explicit factor (Lemma 12.16). The action is now on pairs, by $g * (\tau, w) = \left(g * \tau, \frac{w}{c\tau + d_0}\right)$: it moves the point w of the torus along with the lattice (Figure 11). This is the action of $\mathrm{SL}_2(\mathbb{Z})$ that underlies Jacobi forms [EZ85]. Second, translating w by a q-torsion point $\frac{a\tau+b}{q}$ multiplies Θ by the Heisenberg operator $\mathsf{Z}^b\mathsf{X}^a$ of Definition 10.12, again up to an explicit factor (Lemma 12.15). Both symmetries are instances of the classical transformation laws of theta functions with characteristics [Mum83, Chapter I, §§3 and 7]. Conjugating by the Weil representation permutes the Heisenberg operators up to scalars (Lemma 10.13; see [Wei64, n° 10, Théorème 1], [Gér77, Theorem 2.4(a)] and [GH07, Theorem 1.3.2 and Equation (1.3.1)]), so the second symmetry persists after applying $\rho(g)$. The second symmetry is why we symmetrized the witness: the support condition and the shift-invariance of β make P_β invariant under Z and X (Proposition 12.9), so the scalar function

$$F(\tau, z) := P_\beta\left(\Theta\left(\tau, \frac{z}{q}\right)\right)$$

is quasi-periodic in z for the lattice $\mathbb{Z}\tau + \mathbb{Z}$. Up to roots of unity, F is a Jacobi form of weight $\frac{d}{2}$ in the sense of [EZ85] (Lemmas 12.19 and 12.20): a function of a modulus τ and of a point z on the corresponding torus.



(a) Translations by q-torsion points; the Heisenberg operators



(b) The action of $\mathrm{SL}_2(\mathbb{Z})$ on (τ, w); the Weil representation

Figure 10: The two symmetries of the Jacobi theta function Θ, each holding up to an explicit nonvanishing factor and a root of unity (Lemmas 12.15 and 12.16); both are classical for theta functions with characteristics [Mum83, Chapter I, §§3 and 7]. (a) For fixed τ, translating w by a q-torsion point $\frac{a\tau+b}{q}$ multiplies $\Theta(\tau, w)$ by the Heisenberg operator $\mathsf{Z}^b\mathsf{X}^a$: this is the abelian, Fourier-analytic side of the analogy in Figure 3, now present inside the proof. (b) The action of $g \in \mathrm{SL}_2(\mathbb{Z})$ on (τ, w) multiplies Θ by $\rho(g)$, as for the lattice theta functions of the warm-up. The witness polynomial P_β is invariant under X and Z, so it erases the symmetry (a): the scalar function $P_\beta \circ \Theta$ is quasi-periodic in w.

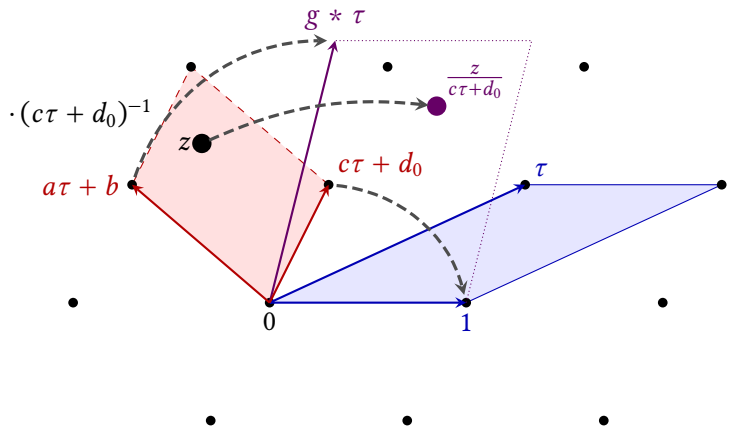


Figure 11: The action of $\mathrm{SL}_2(\mathbb{Z})$ on pairs (τ, z), where z is a point of the torus $\mathbb{C}/(\mathbb{Z}\tau + \mathbb{Z})$. As in Figure 1, the matrix g gives another basis $(c\tau + d_0, a\tau + b)$ of the lattice $\mathbb{Z} + \mathbb{Z}\tau$ (red), and rotating and scaling the plane by $(c\tau + d_0)^{-1}$ (dashed arrows) turns it into the basis $(1, g * \tau)$ (purple). The same similarity sends the point z to $\frac{z}{c\tau+d_0}$, a point of the torus $\mathbb{C}/(\mathbb{Z}(g * \tau) + \mathbb{Z})$. This is the action $g * (\tau, z) = \left(g * \tau, \frac{z}{c\tau+d_0}\right)$ on $\mathcal{H} \times \mathbb{C}$ that underlies Jacobi forms [EZ85].

Two scalarizations. A Jacobi form gives rise to many scalar modular forms [EZ85], and we use two of them (Step 3, Section 12.5). The first is a Taylor coefficient: expanding $F(\tau, z)$ in z around $z = 0$, let $A_m(\tau)$ be the coefficient of z^m for the least m such that this coefficient is not identically zero. Its square A_m^2 is a modular form of weight $d + 2m$ for $\Gamma(q)$ (Lemma 12.29), as for the first nonvanishing Taylor coefficient of any Jacobi form [EZ85, Section 3]. The second is the product $B(\tau)$ of the values of F, suitably normalized, at the three nonzero 2-torsion points $z = \frac{1}{2}, \frac{\tau}{2}, \frac{1+\tau}{2}$ of the torus. The action of $\mathrm{SL}_2(\mathbb{Z})$ permutes these three points (Figure 12), so only their product is modular: B^2 is a modular form of weight $3d$ for $\Gamma(q)$ (Corollary 12.33). Restrictions of Jacobi forms to torsion points are modular for smaller groups [EZ85, Theorem 1.3], and taking the product over an $\mathrm{SL}_2(\mathbb{Z})$-orbit of torsion points is standard [RWR15,LW26]. The two scalarizations see the theta function through different cost functions. The kth term of a coordinate of $\Theta(\tau, z/q)$ has the factor $e^{\pi i \tau (k-1/2)^2/q}$, which at $z = 0$ reads $e^{\pi i \tau (k(k-1)+1/4)/q}$; at the 2-torsion point $z = \frac{\tau}{2}$, the factor $e^{2\pi i (k-1/2)z/q}$ turns it into $e^{\pi i \tau (k^2-1/4)/q}$. So the Taylor coefficient sees the *edge cost* $k(k-1)$, while the 2-torsion evaluations see the *square cost* k^2, which is $w_q(k)$ for $|k| \le \frac{q-1}{2}$ (Definition 5.4).

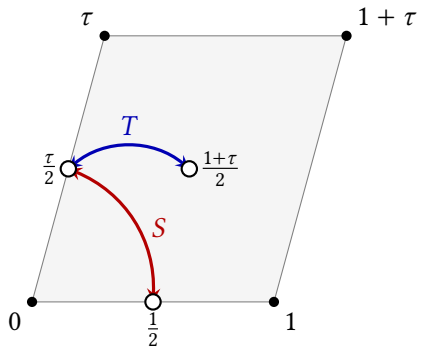


Figure 12: The three nonzero 2-torsion points $\frac{1}{2}$, $\frac{\tau}{2}$ and $\frac{1+\tau}{2}$ of the torus $\mathbb{C}/(\mathbb{Z}\tau+\mathbb{Z})$, in the variable $z = qw$. Rewriting them in the coordinates of a new basis (Figure 11) permutes them: S swaps $\frac{1}{2}$ and $\frac{\tau}{2}$ and fixes $\frac{1+\tau}{2}$, while T swaps $\frac{\tau}{2}$ and $\frac{1+\tau}{2}$ and fixes $\frac{1}{2}$ (all modulo the lattice). As S and T generate $\mathrm{SL}_2(\mathbb{Z})$, no single nonzero 2-torsion point is preserved, which is why the second scalarization uses the product of the evaluations at all three, as is standard [RWR15, LW26].

Root counting at each cusp. At each cusp class $\mathfrak{c}$, with a fixed matrix $g_\mathfrak{c}$ sending ∞ into $\mathfrak{c}$, we record two quantities (Section 12.5.1). The first, $\ell_\mathfrak{c} \geq 0$, records where the expansion of $F_{g_\mathfrak{c}}(\tau, z) := P_\beta(\rho(g_\mathfrak{c})\Theta(\tau, z/q))$ in powers of $e^{\pi i\tau/(4q)}$ starts: its leading terms come from tuples $\mathbf{k} \in \mathbb{Z}^d$ with $\sum_i k_i(k_i-1) = 2\ell_\mathfrak{c}$. The second, $H_\mathfrak{c}$, is the least square cost $\sum_i w_q(x_i)$ of a monomial $v_{x_1}\cdots v_{x_d}$ of $P_\beta \circ \rho(g_\mathfrak{c})$. They bound the orders of our two forms at $\mathfrak{c}$ from below: $\mathrm{ord}_\mathfrak{c}(A_m^2) \geq \frac{d}{4} + 2\ell_\mathfrak{c}$ and $\mathrm{ord}_\mathfrak{c}(B^2) \geq \frac{d}{4} + 2\ell_\mathfrak{c} + 2H_\mathfrak{c}$ (Lemmas 12.29 and 12.34). Summing over all cusp classes, the valence formula (Theorem 10.35; see, e.g., [Ran77, Theorem 4.1.4]) then gives (Lemmas 12.41 and 12.42)

$$\text{(V1)} \quad \sum_\mathfrak{c} \ell_\mathfrak{c} \leq \frac{d(q^2-1)(q-3)}{48} + \frac{m(q^3-q)}{24} \qquad \text{and} \qquad \text{(V2)} \quad \sum_\mathfrak{c} (\ell_\mathfrak{c} + H_\mathfrak{c}) \leq \frac{d(q^2-1)(q-1)}{16}.$$

These take the place of the warm-up's single valence inequality, and both follow a known strategy: (V1) is essentially the case $t = 1$ of [LW26, Lemma 3.1] (see also [Aok22, Lemma 5.1]), applied to a suitable norm of F, and (V2) applies the valence formula to the 2-torsion product in the same spirit as [RWR15, Aok22, LW26]. What is specific to our setting is how the witness controls $\ell_\mathfrak{c}$ and $H_\mathfrak{c}$. It remains to show that the witness forces $H_\mathfrak{c}$ and $\ell_\mathfrak{c}$ to be large at the selected classes.

The square costs: the warm-up argument. At a selected class $\mathfrak{c}_{a,\lambda}$, the coefficients of $P_\beta \circ \rho(g_{a,\lambda})$ are the witness sums for the phase a displayed above, which vanish on the support. So every surviving monomial has square cost more than $\nu\, d$, that is, $H_{\mathfrak{c}_{a,\lambda}} > \nu\, d$ (Lemma 12.35). This is the warm-up's argument at the cusp 0, now run at $s \cdot \frac{q-1}{2}$ cusps.

The edge polynomials. The Taylor order m gives a second lower bound on $\ell_\mathfrak{c}$ at each selected class (Step 4, Section 12.6):

$$\gamma\, d < 2\ell_\mathfrak{c} + \vartheta(d - qm)$$

(Lemma 12.40). The high-level idea of the proof of the above is the following: since F vanishes to order m at $z = 0$, some tuple $\mathbf{k}$ in the leading term of the expansion at $\mathfrak{c}$ has $\sum_i k_i$ within $\frac{d-qm}{2}$ of 0 or of d. (These sums are the exponents of the *edge polynomial* $p_\mathfrak{c}$, which is divisible by $(X^q-1)^m$; see Lemmas 12.36 to 12.38 and Corollary 12.39.) Such a tuple, or its reflection $\mathbf{1}-\mathbf{k}$, has low *tilted edge cost* $w_{q,\vartheta}(j) = \min(k(k-1)+2\vartheta k)$, the minimum over the integers $k \equiv j \pmod q$, so its coefficient vanishes unless $\ell_\mathfrak{c}$ is large.

Combining the inequalities. Summing the bound on $H_{\mathfrak{c}}$ over the selected classes and inserting it into (V2), averaging the edge-polynomial bound over the selected classes, and using (V1) to eliminate m yields

$$\gamma < \left(1 - \frac{6\vartheta s}{q+1}\right)\left(\frac{q^2-1}{4s} - 2\nu\right) + \frac{\vartheta(q-1)}{2},$$

which contradicts the definition of $\gamma = \gamma(\vartheta, \nu)$ in Definition 5.4 (Step 5, Section 12.7). We again work backwards: $\gamma(\vartheta, \nu)$ is the smallest budget for which this combination gives a contradiction, and a smaller budget means a smaller support. The constraints on ϑ and ν ensure that the coefficients in the combination have the right signs. When $\vartheta = 0$, the Taylor coefficient and the edge polynomials are not needed at all, and the 2-torsion product alone already gives a nontrivial exponent (Remark 12.43).

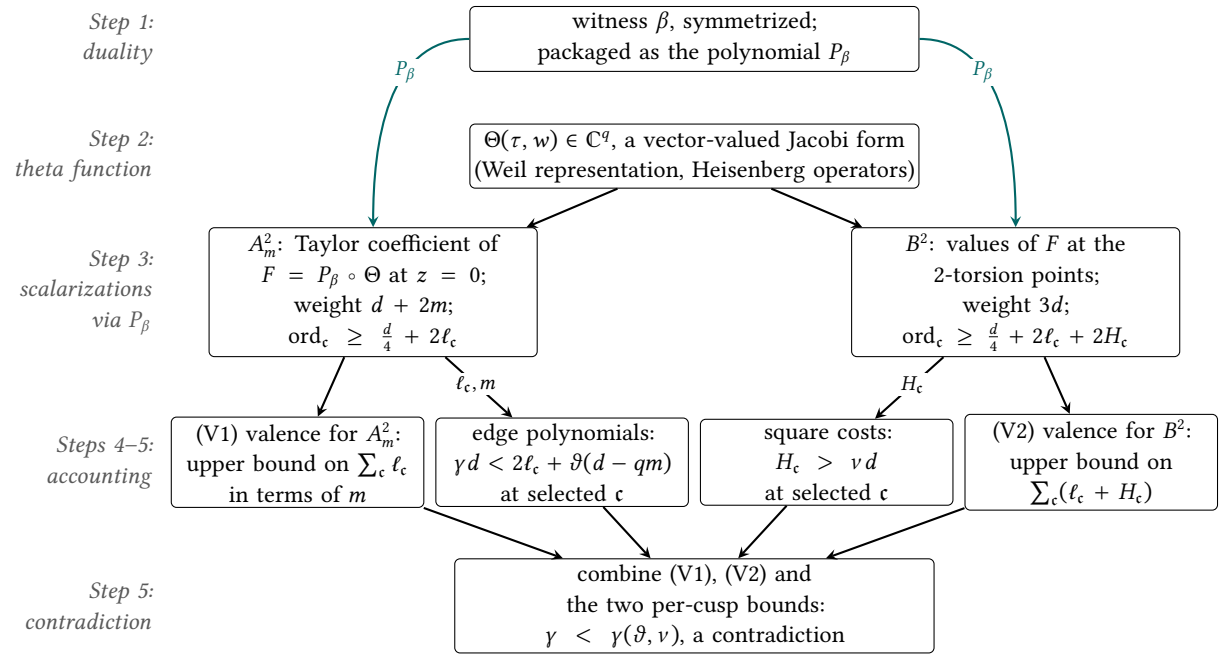


Figure 13: How the proof of Theorem 12.2 fits together, one row per layer of the argument. The witness and the theta function are built independently (Steps 1 and 2). The witness enters through the polynomial P_β (teal, as in Figure 3), which turns Θ into the Jacobi form $F = P_\beta \circ \Theta$ and from it into two scalar modular forms for $\Gamma(q)$ (Step 3). Their orders at each cusp class $\mathfrak{c}$ are bounded below in terms of $\ell_{\mathfrak{c}}$ (where the expansion of F at $\mathfrak{c}$ starts), the Taylor order m, and $H_{\mathfrak{c}}$ (the least square cost of a surviving monomial). In the accounting layer, the valence formula bounds the sums of these quantities over all cusp classes from above (outer boxes), while the witness's orthogonality, which reaches the cusps through P_β, bounds them from below at the $s \cdot \frac{q-1}{2}$ selected classes (inner boxes). Combining the four inequalities gives a contradiction, by the choice of $\gamma(\vartheta, \nu)$ in Definition 12.1. Compare with the warm-up (Figure 8): the right-hand branch (B^2, the square costs and (V2)) is its argument, now run at the $s \cdot \frac{q-1}{2}$ selected classes; for $\vartheta = 0$ it suffices on its own (Remark 12.43), and the left-hand branch is what the Taylor coefficient adds.

The size of the support. The support $S_d(\vartheta, \nu)$ consists of the frequency vectors on which the two arguments need the witness sums to vanish: the set $S_d^{\mathrm{sq}}(\nu)$ of vectors $\mathbf{z}$ of total square cost $\sum_i w_q(z_i)$ at most νd, and the set $S_d^{\mathrm{edge}}(\vartheta, \nu)$ of vectors of total tilted edge cost $\sum_i w_{q,\vartheta}(z_i)$ at most γd. These are closed under the affine maps $\mathbf{z} \mapsto u\mathbf{z} + b\mathbf{1}$ (the rescalings serve the selected classes $\mathfrak{c}_{a,\lambda}$, and the shifts serve the

symmetrization and the reflection above; for $q = 3$, the shifts let the support reach the far end of the weight range, as discussed in Section 8.1) and then thickened by changing one coordinate (to absorb a prefactor from the symmetrization), which costs only a factor $q^{o(d)}$. The same moment-generating-function bound as in the warm-up (Lemma D.1), applied to these two sets, bounds $|S_d(\vartheta, \nu)|$ by $q^{(\eta_{q,s}(\vartheta,\nu)+o(1))d}$ (Lemma E.2), and optimizing over (ϑ, ν) numerically gives the exponents of Corollary 12.4 (Lemma E.3).

12.2 Main Result

We use the costs $w_{q,\vartheta}$ and $w_q = w_{q,1/2}$, their generating functions M_ϑ, the feasible pairs (ϑ, ν), and the budget $\gamma(\vartheta, \nu)$ of Definition 5.4. Explicitly, $w_q(j) = j^2$ for $j \in [0, (q-1)/2]$ and $w_q(j) = (q-j)^2$ for $j \in [(q+1)/2, q-1]$ (Lemma E.1 gives $w_{q,\vartheta}$ in the same way).

Definition 12.1 (The set $S_d(\vartheta, \nu)$)**.** *Let q be an odd prime, $s \in [2, q]$ an integer, $d \geq 1$ an integer dimension parameter that is divisible by q, and (ϑ, ν) a feasible pair, with budget $\gamma = \gamma(\vartheta, \nu)$ (Definition 5.4). Define the sets:*

$$S_d^{\mathrm{sq}}(\nu) := \left\{ \mathbf{z} \in \mathbb{Z}_q^d \,:\, \sum_{i=1}^{d} w_q(z_i) \leq \nu\, d \right\}$$

$$S_d^{\mathrm{edge}}(\vartheta, \nu) := \left\{ \mathbf{z} \in \mathbb{Z}_q^d \,:\, \sum_{i=1}^{d} w_{q,\vartheta}(z_i) \leq \gamma\, d \right\}.$$

Finally, define

$$T_d(\vartheta, \nu) := \left\{ u\mathbf{z} + b\mathbf{1} \,:\, u \in \mathbb{Z}_q^*; b \in \mathbb{Z}_q; \mathbf{z} \in S_d^{\mathrm{sq}}(\nu) \cup S_d^{\mathrm{edge}}(\vartheta, \nu) \right\};$$

$$S_d(\vartheta, \nu) := \left\{ \mathbf{z} + c\mathbf{e}_i \,:\, c \in \mathbb{Z}_q; i \in [d]; \mathbf{z} \in T_d(\vartheta, \nu) \right\}.$$

We can finally state the main theorem of this section:

Theorem 12.2. *For any q, s, d, ϑ, ν as defined in Definition 12.1 (in particular, we require $q \mid d$) and any set $\mathcal{A} \subseteq \mathbb{Z}_q$ of size s, there exists a quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ on $\mathbb{Z}_q^d$, in the sense of Definition 4.3, with support contained in $S_d(\vartheta, \nu) \times \mathcal{A}$. In particular, this expansion is $(|S_d(\vartheta, \nu)|, s)$-sparse.*

The set $S_d(\vartheta, \nu)$ is permutation-invariant. Indeed, the sets $S_d^{\mathrm{sq}}(\nu)$ and $S_d^{\mathrm{edge}}(\vartheta, \nu)$ are, as their definitions depend only on the multiset of coordinates; and permuting the coordinates of $u\mathbf{z}+b\mathbf{1}+c\mathbf{e}_i$ gives $u\mathbf{z}'+b\mathbf{1}+c\mathbf{e}_{i'}$, where $\mathbf{z}'$ is obtained from $\mathbf{z}$ by the same permutation and i' is the new position of the ith coordinate.

Corollary 12.3 (Generalizing to all d)**.** *For any q, s, ϑ, ν as in Definition 12.1 and any positive integer d (not necessarily divisible by q), there exists a $(|S_{q\lfloor d/q \rfloor}(\vartheta, \nu)|, s)$-sparse quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ on $\mathbb{Z}_q^d$ with permutation-invariant support.*

Proof. See Appendix E. □

Bounding the support sizes $|S_d(\vartheta, \nu)|$ gives the following, which we prove in Appendix E.

Corollary 12.4. *Fix an odd prime q, an integer $s \in [2, q]$, and a feasible pair (ϑ, ν), and let d be a growing parameter. Then there exists a (k, s)-sparse quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ on $\mathbb{Z}_q^d$ with permutation-invariant support and with $k \leq q^{(\eta_{q,s}(\vartheta,\nu)+o(1))d}$; the same holds with $\eta_{q,s}$ in place of $\eta_{q,s}(\vartheta, \nu)$. Here $\eta_{q,s}(\vartheta, \nu)$ and $\eta_{q,s}$ are the constants of Definition 5.4. In particular, $\eta_{3,2} \leq 0.820340$, $\eta_{3,3} \leq 0.732672$, $\eta_{5,4} \leq 0.702450$, $\eta_{5,5} \leq 0.653434$, $\eta_{7,6} \leq 0.652057$, $\eta_{7,7} \leq 0.619701$, $\eta_{11,10} \leq 0.607791$, and $\eta_{11,11} \leq 0.590014$ (Lemma E.3). As q grows, we have $\eta_{q,q} \leq \frac{1}{2} + \frac{\frac{1}{2}\ln(\pi e/6)+o(1)}{\ln q} \leq \frac{1}{2} + \frac{0.255+o(1)}{\log_2 q}$ (Lemma E.4).*

12.3 Proof Step 1: Duality and Symmetrization

Throughout this section, we fix an arbitrary subset $\mathcal{A} \subseteq \mathbb{Z}_q$ of size s.

Here, we follow the first part of Section 11.4. Unwinding Definition 4.3 as in the warm-up, Theorem 12.2 follows from the existence of coefficients $\{c_{\mathbf{z},a}\}$ supported on pairs $(\mathbf{z}, a) \in S_d(\vartheta, \nu) \times \mathcal{A}$ with $\sum_{\mathbf{z},a} c_{\mathbf{z},a}\, \zeta_q^{\langle \mathbf{z},\mathbf{x}\rangle + a\langle \mathbf{x},\mathbf{x}\rangle} = \mathbb{1}[\mathbf{x} = \mathbf{0}]$ for every $\mathbf{x} \in \mathbb{Z}_q^d$: such a collection is a $(|S_d(\vartheta, \nu)|, s)$-sparse quadratic-phase delta expansion (its support is $\mathcal{W} \times \mathcal{T}$ in the sense of Definition 4.3, with $\mathcal{W} = S_d(\vartheta, \nu)$ and $\mathcal{T} = \mathcal{A}$), and since our $\mathcal{A} \subseteq \mathbb{Z}_q$ of size s is arbitrary, producing such coefficients for it proves the theorem. Suppose, for the sake of contradiction, that no such coefficients exist. By the duality argument of Section 11.4, applied verbatim in the space $\mathbb{C}^{\mathbb{Z}_q^d}$ with one spanning vector $\left(\zeta_q^{a\langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{z}\rangle}\right)_{\mathbf{x} \in \mathbb{Z}_q^d}$ per pair $(a, \mathbf{z}) \in \mathcal{A} \times S_d(\vartheta, \nu)$, there exists a vector α indexed by elements $\mathbf{x} \in \mathbb{Z}_q^d$ such that:

$$\begin{aligned} \alpha_{\mathbf{0}} &= 1 \\ \sum_{\mathbf{x} \in \mathbb{Z}_q^d} \alpha_{\mathbf{x}} \zeta_q^{a\langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{z}\rangle} &= 0 \quad \forall a \in \mathcal{A}, \mathbf{z} \in S_d(\vartheta, \nu). \end{aligned} \tag{41}$$

It will help us later to replace α by a symmetrized witness β satisfying two additional properties: β is supported on vectors $\mathbf{x}$ with $\sum_{i \in [d]} x_i \equiv 0 \pmod q$, and $\beta_{\mathbf{x}} = \beta_{\mathbf{x}+\mathbf{1}}$ for all $\mathbf{x}$. (These are what will make the polynomial of Proposition 12.9 below invariant under the Heisenberg operators Z and X respectively, which is how translations of the elliptic variable of our theta function will be absorbed, starting with Lemmas 12.15 and 12.20 in Step 2.) These need not be true of α itself, so we will need to make some transformations first to ensure this. We do this using an averaging trick:

Definition 12.5. *For every $J \in \mathbb{Z}_q$ and $\mathbf{x} \in \mathbb{Z}_q^d$, define:*

$$\alpha_{\mathbf{x}}^{(J)} = \mathbb{1}\left[\sum_{i=1}^{d} x_i = 0\right] \cdot \zeta_q^{-Jx_1} \cdot \frac{1}{q} \sum_{b \in \mathbb{Z}_q} \zeta_q^{-Jb}\, \alpha_{\mathbf{x}+b\mathbf{1}}.$$

Each ingredient buys one property: the indicator enforces the support condition (43) below; the weight ζ_q^{-Jb} inside the average provides the one degree of freedom, J, needed to make the normalization $\alpha_{\mathbf{0}}^{(J)} \neq 0$ achievable (it selects a Fourier coefficient along the line $\{b\mathbf{1}\}$, not all of which can vanish); and the matching prefactor $\zeta_q^{-Jx_1}$ is what preserves the shift-invariance (44) in the presence of that weight. The cost of the prefactor is a shift of the frequency vector by $-J\mathbf{e}_1$, which is absorbed by the final thickening in the definition of $S_d(\vartheta, \nu)$ — indeed, only the first-coordinate thickening $c\mathbf{e}_1$ is consumed there; $S_d(\vartheta, \nu)$ thickens every coordinate only for symmetry of the definition.

Proposition 12.6. *There exists a $J \in \mathbb{Z}_q$ such that $\alpha_{\mathbf{0}}^{(J)} \neq 0$.*

Proof. By definition, we have

$$\alpha_{\mathbf{0}}^{(J)} = \frac{1}{q} \sum_{b \in \mathbb{Z}_q} \zeta_q^{-Jb}\, \alpha_{b\mathbf{1}},$$

so this is just a Fourier-transform coefficient of $(\alpha_{\mathbf{0}}, \alpha_{\mathbf{1}}, \alpha_{2\mathbf{1}}, \ldots, \alpha_{(q-1)\mathbf{1}}) \neq 0$. So one of the Fourier coefficients must be nonzero. □

Proposition 12.7. *Fix $J \in \mathbb{Z}_q$ such that $\alpha_{\mathbf{0}}^{(J)} \neq 0$ and define β by $\beta_{\mathbf{x}} = \alpha_{\mathbf{x}}^{(J)} / \alpha_{\mathbf{0}}^{(J)}$. Then we have:*

$$\beta_{\mathbf{0}} = 1 \tag{42}$$

$$\beta_{\mathbf{x}} = 0 \quad \forall \mathbf{x} \in \mathbb{Z}_q^d \text{ s.t. } \sum_{i=1}^{d} x_i \neq 0 \tag{43}$$

$$\beta_{\mathbf{x}} = \beta_{\mathbf{x}+\mathbf{1}} \quad \forall \mathbf{x} \in \mathbb{Z}_q^d. \tag{44}$$

$$\sum_{\mathbf{x} \in \mathbb{Z}_q^d} \beta_{\mathbf{x}} \zeta_q^{a\langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{x},\mathbf{z}\rangle} = 0 \quad \forall a \in \mathcal{A}, \mathbf{z} \in T_d(\vartheta, \nu). \tag{45}$$

Proof. See Appendix E. □

As in the warm-up, we package the dual witness into a homogeneous polynomial. This time there is no average over $\mathbb{Z}_q^*$; instead, the symmetries (43) and (44) will make the polynomial invariant under the Heisenberg operators X and Z of Definition 10.12, which will reappear as the effect of translating the elliptic variable of our theta function (Lemma 12.15 in Step 2).

Definition 12.8. *Define the polynomial $P_\beta : \mathbb{C}^q \to \mathbb{C}$ as follows:*

$$P_\beta(X_0, \ldots, X_{q-1}) = \sum_{\mathbf{x} \in \mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^{d} X_{x_i}.$$

Here X and Z are the Heisenberg operators of Definition 10.12, and their conjugation by the Weil representation (Lemma 10.13) is the engine of the scalarization below.

Proposition 12.9. *We have all of the following:*

- *P_β is homogeneous of degree d.*
- *For all $x \in \mathbb{Z}_q$, $P_\beta(\mathbf{e}_x) = 1$.*
- *For all $\mathbf{y} \in \mathbb{C}^q$, $P_\beta(\mathsf{X}\mathbf{y}) = P_\beta(\mathsf{Z}\mathbf{y}) = P_\beta(\mathbf{y})$.*
- *For all $g \in \mathrm{SL}_2(\mathbb{Z})$ and every product H of powers of X and Z, we have $P_\beta(\rho(g)H\mathbf{y}) \propto_\zeta P_\beta(\rho(g)\mathbf{y})$ as functions of $\mathbf{y} \in \mathbb{C}^q$.*

Proof. See Appendix E. □

12.4 Proof Step 2: The Vector-Valued Jacobi Theta Modular Form

We continue the proof of Theorem 12.2; recall the standing assumption that d is divisible by q.

Definition 12.10 ($\mathrm{SL}_2(\mathbb{Z})$ action on $\mathcal{H} \times \mathbb{C}$)**.** *We define a group action by $\mathrm{SL}_2(\mathbb{Z})$ on $\mathcal{H} \times \mathbb{C}$ as follows:*

$$\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} * (\tau, w) := \left(\frac{a\tau + b}{c\tau + d_0}, \frac{w}{c\tau + d_0} \right).$$

The intuition: the pair (τ, w) should be read as a point w on the complex torus $\mathbb{C}/(\mathbb{Z}\tau + \mathbb{Z})$, together with the modulus τ of that torus. The action transports both consistently: if $\tau' = g * \tau$, then the lattice $\mathbb{Z}\tau' + \mathbb{Z}$ equals $(\mathbb{Z}\tau + \mathbb{Z})/G(g, \tau)$ (the matrix g is a change of basis of the lattice, followed by the rescaling), and dividing w by the same factor $G(g, \tau)$ applies that rescaling to the marked point.

Proposition 12.11. *The above defines a valid group action.*

Proof. The validity on the first coordinate is automatic from the $\mathrm{SL}_2(\mathbb{Z})$ action on $\mathcal{H}$. For the second coordinate, we want to show for any $g_1, g_2 \in \mathrm{SL}_2(\mathbb{Z})$ that:

$$\frac{w}{G(g_1 g_2, \tau)} = \frac{w}{G(g_1, g_2 * \tau) \cdot G(g_2, \tau)},$$

which is in turn immediate from Proposition 10.17. □

For this proof, our vector-valued function is defined as follows:

Definition 12.12. *Define* $\Theta : \mathcal{H} \times \mathbb{C} \to \mathbb{C}^q$ *by:*

$$\Theta(\tau, w) = \sum_{k \in \mathbb{Z}} (-1)^k \exp\left(\frac{\pi i \tau}{q} \left(k - \frac{1}{2}\right)^2 + 2\pi i \left(k - \frac{1}{2}\right) w \right) \mathbf{e}_{k+(q-1)/2 \bmod q}.$$

An analogous argument to that in Lemma 11.15 establishes that this is well-defined and also that it is holomorphic on $\mathcal{H} \times \mathbb{C}$ in the sense of Definition B.5: each term is entire in (τ, w), and on a compact subset of $\mathcal{H} \times \mathbb{C}$ the real part of each exponent is at most $-c\left(k - \frac{1}{2}\right)^2 + C\left|k - \frac{1}{2}\right|$ for constants $c > 0$ and $C \geq 0$ (coming from $\Im\tau$ being bounded below and $|\Im w|$ bounded above), so Lemma B.7 (with $r = 1$ and the coset $\mathbb{Z} - \frac{1}{2}$ of $\mathbb{Z}$) and Fact B.3 apply.

At this point in the warm-up in Section 11.5, we checked a few things about $\Theta_{\mathcal{L}}$:

- that it is weakly modular with respect to the Weil representation; and
- that it is holomorphic at the cusps, and therefore a vector-valued modular form.

We will not do either of these here, because these are not technically true. To begin with, the function now takes an element of $\mathcal{H} \times \mathbb{C}$ rather than just an element of $\mathcal{H}$ as input. Moreover, it will turn out that $\Theta(\tau, w)$ can be thought of as a "pseudo-modular form" of weight $1/2$ (whereas a modular form's weight has to be an integer). This creates cumbersome technicalities in terms of appropriately defining the square root of $c\tau + d_0$ in the transformation law defined in (21). Formally, Θ is a vector-valued Jacobi form of weight $1/2$ for a metaplectic cover of the Jacobi group, and its transformation laws are instances of the classical ones for theta functions with characteristics (see, e.g., [Mum83, Chapter I, §§3 and 7] and [EZ85]); but we will elide this formalism because we do not need it. In particular, we work up to roots of unity (the relation $\propto_\zeta$), much as the classical transformation formula for theta functions under $\mathrm{SL}_2(\mathbb{Z})$ involves an eighth root of unity whose exact value is cumbersome to track [Mum83, Chapter I, Theorem 7.1]. Instead, we will proceed as follows:

- In this subsection, we just explicitly calculate how Θ transforms under the action of $\mathrm{SL}_2(\mathbb{Z})$. We do not address any limiting behavior at cusps.
- In Section 12.5, we mirror Section 11.5 and compose Θ with a suitable homogeneous polynomial. The resulting object will lead us to two ordinary scalar-valued modular forms with integer weight; we will check this using the aforementioned transformation laws and we will also address the holomorphy at the cusps at this point.
- Finally, we will again finish with the valence formula: the scalarized forms' cusp orders, bounded from below in Sections 12.5 and 12.6, feed the valence inequalities of Section 12.7, which together contradict the existence of the witness.

First, we will once again use the Weil representation $\rho : \mathrm{SL}_2(\mathbb{Z}) \to \mathbb{C}^{q\times q}$ of Definition 10.10 and Fact 10.11, specializing to the parameter $\mu = (q+1)/2$ (a valid choice, as $(q+1)/2 \equiv 2^{-1} \not\equiv 0 \pmod q$). Thus ρ_T maps $\mathbf{e}_x \mapsto \zeta_q^{x^2(q+1)/2}\mathbf{e}_x$, and ρ_S maps $\mathbf{e}_x \mapsto \frac{\epsilon_q}{\sqrt{q}}\sum_{y\in\mathbb{Z}_q}\zeta_q^{-xy}\mathbf{e}_y$, where $\epsilon_q := \epsilon_{(q+1)/2,\,q} \in \{\pm 1, \pm i\}$ (Fact 10.9). With this in mind, we can now check the transformation laws. These are the classical transformation laws of theta functions with characteristics, in our normalization; since translating between normalizations is about as long as a direct verification, we include proofs. They closely mirror those in Section 11.5, and we defer them to Appendix E.

Lemma 12.13. *As functions of $(\tau, w) \in \mathcal{H}\times\mathbb{C}$, we have:*

$$\Theta(T * (\tau, w)) = \Theta(\tau+1, w) \propto_\zeta \rho(T)\Theta(\tau, w).$$

Proof. See Appendix E. □

Lemma 12.14. *As functions of $(\tau, w) \in \mathcal{H}\times\mathbb{C}$, we have:*

$$\Theta(S * (\tau, w)) = \Theta(-1/\tau, w/\tau) \propto_\zeta (-i\tau)^{1/2} e^{\pi i q w^2/\tau}\rho(S)\Theta(\tau, w).$$

In the above, $(-i\tau)^{1/2}$ denotes the principal branch, as in Fact B.8; it is well-defined and holomorphic on $\mathcal{H}$ because $\Re(-i\tau) = \Im\tau > 0$.

Proof. See Appendix E. □

Finally, there is one more symmetry that we take a moment to highlight; this symmetry is enabled by the addition of the variable w. It says that Θ is *quasi-periodic* in w under translations by the scaled lattice $\frac{1}{q}(\mathbb{Z}\tau + \mathbb{Z})$: shifting w by $\frac{a\tau+b}{q}$ reproduces Θ up to an explicit nonvanishing factor, a root of unity, and an application of the Heisenberg operators $\mathsf{Z}^b\mathsf{X}^a$, which record the discrete data (a, b) of the translation. This is what justifies regarding w (equivalently $z = qw$) as a point on a torus, and — once composed with the Heisenberg-invariant polynomial P_β of Proposition 12.9 — it is the source of the elliptic transformation law Lemma 12.20 below.

Lemma 12.15. *For any $a, b \in \mathbb{Z}$, we have, as functions of $(\tau, w) \in \mathcal{H}\times\mathbb{C}$:*

$$\Theta\left(\tau, w + \frac{a\tau+b}{q}\right) \propto_\zeta e^{-\pi i a^2\tau/q - 2\pi i a w}\cdot \mathsf{Z}^b\mathsf{X}^a\Theta(\tau, w). \tag{46}$$

Proof. See Appendix E. □

The transformation laws for the generators T and S (Lemmas 12.13 and 12.14) propagate to all of $\mathrm{SL}_2(\mathbb{Z})$, up to a root of unity; this also absorbs the sign ambiguity in composing the square-root factors, whose branches multiply only up to sign:

Lemma 12.16. *For any $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, we have, as functions of $(\tau, w) \in \mathcal{H}\times\mathbb{C}$:*

$$\Theta(g * (\tau, w)) \propto_\zeta G(g, \tau)^{1/2}\exp\left(\frac{\pi i q c w^2}{G(g, \tau)}\right)\rho(g)\Theta(\tau, w), \tag{47}$$

where $G(g, \tau)^{1/2}$ denotes any holomorphic branch of the square root of $\tau \mapsto G(g, \tau)$. (Such a branch exists: the image of $\tau \mapsto c\tau + d_0$ on $\mathcal{H}$ is an open half-plane when $c \neq 0$ and the single nonzero real d_0 when $c = 0$, in either case a simply connected set avoiding 0 on which a branch can be chosen. Any two branches differ by a constant factor ± 1, a root of unity, so the choice does not matter.)

Proof. See Appendix E. □

Definition 12.17. *For any $g \in \mathrm{SL}_2(\mathbb{Z})$, define $F_g : \mathcal{H} \times \mathbb{C} \to \mathbb{C}$ by $F_g(\tau, z) := P_\beta(\rho(g)\Theta(\tau, z/q))$. Define $F := F_I$. Here and throughout, the two coordinates are related by $z = qw$.*

Proposition 12.18. *For $g \in \Gamma(q)$, we have $F_g(\tau, z) = F(\tau, z)$.*

Proof. Immediate from Fact 10.11, which tells us that $\rho(g) = I$ when $g \in \Gamma(q)$. □

Lemma 12.19. *For any $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, we have, as functions of $(\tau, z) \in \mathcal{H} \times \mathbb{C}$:*

$$F(g * (\tau, z)) = F\left(g * \tau, \frac{z}{G(g, \tau)}\right) \propto_\zeta G(g, \tau)^{d/2} \exp\left(\frac{\pi i d c z^2}{q G(g, \tau)}\right) F_g(\tau, z),$$

where $G(g, \tau)^{d/2} := \left(G(g, \tau)^{1/2}\right)^d$ for a holomorphic branch as in Lemma 12.16.

Proof. See Appendix E. □

Lemma 12.20. *For all $a, b \in \mathbb{Z}$ and $g \in \mathrm{SL}_2(\mathbb{Z})$, we have, as functions of $(\tau, z) \in \mathcal{H} \times \mathbb{C}$:*

$$F_g(\tau, z + a\tau + b) \propto_\zeta \exp\left(-\frac{\pi i d}{q}(a^2\tau + 2az)\right) F_g(\tau, z).$$

Proof. See Appendix E. □

12.5 Proof Step 3: Two Scalarizations

12.5.1 Cusp Data

Both scalarizations below follow the same pattern: compose Θ with the polynomial P_β to obtain a scalar function of τ, show that this function is a modular form for $\Gamma(q)$, and lower bound its orders at the cusps. Before carrying either out, we collect what the two constructions share: the fact that there is anything to scalarize at all ($F \not\equiv 0$), the classification of the cusps of $\Gamma(q)$, a matrix $g_\mathfrak{c}$ fixed once and for all per cusp class, and two per-class quantities $\ell_\mathfrak{c}$ and $H_\mathfrak{c}$, in terms of which both cusp-order bounds will be expressed.

Proposition 12.21. *F is not identically zero.*

Proof. See Appendix E. □

Corollary 12.22. *For every $g \in \mathrm{SL}_2(\mathbb{Z})$, the function F_g is not identically zero.*

Proof. By Lemma 12.19, $F(g * (\tau, z)) \propto_\zeta G(g, \tau)^{d/2} \exp\left(\frac{\pi i d c z^2}{q G(g, \tau)}\right) F_g(\tau, z)$, where the constant and the other factors multiplying $F_g(\tau, z)$ never vanish. Since $(\tau, z) \mapsto g * (\tau, z)$ is a bijection of $\mathcal{H} \times \mathbb{C}$ and F is not identically zero (Proposition 12.21), neither is F_g. □

Next, the cusps. We computed the cusp classes of $\Gamma_0(q)$ in Proposition 11.21; the analogous statement for $\Gamma(q)$ is the following.

Proposition 12.23 ([DS05, Proposition 3.8.3 and Section 3.8])**.** *Under the action of $\Gamma(q)$, the cusps $\mathbb{Q} \cup \{\infty\}$ split into $(q^2 - 1)/2$ equivalence classes, each naturally represented by an element of $(\mathbb{Z}_q^2 \backslash \{(0, 0)\})/(\pm 1)$. Moreover, all these cusps have width q.*

By Lemma 10.28, the width of a cusp, holomorphy of a weakly modular function there, and its order there depend only on the $\Gamma(q)$-class of the cusp, and not on the matrix $g_{\mathfrak{c}}$ used to compute them. The per-class quantities $\ell_{\mathfrak{c}}$ and $H_{\mathfrak{c}}$ defined below are different: they are defined *in terms of* such a matrix. We therefore fix, once and for all, for each cusp class $\mathfrak{c}$ a matrix $g_{\mathfrak{c}} \in \mathrm{SL}_2(\mathbb{Z})$ sending ∞ to a cusp in the class $\mathfrak{c}$, and we define $\ell_{\mathfrak{c}}$ and $H_{\mathfrak{c}}$ relative to this fixed choice. (For the selected classes $\mathfrak{c}_{a,\lambda}$ of Section 12.6 we will fix a specific, convenient $g_{\mathfrak{c}}$; for all other classes the choice is arbitrary. No independence from the choice needs to be *proved*: the cusp-order bounds of Lemmas 12.29 and 12.34 bound the choice-independent quantities $\mathrm{ord}_{\mathfrak{c}}(A_m^2)$ and $\mathrm{ord}_{\mathfrak{c}}(B^2)$ in terms of the fixed choice, and the per-class estimates of Section 12.6 (Lemmas 12.35 and 12.40) are proved for the same fixed lifts; using a single fixed choice consistently throughout is all that is required.)

The quantity $\ell_{\mathfrak{c}}$ is extracted from an expansion of $F_{g_{\mathfrak{c}}}$ in powers of $e^{\pi i\tau/(4q)}$, which we now construct. In the definition of Θ, the τ-dependence of the k-th term is through $\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2\right) = \exp\left(\frac{\pi i\tau(2k-1)^2}{4q}\right)$, with an odd square $(2k-1)^2$ in the numerator; the possible τ-frequencies of a degree-d monomial in the coordinates of Θ are therefore sums of d odd squares. Since every odd square is $\equiv 1 \pmod 8$, such a sum is at least d and $\equiv d \pmod 8$. Formally:

Lemma 12.24. *Fix $g \in \mathrm{SL}_2(\mathbb{Z})$. For each integer $\ell \geq d$ with $\ell \equiv d \pmod 8$, there is an entire function $\Phi_{g,\ell} : \mathbb{C} \to \mathbb{C}$ such that:*

$$P_\beta(\rho(g)\,\Theta(\tau, w)) = \sum_{\substack{\ell \geq d \\ \ell \equiv d \pmod 8}} \Phi_{g,\ell}(w)\, e^{\pi i\tau\ell/(4q)}, \tag{48}$$

with the series converging absolutely and locally uniformly on $\mathcal{H} \times \mathbb{C}$. (The functions $\Phi_{g,\ell}$ are unique: for each fixed w, (48) is the Fourier expansion of the $8q$-periodic holomorphic function $\tau \mapsto P_\beta(\rho(g)\Theta(\tau, w))$, whose coefficients are unique by Fact B.4.) Moreover, writing $F_g(\tau, z) = \sum_{j\geq 0} A_{g,j}(\tau)\, z^j$ for the Taylor family of Lemma B.6 (which applies: F_g is holomorphic on $\mathcal{H} \times \mathbb{C}$, being the composition of Θ with the linear map $\rho(g)$, the polynomial P_β, and the rescaling $z \mapsto z/q$), and $\Xi_{g;j,\ell}$ for the j-th Taylor coefficient of $\Phi_{g,\ell}$, we have for every $j \geq 0$:

$$A_{g,j}(\tau) = \sum_{\substack{\ell \geq d \\ \ell \equiv d \pmod 8}} q^{-j}\, \Xi_{g;j,\ell}\, e^{\pi i\tau\ell/(4q)}, \tag{49}$$

again converging absolutely and locally uniformly on $\mathcal{H}$.

Proof. See Appendix E. □

Definition 12.25. *Let $\mathfrak{c}$ be a cusp class, with fixed matrix $g_{\mathfrak{c}}$ as above. By Corollary 12.22 we have $F_{g_{\mathfrak{c}}} \not\equiv 0$, so by (48) not all of the functions $\Phi_{g_{\mathfrak{c}},\ell}$ vanish identically; let ℓ^* be the minimal ℓ with $\Phi_{g_{\mathfrak{c}},\ell} \not\equiv 0$. Since every index ℓ in (48) satisfies $\ell \geq d$ and $\ell \equiv d \pmod 8$, the quantity*

$$\ell_{\mathfrak{c}} := \frac{\ell^* - d}{8}$$

is a nonnegative integer.

Finally, we define the second per-class quantity, which is used in the cusp-order bound of the second scalarization. Recall the weight function $w_q = w_{q,1/2}$ of Definition 5.4, and for a polynomial P in the variables $X_0, \dots, X_{q-1}$, write $[X_{x_1}X_{x_2}\cdots X_{x_d}]P$ for the coefficient of the monomial $X_{x_1}X_{x_2}\cdots X_{x_d}$ in P.

Definition 12.26. *Let $\mathfrak{c}$ be a cusp class, with fixed matrix $g_{\mathfrak{c}}$ as above. Define:*

$$H_{\mathfrak{c}} := \min\left\{\sum_{i=1}^{d} w_q(x_i) \,:\, \mathbf{x}\in\mathbb{Z}_q^d \text{ s.t. } [X_{x_1}X_{x_2}\dots X_{x_d}](P_\beta\circ\rho(g_{\mathfrak{c}}))\neq 0\right\}.$$

(The minimum is over a nonempty finite set: P_β is not the zero polynomial by (42), and $\rho(g_{\mathfrak{c}})$ is invertible, so $P_\beta\circ\rho(g_{\mathfrak{c}})$ is not the zero polynomial either; it is homogeneous of degree d, so all of its monomials have the displayed shape.)

12.5.2 Scalarization via Taylor Coefficients

At long last, we can define the scalarized modular form that we will work with in this step. It is the first nonvanishing Taylor coefficient of F in the elliptic variable, which, for a Jacobi form in the sense of [EZ85], is classically a modular form [EZ85, Section 3]; we verify the analogous statement directly for F, whose transformation laws hold only up to roots of unity.

Definition 12.27. *F is holomorphic on $\mathcal{H}\times\mathbb{C}$ (it is the composition of Θ, which is holomorphic by the discussion above, with the polynomial P_β and the rescaling $z\mapsto z/q$), so by Lemma B.6 there exist unique holomorphic functions $A_0, A_1, \dots$ from $\mathcal{H}\to\mathbb{C}$ such that:*

$$F(\tau,z)=\sum_{j=0}^{\infty}A_j(\tau)z^j.$$

By Proposition 12.21, there exists a minimal integer $m\geq 0$ such that A_m is not identically zero. Our scalarized function is $A_m : \mathcal{H}\to\mathbb{C}$. Equivalently, we can define:

$$A_m(\tau)=\frac{1}{m!}\frac{\partial^m}{\partial z^m}F(\tau,z)\Big|_{z=0}.$$

Proposition 12.28. *For any $g\in \mathrm{SL}_2(\mathbb{Z})$, there exist holomorphic functions $A_{g,m}, A_{g,m+1}, \dots : \mathcal{H}\to\mathbb{C}$ such that:*

$$F_g(\tau,z)=P_\beta(\rho(g)\Theta(\tau,z/q))=\sum_{j=m}^{\infty}A_{g,j}(\tau)z^j.$$

Moreover, we have $A_{g,m}(\tau)\propto_\zeta G(g,\tau)^{-d/2-m}A_m(g*\tau)$ as functions of τ.

Proof. See Appendix E. □

Here and below, for functions holomorphic on $\mathcal{H}\times\mathbb{C}$, the notation $O(z^{m+1})$ denotes a function of the form $z^{m+1}h(\tau,z)$ with h holomorphic on $\mathcal{H}\times\mathbb{C}$ — equivalently, by Lemma B.6, a holomorphic function whose Taylor coefficients (in z) of index at most m vanish identically. The uniqueness in Lemma B.6 then justifies equating the coefficients of z^m between two such expansions.

We show that A_m^2 is a modular form, and bound its orders at the cusps in terms of the quantity $\ell_{\mathfrak{c}}$ of Definition 12.25.

Lemma 12.29. *$A_m^2 : \mathcal{H}\to\mathbb{C}$ is a modular form for $\Gamma(q)$ of weight $d+2m$, with a one-dimensional representation of finite order. Moreover, for all cusps $\mathfrak{c}\in\mathbb{Q}\cup\{\infty\}$, we have $\mathrm{ord}_{\mathfrak{c}}(A_m^2)\geq\frac{d}{4}+2\ell_{\mathfrak{c}}$.*

Proof. See Appendix E. □

12.5.3 Scalarization via 2-Torsion Evaluations

The second scalarization fixes the elliptic variable at the three points $z = \frac{x\tau+y}{2}$ for $(x, y) \in \{(0,1),(1,0),(1,1)\}$: representatives of the three nonzero *2-torsion* points of the torus $\mathbb{C}/(\mathbb{Z}\tau + \mathbb{Z})$, i.e., of the points $z \notin \mathbb{Z}\tau + \mathbb{Z}$ with $2z \in \mathbb{Z}\tau + \mathbb{Z}$. By Lemma 12.20, moving z by a lattice vector changes F_g only by an explicit nonvanishing factor times a root of unity, so these evaluations are canonical up to such factors; and the action of $\mathrm{SL}_2(\mathbb{Z})$ on τ permutes the three nonzero 2-torsion points (as the bijection $(x, y) \mapsto (\tilde{x}, \tilde{y})$ in Lemma 12.31 below makes precise), which is why we work with the *product* B_g of all three evaluations: up to scalars it transforms like a weight-$3d/2$ form. The prefactor $\exp(\pi i d x^2 \tau/(4q))$ in the definition below normalizes the growth of $F_g(\tau, (x\tau + y)/2)$ as $\Im\tau \to \infty$ (see the proof of Lemma 12.34); it is the standard normalizing factor with which restrictions of Jacobi forms to torsion points become modular forms [EZ85, Theorem 1.3] (see also [Aok22, Section 4.2] and [RWR15, Lemma 2.1]). Taking the product over an $\mathrm{SL}_2(\mathbb{Z})$-orbit of torsion points, as we do here, is likewise standard [RWR15, LW26].

Definition 12.30. *For $x, y \in \{0,1\}$ and $g \in \mathrm{SL}_2(\mathbb{Z})$, define:*

$$B_{g;x,y}(\tau) := \exp\left(\frac{\pi i d x^2 \tau}{4q}\right) F_g\left(\tau, \frac{x\tau + y}{2}\right),$$

and then define $B_g(\tau) := B_{g;0,1}(\tau) B_{g;1,0}(\tau) B_{g;1,1}(\tau)$. When g is omitted from notation, it is set to the identity. Note that this is consistent with Proposition 12.18.

Lemma 12.31. *Let $g \in \mathrm{SL}_2(\mathbb{Z})$ and $(x, y) \in \{(0,1),(1,0),(1,1)\}$. There exists $(\tilde{x}, \tilde{y}) \in \{(0,1),(1,0),(1,1)\}$ — explicitly, $(\tilde{x}, \tilde{y}) \equiv (ax + cy,\ bx + d_0 y) \pmod 2$ — such that, as functions of $\tau \in \mathcal{H}$:*

$$B_{x,y}(g * \tau) \propto_\zeta G(g, \tau)^{d/2} B_{g;\tilde{x},\tilde{y}}(\tau),$$

with $G(g, \tau)^{d/2}$ as in Lemma 12.19. Moreover, for each g the mapping $(x, y) \mapsto (\tilde{x}, \tilde{y})$ is bijective.

Proof. See Appendix E. □

Proposition 12.32. *B is not identically zero.*

Proof. First, each $B_{x,y}$ is holomorphic on $\mathcal{H}$: writing $F(\tau, z) = \sum_j A_j(\tau) z^j$ (Lemma B.6), we have $F(\tau, \varphi(\tau)) = \sum_j A_j(\tau)\varphi(\tau)^j$ for the holomorphic function $\varphi(\tau) := \frac{x\tau+y}{2}$, a locally uniform limit of holomorphic functions, which is holomorphic by Fact B.3; multiplying by the nonvanishing factor $\exp(\pi i d x^2 \tau/(4q))$ preserves this. Since $\mathcal{H}$ is connected and the zeros of a nonzero holomorphic function are isolated (Fact B.2), a product of finitely many holomorphic functions, none identically zero, is not identically zero; so it suffices to show $B_{x,y} \not\equiv 0$ for each of the three pairs.

For $(x, y) = (1, 0)$: by definition, $B_{1,0}(\tau) = e^{\pi i d\tau/(4q)} F(\tau, \tau/2) = e^{\pi i d\tau/(4q)} P_\beta(\Theta(\tau, \tau/(2q)))$. The proof of Proposition 12.21 shows precisely that $P_\beta(\Theta(iy, iy/(2q))) \neq 0$ for all large enough $y > 0$, so $B_{1,0} \not\equiv 0$.

For the other two pairs, we use Lemma 12.31 as a pullback. In each case we exhibit a matrix $g \in \mathrm{SL}_2(\mathbb{Z})$ with $g \equiv I \pmod q$ — so that $\rho(g) = I$ (Fact 10.11), hence $F_g = F$ (Proposition 12.18) and $B_{g;\tilde{x},\tilde{y}} = B_{\tilde{x},\tilde{y}}$ — for which the lemma's bijection — explicitly, $(\tilde{x}, \tilde{y}) \equiv (ax + cy,\ bx + d_0 y) \pmod 2$ — sends the pair at hand to $(\tilde{x}, \tilde{y}) = (1, 0)$. The lemma then gives $B_{x,y}(g * \tau) \propto_\zeta G(g, \tau)^{d/2} B_{1,0}(\tau)$; the right-hand side is not identically zero by the previous paragraph, and $\tau \mapsto g * \tau$ is a bijection of $\mathcal{H}$, so $B_{x,y} \not\equiv 0$.

- For $(x, y) = (1, 1)$: take $g = T^q = \begin{bmatrix} 1 & q \\ 0 & 1 \end{bmatrix} \equiv I \pmod q$. Then $(\tilde{x}, \tilde{y}) \equiv (1 + 0,\ q + 1) \equiv (1, 0) \pmod 2$, since q is odd.

- For $(x, y) = (0, 1)$: take $g = \begin{bmatrix} q^2 - q + 1 & q^2 \\ q & q + 1 \end{bmatrix}$, which has determinant $(q^2 - q + 1)(q + 1) - q^3 = 1$ and is $\equiv I \pmod{q}$. Then $(\tilde{x}, \tilde{y}) \equiv (c, d_0) = (q,\ q + 1) \equiv (1, 0) \pmod{2}$. □

Corollary 12.33. *The function $\tau \mapsto (B(\tau))^2$ is a weakly modular function for $\Gamma(q)$ with weight $3d$, with a one-dimensional representation of finite order.*

Proof. Fix $g \in \Gamma(q)$. By Lemma 12.31 and the bijectivity of $(x, y) \mapsto (\tilde{x}, \tilde{y})$:

$$B(g * \tau) = \prod_{(x,y)} B_{x,y}(g * \tau) \propto_\zeta G(g, \tau)^{3d/2} \prod_{(\tilde{x},\tilde{y})} B_{g;\tilde{x},\tilde{y}}(\tau) = G(g, \tau)^{3d/2}\, B_g(\tau),$$

where all products range over $\{(0, 1), (1, 0), (1, 1)\}$ and $G(g, \tau)^{3d/2} := \left(G(g, \tau)^{1/2}\right)^{3d}$. Since $g \in \Gamma(q)$, we have $F_g = F$ (Proposition 12.18) and hence $B_g = B$. Squaring gives $B^2(g * \tau) \propto_\zeta G(g, \tau)^{3d} B^2(\tau)$, now with an integer power of $G(g, \tau)$; since $B^2 \not\equiv 0$ (Proposition 12.32), Lemma 10.21 shows that B^2 is weakly modular of weight $3d$ for $\Gamma(q)$, with a one-dimensional representation of finite order. □

Lemma 12.34. *For all cusps $\mathfrak{c} \in \mathbb{Q} \cup \{\infty\}$, B^2 is holomorphic at $\mathfrak{c}$ and moreover we have $\mathrm{ord}_{\mathfrak{c}}(B^2) \geq \frac{d}{4} + 2\ell_{\mathfrak{c}} + 2H_{\mathfrak{c}}$, with $\ell_{\mathfrak{c}}$ and $H_{\mathfrak{c}}$ as in Definitions 12.25 and 12.26.*

Proof. See Appendix E. □

Thus we have shown that B^2 is a modular form for $\Gamma(q)$ of weight $3d$, with a one-dimensional representation of finite order.

12.6 Proof Step 4: Lower Bounding Cusp Orders

Recall the classification of the cusp classes of $\Gamma(q)$ (Proposition 12.23) and the per-class matrices $g_{\mathfrak{c}}$ and quantities $\ell_{\mathfrak{c}}, H_{\mathfrak{c}}$ fixed in Section 12.5.1. In this step we single out a family of *selected* cusp classes and lower bound $H_{\mathfrak{c}}$ on them, as well as upper bound the degrees to which the relevant expansions can be supported; these are the inputs that the final combination of inequalities in Section 12.7 will consume, alongside the valence bounds themselves.

By Proposition 12.23, for each $a \in \mathcal{A}$ and $\lambda \in \{1, \ldots, (q-1)/2\}$ we may speak of the cusp class $\pm(2a\lambda, \lambda)$; we write $\mathfrak{c}_{a,\lambda} := \pm\lambda(2a, 1)$ for this class, and we freely identify a class with any cusp representing it. We let $g_{a,\lambda} = \begin{bmatrix} 2a\lambda & -\lambda^{-1} \\ \lambda & 0 \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}_q)$ (a matrix of determinant 1), and — making the choice promised in Section 12.5.1 specific for these classes — we take the fixed matrix $g_{\mathfrak{c}_{a,\lambda}}$ to be an integral lift $g \in \mathrm{SL}_2(\mathbb{Z})$ with $g \equiv g_{a,\lambda} \pmod{q}$, which exists by the surjectivity in Lemma 10.5. Then $g * \infty$ is indeed a cusp in the class $\mathfrak{c}_{a,\lambda}$: it is the ratio of the entries of the first column of g, which reduces to $(2a\lambda, \lambda)$ modulo q. By Fact 10.11, the matrix $\rho(g)$ is determined by $g_{a,\lambda}$ alone. To compute it, factor $g_{a,\lambda} = p_{a,\lambda}\bar{S}$ in $\mathrm{SL}_2(\mathbb{Z}_q)$, where $\bar{S}$ is the reduction of S and $p_{a,\lambda} := \begin{bmatrix} \lambda^{-1} & 2a\lambda \\ 0 & \lambda \end{bmatrix}$ (direct multiplication). Since $p_{a,\lambda}$ has lower-left entry 0, any lift of it lies in $\Gamma_0(q)$, so (20) (with $\mu = (q + 1)/2 \equiv 2^{-1}$, so that $\mu \cdot 2a\lambda \cdot \lambda \equiv a\lambda^2$) gives $\rho(p_{a,\lambda})\mathbf{e}_y = \left(\frac{\lambda^{-1}}{q}\right) \zeta_q^{a\lambda^2 y^2} \mathbf{e}_{\lambda y}$. Combining this with $\rho(S)\mathbf{e}_x = \frac{\epsilon_q}{\sqrt{q}} \sum_y \zeta_q^{-xy} \mathbf{e}_y$, re-indexing $y \mapsto \lambda^{-1}y$, and noting $\left(\frac{\lambda^{-1}}{q}\right) = \left(\frac{\lambda}{q}\right)$ (Fact 10.8), we obtain, for every $x \in \mathbb{Z}_q$:

$$\rho(g_{a,\lambda})\mathbf{e}_x = c_{a,\lambda} \sum_{y \in \mathbb{Z}_q} \zeta_q^{ay^2 - \lambda^{-1}xy}\, \mathbf{e}_y, \qquad \text{where } c_{a,\lambda} := \frac{\epsilon_q}{\sqrt{q}} \left(\frac{\lambda}{q}\right) \neq 0. \tag{50}$$

We define the set of selected cusp classes

$$\mathfrak{C} := \left\{ \mathfrak{c}_{a,\lambda} \,:\, a \in \mathcal{A}, \lambda \in \left\{ 1, \dots, \frac{q-1}{2} \right\} \right\}.$$

These classes are pairwise distinct: if $\pm(2a\lambda, \lambda) = \pm(2a'\lambda', \lambda')$, then comparing second coordinates gives $\lambda' = \pm\lambda$, and since $\lambda, \lambda' \in \{1, \dots, (q-1)/2\}$ while $-\lambda \equiv q - \lambda \in \{(q+1)/2, \dots, q-1\} \pmod q$, the sign must be $+$; so $\lambda' = \lambda$, and then $2a\lambda \equiv 2a'\lambda$ forces $a' = a$. Hence $|\mathfrak{C}| = s \cdot \frac{q-1}{2}$.

In the following, we use the notation ϑ, ν introduced in Section 12.2.

Lemma 12.35. *For every $a \in \mathcal{A}$ and $\lambda \in \{1, \dots, (q-1)/2\}$, we have $H_{\mathfrak{c}_{a,\lambda}} > \nu d$.*

Proof. By Fact 10.11, $\rho(g_{\mathfrak{c}_{a,\lambda}}) = \rho(g_{a,\lambda})$. Substituting (50) and using the degree-d homogeneity of P_β to pull out the nonzero constant $c_{a,\lambda}^d$, we have:

$$\begin{aligned} P_\beta \circ \rho(g_{a,\lambda}) \left(\sum_{x \in \mathbb{Z}_q} X_x \mathbf{e}_x \right) &= c_{a,\lambda}^d \, P_\beta \left(\sum_{x \in \mathbb{Z}_q} \sum_{y \in \mathbb{Z}_q} X_x \zeta_q^{ay^2 - \lambda^{-1}xy} \mathbf{e}_y \right) \\ &= c_{a,\lambda}^d \sum_{\mathbf{y} \in \mathbb{Z}_q^d} \beta_{\mathbf{y}} \prod_{i=1}^{d} \left(\sum_{x \in \mathbb{Z}_q} X_x \zeta_q^{ay_i^2 - \lambda^{-1}xy_i} \right) \\ &= c_{a,\lambda}^d \sum_{\mathbf{x} \in \mathbb{Z}_q^d} X_{x_1} \dots X_{x_d} \left(\sum_{\mathbf{y} \in \mathbb{Z}_q^d} \beta_{\mathbf{y}} \zeta_q^{a\langle \mathbf{y}, \mathbf{y} \rangle - \lambda^{-1}\langle \mathbf{x}, \mathbf{y} \rangle} \right). \end{aligned}$$

By (45) (applied with $\mathbf{z} = -\lambda^{-1}\mathbf{x}$, noting $a \in \mathcal{A}$), the bracketed coefficient vanishes whenever $-\lambda^{-1}\mathbf{x} \in T_d(\vartheta, \nu)$; and since $T_d(\vartheta, \nu)$ is closed under coordinatewise multiplication by units (multiplying $u\mathbf{z} + b\mathbf{1}$ by a unit μ gives $(\mu u)\mathbf{z} + (\mu b)\mathbf{1}$), this happens if and only if $\mathbf{x} \in T_d(\vartheta, \nu)$. In particular, the coefficient vanishes for every $\mathbf{x}$ with $\sum_{i=1}^d w_q(x_i) \le \nu d$, as such $\mathbf{x}$ lie in $S_d^{\mathrm{sq}}(\nu) \subseteq T_d(\vartheta, \nu)$. The conclusion is now immediate from Definition 12.26. □

Lemma 12.36. *Let $\mathfrak{c}$ be a cusp class, with fixed matrix $g_{\mathfrak{c}}$ and expansion functions $\Phi_{g_{\mathfrak{c}},\ell}$ as in Section 12.5.1. There exists a polynomial $p_{\mathfrak{c}} \in \mathbb{C}[X]$ of degree $\le d$ such that:*

$$F_{g_{\mathfrak{c}}}(\tau, z) = e^{-2\pi i z d/(2q)} \, p_{\mathfrak{c}}\left(e^{2\pi i z/q}\right) \exp\left(\frac{\pi i \tau (d + 8\ell_{\mathfrak{c}})}{4q} \right) + \sum_{\substack{\ell > d + 8\ell_{\mathfrak{c}} \\ \ell \equiv d \pmod 8}} \Phi_{g_{\mathfrak{c}},\ell}(z/q) \exp\left(\frac{\pi i \tau \ell}{4q} \right). \tag{51}$$

Proof. See Appendix E. □

Lemma 12.37. *In the setting of Lemma 12.36, the exponents of the monomials with nonzero coefficient in $p_{\mathfrak{c}}$ all lie in a single residue class modulo q. Consequently, there exist an integer $r \in \{0, \dots, q-1\}$ and a univariate polynomial $f_{\mathfrak{c}}$ of degree $\le (d-r)/q$ such that $p_{\mathfrak{c}}(X) = X^r f_{\mathfrak{c}}(X^q)$.*

Proof. By Lemma 12.20 (with $(a, b) = (0, 1)$), there is a nonzero constant ω, independent of τ and z, such that:

$$F_{g_{\mathfrak{c}}}(\tau, z+1) = \omega F_{g_{\mathfrak{c}}}(\tau, z).$$

Substituting (51) into both sides and matching the coefficients of $\exp(\pi i \tau (d + 8\ell_{\mathfrak{c}})/(4q))$ — which are determined, for each fixed z, by the uniqueness in Fact B.4 — we obtain, for every $z \in \mathbb{C}$:

$$e^{-2\pi i (z+1) d/(2q)} p_{\mathfrak{c}}(e^{2\pi i (z+1)/q}) = \omega e^{-2\pi i z d/(2q)} p_{\mathfrak{c}}(e^{2\pi i z/q})$$

$$\Leftrightarrow p_{\mathfrak{c}}(\zeta_q e^{2\pi i z/q}) = \omega' p_{\mathfrak{c}}(e^{2\pi i z/q}), \qquad \text{where } \omega' := \omega e^{\pi i d/q}.$$

Since $e^{2\pi i z/q}$ ranges over all of $\mathbb{C}^*$ as z does, we conclude $p_{\mathfrak{c}}(\zeta_q X) = \omega' p_{\mathfrak{c}}(X)$ identically. Comparing coefficients, $\zeta_q^R = \omega'$ whenever the coefficient of X^R is nonzero; so any two such exponents R, R' satisfy $\zeta_q^{R-R'} = 1$, i.e., $R \equiv R' \pmod q$. Letting $r \in \{0, \dots, q-1\}$ be their common residue (or $r := 0$ if $p_{\mathfrak{c}} = 0$), and recalling $\deg p_{\mathfrak{c}} \le d$, this is exactly the displayed conclusion. □

Lemma 12.38. *In the setting of Lemmas 12.36 and 12.37, $f_{\mathfrak{c}}$ is divisible by $(X-1)^m$. Consequently, $p_{\mathfrak{c}}$ is divisible by $(X^q-1)^m$.*

Proof. By Proposition 12.28, the Taylor coefficients of $F_{g_{\mathfrak{c}}}$ in z of index below m vanish identically; by the uniqueness of the coefficients in (49) (Fact B.4), this forces $\Xi_{g_{\mathfrak{c}};j,\ell} = 0$ for all $j < m$ and all ℓ. In particular, the entire function $\Phi_{g_{\mathfrak{c}},\ell^*}$ (whose Taylor coefficients are the $\Xi_{g_{\mathfrak{c}};j,\ell^*}$) vanishes to order at least m at 0; and by the construction of $p_{\mathfrak{c}}$ (see Appendix E), $\Phi_{g_{\mathfrak{c}},\ell^*}(z/q) = e^{-2\pi i z d/(2q)}\, p_{\mathfrak{c}}(e^{2\pi i z/q})$. Consequently, we have:

$$\begin{aligned}
e^{-2\pi i z d/(2q)} p_{\mathfrak{c}}(e^{2\pi i z/q}) &= O(z^m) \\
\Rightarrow p_{\mathfrak{c}}(e^{2\pi i z/q}) &= e^{2\pi i z d/(2q)} \cdot O(z^m) \\
&= O(z^m) \\
\Rightarrow e^{2\pi i r z/q} f_{\mathfrak{c}}(e^{2\pi i z}) &= O(z^m) \quad \text{(Lemma 12.37: } p_{\mathfrak{c}}(X) = X^r f_{\mathfrak{c}}(X^q)\text{)} \\
\Rightarrow f_{\mathfrak{c}}(e^{2\pi i z}) &= O(z^m).
\end{aligned}$$

To see that this implies the conclusion, first note that if $f_{\mathfrak{c}} \equiv 0$ then the conclusion holds trivially; so assume $f_{\mathfrak{c}} \not\equiv 0$ and write $f_{\mathfrak{c}}(X) = (X-1)^t h(X)$ for some $t \ge 0$ and a polynomial h with $h(1) \neq 0$. Then we have:

$$\begin{aligned}
f_{\mathfrak{c}}(e^{2\pi i z}) &= (e^{2\pi i z} - 1)^t h(e^{2\pi i z}) \\
&= (2\pi i z + O(z^2))^t (h(1) + O(z)).
\end{aligned}$$

The lowest-order term on the RHS is $(2\pi i)^t h(1) z^t$, so since $h(1) \neq 0$ this forces $t \ge m$ as desired. Finally, substituting X^q for X, the polynomial $(X^q-1)^m$ divides $f_{\mathfrak{c}}(X^q)$, and hence divides $p_{\mathfrak{c}}(X) = X^r f_{\mathfrak{c}}(X^q)$. □

Corollary 12.39. *In the setting of Lemma 12.36, we have $qm \le d$. Moreover, there exists an integer $R \in [0, d]$ such that:*

- *the coefficient of X^R in $p_{\mathfrak{c}}(X)$ is nonzero; and*
- $\min(R, d-R) \le \frac{d-qm}{2}$.

Proof. First, $p_{\mathfrak{c}}$ is not the zero polynomial: by the construction of $p_{\mathfrak{c}}$ (see Appendix E), $e^{-2\pi i z d/(2q)}\, p_{\mathfrak{c}}(e^{2\pi i z/q}) = \Phi_{g_{\mathfrak{c}},\ell^*}(z/q)$, which is not identically zero by the minimality of ℓ^* in Definition 12.25. Call the *spread* of a nonzero univariate polynomial the difference of its largest and smallest exponents with a nonzero coefficient. Then the spread of $f_1 f_2$ is the sum of the spreads of f_1 and f_2, as long as neither f_1 nor f_2 is the zero polynomial: the lowest terms of f_1 and f_2 multiply to the (nonzero) lowest term of $f_1 f_2$, and likewise for the highest terms. Since $(X^q-1)^m$ divides $p_{\mathfrak{c}}$ (Lemma 12.38) and has spread qm, the spread of $p_{\mathfrak{c}}$ is at least qm. As the exponents of $p_{\mathfrak{c}}$ lie in $[0, d]$ (Lemma 12.36), this already forces $qm \le d$. Moreover, the exponents that appear in $p_{\mathfrak{c}}$ with nonzero coefficient cannot all lie in the open interval $\left(\frac{d-qm}{2}, \frac{d+qm}{2}\right)$, whose width is exactly qm; so some exponent R satisfies $R \le \frac{d-qm}{2}$ or $R \ge \frac{d+qm}{2}$, i.e., $\min(R, d-R) \le \frac{d-qm}{2}$. □

Lemma 12.40. *For every cusp $\mathfrak{c} \in \mathfrak{C}$, we have:*

$$\gamma d < 2\ell_{\mathfrak{c}} + \vartheta(d - qm).$$

Proof. Write $\mathfrak{c} = \mathfrak{c}_{a,\lambda}$, and let $R \in [0, d]$ be the integer guaranteed by Corollary 12.39 for this cusp class. Let us unwrap the definition of $F_{g_{\mathfrak{c}}}(\tau, z)$ a bit, using $\rho(g_{\mathfrak{c}}) = \rho(g_{a,\lambda})$ (Fact 10.11):

$$\begin{aligned}
&F_{g_{\mathfrak{c}}}(\tau, z) \\
&= P_\beta(\rho(g_{\mathfrak{c}})\Theta(\tau, z/q)) \\
&= P_\beta\left(\rho(g_{a,\lambda}) \sum_{k\in\mathbb{Z}} (-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{z}{q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\right) \\
&\propto_\zeta P_\beta\left(\rho(g_{a,\lambda}) \sum_{k\in\mathbb{Z}} (-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{z}{q}\right)\mathbf{e}_{k \bmod q}\right) \quad \text{(Proposition 12.9, with } \mathsf{X}^{(q+1)/2}\text{)} \\
&\propto P_\beta\left(\sum_{k\in\mathbb{Z}}\sum_{y\in\mathbb{Z}_q} (-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{z}{q}\right)\exp\left(\frac{2\pi i}{q}\left(ay^2 - \lambda^{-1}ky\right)\right)\mathbf{e}_y\right) \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^{d}\left(\sum_{k\in\mathbb{Z}} (-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{z}{q}\right)\exp\left(\frac{2\pi i}{q}\left(ax_i^2 - \lambda^{-1}kx_i\right)\right)\right) \\
&= \sum_{k_1,\ldots,k_d\in\mathbb{Z}} (-1)^{k_1+\ldots+k_d} \exp\left(\frac{\pi i\tau}{q}\sum_i\left(k_i-\frac{1}{2}\right)^2 + 2\pi i\sum_i\left(k_i-\frac{1}{2}\right)\frac{z}{q}\right)\left[\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\beta_{\mathbf{x}}\zeta_q^{a\langle\mathbf{x},\mathbf{x}\rangle - \lambda^{-1}\langle\mathbf{k},\mathbf{x}\rangle}\right].
\end{aligned}$$

(In the third line we used the X-invariance, up to a root of unity, of $P_\beta \circ \rho(g)$ — the fourth bullet of Proposition 12.9; in the fourth we substituted (50) and pulled out the nonzero constant $c_{a,\lambda}^d$ by homogeneity — this constant, like the root of unity hidden in $\propto_\zeta$, is immaterial for what follows; the rearrangements are justified by absolute convergence, exactly as in Appendix E.)

A tuple $(k_1, \ldots, k_d)$ above contributes to the coefficient of X^R in $p_{\mathfrak{c}}$ exactly when $\sum_i (k_i - 1/2)^2 = d/4 + 2\ell_{\mathfrak{c}}$ and $\sum_i k_i = R$: such a term has the τ-exponential of level $d + 8\ell_{\mathfrak{c}}$ and the z-exponential $e^{2\pi i(R-d/2)z/q}$, matching the leading part of (51). Moreover, by (45) (with $\mathbf{z} = -\lambda^{-1}\mathbf{k} \bmod q$, noting $a \in \mathcal{A}$), the bracketed factor vanishes whenever $-\lambda^{-1}\mathbf{k} \bmod q \in T_d(\vartheta, \nu)$, which — $T_d(\vartheta, \nu)$ being closed under multiplication by units, as in the proof of Lemma 12.35 — happens if and only if $\mathbf{k} \bmod q \in T_d(\vartheta, \nu)$. Since the coefficient of X^R in $p_{\mathfrak{c}}$ is *nonzero* (Corollary 12.39), at least one contributing tuple has a nonvanishing bracket. Fix such a tuple $\mathbf{k} \in \mathbb{Z}^d$; it satisfies:

- $\mathbf{k} \bmod q \notin T_d(\vartheta, \nu)$;
- $\sum_{i=1}^d (k_i - 1/2)^2 = d/4 + 2\ell_{\mathfrak{c}}$ — equivalently, $\sum_{i=1}^d k_i(k_i - 1) = 2\ell_{\mathfrak{c}}$;
- $\sum_{i=1}^d k_i = R$.

At this point, we use the fact from Corollary 12.39 that either R or $d - R$ is $\leq \frac{d-qm}{2}$, and address these two cases separately:

- If $R \leq \frac{d-qm}{2}$: taking $(u, b) = (1, 0)$ in the definition of $T_d(\vartheta, \nu)$ (Definition 12.1), the first bullet implies $\mathbf{k} \bmod q \notin S_d^{\text{edge}}(\vartheta, \nu)$, i.e., $\sum_i w_{q,\vartheta}(k_i \bmod q) > \gamma d$. Bounding each summand by its integer representative k_i (Definition 5.4), we have:

$$\gamma d < \sum_{i=1}^{d} w_{q,\vartheta}(k_i \bmod q)$$

$$\leq \sum_{i=1}^{d} (k_i(k_i-1) + 2\vartheta k_i)$$
$$= 2\ell_{\mathfrak{c}} + 2\vartheta R$$
$$\leq 2\ell_{\mathfrak{c}} + \vartheta(d - qm).$$

- If $d - R \leq \frac{d-qm}{2}$: taking $(u, b) = (-1, 1)$ instead, the first bullet implies $(\mathbf{1} - \mathbf{k}) \bmod q \notin S_d^{\text{edge}}(\vartheta, v)$ (if $\mathbf{1} - \mathbf{k} \equiv \mathbf{z}$ for some $\mathbf{z} \in S_d^{\text{edge}}(\vartheta, v)$, then $\mathbf{k} \equiv -\mathbf{z} + \mathbf{1} \in T_d(\vartheta, v)$), i.e., $\sum_i w_{q,\vartheta}((1 - k_i) \bmod q) > \gamma d$. Bounding each summand by the integer representative $1 - k_i$, whose value is $(1 - k_i)((1 - k_i) - 1) + 2\vartheta(1 - k_i) = k_i(k_i - 1) + 2\vartheta(1 - k_i)$, we have:

$$\gamma d < \sum_{i=1}^{d} w_{q,\vartheta}((1 - k_i) \bmod q)$$
$$\leq \sum_{i=1}^{d} (k_i(k_i - 1) + 2\vartheta(1 - k_i))$$
$$= 2\ell_{\mathfrak{c}} + 2\vartheta(d - R)$$
$$\leq 2\ell_{\mathfrak{c}} + \vartheta(d - qm).$$

Either way, we are done. □

12.7 Proof Step 5: Applying the Valence Formula

We now apply the valence formula to the two modular forms constructed in Step 3 and combine the resulting inequalities with Step 4's per-class bounds. Recall the standing hypothesis of this proof (Section 12.3): for the sake of contradiction, we assumed that the coefficients required by Theorem 12.2 do not exist, and from this assumption constructed the symmetrized witness β; the function F, the integer m, the modular forms A_m^2 and B^2, and the per-class quantities $\ell_{\mathfrak{c}}$ and $H_{\mathfrak{c}}$ were all built from that β. In the two lemmas below, the sums range over all $(q^2 - 1)/2$ cusp classes of $\Gamma(q)$ (Proposition 12.23).

Lemma 12.41. *Summing over all cusp classes of $\Gamma(q)$, we have:*

$$\sum_{\mathfrak{c}} \ell_{\mathfrak{c}} \leq \frac{d(q^2-1)(q-3)}{48} + \frac{m(q^3-q)}{24}.$$

Proof. We apply Corollary 10.36 to A_m^2, which is a nonzero modular form of weight $d + 2m$ for $\Gamma(q)$ (Lemma 12.29; it is nonzero since $A_m \not\equiv 0$ by the definition of m) whose one-dimensional representation has finite order (Lemma 12.29), as Corollary 10.36 requires. Since A_m^2 is holomorphic, its interior orders are nonnegative, so dropping the interior sum in (29) and inserting the cusp-order bound of Lemma 12.29 gives:

$$\frac{d + 2m}{24} \cdot \frac{q^3 - q}{2} \geq \frac{1}{2} \sum_{\mathfrak{c}} \operatorname{ord}_{\mathfrak{c}}(A_m^2)$$
$$\geq \frac{1}{2} \sum_{\mathfrak{c}} \left(2\ell_{\mathfrak{c}} + \frac{d}{4} \right)$$
$$= \sum_{\mathfrak{c}} \ell_{\mathfrak{c}} + \frac{d}{8} \cdot \frac{q^2 - 1}{2}$$

$$\Leftrightarrow \sum_{\mathfrak{c}} \ell_{\mathfrak{c}} \le \frac{d(q^2-1)(q-3)}{48} + \frac{m(q^3-q)}{24},$$

as desired. □

Lemma 12.42. *Summing over all cusp classes of* $\Gamma(q)$*, we have:*

$$\sum_{\mathfrak{c}} \ell_{\mathfrak{c}} + \sum_{\mathfrak{c}} H_{\mathfrak{c}} \le \frac{d(q^2-1)(q-1)}{16}.$$

Proof. We apply Corollary 10.36 to B^2, which is a nonzero modular form of weight $3d$ for $\Gamma(q)$ (Corollary 12.33, Lemma 12.34, and Proposition 12.32) whose one-dimensional representation has finite order (Corollary 12.33), as Corollary 10.36 requires. Since B^2 is holomorphic, its interior orders are nonnegative, so dropping the interior sum in (29) and inserting the cusp-order bound of Lemma 12.34 gives:

$$\begin{aligned}
\frac{3d}{24} \cdot \frac{q^3-q}{2} &\ge \frac{1}{2} \sum_{\mathfrak{c}} \operatorname{ord}_{\mathfrak{c}}(B^2) \\
&\ge \frac{1}{2} \sum_{\mathfrak{c}} \left(\frac{d}{4} + 2\ell_{\mathfrak{c}} + 2H_{\mathfrak{c}} \right) \\
&= \sum_{\mathfrak{c}} \ell_{\mathfrak{c}} + \sum_{\mathfrak{c}} H_{\mathfrak{c}} + \frac{d}{8} \cdot \frac{q^2-1}{2} \\
\Leftrightarrow \sum_{\mathfrak{c}} \ell_{\mathfrak{c}} + \sum_{\mathfrak{c}} H_{\mathfrak{c}} &\le \frac{d(q^2-1)(q-1)}{16}.
\end{aligned}$$

□

Both valence bounds are instances of a known strategy. Lemma 12.41 applies the valence formula to (a power of) the first nonvanishing Taylor coefficient, which is the approach of [LW26, Lemma 3.1] (see also [Aok22, Lemma 5.1]); indeed, it is essentially the case $t = 1$ of [LW26, Lemma 3.1], applied to a suitable norm of F. Lemma 12.42 applies it to the product of the restrictions to the nonzero 2-torsion points, in the same spirit as [RWR15, Aok22, LW26]. What is particular to our setting is how the cusp orders entering these bounds are controlled by the witness, through $\ell_{\mathfrak{c}}$ and $H_{\mathfrak{c}}$ (Lemmas 12.29, 12.34, 12.35 and 12.40).

We can finally complete the proof of Theorem 12.2:

Proof of Theorem 12.2. We collect here the four inequalities we will use, from Lemmas 12.35 and 12.40 to 12.42 respectively. In the first two, the sums range over all cusp classes; the third holds for every selected class $\mathfrak{c}_{a,\lambda}$; the fourth holds for every $\mathfrak{c} \in \mathfrak{C}$. (We will also freely use that $\ell_{\mathfrak{c}} \ge 0$ and $H_{\mathfrak{c}} \ge 0$ for *every* class — the former by Definition 12.25, the latter from Definition 12.26 since $w_q \ge 0$ — and that $|\mathfrak{C}| = s \cdot \frac{q-1}{2}$.)

$$\sum_{\mathfrak{c}} \ell_{\mathfrak{c}} \le \frac{d(q^2-1)(q-3)}{48} + \frac{m(q^3-q)}{24}$$

$$\Leftrightarrow d - qm \le \frac{d(q-1)}{2} - \frac{24}{q^2-1} \sum_{\mathfrak{c}} \ell_{\mathfrak{c}} \quad \text{(multiply by } \tfrac{24}{q^2-1} \text{ and rearrange)} \tag{52}$$

$$\sum_{\mathfrak{c}} \ell_{\mathfrak{c}} + \sum_{\mathfrak{c}} H_{\mathfrak{c}} \le \frac{d(q^2-1)(q-1)}{16} \tag{53}$$

$$H_{\mathfrak{c}_{a,\lambda}} > \nu\, d \tag{54}$$

$$\gamma d < 2\ell_{\mathfrak{c}} + \vartheta(d - qm). \tag{55}$$

First, summing (54) over the $s \cdot \frac{q-1}{2}$ selected classes and using $H_{\mathfrak{c}} \geq 0$ for the others gives $\sum_{\mathfrak{c}} H_{\mathfrak{c}} \geq \sum_{\mathfrak{c}\in\mathfrak{C}} H_{\mathfrak{c}} > \frac{\nu d s(q-1)}{2}$; plugging this into (53) yields:

$$\sum_{\mathfrak{c}} \ell_{\mathfrak{c}} + \frac{\nu d s(q-1)}{2} < \frac{d(q^2-1)(q-1)}{16}. \tag{56}$$

Now summing (55) over all $\mathfrak{c} \in \mathfrak{C}$ and dividing by $|\mathfrak{C}| = s \cdot \frac{q-1}{2}$ yields the first line below:

$$\begin{aligned}
\gamma d &< \frac{4}{s(q-1)} \sum_{\mathfrak{c}\in\mathfrak{C}} \ell_{\mathfrak{c}} + \vartheta(d - qm) \\
&\leq \frac{4}{s(q-1)} \sum_{\mathfrak{c}\in\mathfrak{C}} \ell_{\mathfrak{c}} + \frac{\vartheta d(q-1)}{2} - \frac{24\vartheta}{q^2-1} \sum_{\mathfrak{c}} \ell_{\mathfrak{c}} \quad \text{(plugging in (52), using } \vartheta \geq 0) \\
&\leq \frac{4}{s(q-1)} \underbrace{\left(1 - \frac{6\vartheta s}{q+1}\right)}_{\geq 0} \sum_{\mathfrak{c}\in\mathfrak{C}} \ell_{\mathfrak{c}} + \frac{\vartheta d(q-1)}{2} \\
&\leq \frac{4}{s(q-1)} \left(1 - \frac{6\vartheta s}{q+1}\right) \left(\frac{d(q^2-1)(q-1)}{16} - \frac{\nu d s(q-1)}{2}\right) + \frac{\vartheta d(q-1)}{2} \quad \text{(plugging in (56))} \\
&= \left(1 - \frac{6\vartheta s}{q+1}\right) \left(\frac{d(q^2-1)}{4s} - 2\nu d\right) + \frac{\vartheta d(q-1)}{2} \\
\Leftrightarrow\ \gamma &< \left(1 - \frac{6\vartheta s}{q+1}\right) \left(\frac{q^2-1}{4s} - 2\nu\right) + \frac{\vartheta(q-1)}{2}.
\end{aligned}$$

Here the third line drops the non-selected classes from $\sum_{\mathfrak{c}} \ell_{\mathfrak{c}}$ — valid since $\ell_{\mathfrak{c}} \geq 0$ and the coefficient $-\frac{24\vartheta}{q^2-1}$ is nonpositive — and then uses the identity $\frac{24\vartheta}{q^2-1} = \frac{4}{s(q-1)} \cdot \frac{6\vartheta s}{q+1}$; the underbraced factor is nonnegative because $\vartheta \leq \frac{q+1}{6s}$ (Definition 12.1), which also justifies the fourth line, together with $\sum_{\mathfrak{c}\in\mathfrak{C}} \ell_{\mathfrak{c}} \leq \sum_{\mathfrak{c}} \ell_{\mathfrak{c}}$ and (56); and the final line divides by $d > 0$. But the right-hand side of the final line is *exactly* the definition of $\gamma = \gamma(\vartheta, \nu)$ in Definition 12.1, so we have derived $\gamma < \gamma$ — a contradiction. Hence the assumed coefficients exist after all, proving Theorem 12.2. □

Remark 12.43 (The case $\vartheta = 0$). *For the boundary parameters $\vartheta = 0$ and $\nu = \frac{q^2-1}{8s}$, the proof above simplifies considerably. Here $\gamma(0, \nu) = 0$, and summing* (54) *over the $s \cdot \frac{q-1}{2}$ selected classes gives $\sum_{\mathfrak{c}} H_{\mathfrak{c}} > \frac{d(q^2-1)(q-1)}{16}$, which, together with $\ell_{\mathfrak{c}} \geq 0$, already contradicts* (53). *So in this case only the scalarization via 2-torsion evaluations and a single application of the valence formula are needed — not the scalarization via Taylor coefficients, Lemma 12.41, or the edge-polynomial analysis of Step 4 (Lemma 12.36 through Lemma 12.40). This case already gives a nontrivial support exponent for every odd prime q and every $s \in [2, q]$: the edge set $S_d^{\mathrm{edge}}(0, \nu)$ is $\{0,1\}^d$ (as $w_{q,0}(j) = 0$ exactly when $j \in \{0,1\}$), so $E_{q,s}^{\mathrm{edge}}(0, \nu) = \log_q 2$, while $E_q^{\mathrm{sq}}(\nu) < 1$ because ν is smaller than the average $\frac{q^2-1}{12}$ of w_q over $\mathbb{Z}_q$ (for $s \geq 2$). It is, however, weaker than the full result: for $(q, s) = (5, 4), (7, 6), (7, 7), (11, 10)$, the case $\vartheta = 0$ gives exponents of approximately* 0.791, 0.729, 0.690, *and* 0.676, *respectively, compared with the bounds of Lemma E.3. The scalarization via Taylor coefficients and the edge-polynomial analysis of Step 4 make positive tilts $\vartheta > 0$ available, and all of the bounds in Lemma E.3 use such tilts.*

Statement on AI Use

Activity	AI used
Asking the research question	No
Coming up with the approach	Yes
Proof development	Yes
Writing and exposition	Yes
Checking for bugs	Yes
Other supporting tasks	Yes

Details (elaborating on the points above). The results of this paper were obtained via months-long conversations with ChatGPT (5.5–6 Astra), seeded with ideas of the authors and steered by the authors in its exploration of the search space. Initially, we gave prompts asking for any improvement on the tradeoffs of [HR26], uploading our pre-existing research notes and specifying concrete ideas towards adapting matching-vector codes. After some back-and-forth, this led to the Track 1 results (Main Theorem 5.2) and a different ad-hoc proof of the Track 3 result (Main Theorem 5.5) in the 3-server case. We identified Fourier-sparse cone-delta functions as the key ingredient underlying these protocols.

Generalizing these constructions of Fourier-sparse cone-delta functions to infinitely many server counts (in fact, even one server count > 3) took several more weeks of iteration with ChatGPT. The first general constructions were along the lines of Track 2 (Main Theorem 5.3), but their initial presentations were more mathematically involved than anything in this paper. We spent several more weeks steering ChatGPT towards results that were stronger and had less mathematically involved proofs, culminating in Track 3 (Main Theorem 5.5) which only uses machinery from complex analysis and some number theory.

The aforementioned GPT models and Claude (Opus 5, Opus 5.5, and Fable 5) were all used for writing, in particular to draft and verify proofs. The authors are fully accountable for the contents and correctness of this paper.

Acknowledgements

We thank Vinod Vaikuntanathan, Henry Corrigan-Gibbs, and Dan Boneh for encouraging this work and for numerous useful discussions on PIR. We also thank Assaf Naor, Peter Sarnak, and Andrew Sutherland for teaching beautiful courses that helped with the presentation of this work.

References

[ABL25] Bar Alon, Amos Beimel, and Or Lasri. Simplified PIR and CDS protocols and improved linear secret-sharing schemes. In *TCC*, 2025. 15, 27

[AL07] Noga Alon and Eyal Lubetzky. Graph powers, Delsarte, Hoffman, Ramsey, and Shannon. *SIAM Journal on Discrete Mathematics*, 21(2):329–348, 2007. 8, 29, 30, 49

[Aok22] Hiroki Aoki. On the upper bound of the orders of Jacobi forms. *International Journal of Mathematics*, 33(8):2250056, 2022. 14, 17, 38, 60, 64, 74, 80

[App05] D. M. Appleby. SIC-POVMs and the extended Clifford group. *Journal of Mathematical Physics*, 46(5):052107, 2005. arXiv:quant-ph/0412001. 9

[Ban93] Wojciech Banaszczyk. New bounds in some transference theorems in the geometry of numbers. *Math. Ann.*, 296(4):625–635, 1993. 17, 91

[BE72] Michel Broué and Michel Enguehard. Polynômes des poids de certains codes et fonctions thêta de certains réseaux. *Annales Scientifiques de l'École Normale Supérieure (4)*, 5:157–181, 1972. 12, 17, 38, 47, 50

[BHMW21] Elette Boyle, Justin Holmgren, Fermi Ma, and Mor Weiss. On the security of doubly efficient PIR. Cryptology ePrint Archive, Report 2021/1113, 2021. 15

[BHW19] Elette Boyle, Justin Holmgren, and Mor Weiss. Permuted puzzles and cryptographic hardness. In *TCC*. Springer, 2019. 15

[BIM00] Amos Beimel, Yuval Ishai, and Tal Malkin. Reducing the servers computation in private information retrieval: PIR with preprocessing. In Mihir Bellare, editor, *Advances in Cryptology - CRYPTO 2000, 20th Annual International Cryptology Conference, Santa Barbara, California, USA, August 20-24, 2000, Proceedings*, volume 1880 of *Lecture Notes in Computer Science*, pages 55–73. Springer, 2000. 1, 16

[BIM04] Amos Beimel, Yuval Ishai, and Tal Malkin. Reducing the servers' computation in private information retrieval: PIR with preprocessing. *Journal of Cryptology*, 17(2):125–151, February 2004. 1, 4, 15, 18

[BIPW17] Elette Boyle, Yuval Ishai, Rafael Pass, and Mary Wootters. Can we access a database both locally and privately? In *TCC*, pages 662–693. Springer, 2017. 1, 15, 16

[BL26] Amos Beimel and Or Lasri. Private Information Retrieval: Share conversions vs decoding polynomials. In *EUROCRYPT*, 2026. 27

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. 94

[Bor98] Richard E. Borcherds. Automorphic forms with singularities on Grassmannians. *Inventiones Mathematicae*, 132(3):491–562, 1998. 8, 12, 13, 37, 38, 41, 50, 55, 56

[BW21] Keller Blackwell and Mary Wootters. A note on the permuted puzzles toy conjecture. *arXiv preprint arXiv:2108.07885*, 2021. 15

[CE03] Henry Cohn and Noam Elkies. New upper bounds on sphere packings I. *Annals of Mathematics*, 157(2):689–714, 2003. 17

[CFL+13] Yeow Meng Chee, Tao Feng, San Ling, Huaxiong Wang, and Liang Feng Zhang. Query-efficient locally decodable codes of subexponential length. *Computational Complexity*, 22(1):159–189, 2013. 15

[CGKS95] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In *FOCS*, pages 41–50. IEEE Computer Society, 1995. 1, 10

[CHR17] Ran Canetti, Justin Holmgren, and Silas Richelson. Towards doubly efficient private information retrieval. In *TCC*, pages 694–726. Springer, 2017. 1, 15, 16

[CIJ+26] Caicai Chen, Yuval Ishai, Aayush Jain, Tamer Mour, Alon Rosen, and Chaoping Xing. Doubly-efficient secret-key PIR with low storage overhead. *IACR Cryptol. ePrint Arch.*, 2026:1480, 2026. 15

[CIMR26] Caicai Chen, Yuval Ishai, Tamer Mour, and Alon Rosen. Secret-key PIR from random linear codes. In *STOC*, 2026. Full version: Cryptology ePrint Archive, Paper 2025/646. 15

[CKM+17] Henry Cohn, Abhinav Kumar, Stephen D. Miller, Danylo Radchenko, and Maryna Viazovska. The sphere packing problem in dimension 24. *Annals of Mathematics*, 185(3):1017–1033, 2017. 17

[Dav80] Harold Davenport. *Multiplicative Number Theory*, volume 74 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1980. Revised by Hugh L. Montgomery. 35, 89

[Del73] Philippe Delsarte. An algebraic approach to the association schemes of coding theory. *Philips Research Reports Supplements*, 10, 1973. 17

[DG16] Zeev Dvir and Sivakanth Gopi. 2-server PIR with subpolynomial communication. *J. ACM*, 2016. 1, 4, 15, 25

[DPC23] Alex Davidson, Gonçalo Pestana, and Sofía Celi. FrodoPIR: Simple, scalable, single-server private information retrieval. *PETS*, 2023. 16

[DPT10] Yevgeniy Dodis, Mihai Pătraşcu, and Mikkel Thorup. Changing base without losing space. In Leonard J. Schulman, editor, *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 593–602. ACM, 2010. 7, 22, 26, 35

[DS05] Fred Diamond and Jerry Shurman. *A First Course in Modular Forms*, volume 228 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2005. 10, 37, 38, 40, 42, 48, 57, 61, 71, 90

[Efr09] Klim Efremenko. 3-query locally decodable codes of subexponential length. In *STOC*, 2009. 1, 4, 5, 15, 27

[EZ85] Martin Eichler and Don Zagier. *The Theory of Jacobi Forms*, volume 55 of *Progress in Mathematics*. Birkhäuser, Boston, 1985. 14, 17, 38, 60, 62, 63, 69, 73, 74

[Fre86] Gerhard Frey. Links between stable elliptic curves and certain Diophantine equations. *Annales Universitatis Saraviensis. Series Mathematicae*, 1(1):1–40, 1986. 17

[Gér77] Paul Gérardin. Weil representations associated to finite fields. *Journal of Algebra*, 46:54–101, 1977. 9, 37, 41, 62

[GH07] Shamgar Gurevich and Ronny Hadani. The geometric Weil representation. *Selecta Mathematica (N.S.)*, 13:465–481, 2007. 62

[GK26] Fatemeh Ghasemi and Swastik Kopparty. Fourier sparsity of delta functions and matching vector PIRs. In Shubhangi Saraf, editor, *17th Innovations in Theoretical Computer Science Conference, ITCS 2026, Bocconi University, Milan, Italy, January 27-30, 2026*, volume 362 of *LIPIcs*, pages 68:1–68:22. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2026. 2, 4, 5, 16, 19, 33

[GKS25] Fatemeh Ghasemi, Swastik Kopparty, and Madhu Sudan. Improved PIR schemes using matching vectors and derivatives. In *STOC*, 2025. 1, 15, 27

[GLM+25] Ashrujit Ghoshal, Baitian Li, Yaohua Ma, Chenxin Dai, and Elaine Shi. Scalable multi-server private information retrieval. In *TCC*. Springer, 2025. 1, 15, 16

[GR27] Aparna Gupte and Seyoon Ragavan. Exponentially fewer-server PIR from sparser s-decoding polynomials. In *SODA*, 2027. 15, 16, 27

[HHCG+23] Alexandra Henzinger, Matthew M. Hong, Henry Corrigan-Gibbs, Sarah Meiklejohn, and Vinod Vaikuntanathan. One server for the price of two: Simple and fast single-server private information retrieval. In *USENIX Security*, 2023. 16

[HK71] Kenneth Hoffman and Ray Kunze. *Linear Algebra*. Prentice-Hall, Englewood Cliffs, NJ, second edition, 1971. 45

[HR26] Alexandra Henzinger and Seyoon Ragavan. Two-server private information retrieval in sublinear time and quasilinear space. In *EUROCRYPT*, 2026. 1, 2, 3, 4, 14, 15, 16, 18, 22, 36, 82

[IR90] Kenneth Ireland and Michael Rosen. *A Classical Introduction to Modern Number Theory*, volume 84 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1990. 40

[KMV00] Neal Koblitz, Alfred Menezes, and Scott Vanstone. The state of elliptic curve cryptography. *Designs, codes and cryptography*, 19(2), 2000. 17

[KO97] Eyal Kushilevitz and Rafail Ostrovsky. Replication is not needed: Single database, computationally-private information retrieval. In *FOCS*, pages 364–373. IEEE, 1997. 1

[Kob87] Neal Koblitz. Elliptic curve cryptosystems. *Mathematics of computation*, 48(177), 1987. 17

[Kra99] Christian Krattenthaler. Advanced determinant calculus. *Séminaire Lotharingien de Combinatoire*, 42:B42q, 1999. 31

[Kra01] Steven G. Krantz. *Function Theory of Several Complex Variables*. AMS Chelsea Publishing, Providence, RI, second edition, 2001. 90

[Lee26a] Keewoo Lee. Barely doubly-efficient PIR from quadratic residuosity. Cryptology ePrint Archive, Paper 2026/2034, 2026. 1, 16

[Lee26b] Keewoo Lee. Barely doubly-efficient SimplePIR. In Nadia Heninger and Mike Rosulek, editors, *Advances in Cryptology - CRYPTO 2026 - 46th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 17-20, 2026, Proceedings, Part X*, volume 16809 of *Lecture Notes in Computer Science*, pages 499–513. Springer, 2026. 1, 16

[Lip09] Helger Lipmaa. First CPIR protocol with data-dependent computation. In *ICISC*, 2009. 1

[LLF+25] Arthur Lazzaretti, Zeyu Liu, Ben Fisch, Peihan Miao, and Charalampos Papamanthou. Multi-server doubly efficient PIR in the classical model and beyond. In *TCC*. Springer, 2025. 1, 15, 16

[LMW23] Wei-Kai Lin, Ethan Mook, and Daniel Wichs. Doubly efficient private information retrieval and fully homomorphic RAM computation from ring LWE. In *STOC*, 2023. 1, 15, 16, 18, 19

[LPS88] Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988. 16

[LS77] Roger C. Lyndon and Paul E. Schupp. *Combinatorial Group Theory*, volume 89 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, Berlin, 1977. 43

[LVW17] Tianren Liu, Vinod Vaikuntanathan, and Hoeteck Wee. Conditional disclosure of secrets via non-linear reconstruction. In *Advances in Cryptology - CRYPTO 2017*, Lecture Notes in Computer Science. Springer, 2017. Full version: Cryptology ePrint Archive, Paper 2017/359. 7

[LW26] Jialin Li and Haowu Wang. On the vanishing order of Jacobi forms at infinity. *Forum Mathematicum*, 2026. `arXiv:2506.16083`; we cite the numbering of its first arXiv version. 14, 17, 38, 60, 63, 64, 74, 80

[Mar88] Grigorii A. Margulis. Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators. *Problems of Information Transmission*, 24(1):39–46, 1988. 16

[Men93] Alfred J Menezes. *Elliptic curve public key cryptosystems*, volume 234. Springer Science & Business Media, 1993. 17

[MH73] John Milnor and Dale Husemoller. *Symmetric Bilinear Forms*, volume 73 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, New York–Heidelberg, 1973. 53

[Mil85] Victor S Miller. Use of elliptic curves in cryptography. In *CRYPTO*, 1985. 17

[MR07] Daniele Micciancio and Oded Regev. Worst-case to average-case reductions based on Gaussian measures. *SIAM J. Comput.*, 37(1):267–302, 2007. 17, 91

[MS73] Colin L. Mallows and Neil J. A. Sloane. An upper bound for self-dual codes. *Information and Control*, 22(2):188–200, 1973. 17, 38

[MS77] F. Jessie MacWilliams and Neil J. A. Sloane. *The Theory of Error-Correcting Codes*, volume 16 of *North-Holland Mathematical Library*. North-Holland, Amsterdam, 1977. 29

[Mum83] David Mumford. *Tata Lectures on Theta I*, volume 28 of *Progress in Mathematics*. Birkhäuser, Boston, 1983. 14, 37, 59, 62, 69, 91

[NRS06] Gabriele Nebe, Eric M. Rains, and Neil J. A. Sloane. *Self-Dual Codes and Invariant Theory*, volume 17 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 2006. 12, 17, 38, 47, 50

[OPPW24] Hiroki Okada, Rachel Player, Simon Pohmann, and Christian Weinert. Towards practical doubly-efficient private information retrieval. In *International Conference on Financial Cryptography and Data Security*, 2024. 15, 16

[OPPW25] Hiroki Okada, Rachel Player, Simon Pohmann, and Christian Weinert. On algebraic homomorphic encryption and its applications to doubly-efficient PIR. In *EUROCRYPT*, 2025. 1, 15

[Ran77] Robert A. Rankin. *Modular Forms and Functions*. Cambridge University Press, Cambridge, 1977. 8, 13, 47, 48, 51, 64

[Reg09] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*, 56(6):34:1–34:40, 2009. 17, 91

[Rib90] Kenneth A. Ribet. On modular representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ arising from modular forms. *Inventiones Mathematicae*, 100(2):431–476, 1990. 17

[RWR15] Olav K. Richter and Martin Westerholt-Raum. Sturm bounds for Siegel modular forms. *Research in Number Theory*, 1:Art. 5, 2015. 14, 17, 38, 60, 63, 64, 74, 80

[Sch09] Nils R. Scheithauer. The Weil representation of $\mathrm{SL}_2(\mathbb{Z})$ and some applications. *International Mathematics Research Notices*, 2009(8):1488–1545, 2009. 41

[Ser73] Jean-Pierre Serre. *A Course in Arithmetic*, volume 7 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1973. 10, 11, 37, 38, 39

[Ser87] Jean-Pierre Serre. Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. *Duke Mathematical Journal*, 54(1):179–230, 1987. 17

[Shi71] Goro Shimura. *Introduction to the Arithmetic Theory of Automorphic Functions*, volume 11 of *Publications of the Mathematical Society of Japan*. Princeton University Press, Princeton, 1971. 10, 37, 38

[SS03] Elias M. Stein and Rami Shakarchi. *Complex Analysis*, volume 2 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, 2003. 89

[Str13] Fredrik Strömberg. Weil representations associated with finite quadratic modules. *Mathematische Zeitschrift*, 275(1–2):509–527, 2013. `arXiv:1108.0202`. 9, 37, 41, 61

[TW95] Richard Taylor and Andrew Wiles. Ring-theoretic properties of certain Hecke algebras. *Annals of Mathematics*, 141(3):553–572, 1995. 17

[Via17] Maryna S. Viazovska. The sphere packing problem in dimension 8. *Annals of Mathematics*, 185(3):991–1015, 2017. 17

[Wei64] André Weil. Sur certains groupes d'opérateurs unitaires. *Acta Mathematica*, 111:143–211, 1964. 9, 37, 41, 62

[Wil95] Andrew Wiles. Modular elliptic curves and Fermat's Last Theorem. *Annals of Mathematics*, 141(3):443–551, 1995. 17

[WY05] David Woodruff and Sergey Yekhanin. A geometric approach to information-theoretic private information retrieval. In *CCC*. IEEE, 2005. 1, 10, 16

[Yek08] Sergey Yekhanin. Towards 3-query locally decodable codes of subexponential length. *J. ACM*, 2008. 1, 4

A Proofs of Formal Theorem Statements

Proof of Main Theorem 5.2.

Proof. Let n be sufficiently large, and let d be the largest integer with $16^d \le n$. By Lemma 9.5 with $q = 2$ and $M = n$, there are an integer $B = \mathsf{polylog}(n)$ and an odd prime $p > n$ with $p \ge n^4$ and $2^B < p \le 2^B(1 + \frac{1}{4n})$. Let $j \in [d]$. A database of 16^j records in $\mathbb{F}_p$ splits into four databases of 2^{4j-2} records each, where $4j - 2 \equiv 2 \pmod 4$ and $4j - 2 < n < p$. Running the scheme of Theorem 6.1 on each of the four, with independent randomness and with a query for an arbitrary fixed record to each of the three that do not contain the desired record, gives a 2-server scheme over databases of 16^j records in $\mathbb{F}_p$. In it, each server stores 16^j elements of $\mathbb{F}_p$, the user uploads $4(4j - 2)$ bits, and each server reads and sends back at most $4 \cdot 2^{H_2(1/4)(4j-1)} \le 4 \cdot 16^{H_2(1/4)d} \le 4n^{H_2(1/4)}$ elements of $\mathbb{F}_p$. So Lemma 9.6, applied with $b = 16$, gives a 2-server scheme with exponent $H_2(1/4)$.

Three servers. Let n be sufficiently large, and let d be the largest integer with $3^d \le n$. By Lemma 9.5 with $q = 3$ and $M = n$, there are an integer $B = \mathsf{polylog}(n)$ and a prime $p \equiv 1 \pmod 3$ with $p > n$, $p \ge n^4$ and $2^B < p \le 2^B(1 + \frac{1}{4n})$. Then $\mathbb{F}_p$ contains a primitive cube root of unity, and its characteristic exceeds $\max\{\lfloor d/3 \rfloor, 2\}$. So Lemma 8.3, with $\ell = \lfloor \alpha^* d \rfloor$, gives a cone-delta function on $\mathbb{Z}_3^d$ over $\mathbb{F}_p$ with Fourier sparsity $k \le 3^{(\eta_{\text{ends}} + o(1))d} \le n^{\eta_{\text{ends}} + o(1)}$ (Corollary 8.4). By Lemma 4.4, this yields a $(k, 3)$-sparse quadratic-phase delta expansion over $\mathbb{F}_p$, which Theorem 7.1 (with $s = q = 3$) and Remark 7.2 turn into 3-server schemes over databases of 3^j records in $\mathbb{F}_p$, for every $j \in [d]$, in which each server reads and sends back at most k elements of $\mathbb{F}_p$. Applying Lemma 9.6 with $b = 3$ gives a 3-server scheme with exponent η_{ends}. □

Proof of Main Theorem 5.3.

Proof. Fix a generator $\mathbf{a} + \mathcal{L}$ of $\mathcal{L}^*/\mathcal{L}$. Let n be sufficiently large, let d be the largest integer with $q^d \le n$, and let $S_d \subseteq \mathbb{Z}_q^d$ be the set of Theorem 11.9. That theorem gives a cone-delta function on $\mathbb{Z}_q^d$ over $\mathbb{Q}(\zeta_q)$ whose Fourier support is contained in S_d, and $|S_d| \le q^{(\eta_{\mathcal{L}} + o(1))d} \le n^{\eta_{\mathcal{L}} + o(1)}$ by Corollary 11.10. By Lemma 4.4, this yields a quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ with support $S_d \times \mathbb{Z}_q$. The set S_d is invariant under permutations of the d coordinates, since its definition depends only on the multiset of coordinates, so by Lemma 9.1, there is also a symmetric such expansion. By Lemma 9.5 with $M = M_d$, the explicit bound of Lemma 9.3, there are an integer $B = \mathsf{polylog}(n)$ and a prime $p \equiv 1 \pmod q$ with $p > M$, $p \ge n^4$ and $2^B < p \le 2^B(1 + \frac{1}{4n})$. Then Lemma 9.3 gives a quadratic-phase delta expansion over $\mathbb{F}_p$ supported on $S_d \times \mathbb{Z}_q$, whose coefficients the user can compute in $\mathsf{polylog}(n)$ time, and which Theorem 7.1 (with $s = q$) and Remark 7.2 turn into q-server schemes over databases of q^j records in $\mathbb{F}_p$, for every $j \in [d]$, in which each server reads and sends back at most $|S_d|$ elements of $\mathbb{F}_p$. Applying Lemma 9.6 with $b = q$ gives a q-server scheme with exponent $\eta_{\mathcal{L}}$.

For primes $q \equiv 3 \pmod 4$, the rank-2 lattice of Definition 11.11 has $\eta_{\mathcal{L}} = \eta_q$, as shown in the proof of Corollary 11.13, and the bounds on η_q are Corollary 11.13. Finally, given $s \ge 3$, let q be the largest prime

with $q \equiv 3 \pmod 4$ and $q \le s$, and run the q-server scheme on q of the s servers. By the prime number theorem in arithmetic progressions [Dav80, Chapter 22, p. 133], $q \ge s/2$ for all sufficiently large s, so $\log_2 q \ge \log_2 s - 1$ and $\frac{0.467+o(1)}{\log_2 q} = \frac{0.467+o(1)}{\log_2 s}$. □

Proof of Main Theorem 5.5.

Proof. Let η be either $\eta_{q,s}(\vartheta, \nu)$ for a fixed feasible pair (ϑ, ν), or $\eta_{q,s}$. Let n be sufficiently large, and let d be the largest integer with $q^d \le n$. By Corollary 12.4, there is a (k, s)-sparse quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ on $\mathbb{Z}_q^d$ with permutation-invariant support, where $k \le q^{(\eta+o(1))d} \le n^{\eta+o(1)}$; by Lemma 9.1, there is also a symmetric such expansion, with the same support. By Lemma 9.5 with $M = M_d$, the explicit bound of Lemma 9.3, there are an integer $B = \mathsf{polylog}(n)$ and a prime $p \equiv 1 \pmod q$ with $p > M$, $p \ge n^4$ and $2^B < p \le 2^B(1 + \frac{1}{4n})$. Then Lemma 9.3 gives a (k, s)-sparse quadratic-phase delta expansion over $\mathbb{F}_p$, whose coefficients the user can compute in $\mathsf{polylog}(n)$ time, and which Theorem 7.1 and Remark 7.2 turn into s-server schemes over databases of q^j records in $\mathbb{F}_p$, for every $j \in [d]$, in which each server reads and sends back at most k elements of $\mathbb{F}_p$. Applying Lemma 9.6 with $b = q$ gives an s-server scheme with exponent η.

The bounds on $\eta_{q,s}$ are Corollary 12.4. Finally, given $s \ge 3$, let q be the largest prime with $q \le s$, and run the scheme with exponent $\eta_{q,q}$ on q of the s servers. By the prime number theorem [Dav80, Chapter 18], $q \ge s/2$ for all sufficiently large s, so $\log_2 q \ge \log_2 s - 1$ and $\frac{0.255+o(1)}{\log_2 q} = \frac{0.255+o(1)}{\log_2 s}$. □

B Background for Part II

B.1 Complex Analysis

This subsection collects the facts from complex analysis that we use, in the forms in which we use them. Throughout, "locally uniformly" means uniformly on every compact subset of the domain in question. For the one-variable facts, see, e.g., [SS03, Chapter 2].

Definition B.1 (Holomorphic functions)**.** *Let $U \subseteq \mathbb{C}$ be open. A function $f : U \to \mathbb{C}$ is* holomorphic *if it is complex differentiable at every point of U, i.e., for every $z_0 \in U$ the limit*

$$f'(z_0) := \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. A function $F : U \to \mathbb{C}^N$ is holomorphic if each of its coordinates is. Equivalently, f is holomorphic if and only if, near each point $z_0 \in U$, it is given by a convergent power series $f(z) = \sum_{n \ge 0} a_n (z - z_0)^n$, its Taylor expansion *at z_0; see, e.g., [SS03, Chapter 2].*

Fact B.2 (Identity theorem)**.** *Let $U \subseteq \mathbb{C}$ be a connected open set and let $f : U \to \mathbb{C}$ be holomorphic and not identically zero. Then the zeros of f are isolated, and the Taylor expansion of f at every point of U has a first nonzero coefficient. Consequently, if two holomorphic functions on U agree on a subset of U with an accumulation point in U, then they agree on all of U.*

Fact B.3 (Weierstrass convergence theorem)**.** *Let $U \subseteq \mathbb{C}$ be open and let $f_1, f_2, \ldots : U \to \mathbb{C}^N$ be holomorphic functions converging locally uniformly to $f : U \to \mathbb{C}^N$. Then f is holomorphic, and $f_n' \to f'$ locally uniformly. In particular, a series of holomorphic functions that converges locally uniformly (e.g., because the terms admit a convergent series of uniform bounds on each compact set) has a holomorphic sum.*

Fact B.4 (Fourier expansion of periodic holomorphic functions)**.** *Let $h > 0$ and let $G : \mathcal{H} \to \mathbb{C}^N$ be a holomorphic function satisfying $G(\tau + h) = G(\tau)$ for all $\tau \in \mathcal{H}$. Then there exist unique $\mathbf{c}_n \in \mathbb{C}^N$ (for $n \in \mathbb{Z}$) such that*

$$G(\tau) = \sum_{n=-\infty}^{\infty} \mathbf{c}_n e^{2\pi i \tau n/h},$$

with the series converging absolutely and locally uniformly on $\mathcal{H}$. (Applying the scalar case coordinatewise gives the vector-valued statement, and the scalar case is standard: G descends through the holomorphic covering map $\tau \mapsto e^{2\pi i\tau/h}$ from $\mathcal{H}$ onto the punctured unit disk, where the expansion above is the Laurent expansion. See, e.g., [DS05, Section 1.1] (stated there for $h = 1$; the general case follows by rescaling $\tau \mapsto \tau/h$).)

We will also deal with holomorphic functions of two complex variables, for which the following definition suffices for our purposes.

Definition B.5. *Let $U \subseteq \mathbb{C} \times \mathbb{C}$ be open. A function $F : U \to \mathbb{C}^N$ is* holomorphic *if it is continuous and, for each fixed value of either variable, holomorphic in the other. (By a theorem of Hartogs, the continuity requirement is automatic, but we will not need this.)*

With this definition, sums, products, and compositions with polynomials of holomorphic functions are holomorphic, and the convergence statement of Fact B.3 extends verbatim to two variables: continuity passes to locally uniform limits, and holomorphy in each variable follows from the one-variable case applied on slices. The two-variable fact we need is that Taylor expansion in one variable behaves well in the other:

Lemma B.6 (Taylor coefficients of a holomorphic family [Kra01, Theorem 1.2.2 and Corollary 1.2.4])**.** *Let $U \subseteq \mathbb{C}$ be open and let $F : U \times \mathbb{C} \to \mathbb{C}^N$ be holomorphic. Then there exist unique holomorphic functions $A_0, A_1, \ldots : U \to \mathbb{C}^N$ such that:*

$$F(\tau, z) = \sum_{j=0}^{\infty} A_j(\tau)\, z^j,$$

with the series converging absolutely and locally uniformly on $U \times \mathbb{C}$. Moreover, $A_j(\tau) = \frac{1}{j!}\frac{\partial^j}{\partial z^j} F(\tau, z)\big|_{z=0}$.

B.2 Poisson Summation Formula

We use the Poisson summation formula only through the following identity for Gaussian sums over a lattice (Fact B.8). Throughout this subsection, r is an arbitrary positive integer (we do *not* impose the evenness convention of Section 11.2; one application below has $r = 1$), $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^r$ together with its bilinear extension to $\mathbb{C}^r$, $\|\mathbf{x}\| := \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$ for $\mathbf{x} \in \mathbb{R}^r$, and $\mathcal{L} = \mathbf{B}\mathbb{Z}^r \subseteq \mathbb{R}^r$ is a full-rank lattice with dual $\mathcal{L}^*$ and determinant $\det \mathcal{L} = |\det \mathbf{B}|$ as in Section 11.2 (so $\det \mathcal{L}^* = 1/\det \mathcal{L}$).

We first record a crude bound that makes every sum in this subsection (and our theta functions) converge:

Lemma B.7 (Gaussian-dominated lattice sums)**.** *For any $\mathbf{v} \in \mathbb{R}^r$ and constants $c > 0$ and $C \geq 0$:*

$$\sum_{\mathbf{y} \in \mathbf{v} + \mathcal{L}} \exp\left(-c\|\mathbf{y}\|^2 + C\|\mathbf{y}\|\right) < \infty.$$

Proof. Distinct points of $\mathbf{v} + \mathcal{L}$ differ by a nonzero element of $\mathcal{L}$, and every nonzero $\mathbf{y} = \mathbf{B}\mathbf{x} \in \mathcal{L}$ (where $\mathbf{x} \in \mathbb{Z}^r \setminus \{\mathbf{0}\}$) satisfies $\|\mathbf{y}\| \geq \sigma\|\mathbf{x}\| \geq \sigma$, where $\sigma > 0$ is the smallest singular value of $\mathbf{B}$. So the open balls of radius $\sigma/2$ centered at the points of $\mathbf{v} + \mathcal{L}$ are pairwise disjoint. For an integer $n \geq 0$, each such ball whose

center has norm less than $n+1$ lies inside the ball of radius $n+1+\sigma/2$ around the origin, so comparing volumes, the number of $\mathbf{y} \in \mathbf{v}+\mathcal{L}$ with $n \le \|\mathbf{y}\| < n+1$ is at most $\left(\frac{2(n+1)}{\sigma}+1\right)^r$. Each such $\mathbf{y}$ contributes at most $\exp(-cn^2+C(n+1))$, so the sum is at most $\sum_{n=0}^{\infty}\left(\frac{2(n+1)}{\sigma}+1\right)^r \exp(-cn^2+C(n+1))$, which converges because the summand decays superexponentially in n. □

Fact B.8 ([Mum83, Chapter II, Equation (5.6)]). *For any $\mathbf{u} \in \mathbb{C}^r$ (the shift may be complex) and $\tau \in \mathcal{H}$, we have:*

$$\sum_{\mathbf{y}\in\mathcal{L}} \exp\left(-\pi i\langle \mathbf{y}+\mathbf{u}, \mathbf{y}+\mathbf{u}\rangle/\tau\right) = (-i\tau)^{r/2}\det(\mathcal{L}^*)\sum_{\mathbf{y}\in\mathcal{L}^*}\exp(\pi i\tau\langle \mathbf{y},\mathbf{y}\rangle - 2\pi i\langle \mathbf{y},\mathbf{u}\rangle),$$

with both sums converging absolutely. Here and throughout, $(-i\tau)^{r/2} := \exp\left(\frac{r}{2}\operatorname{Log}(-i\tau)\right)$ with Log *the principal branch of the logarithm, which is well-defined and holomorphic on $\mathcal{H}$ because $\Re(-i\tau) = \Im\tau > 0$.*

This is the functional equation $\vartheta(\Omega^{-1}z, -\Omega^{-1}) = \det(\Omega/i)^{1/2}\exp(\pi i\, z^\top\Omega^{-1}z)\,\vartheta(z,\Omega)$ of the Riemann theta function [Mum83, Chapter II, Equation (5.6) and the paragraph after it], with $\Omega = \tau(\mathbf{B}^\top\mathbf{B})^{-1}$ and $z = -\mathbf{B}^{-1}\mathbf{u}$, after cancelling the factor $\exp(\pi i\langle \mathbf{u},\mathbf{u}\rangle/\tau)$ from both sides. Mumford's branch of $\det(\Omega/i)^{1/2}$ is positive for purely imaginary Ω, so it agrees with $(-i\tau)^{r/2}\det(\mathcal{L}^*)$. Both sums converge absolutely by Lemma B.7.

Remark B.9. *For purely imaginary $\tau = is^2$, with $s > 0$, and real $\mathbf{u}$, Fact B.8 reads*

$$\sum_{\mathbf{x}\in\mathcal{L}+\mathbf{u}} e^{-\pi\|\mathbf{x}\|^2/s^2} = s^r\det(\mathcal{L}^*)\sum_{\mathbf{y}\in\mathcal{L}^*} e^{-\pi s^2\|\mathbf{y}\|^2 - 2\pi i\langle \mathbf{y},\mathbf{u}\rangle}.$$

This is the form of the Poisson summation formula behind Banaszczyk's Gaussian measures on lattices [Ban93, Lemma 1.1(i)] and the analysis of discrete Gaussians in lattice-based cryptography (the proofs of [MR07, Lemma 4.1] and [Reg09, Claim 3.8]).

B.3 Gaussian Riemann Sums

Lemma B.10 (Gaussian Riemann sums). *Let N be a positive integer, $h, \lambda > 0$ real numbers, and $\alpha \in [0,1]$ some fixed constant offset. Then:*

$$\left| h\sum_{j=1}^{N} e^{-\lambda(j-\alpha)^2h^2} - \frac{1}{2}\sqrt{\frac{\pi}{\lambda}}\right| = \left| h\sum_{j=1}^{N} e^{-\lambda(j-\alpha)^2h^2} - \int_0^\infty e^{-\lambda x^2}\,dx\right| \le h + \frac{e^{-\lambda h^2N^2}}{2\lambda hN}.$$

Proof. The equality is the Gaussian integral $\int_0^\infty e^{-\lambda x^2}\,dx = \frac{1}{2}\sqrt{\pi/\lambda}$. For the inequality, split the integral as $\int_0^{Nh} + \int_{Nh}^\infty$. The function $g(x) := e^{-\lambda x^2}$ is nonincreasing on $[0,\infty)$, and $h\sum_{j=1}^N g((j-\alpha)h)$ is a Riemann sum for $\int_0^{Nh} g(x)\,dx$ with one sample point $(j-\alpha)h \in [(j-1)h,\ jh]$ in each cell; for a nonincreasing function, the error in each cell is at most $h\,(g((j-1)h) - g(jh))$, and these telescope:

$$\left| h\sum_{j=1}^{N} g((j-\alpha)h) - \int_0^{Nh} g(x)\,dx\right| \le h\,(g(0) - g(Nh)) \le h.$$

For the tail, $\int_{Nh}^\infty e^{-\lambda x^2}\,dx \le \int_{Nh}^\infty \frac{x}{Nh}\,e^{-\lambda x^2}\,dx = \frac{e^{-\lambda h^2N^2}}{2\lambda hN}$. □

C Deferred Proofs from Section 10

Proof of Lemma 10.28

First consider two choices $g_{\mathfrak{a}}, g'_{\mathfrak{a}} \in \mathrm{SL}_2(\mathbb{Z})$ with $g_{\mathfrak{a}} * \infty = g'_{\mathfrak{a}} * \infty = \mathfrak{a}$. Then $g_{\mathfrak{a}}^{-1} g'_{\mathfrak{a}}$ fixes ∞, and the stabilizer of ∞ in $\mathrm{SL}_2(\mathbb{Z})$ is $\{\pm T^n : n \in \mathbb{Z}\}$: a matrix fixes ∞ if and only if its lower-left entry is 0, and then its diagonal entries are integers with product 1, hence both 1 or both -1. So $g'_{\mathfrak{a}} = \pm g_{\mathfrak{a}} T^n$ for some choice of sign and some $n \in \mathbb{Z}$. The width is unaffected: for every t we have $g'_{\mathfrak{a}} T^t (g'_{\mathfrak{a}})^{-1} = (\pm g_{\mathfrak{a}} T^n) T^t (\pm g_{\mathfrak{a}} T^n)^{-1} = g_{\mathfrak{a}} T^t g_{\mathfrak{a}}^{-1}$ (the signs cancel and powers of T commute), so both choices yield the same set $\{t : g_{\mathfrak{a}} T^t g_{\mathfrak{a}}^{-1} \in \Gamma\}$ and hence the same width h, as well as the same matrix $A := \rho(g_{\mathfrak{a}} T^h g_{\mathfrak{a}}^{-1})$ and thus the same matrix M from Equation (23). Next, writing $F'_{\mathfrak{a}}(\tau) := G(g'_{\mathfrak{a}}, \tau)^{-k} F(g'_{\mathfrak{a}} * \tau)$, we claim that $F'_{\mathfrak{a}}(\tau) = (\pm 1)^k F_{\mathfrak{a}}(\tau + n)$: indeed, $(\pm g_{\mathfrak{a}} T^n) * \tau = g_{\mathfrak{a}} * (\tau + n)$, while $G(\pm g_{\mathfrak{a}} T^n, \tau) = \pm G(g_{\mathfrak{a}} T^n, \tau) = \pm G(g_{\mathfrak{a}}, \tau + n)$ by Proposition 10.17 and $G(T^n, \tau) = 1$ (and $(\pm 1)^{-k} = (\pm 1)^k$). Now let $\mathbf{c}_m$ and $\mathbf{c}'_m$ be the expansion coefficients (Equation (24)) of $\exp(-2\pi i M\tau/h) F_{\mathfrak{a}}(\tau)$ and $\exp(-2\pi i M\tau/h) F'_{\mathfrak{a}}(\tau)$ respectively. We compute:

$$\exp(-2\pi i M\tau/h) F'_{\mathfrak{a}}(\tau) = (\pm 1)^k \exp(2\pi i M n/h) \cdot \exp(-2\pi i M(\tau + n)/h) F_{\mathfrak{a}}(\tau + n),$$

so by the uniqueness in Fact B.4, $\mathbf{c}'_m = (\pm 1)^k e^{2\pi i m n/h} \exp(2\pi i M n/h)\, \mathbf{c}_m$. The matrix $\exp(2\pi i M n/h)$ acts on the range of each spectral projector P_α of A as the nonzero scalar $e^{2\pi i \alpha n/h}$, so $P_\alpha \mathbf{c}'_m \neq 0$ if and only if $P_\alpha \mathbf{c}_m \neq 0$. By Remark 10.26, holomorphy at $\mathfrak{a}$ and $\mathrm{ord}_{\mathfrak{a}}(F)$ are determined by exactly these conditions, so both choices give the same answers.

Now consider $\mathfrak{b} = \gamma * \mathfrak{a}$ with $\gamma \in \Gamma$. By the first part we are free to choose the lifts, so take $g_{\mathfrak{b}} := \gamma g_{\mathfrak{a}}$ (indeed $g_{\mathfrak{b}} * \infty = \mathfrak{b}$). The widths agree because $g_{\mathfrak{b}} T^t g_{\mathfrak{b}}^{-1} = \gamma (g_{\mathfrak{a}} T^t g_{\mathfrak{a}}^{-1}) \gamma^{-1}$, and conjugation by $\gamma \in \Gamma$ preserves membership in Γ. Next, by Equation (26), Definition 10.19, and Proposition 10.17:

$$\begin{aligned} F_{\mathfrak{b}}(\tau) &= G(\gamma g_{\mathfrak{a}}, \tau)^{-k} F(\gamma g_{\mathfrak{a}} * \tau) \\ &= G(\gamma g_{\mathfrak{a}}, \tau)^{-k} G(\gamma, g_{\mathfrak{a}} * \tau)^k \rho(\gamma) F(g_{\mathfrak{a}} * \tau) \\ &= G(\gamma g_{\mathfrak{a}}, \tau)^{-k} G(\gamma, g_{\mathfrak{a}} * \tau)^k G(g_{\mathfrak{a}}, \tau)^k \rho(\gamma) F_{\mathfrak{a}}(\tau) \\ &= \rho(\gamma) F_{\mathfrak{a}}(\tau). \end{aligned}$$

The matrices A at the two cusps are conjugate: $A_{\mathfrak{b}} := \rho(g_{\mathfrak{b}} T^h g_{\mathfrak{b}}^{-1}) = \rho(\gamma) A_{\mathfrak{a}} \rho(\gamma)^{-1}$. Hence $A_{\mathfrak{b}}$ has the same eigenvalues as $A_{\mathfrak{a}}$, with spectral projectors $P^{\mathfrak{b}}_\alpha = \rho(\gamma) P^{\mathfrak{a}}_\alpha \rho(\gamma)^{-1}$, and $M_{\mathfrak{b}} = \rho(\gamma) M_{\mathfrak{a}} \rho(\gamma)^{-1}$. Therefore:

$$\exp(-2\pi i M_{\mathfrak{b}} \tau/h)\, F_{\mathfrak{b}}(\tau) = \rho(\gamma) \exp(-2\pi i M_{\mathfrak{a}} \tau/h)\, F_{\mathfrak{a}}(\tau),$$

so the expansion coefficients satisfy $\mathbf{c}^{\mathfrak{b}}_m = \rho(\gamma)\, \mathbf{c}^{\mathfrak{a}}_m$ by uniqueness, and $P^{\mathfrak{b}}_\alpha \mathbf{c}^{\mathfrak{b}}_m = \rho(\gamma)\, P^{\mathfrak{a}}_\alpha \mathbf{c}^{\mathfrak{a}}_m$, which is nonzero if and only if $P^{\mathfrak{a}}_\alpha \mathbf{c}^{\mathfrak{a}}_m$ is. As before, by Remark 10.26, holomorphy and order at the two cusps coincide. □

Proof of Lemma 10.30: Holomorphy at the Cusps

Let F, P, χ and $f = P \circ F$ be as in Lemma 10.30. Recall from the main-text part of the proof that χ has finite order and that f is weakly modular of weight kD for Γ with representation χ. It remains to show that f is holomorphic at every cusp $\mathfrak{a} \in \mathbb{Q} \cup \{\infty\}$. Fix a lift $g_{\mathfrak{a}}$ (taking $g_\infty := I$ for the cusp at infinity), and let h be the width of the cusp — the width is a property of Γ and $\mathfrak{a}$ alone, so it is the same for f and F. Homogeneity gives:

$$f_{\mathfrak{a}}(\tau) = G(g_{\mathfrak{a}}, \tau)^{-kD} P(F(g_{\mathfrak{a}} * \tau)) = P\left(G(g_{\mathfrak{a}}, \tau)^{-k} F(g_{\mathfrak{a}} * \tau)\right) = P(F_{\mathfrak{a}}(\tau)).$$

Let $A := \rho(g_{\mathfrak{a}} T^h g_{\mathfrak{a}}^{-1})$ be the matrix describing how $F_{\mathfrak{a}}$ transforms over one period, with eigenbasis $\mathbf{v}_j$, dual basis $\mathbf{u}_j$, and eigenvalues $e^{2\pi i\alpha_j}$ ($\alpha_j \in [0,1)$) as in Section 10.4.2; and let $e^{2\pi i\beta}$ with $\beta \in [0,1)$ be the scalar $\chi(g_{\mathfrak{a}} T^h g_{\mathfrak{a}}^{-1})$ (well-defined this way since χ has finite order), which is the corresponding phase picked up by $f_{\mathfrak{a}}$ over one period.

Write the coordinates of $F_{\mathfrak{a}}$ in the eigenbasis as $y_j(\tau) := \mathbf{u}_j^* F_{\mathfrak{a}}(\tau)$. Since F is holomorphic at $\mathfrak{a}$, the expansion from Section 10.4.2 gives:

$$y_j(\tau) = \sum_{n \geq 0} (\mathbf{u}_j^* \mathbf{c}_n)\, e^{2\pi i(n+\alpha_j)\tau/h},$$

converging absolutely and locally uniformly. Also expand P in these coordinates: there are coefficients $\tilde{p}_\kappa$, indexed by multi-indices $\kappa = (\kappa_1, \ldots, \kappa_N)$ with $\kappa_1 + \ldots + \kappa_N = D$, such that $P(\mathbf{w}) = \sum_\kappa \tilde{p}_\kappa \prod_j (\mathbf{u}_j^* \mathbf{w})^{\kappa_j}$ for all $\mathbf{w} \in \mathbb{C}^N$. The covariance hypothesis (27), applied to the element $g_{\mathfrak{a}} T^h g_{\mathfrak{a}}^{-1} \in \Gamma$, pins down which κ can occur: since A scales the coordinate $\mathbf{u}_j^* \mathbf{w}$ by $e^{2\pi i\alpha_j}$, the identity $P(A\mathbf{w}) = e^{2\pi i\beta} P(\mathbf{w})$ reads

$$\sum_\kappa \tilde{p}_\kappa\, e^{2\pi i\langle \kappa, \alpha\rangle} \prod_j (\mathbf{u}_j^* \mathbf{w})^{\kappa_j} = \sum_\kappa \tilde{p}_\kappa\, e^{2\pi i\beta} \prod_j (\mathbf{u}_j^* \mathbf{w})^{\kappa_j} \qquad \text{for all } \mathbf{w} \in \mathbb{C}^N,$$

where $\langle \kappa, \alpha\rangle := \sum_j \kappa_j \alpha_j$. Distinct monomials are linearly independent polynomials of $\mathbf{w}$, so every κ with $\tilde{p}_\kappa \neq 0$ satisfies $\langle \kappa, \alpha\rangle \equiv \beta \pmod 1$.

Now substitute the series for the y_j into $f_{\mathfrak{a}}(\tau) = \sum_\kappa \tilde{p}_\kappa \prod_j y_j(\tau)^{\kappa_j}$, and expand each product of (finitely many) absolutely convergent series into an absolutely convergent sum of products. Every resulting term is a constant times $e^{2\pi i t\tau/h}$, where t is a sum of D exponents $n_i + \alpha_{j_i}$ with each $n_i \geq 0$ and with the multiset $\{j_i\}$ prescribed by some κ with $\tilde{p}_\kappa \neq 0$. Hence every occurring t satisfies:

$$t = \sum_i n_i + \langle \kappa, \alpha\rangle \;\in\; \beta + \mathbb{Z} \qquad \text{and} \qquad t \geq 0,$$

so $t = m + \beta$ for an integer m with $m \geq -\beta > -1$, i.e., $m \geq 0$. Collecting terms (justified by absolute convergence), we conclude that $e^{-2\pi i\beta\tau/h} f_{\mathfrak{a}}(\tau) = \sum_{m \geq 0} b_m\, e^{2\pi i m\tau/h}$ for some $b_m \in \mathbb{C}$; the collected series converges absolutely at every point of $\mathcal{H}$, so $\sum_m |b_m| r^m < \infty$ for every $r < 1$, and the convergence is therefore also locally uniform. By the uniqueness in Fact B.4, these are precisely the cusp-expansion coefficients of f at $\mathfrak{a}$, and they vanish for all $m < 0$; so f is holomorphic at $\mathfrak{a}$. □

Proof of Lemma 10.32

Weak modularity of F for Γ' with $\rho|_{\Gamma'}$ is immediate, since (21) holds for all $g \in \Gamma \supseteq \Gamma'$. Fix a cusp $\mathfrak{a}$ with lift $g_{\mathfrak{a}}$ (taking $g_\infty := I$ as before), and let h, h' be its widths with respect to Γ, Γ' respectively. The sets $\{t \in \mathbb{Z} : g_{\mathfrak{a}} T^t g_{\mathfrak{a}}^{-1} \in \Gamma'\} \subseteq \{t \in \mathbb{Z} : g_{\mathfrak{a}} T^t g_{\mathfrak{a}}^{-1} \in \Gamma\}$ are subgroups of $\mathbb{Z}$ (closure follows from $g_{\mathfrak{a}} T^s g_{\mathfrak{a}}^{-1} \cdot g_{\mathfrak{a}} T^t g_{\mathfrak{a}}^{-1} = g_{\mathfrak{a}} T^{s+t} g_{\mathfrak{a}}^{-1}$), and they are nonzero because $g_{\mathfrak{a}} T^q g_{\mathfrak{a}}^{-1} \in \Gamma(q) \subseteq \Gamma'$ (by normality of $\Gamma(q)$ and because Γ' has level q); so they equal $h'\mathbb{Z} \subseteq h\mathbb{Z}$, and $h \mid h'$. Write $e := h'/h \in \mathbb{Z}_{\geq 1}$.

The matrix A of F at $\mathfrak{a}$ with respect to Γ' is the eth power of the one with respect to Γ: $A' := \rho(g_{\mathfrak{a}} T^{h'} g_{\mathfrak{a}}^{-1}) = \rho\left((g_{\mathfrak{a}} T^h g_{\mathfrak{a}}^{-1})^e\right) = A^e$. In particular, the eigenbasis $\mathbf{v}_j$ of A (with dual basis $\mathbf{u}_j$) also diagonalizes A', with eigenvalues $e^{2\pi i e\alpha_j}$; the representatives in $[0,1)$ are $\alpha'_j := e\alpha_j - \lfloor e\alpha_j \rfloor$, and $M' := \sum_j \alpha'_j \mathbf{v}_j \mathbf{u}_j^*$.

Let $\sum_n \mathbf{c}_n e^{2\pi i n\tau/h}$ be the expansion of $\exp(-2\pi i M\tau/h) F_{\mathfrak{a}}(\tau)$ from Equation (24). We compute the corresponding h'-periodic function directly:

$$\exp(-2\pi i M'\tau/h')\, F_{\mathfrak{a}}(\tau) = \sum_j e^{-2\pi i\alpha'_j\tau/h'}\, \mathbf{v}_j \mathbf{u}_j^* F_{\mathfrak{a}}(\tau) = \sum_{n,j} (\mathbf{u}_j^* \mathbf{c}_n)\, e^{2\pi i(e(n+\alpha_j)-\alpha'_j)\tau/h'}\, \mathbf{v}_j$$

$$= \sum_{n,j} (\mathbf{u}_j^* \mathbf{c}_n)\, e^{2\pi i m_{n,j}\tau/h'}\, \mathbf{v}_j, \qquad \text{where } m_{n,j} := en + \lfloor e\alpha_j \rfloor \in \mathbb{Z}.$$

By the uniqueness in Fact B.4, the Γ'-expansion coefficients of $F_{\mathfrak{a}}$ are obtained by grouping this (absolutely convergent) sum by the value of $m_{n,j}$: $\mathbf{c}'_m = \sum_{(n,j):\ m_{n,j}=m} (\mathbf{u}_j^* \mathbf{c}_n)\, \mathbf{v}_j$. Now suppose F is holomorphic at $\mathfrak{a}$ with respect to Γ, i.e., $\mathbf{c}_n = 0$ for all $n < 0$. Then every nonzero term has $n \geq 0$, hence $m_{n,j} = en + \lfloor e\alpha_j \rfloor \geq 0$ (as $\lfloor e\alpha_j \rfloor \geq 0$), so $\mathbf{c}'_m = 0$ for all $m < 0$: F is holomorphic at $\mathfrak{a}$ with respect to Γ'. Since $\mathfrak{a}$ was arbitrary, F is a modular form for Γ'. □

D Deferred Proofs from Section 11

Bounding the Support Size

In this subsection, we derive the two corollaries of Theorem 11.9 stated in Section 11: first Corollary 11.10, for an arbitrary lattice, and then Corollary 11.13, by instantiating the lattice of Definition 11.11. The restriction to primes $q \equiv 3 \pmod 4$ in Corollary 11.13 enters only in the second step: Theorem 11.9 and Corollary 11.10 hold for every odd prime q and every even lattice of determinant $\sqrt{q}$, but the rank-2 instantiation below exists only in this residue class (Remark 11.12).

The first step is the Chernoff bound, in the form of a bound on the number of vectors $\mathbf{z} \in \mathbb{Z}_q^d$ whose total weight $\sum_i w(z_i)$ is at most a given budget, which we will use again in Appendix E.

Lemma D.1 (Chernoff bound [BLM13, Section 2.2])**.** *Let $w : \mathbb{Z}_q \to \mathbb{R}$ be a function, let $B \in \mathbb{R}$, and let d be a positive integer. Then for every $0 < t \leq 1$, we have:*

$$\left| \left\{ \mathbf{z} \in \mathbb{Z}_q^d \,:\, \sum_{i=1}^{d} w(z_i) \leq Bd \right\} \right| \leq \left(t^{-B} \sum_{j \in \mathbb{Z}_q} t^{w(j)} \right)^d .$$

(This is the bound $\Pr\{Z \geq a\} \leq e^{-\lambda a}\, \mathbb{E}\, e^{\lambda Z}$ for $\lambda \geq 0$ of [BLM13, Section 2.2], applied to $Z = -\sum_i w(Z_i)$ for $(Z_1, \ldots, Z_d)$ uniform on $\mathbb{Z}_q^d$, with $a = -Bd$ and $t = e^{-\lambda}$, and multiplied by q^d.)

Proof of Corollary 11.10. Fix a generator $\mathbf{a} + \mathcal{L}$ of $\mathcal{L}^*/\mathcal{L}$. Theorem 11.9 provides a cone-delta function $f : \mathbb{Z}_q^d \to \mathbb{Q}(\zeta_q)$ with $\hat{f}$ supported on S_d, so its Fourier sparsity is at most $|S_d|$, and it suffices to show that $|S_d| \leq q^{(\eta_{\mathcal{L}} + o(1))d}$. By definition, S_d is the union, over $u \in \mathbb{Z}_q \backslash \{0\}$, of the sets $\left\{ \mathbf{z} \in \mathbb{Z}_q^d \,:\, \sum_{i=1}^d w_u(z_i) \leq \frac{r(q+1)}{24} d \right\}$, where $w_u(x) := w_{\mathcal{L}}(ux\mathbf{a})$. For each such u, the map $x \mapsto ux\mathbf{a} + \mathcal{L}$ is a bijection from $\mathbb{Z}_q$ onto $\mathcal{L}^*/\mathcal{L}$ (as u is invertible modulo q, and $\mathbf{a} + \mathcal{L}$ generates this group of order q), so $\sum_{x \in \mathbb{Z}_q} t^{w_u(x)} = M_{\mathcal{L}}(t)$ for every $0 < t \leq 1$. Applying Lemma D.1 to each of these $q - 1$ sets, with $B = \frac{r(q+1)}{24}$, thus gives:

$$|S_d| \leq (q-1) \left(t^{-r(q+1)/24} M_{\mathcal{L}}(t) \right)^d \qquad \text{for every } 0 < t \leq 1.$$

Taking the infimum over t gives $|S_d| \leq (q-1)\, q^{\eta_{\mathcal{L}} d}$, by the definition (34) of $\eta_{\mathcal{L}}$; and $q - 1 = q^{o(d)}$, as q is fixed while d grows. □

The second step uses the rank-2 lattice of Definition 11.11.

Lemma D.2. *Fix a prime $q \equiv 3 \pmod 4$, and define the lattice $\mathcal{L}$ and generator $\mathbf{a}$ as in Definition 11.11. Then $w_{\mathcal{L}}(u\mathbf{a}) = w_{\mathcal{L}}(-u\mathbf{a})$ for every $u \in \mathbb{Z}_q$, and for $u \in [0,(q-1)/2]$, we have:*

$$w_{\mathcal{L}}(u\mathbf{a}) = \begin{cases} 0, \ \text{if } u = 0; \\ u^2, \ \text{if } 1 \le u \le K; \\ u^2 - qu + qK, \ \text{if } K+1 \le u \le (q-1)/2. \end{cases}$$

Consequently, for $0 < t \le 1$, we have $M_{\mathcal{L}}(t) = 1 + 2\sum_{u=1}^{K} t^{u^2} + 2\sum_{u=K+1}^{(q-1)/2} t^{u^2-qu+qK}$, and the exponent $\eta_{\mathcal{L}}$ of (34) equals the exponent η_q of (35).

Proof. We first compute the prices; throughout, $K = (q+1)/4$, as in Corollary 11.13. If $u = 0$, then $u\mathbf{a} + \mathcal{L}$ contains 0 so the assertion is immediate. We also have $u\mathbf{a} + \mathcal{L} = -(-u\mathbf{a} + \mathcal{L})$ so the symmetry claim is immediate. It remains to handle $u \in [1,(q-1)/2]$. For such u, recall from Definition 11.5 that:

$$\begin{aligned} w_{\mathcal{L}}(u\mathbf{a}) &= \frac{q}{2} \min_{\mathbf{x} \in u\mathbf{a}+\mathcal{L}} \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \frac{q}{2} \min_{a,b \in \mathbb{Z}} \| u(0, \sqrt{2/q}) + a(\sqrt{2}, 0) + b(1/\sqrt{2}, \sqrt{q/2}) \|^2 \\ &= \frac{q}{2} \min_{a,b \in \mathbb{Z}} \left((a\sqrt{2} + b/\sqrt{2})^2 + (u\sqrt{2/q} + b\sqrt{q/2})^2 \right) \\ &= \min_{a,b \in \mathbb{Z}} \left(q(a + b/2)^2 + (u + bq/2)^2 \right). \end{aligned}$$

Now we have two cases:

- If b is even, then minimizing over a will set $a + b/2 = 0$, so the value is $(u + bq/2)^2$. For $u \in [1,(q-1)/2]$, the choice of even b that minimizes this is $b = 0$. Thus the minimum when we restrict to b being even is u^2.
- If b is odd, then minimizing over a will set $a + b/2 = \pm 1/2$, so the value is $q/4 + (u + bq/2)^2$. For $u \in [1,(q-1)/2]$, the choice of odd b that minimizes this is $b = -1$. Thus the minimum when we restrict to b being odd is $q/4 + (u - q/2)^2 = u^2 - uq + (q + q^2)/4 = u^2 - uq + qK$.

To minimize between these two branches, we simply note that $u^2 \le u^2 - qu + qK$ if and only if $u \le K$.

It remains to deduce the consequences. Since $\mathbf{a} + \mathcal{L}$ generates $\mathcal{L}^*/\mathcal{L}$, the map $x \mapsto x\mathbf{a} + \mathcal{L}$ is a bijection from $\mathbb{Z}_q$ onto $\mathcal{L}^*/\mathcal{L}$, so $M_{\mathcal{L}}(t) = \sum_{x \in \mathbb{Z}_q} t^{w_{\mathcal{L}}(x\mathbf{a})}$. In this sum, the term $x = 0$ contributes $t^0 = 1$, and the symmetry $w_{\mathcal{L}}(x\mathbf{a}) = w_{\mathcal{L}}(-x\mathbf{a})$ pairs each $x \in [1,(q-1)/2]$ with $-x$; so the formula above gives the stated expression for $M_{\mathcal{L}}(t)$. Finally, here $r = 2$, so the coefficient $\frac{r(q+1)}{24}$ in (34) equals $\frac{q+1}{12}$, and $\ln M_{\mathcal{L}}(t) - \frac{q+1}{12} \ln t = \ln \left(t^{-(q+1)/12} M_{\mathcal{L}}(t) \right)$ is the logarithm of the quantity whose infimum over $0 < t \le 1$ defines η_q in (35). Hence $\eta_{\mathcal{L}} = \eta_q$. □

Lemma D.3. *As $q \to \infty$ along primes $q \equiv 3 \pmod 4$ (so that η_q is defined), we have*

$$\eta_q \le \frac{1}{2} + \frac{C_{\text{warmup}}}{\ln q} + o\left(\frac{1}{\ln q} \right),$$

where we define:

$$C_{\text{warmup}} := \inf_{\lambda > 0} \left(\lambda/12 + \ln(1 + e^{-\lambda/4}) + \ln(\sqrt{\pi/\lambda}) \right) \le 0.3234.$$

Proof. In (35), we will set $t = e^{-\lambda/q}$ for a constant $\lambda > 0$. The rough idea is to show using a Riemann sum argument that for fixed λ, the expression in (35) scales like $1/2 + C_\lambda/\ln q + o(1/\ln q)$ as $q \to \infty$. Then we will minimize over λ at the end.

Plugging in this setting of t and replacing u with $(q-1)/2 - u$ in the second sum, we have:

$$\begin{aligned}
&t^{-(q+1)/12}\left(1 + 2\sum_{u=1}^{K} t^{u^2} + 2\sum_{u=K+1}^{(q-1)/2} t^{u^2-qu+qK}\right)\\
=&e^{\lambda(q+1)/(12q)}\left(1 + 2\sum_{u=1}^{K} e^{-\lambda u^2/q} + 2\sum_{u=K+1}^{(q-1)/2} e^{-\lambda(u^2-qu+qK)/q}\right)\\
=&e^{\lambda(q+1)/(12q)}\left(1 + 2\sum_{u=1}^{(q+1)/4} e^{-\lambda u^2/q} + 2\sum_{u=0}^{(q-7)/4} e^{-\lambda((2u+1)^2+q)/(4q)}\right)\\
=&e^{\lambda(q+1)/(12q)}\left(1 + 2\sum_{u=1}^{(q+1)/4} e^{-\lambda u^2/q} + 2e^{-\lambda/4}\sum_{u=0}^{(q-7)/4} e^{-\lambda(u+1/2)^2/q}\right).
\end{aligned}$$

Now by Lemma B.10 — applied with $h = 1/\sqrt{q}$, the same λ, and $(N,\alpha) = \left(\frac{q+1}{4}, 0\right)$ for the first sum, respectively $(N,\alpha) = \left(\frac{q-3}{4}, \frac{1}{2}\right)$ for the second (reindexing $j = u+1$; here $q \geq 7$, as we may since $q \to \infty$) — we have:

$$\begin{aligned}
\sum_{u=1}^{(q+1)/4} e^{-\lambda u^2/q} &= \frac{1}{2}\sqrt{\frac{q\pi}{\lambda}} + O(1)\\
\sum_{u=0}^{(q-7)/4} e^{-\lambda(u+1/2)^2/q} &= \frac{1}{2}\sqrt{\frac{q\pi}{\lambda}} + O(1).
\end{aligned}$$

(Both $O(1)$'s hide constants depending only on λ: multiplying the lemma's error bound by $\sqrt{q}$ gives at most $1 + \frac{q\, e^{-\lambda N^2/q}}{2\lambda N}$, which is bounded as $q \to \infty$ with λ fixed, since $N \geq (q-3)/4$.) Plugging these in, we may continue as follows:

$$\begin{aligned}
&e^{\lambda(q+1)/(12q)}\left(1 + 2\sum_{u=1}^{(q+1)/4} e^{-\lambda u^2/q} + 2e^{-\lambda/4}\sum_{u=0}^{(q-7)/4} e^{-\lambda(u+1/2)^2/q}\right)\\
=&e^{\lambda(q+1)/(12q)}\left(\sqrt{\frac{q\pi}{\lambda}} + e^{-\lambda/4}\sqrt{\frac{q\pi}{\lambda}} + O(1)\right)\\
=&\exp\left(\frac{\lambda}{12} + \ln(1+e^{-\lambda/4}) + \ln(\sqrt{\pi/\lambda}) + \frac{1}{2}\ln q + o(1)\right),
\end{aligned}$$

so the expression in (35) for this particular setting of t is at most:

$$\frac{1}{2} + \frac{1}{\ln q}\left(\lambda/12 + \ln(1+e^{-\lambda/4}) + \ln(\sqrt{\pi/\lambda})\right) + o\left(\frac{1}{\ln q}\right).$$

Since this holds for every fixed $\lambda > 0$, and the infimum defining C_{warmup} is attained (the expression is continuous in λ and tends to $+\infty$ as $\lambda \to 0^+$ and as $\lambda \to \infty$), we may run the argument at a minimizing λ, which gives the statement. Finally, the numerical bound $C_{\text{warmup}} \leq 0.3234$ is witnessed by $\lambda = 8.67$: numerically, $\frac{8.67}{12} + \ln\left(1 + e^{-8.67/4}\right) + \ln\sqrt{\pi/8.67} = 0.32330\ldots \leq 0.3234$. □

Lemma D.4. *We have $\eta_3 \leq 0.789691$, $\eta_7 \leq 0.709410$, and $\eta_{11} \leq 0.660807$.*

Proof. Each bound follows by evaluating the expression inside the infimum in (35) at one point t: at $t = 1/4$ for $q = 3$ (giving 0.7896901), at $t = 0.351795$ for $q = 7$ (giving 0.7094098), and at $t = 0.481707$ for $q = 11$ (giving 0.6608066); rounding up in the sixth decimal gives the stated values. □

With these two lemmas, Corollary 11.13 reduces to Corollary 11.10:

Proof of Corollary 11.13. Instantiate Corollary 11.10 with the rank-2 lattice $\mathcal{L}$ of Definition 11.11 (available exactly for the primes $q \equiv 3 \pmod 4$ under consideration, Remark 11.12). This produces a cone-delta function $f : \mathbb{Z}_q^d \to \mathbb{Q}(\zeta_q)$ with Fourier sparsity at most $q^{(\eta_{\mathcal{L}}+o(1))d}$, and $\eta_{\mathcal{L}} = \eta_q$ by Lemma D.2. Note the order of the limits: q is held fixed while $d \to \infty$ (this is the $o(1)$ in the exponent); the final claim of the corollary concerns the subsequent limit $q \to \infty$ along primes $q \equiv 3 \pmod 4$, and is Lemma D.3, together with $\frac{C_{\text{warmup}}}{\ln q} = \frac{C_{\text{warmup}}/\ln 2}{\log_2 q}$ and $0.3234/\ln 2 < 0.467$. The numerical bounds on η_3, η_7 and η_{11} are Lemma D.4. □

Proof of Proposition 11.4

The first bullet is immediate by direct expansion: the difference between the two sides is $\langle \mathbf{x}, \mathbf{y}\rangle + \frac{1}{2}\langle \mathbf{y}, \mathbf{y}\rangle$. The first term is an integer by definition of $\mathcal{L}^*$ and the second term is an integer since $\mathbf{y} \in \mathcal{L}$ and $\mathcal{L}$ is even.

Now for the second bullet, start by noting that we have $|\mathcal{L}^*/\mathcal{L}| = q$, so for any $\mathbf{x} \in \mathcal{L}^*$ we have $q\mathbf{x} \in \mathcal{L}$. Since $q\mathbf{x} \in \mathcal{L}$ and $\mathbf{x} \in \mathcal{L}^*$, their inner product must be an integer, hence $q\langle \mathbf{x}, \mathbf{x}\rangle \in \mathbb{Z} \Rightarrow \langle \mathbf{x}, \mathbf{x}\rangle \in \frac{1}{q}\mathbb{Z}$. Secondly, by the evenness of $\mathcal{L}$ we have $\langle q\mathbf{x}, q\mathbf{x}\rangle \in 2\mathbb{Z} \Rightarrow \langle \mathbf{x}, \mathbf{x}\rangle \in \frac{2}{q^2}\mathbb{Z}$. Therefore we have $\langle \mathbf{x}, \mathbf{x}\rangle \in \frac{1}{q}\mathbb{Z} \cap \frac{2}{q^2}\mathbb{Z} = \frac{2}{q}\mathbb{Z}$, as claimed. □

Proof of Proposition 11.7

Suppose instead that $q \mid \mu$, so that $\langle \mathbf{a}, \mathbf{a}\rangle = 2\mu/q$ is an integer. Every element of $\mathcal{L}^*$ has the form $x\mathbf{a} + \mathbf{v}$ with $x \in \mathbb{Z}$ and $\mathbf{v} \in \mathcal{L}$ (as $\mathbf{a} + \mathcal{L}$ generates $\mathcal{L}^*/\mathcal{L}$), and expanding

$$\langle x\mathbf{a} + \mathbf{v},\ y\mathbf{a} + \mathbf{w}\rangle = xy\langle \mathbf{a}, \mathbf{a}\rangle + x\langle \mathbf{a}, \mathbf{w}\rangle + y\langle \mathbf{a}, \mathbf{v}\rangle + \langle \mathbf{v}, \mathbf{w}\rangle$$

shows that any two elements of $\mathcal{L}^*$ would then have integer inner product: the first term is an integer by assumption, and the remaining three because $\mathbf{a} \in \mathcal{L}^*$ and $\mathbf{v} \in \mathcal{L} \subseteq \mathcal{L}^*$ (Proposition 11.3) pair integrally with the lattice vectors $\mathbf{w}, \mathbf{v} \in \mathcal{L}$. In other words, we would have $\mathcal{L}^* \subseteq (\mathcal{L}^*)^*$. But writing $\mathcal{L} = \mathbf{B}\mathbb{Z}^r$, we have $(\mathcal{L}^*)^* = (\mathbf{B}^{-\top})^{-\top}\mathbb{Z}^r = \mathbf{B}\mathbb{Z}^r = \mathcal{L}$, so this would give $\mathcal{L}^* \subseteq \mathcal{L}$ — contradicting $|\mathcal{L}^*/\mathcal{L}| = q > 1$. □

Proof of Lemma 11.15

Fix a coordinate $x \in \mathbb{Z}_q$. Then it suffices to check that the sum $\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(\pi i\langle \mathbf{y}, \mathbf{y}\rangle\tau)$ is locally uniformly absolutely convergent: each term therein is holomorphic, so the sum is then holomorphic by Fact B.3. To do this, note that:

$$|\exp(\pi i\langle \mathbf{y}, \mathbf{y}\rangle\tau)| = \exp(-\pi\langle \mathbf{y}, \mathbf{y}\rangle\Im\tau).$$

Now consider any compact subset Ω of $\mathcal{H}$; we want to show that $\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(-\pi\langle \mathbf{y}, \mathbf{y}\rangle\Im\tau)$ converges uniformly on Ω. On Ω, any τ has imaginary part at least some $h_0 > 0$. Consequently, we just need to check that

$$\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(-\pi h_0\langle \mathbf{y}, \mathbf{y}\rangle)$$

converges. This is exactly Lemma B.7, applied with $\mathbf{v}$ any representative of the coset $x\mathbf{a} + \mathcal{L}$, $c = \pi h_0$, and $C = 0$. □

Proof of Lemma 11.16

Expanding the RHS, we have:

$$\begin{aligned}\rho_T \Theta_{\mathcal{L}}(\tau) &= \sum_{x \in \mathbb{Z}_q} \left(\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp\left(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau\right) \right) \rho_T \mathbf{e}_x \\ &= \sum_{x \in \mathbb{Z}_q} \left(\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp\left(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau\right) \right) \exp\left(\frac{2\pi i \mu x^2}{q}\right) \mathbf{e}_x.\end{aligned}$$

As we argued in Definition 11.6, every $\mathbf{y} \in x\mathbf{a}+\mathcal{L}$ satisfies $\frac{1}{2}\langle \mathbf{y}, \mathbf{y} \rangle \equiv \phi(x) \equiv \frac{\mu x^2}{q} \pmod 1$, so $\exp\left(\frac{2\pi i \mu x^2}{q}\right) = \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle)$ for every $\mathbf{y}$ in the coset. Moving this factor inside the inner sum, we have:

$$\begin{aligned}\rho_T \Theta_{\mathcal{L}}(\tau) &= \sum_{x \in \mathbb{Z}_q} \left(\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau) \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle) \right) \mathbf{e}_x \\ &= \sum_{x \in \mathbb{Z}_q} \left(\sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle (\tau + 1)) \right) \mathbf{e}_x \\ &= \Theta_{\mathcal{L}}(\tau + 1),\end{aligned}$$

as claimed. □

Proof of Lemma 11.17

We have:

$$\begin{aligned}\Theta_{\mathcal{L}}(-1/\tau) &= \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(-\pi i \langle \mathbf{y}, \mathbf{y} \rangle / \tau) \mathbf{e}_x, \text{ and} \\ \tau^{r/2} \rho_S \Theta_{\mathcal{L}}(\tau) &= \tau^{r/2} \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau) \rho_S \mathbf{e}_x \\ &= \frac{\epsilon_{\mu,q} \tau^{r/2}}{\sqrt{q}} \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \sum_{z \in \mathbb{Z}_q} \exp\left(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau - \frac{4\pi i \mu x z}{q}\right) \mathbf{e}_z.\end{aligned}$$

Consequently, our goal is to show for every $z \in \mathbb{Z}_q$ that:

$$\sum_{\mathbf{y} \in z\mathbf{a}+\mathcal{L}} \exp\left(-\pi i \langle \mathbf{y}, \mathbf{y} \rangle / \tau\right) = \frac{\epsilon_{\mu,q} \tau^{r/2}}{\sqrt{q}} \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp\left(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau - \frac{4\pi i \mu x z}{q}\right). \tag{57}$$

We further unpack each side of (57) as follows:

$$\begin{aligned}\sum_{\mathbf{y} \in z\mathbf{a}+\mathcal{L}} \exp(-\pi i \langle \mathbf{y}, \mathbf{y} \rangle / \tau) &= \sum_{\mathbf{y} \in \mathcal{L}} \exp(-\pi i \langle z\mathbf{a}+\mathbf{y}, z\mathbf{a}+\mathbf{y} \rangle / \tau), \text{ and} \\ \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in x\mathbf{a}+\mathcal{L}} \exp\left(\pi i \langle \mathbf{y}, \mathbf{y} \rangle \tau - \frac{4\pi i \mu x z}{q}\right) &= \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in \mathcal{L}} \exp\left(\pi i \langle \mathbf{y}+x\mathbf{a}, \mathbf{y}+x\mathbf{a} \rangle \tau - \frac{4\pi i \mu x z}{q}\right) \\ &= \sum_{x \in \mathbb{Z}_q} \sum_{\mathbf{y} \in \mathcal{L}} \exp\left(\pi i \langle \mathbf{y}+x\mathbf{a}, \mathbf{y}+x\mathbf{a} \rangle \tau - 2\pi i \langle x\mathbf{a}+\mathbf{y}, z\mathbf{a} \rangle\right)\end{aligned}$$

$$= \sum_{\mathbf{y}\in\mathcal{L}^*} \exp(\pi i\langle \mathbf{y},\mathbf{y}\rangle\tau - 2\pi i\langle \mathbf{y}, z\mathbf{a}\rangle).$$

In the above, we have used: a change of variables to sum over $\mathcal{L}$ instead of one of its cosets; the observation from Definition 11.6 that $2\mu xz/q \equiv \langle x\mathbf{a}+\mathbf{y}, z\mathbf{a}\rangle \pmod 1$; and the observation that as $x, \mathbf{y}$ range over $\mathbb{Z}_q$ and $\mathcal{L}$ respectively, the vector $\mathbf{y} + x\mathbf{a}$ ranges over $\mathcal{L}^*$ (each element exactly once). All the series here converge absolutely (Lemma B.7), so these rearrangements are valid. In summary, (57) is equivalent to:

$$\sum_{\mathbf{y}\in\mathcal{L}} \exp(-\pi i\langle z\mathbf{a}+\mathbf{y}, z\mathbf{a}+\mathbf{y}\rangle/\tau) = \frac{\epsilon_{\mu,q}\tau^{r/2}}{\sqrt{q}} \sum_{\mathbf{y}\in\mathcal{L}^*} \exp(\pi i\langle \mathbf{y},\mathbf{y}\rangle\tau - 2\pi i\langle \mathbf{y}, z\mathbf{a}\rangle). \tag{58}$$

This in turn follows directly from the Poisson summation formula (Fact B.8), setting $\mathbf{u} = z\mathbf{a}$. To explain the normalization factor on the RHS, note that:

$$(-i\tau)^{r/2}\det(\mathcal{L}^*) = \frac{(-i\tau)^{r/2}}{\sqrt{q}} = \frac{\epsilon_{\mu,q}\tau^{r/2}}{\sqrt{q}},$$

where $\det(\mathcal{L}^*) = 1/\det(\mathcal{L}) = 1/\sqrt{q}$, the factor splits as $(-i\tau)^{r/2} = (-i)^{r/2}\tau^{r/2}$ (these are ordinary integer powers, as r is even), and $\epsilon_{\mu,q} = (-i)^{r/2}$ by Milgram's formula (Fact 11.8). This completes the proof. □

Proof of Lemma 11.18

First consider the cusp at ∞. Its width with respect to $\mathrm{SL}_2(\mathbb{Z})$ is 1 (we have $T \in \mathrm{SL}_2(\mathbb{Z})$), and the associated matrix of Section 10.4.2 is $A = \rho(T) = \rho_T$, which is diagonal with entries $\zeta_q^{\mu x^2} = e^{2\pi i\alpha_x}$ for $\alpha_x := (\mu x^2 \bmod q)/q \in [0,1)$, indexed by $x \in \mathbb{Z}_q$. We may therefore apply the results of Section 10.4.2 in the standard basis $\{\mathbf{e}_x\}$, with M the diagonal matrix with entries α_x. The x-th coordinate of $\exp(-2\pi iM\tau)\,\Theta_{\mathcal{L}}(\tau)$ is:

$$e^{-2\pi i\alpha_x\tau} \sum_{\mathbf{y}\in x\mathbf{a}+\mathcal{L}} e^{\pi i\langle \mathbf{y},\mathbf{y}\rangle\tau} = \sum_{\mathbf{y}\in x\mathbf{a}+\mathcal{L}} e^{2\pi i\left(\frac{1}{2}\langle \mathbf{y},\mathbf{y}\rangle - \alpha_x\right)\tau}.$$

By Definition 11.6, every $\mathbf{y} \in x\mathbf{a} + \mathcal{L}$ has $\frac{1}{2}\langle \mathbf{y},\mathbf{y}\rangle \equiv \phi(x) \equiv \alpha_x \pmod 1$, so every exponent $\frac{1}{2}\langle \mathbf{y},\mathbf{y}\rangle - \alpha_x$ above is an integer; and since $\frac{1}{2}\langle \mathbf{y},\mathbf{y}\rangle \geq 0$ and $\alpha_x < 1$, every exponent exceeds -1, hence is nonnegative. Collecting terms by the value of the exponent (justified by absolute convergence, exactly as in the proof of Lemma 10.30) and invoking the uniqueness in Fact B.4, we conclude that the coefficients $\mathbf{c}_n$ of (24) vanish for all $n < 0$, i.e., $\Theta_{\mathcal{L}}$ is holomorphic at ∞ (Definition 10.25).

Next, every cusp $\mathfrak{a} \in \mathbb{Q}$ is $\mathrm{SL}_2(\mathbb{Z})$-equivalent to ∞: by Definition 10.18, there is $g_{\mathfrak{a}} \in \mathrm{SL}_2(\mathbb{Z})$ with $g_{\mathfrak{a}} * \infty = \mathfrak{a}$. So holomorphy at every other cusp follows from Lemma 10.28, and $\Theta_{\mathcal{L}}$ is a modular form.

Finally, we identify the behavior at ∞. The exponent $\frac{1}{2}\langle \mathbf{y},\mathbf{y}\rangle$ vanishes if and only if $\mathbf{y} = \mathbf{0}$, which lies in the coset $x\mathbf{a} + \mathcal{L}$ exactly when $x = 0$; so the unique term with $n + \alpha_x = \frac{1}{2}\langle \mathbf{y},\mathbf{y}\rangle = 0$ is the $\mathbf{y} = \mathbf{0}$ term, which occurs in the coordinate $x = 0$ with coefficient 1; since all other exponents $n + \alpha_x$ are positive, this gives $\mathrm{ord}_\infty(\Theta_{\mathcal{L}}) = 0$. For the limit, let $2\delta := \min\{\langle \mathbf{y},\mathbf{y}\rangle : \mathbf{y} \in \mathcal{L}^* \setminus \{\mathbf{0}\}\}$, which is attained and positive because every ball contains only finitely many points of $\mathcal{L}^*$ (as in the proof of Lemma B.7). For $\Im\tau \geq 1$, every coordinate of $\Theta_{\mathcal{L}}(\tau) - \mathbf{e}_0$ is bounded in absolute value by:

$$\sum_{\mathbf{y}\in\mathcal{L}^*\setminus\{\mathbf{0}\}} e^{-\pi\langle \mathbf{y},\mathbf{y}\rangle\Im\tau} \leq e^{-2\pi\delta(\Im\tau-1)} \sum_{\mathbf{y}\in\mathcal{L}^*\setminus\{\mathbf{0}\}} e^{-\pi\langle \mathbf{y},\mathbf{y}\rangle},$$

using $e^{-\pi\langle \mathbf{y},\mathbf{y}\rangle\Im\tau} = e^{-\pi\langle \mathbf{y},\mathbf{y}\rangle(\Im\tau-1)}e^{-\pi\langle \mathbf{y},\mathbf{y}\rangle} \leq e^{-2\pi\delta(\Im\tau-1)}e^{-\pi\langle \mathbf{y},\mathbf{y}\rangle}$. The remaining sum is finite by Lemma B.7, so the bound tends to 0 as $\Im\tau \to \infty$, and it does not depend on $\Re\tau$. Hence $\Theta_{\mathcal{L}}(\tau) \to \mathbf{e}_0$, uniformly in $\Re\tau$. □

Proof of Lemma 11.20

Write $g = \begin{bmatrix} a & b \\ c & d_0 \end{bmatrix}$ and $\mathbf{v} = (v_0, \dots, v_{q-1}) = \sum_{x \in \mathbb{Z}_q} v_x \mathbf{e}_x$. Since $g \in \Gamma_0(q)$, Fact 10.11 (specifically, (20)) gives $\rho(g)\mathbf{e}_x = \left(\frac{a}{q}\right) \zeta_q^{\mu b d_0 x^2} \mathbf{e}_{d_0 x \bmod q}$. Hence, as functions of $\mathbf{v}$:

$$\begin{aligned}
\rho(g)\mathbf{v} &= \left(\frac{a}{q}\right) \sum_{x \in \mathbb{Z}_q} v_x \zeta_q^{\mu b d_0 x^2} \mathbf{e}_{d_0 x \bmod q} \\
&= \left(\frac{a}{q}\right) \sum_{x \in \mathbb{Z}_q} v_{d_0^{-1} x \bmod q} \zeta_q^{\mu b d_0^{-1} x^2} \mathbf{e}_x \\
\Rightarrow P_\beta(\rho(g)\mathbf{v}) &\propto \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \prod_{i=1}^{d} \left(v_{d_0^{-1} u x_i \bmod q} \zeta_q^{\mu b d_0^{-1} (u x_i)^2} \right) \\
&= \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \underbrace{\zeta_q^{\mu b d_0^{-1} u^2 \langle \mathbf{x}, \mathbf{x} \rangle}}_{=1 \text{ since } \mathbf{x} \in Q \Rightarrow \langle \mathbf{x}, \mathbf{x} \rangle \equiv 0 \pmod q} \prod_{i=1}^{d} v_{d_0^{-1} u x_i \bmod q} \\
&= \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \prod_{i=1}^{d} v_{d_0^{-1} u x_i \bmod q} \\
&= \sum_{u \in \mathbb{Z}_q^*} \sum_{\mathbf{x} \in Q} \beta_{\mathbf{x}} \prod_{i=1}^{d} v_{u x_i \bmod q} \\
&= P_\beta(\mathbf{v}),
\end{aligned}$$

as desired. (The implication uses the degree-d homogeneity of P_β to pull out the sign $\left(\frac{a}{q}\right)^d$, and the penultimate step substitutes $u \mapsto d_0 u$, a bijection of $\mathbb{Z}_q^*$.) □

E Deferred Proofs from Section 12

Calculating the Work Exponent

In this subsection, we bound the support sizes $|S_d(\vartheta, \nu)|$ that Theorem 12.2 delivers, and so prove Corollary 12.4. We emphasize that this subsection bounds a *support size* only; the conversion of such a bound into PIR server work and communication is the proof of Main Theorem 5.5, which combines Theorem 7.1 with Lemma 9.6: a (k, s)-sparse quadratic-phase delta expansion on $\mathbb{Z}_q^d$ with $k \le q^{(\eta + o(1))d}$ gives an s-server scheme with exponent η on n-bit databases. Throughout this subsection, q and s are fixed and $d \to \infty$ along multiples of q; all $o(1)$ terms refer to this regime.

The support-size bound is again the moment-generating-function bound of Lemma D.1, as in the proof of Corollary 11.10; the new ingredients are the two sets $S_d^{\mathrm{sq}}(\nu)$ and $S_d^{\mathrm{edge}}(\vartheta, \nu)$ of vectors of small total cost, their union, and the affine closure and coordinate thickening in the definitions of T_d and S_d, which cost only a $q^{o(d)}$ factor.

Lemma E.1. *Let $\vartheta \in [0, 1]$. For $j \in \mathbb{Z}_q$ identified with its representative in $[0, q-1]$:*

$$w_{q,\vartheta}(j) = \begin{cases} j(j-1) + 2\vartheta j, & \text{if } 0 \le j \le \frac{q-1}{2}; \\ (q-j)(q-j+1) - 2\vartheta(q-j), & \text{if } \frac{q+1}{2} \le j \le q-1. \end{cases}$$

In particular, $w_{q,\vartheta} \geq 0$ and $w_{q,\vartheta}(0) = 0$, and the generating function M_ϑ of Definition 5.4 is given, for $t \in (0,1]$, by:

$$M_\vartheta(t) = 1 + \sum_{u=1}^{(q-1)/2} \left(t^{u(u-1)+2\vartheta u} + t^{u(u+1)-2\vartheta u} \right); \qquad \text{in particular, } M_{1/2}(t) = 1 + 2\sum_{u=1}^{(q-1)/2} t^{u^2}.$$

Proof. The function $k \mapsto k(k-1)+2\vartheta k = k^2+(2\vartheta-1)k$ is a parabola in k with vertex at $k_0 = \frac{1-2\vartheta}{2} \in [-\frac{1}{2}, \frac{1}{2}]$, so among the representatives $\{j + aq : a \in \mathbb{Z}\}$ of j, its minimum is attained at the representative closest to k_0 (a tie gives equal values, by the symmetry of the parabola about k_0). For $0 \leq j \leq \frac{q-1}{2}$ that representative is j itself: $|j - k_0| \leq \frac{q-1}{2} + \frac{1}{2} = \frac{q}{2}$, while every other representative is at distance at least $q - \frac{q}{2} = \frac{q}{2}$, with both equalities only in the tie case. For $\frac{q+1}{2} \leq j \leq q-1$ the closest representative is $j - q \in [-\frac{q-1}{2}, -1]$, and $(j-q)(j-q-1) + 2\vartheta(j-q) = (q-j)(q-j+1) - 2\vartheta(q-j)$. Nonnegativity: for any integer k, the factors k and $k-1+2\vartheta$ are both ≥ 0 (if $k \geq 1$) or both ≤ 0 (if $k \leq -1$, using $\vartheta \leq 1$), while for $k = 0$ the product vanishes. The formula for M_ϑ substitutes $u := j$ in the first branch and $u := q - j$ in the second, and the special case sets $\vartheta = \frac{1}{2}$ (both exponents become u^2). □

Lemma E.2. *Fix an odd prime q, an integer $s \in [2, q]$, and a feasible pair (ϑ, ν) (Definition 5.4). Then, as $d \to \infty$ along multiples of q, we have $|S_d(\vartheta, \nu)| \leq q^{(\eta_{q,s}(\vartheta,\nu)+o(1))\,d}$, with $\eta_{q,s}(\vartheta, \nu)$ as defined in* (14).

Proof. By Definition 12.1, every element of $S_d(\vartheta, \nu)$ has the form $u\mathbf{z} + b\mathbf{1} + c\mathbf{e}_i$ with $u \in \mathbb{Z}_q^*$, $b, c \in \mathbb{Z}_q$, $i \in [d]$, and $\mathbf{z} \in S_d^{\mathrm{sq}}(\nu) \cup S_d^{\mathrm{edge}}(\vartheta, \nu)$; hence:

$$|S_d(\vartheta, \nu)| \leq q(q-1) \cdot qd \cdot \left(|S_d^{\mathrm{sq}}(\nu)| + |S_d^{\mathrm{edge}}(\vartheta, \nu)| \right),$$

and the prefactor $q^2(q-1)d$ is $q^{o(d)}$. It remains to bound $|S_d^{\mathrm{sq}}(\nu)|$ and $|S_d^{\mathrm{edge}}(\vartheta, \nu)|$. Both sets have the form $\{\mathbf{z} \in \mathbb{Z}_q^d : \sum_i w(z_i) \leq Bd\}$ for a nonnegative weight function w on $\mathbb{Z}_q$ and a budget $B \geq 0$: take $(w, B) = (w_{q,1/2}, \nu)$ for the square set and $(w, B) = (w_{q,\vartheta}, \gamma(\vartheta, \nu))$ for the edge set. (Here $\gamma(\vartheta, \nu) \geq 0$ under feasibility: both summands in its definition are products of nonnegative factors, since $\vartheta \leq \frac{q+1}{6s}$ and $\nu \leq \frac{q^2-1}{8s}$.) By Lemma D.1, for any fixed $0 < t \leq 1$, each of them satisfies:

$$\left| \left\{ \mathbf{z} \in \mathbb{Z}_q^d : \sum_{i=1}^{d} w(z_i) \leq Bd \right\} \right| \leq \left(t^{-B} \sum_{j \in \mathbb{Z}_q} t^{w(j)} \right)^d.$$

Taking the infimum over t for each set, and bounding the sum of the two cardinalities by twice the larger, gives the claim. □

Optimizing this bound over the feasible parameters is a finite-dimensional problem, which we solve numerically:

Lemma E.3. *We have $\eta_{3,2} \leq 0.820340$, $\eta_{3,3} \leq 0.732672$, $\eta_{5,4} \leq 0.702450$, $\eta_{5,5} \leq 0.653434$, $\eta_{7,6} \leq 0.652057$, $\eta_{7,7} \leq 0.619701$, $\eta_{11,10} \leq 0.607791$, and $\eta_{11,11} \leq 0.590014$.*

Proof. Any feasible pair (ϑ, ν), together with one evaluation point t for each of the two infima in (14), certifies an upper bound on $\eta_{q,s}$ by direct evaluation. The following witnesses certify the eight bounds, which are the exponents of Main Theorem 5.5; rounding the last column up in the sixth decimal gives the stated values.

q	s	ϑ	ν	t^{sq}	t^{edge}	$\eta_{q,s} \le$
3	2	$1/3$	0.358650	0.279606	0.286169	0.8203393
3	3	0.169085	0.291108	0.205326	0.201386	0.7326712
5	4	$1/4$	0	—	0.414520	0.7024495
5	5	0.100050	0.479668	0.354308	0.354830	0.6534337
7	6	0.166638	0.740704	0.509311	0.509527	0.6520565
7	7	0.071420	0.653042	0.465067	0.465248	0.6197008
11	10	0.1	1.08	0.629376	0.629376	0.6077910
11	11	0.045454	0.991735	0.604009	0.604010	0.5900131

(For $(q,s) = (5,4)$ the optimum lies on the boundary of the feasible region: $\nu = 0$ makes $S_d^{\mathrm{sq}}(0) = \{\mathbf{0}\}$, so $E_q^{\mathrm{sq}}(0) = 0$ and no t^{sq} is needed, while $\vartheta = \frac{q+1}{6s} = \frac{1}{4}$ zeroes the second term of γ, leaving $\gamma = \frac{(q-1)\vartheta}{2} = \frac{1}{2}$.) In each row, the pair (ϑ, ν) is feasible for Definition 12.1, and evaluating the two expressions of (14) at the listed t's and taking the maximum yields a value at most the stated bound; finer numerical optimization does not improve these bounds in the sixth decimal, so we believe them to be essentially optimal for this analysis. The row $(q,s) = (7,7)$ is the "seven-server" exponent that Part I also deploys for $s \in \{8,9\}$ (running the seven-server construction with idle servers). □

Lemma E.4. *We have* $\eta_{q,q} \le \frac{1}{2} + \frac{1}{2}\frac{\ln(\pi e/6)}{\ln q} + o\left(\frac{1}{\ln q}\right)$ *as* $q \to \infty$.

Proof. We will take $\vartheta = 1/6$ and $\nu = 0$ as our feasible values, so that:

$$\begin{aligned}\gamma &= \frac{q-1}{2}\cdot\frac{1}{6} + \left(1 - \frac{q}{q+1}\right)\left(\frac{q^2-1}{4q}\right)\\ &= \frac{q-1}{12} + \frac{q-1}{4q}\\ &= \frac{q^2+2q-3}{12q}.\end{aligned}$$

Now note that $E_q^{\mathrm{sq}}(\nu) = 0$, as can be seen by taking the limit as $t \to 0$. So we need only address $E_{q,q}^{\mathrm{edge}}(1/6, 0)$. To this end, write $t = e^{-\lambda/q}$ for $\lambda > 0$, then we have:

$$\begin{aligned}M_{1/6}(t) &= 1 + \sum_{u=1}^{(q-1)/2} \exp\left(-\frac{\lambda}{q}(u^2 - 2u/3)\right) + \sum_{u=1}^{(q-1)/2} \exp\left(-\frac{\lambda}{q}(u^2 + 2u/3)\right)\\ &= 1 + e^{\lambda/(9q)} \sum_{u=1}^{(q-1)/2} \exp\left(-\frac{\lambda}{q}(u - 1/3)^2\right) + e^{\lambda/(9q)} \sum_{u=1}^{(q-1)/2} \exp\left(-\frac{\lambda}{q}(u + 1/3)^2\right).\end{aligned}$$

Now by Lemma B.10, applied with $h = 1/\sqrt{q}$, we have:

$$\begin{aligned}\sum_{u=1}^{(q-1)/2} \exp\left(-\frac{\lambda}{q}(u - 1/3)^2\right) &= \frac{1}{2}\sqrt{\frac{q\pi}{\lambda}} + O(1)\\ \sum_{u=1}^{(q-1)/2} \exp\left(-\frac{\lambda}{q}(u + 1/3)^2\right) &= -e^{-\lambda/(9q)} + \sum_{u=1}^{(q+1)/2} \exp\left(-\frac{\lambda}{q}(u - 2/3)^2\right)\\ &= \frac{1}{2}\sqrt{\frac{q\pi}{\lambda}} + O(1).\end{aligned}$$

Plugging these in, we get:

$$M_{1/6}(e^{-\lambda/q}) = e^{\lambda/(9q)}\sqrt{\frac{q\pi}{\lambda}} + O(1)$$
$$= \sqrt{\frac{q\pi}{\lambda}}(1 + o(1)).$$

Thus we have:

$$\begin{aligned} E^{\mathrm{edge}}_{q,q}(\vartheta, \nu) &= \frac{1}{\ln q}\inf_{\lambda>0}(\ln M_{1/6}(e^{-\lambda/q}) + \gamma\lambda/q) \\ &= \frac{1}{\ln q}\inf_{\lambda>0}\left(\ln\sqrt{\frac{q\pi}{\lambda}} + \frac{\lambda}{12} + o(1)\right) \\ &\le \frac{1}{\ln q}\left(\ln\sqrt{\frac{q\pi e}{6}} + o(1)\right) \qquad \text{(setting } \lambda = 6\text{)} \\ &= \frac{1}{2} + \frac{1}{2}\frac{\ln(\pi e/6)}{\ln q} + o\left(\frac{1}{\ln q}\right), \end{aligned}$$

as desired. □

With these lemmas, the corollary reduces to the theorem:

Proof of Corollary 12.4. First fix a feasible pair (ϑ, ν). For any d, Corollary 12.3 (which rests on Theorem 12.2) provides a $(|S_D(\vartheta, \nu)|, s)$-sparse quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ on $\mathbb{Z}_q^d$ with permutation-invariant support, where $D := q\lceil d/q\rceil$ satisfies $d \le D < d+q$. By Lemma E.2, $|S_D(\vartheta, \nu)| \le q^{(\eta_{q,s}(\vartheta,\nu)+o(1))D} = q^{(\eta_{q,s}(\vartheta,\nu)+o(1))d}$, which proves the first statement.

For the second, fix $\epsilon > 0$; it suffices to show that for all sufficiently large d there is a (k, s)-sparse quadratic-phase delta expansion on $\mathbb{Z}_q^d$ with $k \le q^{(\eta_{q,s}+\epsilon)d}$. Pick feasible (ϑ, ν) with $\eta_{q,s}(\vartheta, \nu) \le \eta_{q,s} + \epsilon/3$. For any d, Corollary 12.3 (which rests on Theorem 12.2) provides a $(|S_D(\vartheta, \nu)|, s)$-sparse quadratic-phase delta expansion over $\mathbb{Q}(\zeta_q)$ on $\mathbb{Z}_q^d$ with permutation-invariant support, where $D := q\lceil d/q\rceil$ satisfies $d \le D < d + q$. By Lemma E.2, $|S_D(\vartheta, \nu)| \le q^{(\eta_{q,s}+2\epsilon/3)D} \le q^{(\eta_{q,s}+2\epsilon/3)(d+q)}$ for all sufficiently large d, and the right-hand side is at most $q^{(\eta_{q,s}+\epsilon)d}$ once $(\eta_{q,s} + 2\epsilon/3)\, q \le \epsilon d/3$. The particular numerical bounds are Lemma E.3, and the asymptotic bound is given by Lemma E.4, together with $\frac{\frac{1}{2}\ln(\pi e/6)}{\ln q} = \frac{\frac{1}{2}\log_2(\pi e/6)}{\log_2 q}$ and $\frac{1}{2}\log_2(\pi e/6) = 0.2546\ldots \le 0.255$. □

Proof of Corollary 12.3

Let $D := q\lceil d/q\rceil \ge d$, so that $q \mid D$ and $D - d < q$. By Theorem 12.2, there exist coefficients $\{c_{\mathbf{z},a}\}$ supported on $\mathcal{W} \times \mathcal{A}$, where $\mathcal{W} := S_D(\vartheta, \nu) \subseteq \mathbb{Z}_q^D$ and $\mathcal{A} \subseteq \mathbb{Z}_q$ has size s, satisfying:

$$\sum_{a\in\mathcal{A}}\sum_{\mathbf{z}\in\mathbb{Z}_q^D} c_{\mathbf{z},a}\,\zeta_q^{\langle\mathbf{z},\mathbf{x}\rangle + a\langle\mathbf{x},\mathbf{x}\rangle} = \mathbb{1}[\mathbf{x} = \mathbf{0}] \qquad \text{for all } \mathbf{x} \in \mathbb{Z}_q^D.$$

Restrict this identity to the padded vectors $\mathbf{x} = (\mathbf{x}', \mathbf{0})$ with $\mathbf{x}' \in \mathbb{Z}_q^d$: then $\langle\mathbf{x},\mathbf{x}\rangle = \langle\mathbf{x}',\mathbf{x}'\rangle$, $\langle\mathbf{z},\mathbf{x}\rangle = \langle\pi(\mathbf{z}),\mathbf{x}'\rangle$ for the projection $\pi : \mathbb{Z}_q^D \to \mathbb{Z}_q^d$ onto the first d coordinates, and $\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x}' = \mathbf{0}$. Grouping the left-hand side by the value of $\mathbf{z}' := \pi(\mathbf{z})$, the coefficients:

$$c'_{\mathbf{z}',a} := \sum_{\mathbf{z}\in\mathbb{Z}_q^D:\ \pi(\mathbf{z})=\mathbf{z}'} c_{\mathbf{z},a}$$

form a quadratic-phase delta expansion on $\mathbb{Z}_q^d$ (Definition 4.3). Their support satisfies $\mathsf{supp}\{c'_{\mathbf{z}',a}\} \subseteq \pi(\mathcal{W}) \times \mathcal{A}$, and $|\pi(\mathcal{W})| \le |\mathcal{W}| = |S_D(\vartheta, \nu)|$: distinct frequencies may merge under the projection π, which only shrinks the support, never grows it. So this expansion is $(|S_D(\vartheta, \nu)|, s)$-sparse. Its support $\pi(\mathcal{W})$ is permutation-invariant: a permutation σ of the first d coordinates extends to a permutation $\tilde{\sigma}$ of the D coordinates that fixes the last $D-d$, and $\sigma(\pi(\mathcal{W})) = \pi(\tilde{\sigma}(\mathcal{W})) = \pi(\mathcal{W})$, as $\mathcal{W}$ is permutation-invariant (Theorem 12.2). □

Proof of Proposition 12.7

(42) and (43) are apparent from definition. To see (44), note that:

$$\begin{aligned}
\beta_{\mathbf{x}+\mathbf{1}}\alpha_0^{(J)} &= \alpha_{\mathbf{x}+\mathbf{1}}^{(J)} \\
&= \mathbb{1}\left[\sum_{i=1}^{d}(x_i+1)=0\right]\cdot\zeta_q^{-J(x_1+1)}\cdot\frac{1}{q}\sum_{b\in\mathbb{Z}_q}\zeta_q^{-Jb}\alpha_{\mathbf{x}+(b+1)\mathbf{1}} \\
&= \mathbb{1}\left[\sum_{i=1}^{d}x_i=0\right]\cdot\zeta_q^{-Jx_1}\cdot\frac{1}{q}\sum_{b\in\mathbb{Z}_q}\zeta_q^{-J(b+1)}\alpha_{\mathbf{x}+(b+1)\mathbf{1}} \\
&= \alpha_{\mathbf{x}}^{(J)} \\
&= \beta_{\mathbf{x}}\alpha_0^{(J)}.
\end{aligned}$$

In the above, we have used the fact that $d \equiv 0 \pmod q$, so that $\sum_i (x_i+1) \equiv \sum_i x_i \pmod q$. Finally, we check (45):

$$\begin{aligned}
\alpha_0^{(J)}\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\beta_{\mathbf{x}}\zeta_q^{a\langle\mathbf{x},\mathbf{x}\rangle+\langle\mathbf{x},\mathbf{z}\rangle} &= \sum_{\mathbf{x}\in\mathbb{Z}_q^d}\alpha_{\mathbf{x}}^{(J)}\zeta_q^{a\langle\mathbf{x},\mathbf{x}\rangle+\langle\mathbf{x},\mathbf{z}\rangle} \\
&= \frac{1}{q}\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\sum_{b\in\mathbb{Z}_q}\mathbb{1}\left[\sum_{i=1}^{d}x_i=0\right]\cdot\zeta_q^{-Jx_1}\cdot\zeta_q^{-Jb}\alpha_{\mathbf{x}+b\mathbf{1}}\zeta_q^{a\langle\mathbf{x},\mathbf{x}\rangle+\langle\mathbf{x},\mathbf{z}\rangle} \\
&= \frac{1}{q}\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\mathbb{1}\left[\sum_{i=1}^{d}x_i=0\right]\cdot\left(\sum_{b\in\mathbb{Z}_q}\zeta_q^{-Jx_1}\alpha_{\mathbf{x}}\zeta_q^{a\langle\mathbf{x}-b\mathbf{1},\mathbf{x}-b\mathbf{1}\rangle+\langle\mathbf{x}-b\mathbf{1},\mathbf{z}\rangle}\right) \\
&= \frac{1}{q^2}\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\sum_{c\in\mathbb{Z}_q}\zeta_q^{c\langle\mathbf{x},\mathbf{1}\rangle}\cdot\left(\sum_{b\in\mathbb{Z}_q}\zeta_q^{-Jx_1}\alpha_{\mathbf{x}}\zeta_q^{a\langle\mathbf{x}-b\mathbf{1},\mathbf{x}-b\mathbf{1}\rangle+\langle\mathbf{x}-b\mathbf{1},\mathbf{z}\rangle}\right).
\end{aligned}$$

(The substitution $\mathbf{x} \mapsto \mathbf{x} - b\mathbf{1}$ in the third line uses $q \mid d$ once more: the indicator $\mathbb{1}[\sum_i x_i = 0]$ is unchanged by it because $\sum_i b = bd \equiv 0$, and the factor ζ_q^{Jb} it produces cancels the ζ_q^{-Jb} of the average. The final line replaced the indicator by the character sum $\frac{1}{q}\sum_c \zeta_q^{c\langle\mathbf{x},\mathbf{1}\rangle}$.) To unpack this, note that:

$$\begin{aligned}
a\langle\mathbf{x}-b\mathbf{1},\mathbf{x}-b\mathbf{1}\rangle + \langle\mathbf{x}-b\mathbf{1},\mathbf{z}\rangle + c\langle\mathbf{x},\mathbf{1}\rangle - Jx_1 &= a\langle\mathbf{x},\mathbf{x}\rangle + \langle\mathbf{x}, -2ab\mathbf{1}+\mathbf{z}+c\mathbf{1}-J\mathbf{e}_1\rangle + ab^2\langle\mathbf{1},\mathbf{1}\rangle - b\langle\mathbf{1},\mathbf{z}\rangle \\
&= a\langle\mathbf{x},\mathbf{x}\rangle + \langle\mathbf{x}, \mathbf{z}+(c-2ab)\mathbf{1}-J\mathbf{e}_1\rangle - b\langle\mathbf{1},\mathbf{z}\rangle,
\end{aligned}$$

where the second equality uses $\langle\mathbf{1},\mathbf{1}\rangle = d \equiv 0 \pmod q$ — the third and final place where $q \mid d$ enters this step. Note that $\mathbf{z} \in T_d(\vartheta,\nu) \Rightarrow \mathbf{z}+(c-2ab)\mathbf{1} \in T_d(\vartheta,\nu) \Rightarrow \mathbf{z}+(c-2ab)\mathbf{1}-J\mathbf{e}_1 \in S_d(\vartheta,\nu)$. Thus we have:

$$\frac{1}{q^2}\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\sum_{c\in\mathbb{Z}_q}\zeta_q^{c\langle\mathbf{x},\mathbf{1}\rangle}\cdot\left(\sum_{b\in\mathbb{Z}_q}\zeta_q^{-Jx_1}\alpha_{\mathbf{x}}\zeta_q^{a\langle\mathbf{x}-b\mathbf{1},\mathbf{x}-b\mathbf{1}\rangle+\langle\mathbf{x}-b\mathbf{1},\mathbf{z}\rangle}\right) = \frac{1}{q^2}\sum_{b,c\in\mathbb{Z}_q}\zeta_q^{-b\langle\mathbf{1},\mathbf{z}\rangle}\underbrace{\left(\sum_{\mathbf{x}\in\mathbb{Z}_q^d}\alpha_{\mathbf{x}}\zeta_q^{a\langle\mathbf{x},\mathbf{x}\rangle+\langle\mathbf{x},\mathbf{z}+(c-2ab)\mathbf{1}-J\mathbf{e}_1\rangle}\right)}_{=0 \text{ by } (41)}.$$

This is hence zero as claimed. □

Proof of Proposition 12.9

Homogeneity is clear. For the second bullet, we have $P_\beta(\mathbf{e}_x) = \beta_{x\mathbf{1}} = \beta_0 = 1$ by (42) and (44). For the invariance properties under X and Z, recall the coordinate descriptions $(\mathsf{X}\mathbf{y})_j = y_{j+1}$ and $(\mathsf{Z}\mathbf{y})_j = \zeta_q^j y_j$ (indices modulo q). Then:

$$\begin{aligned}
P_\beta(\mathsf{X}\mathbf{y}) &= P_\beta(y_1, y_2, \dots, y_{q-1}, y_0) \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^{d} y_{x_i+1} \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}-\mathbf{1}} \prod_{i=1}^{d} y_{x_i} \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^{d} y_{x_i} \text{ (by (44))} \\
&= P_\beta(\mathbf{y}), \text{ and} \\
P_\beta(\mathsf{Z}\mathbf{y}) &= P_\beta(y_0, \zeta_q y_1, \zeta_q^2 y_2, \dots, \zeta_q^{q-1} y_{q-1}) \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^{d} \zeta_q^{x_i} y_{x_i} \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \zeta_q^{\langle \mathbf{x},\mathbf{1}\rangle} \prod_{i=1}^{d} y_{x_i} \\
&= \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^{d} y_{x_i} \text{ (by (43), } \zeta_q^{\langle \mathbf{x},\mathbf{1}\rangle} = 1 \text{ on the support of } \beta) \\
&= P_\beta(\mathbf{y}),
\end{aligned}$$

as desired. For the final assertion, it suffices to treat $H \in \{\mathsf{X}, \mathsf{Z}\}$: a general H is a product of such factors (negative powers are positive powers, as $\mathsf{X}^q = \mathsf{Z}^q = I$), which we may strip off one at a time, as $\propto_\zeta$ is transitive. For $H \in \{\mathsf{X}, \mathsf{Z}\}$, Lemma 10.13 (with $(m, n) = (1, 0)$ or $(0, 1)$) gives integers a, b, c such that $\rho(g)H = \zeta_q^c \mathsf{X}^a \mathsf{Z}^b \rho(g)$. This then tells us that:

$$\begin{aligned}
P_\beta(\rho(g)H\mathbf{y}) &= P_\beta(\zeta_q^c \mathsf{X}^a \mathsf{Z}^b \rho(g)\mathbf{y}) \\
&= \zeta_q^{cd} P_\beta(\mathsf{X}^a \mathsf{Z}^b \rho(g)\mathbf{y}) \\
&\propto_\zeta P_\beta(\rho(g)\mathbf{y}).
\end{aligned}$$

In the last step, we used the invariance $P_\beta(\mathsf{X}^a\mathsf{Z}^b\mathbf{w}) = P_\beta(\mathbf{w})$ (for all $\mathbf{w} \in \mathbb{C}^q$) that follows from the third bullet, and that ζ_q^{cd} is a root of unity. □

Proof of Lemma 12.13

We prove the exact identity $\Theta(\tau + 1, w) = e^{\pi i q/4} \rho(T) \Theta(\tau, w)$; since $e^{\pi i q/4}$ is a root of unity, this gives the statement.

Firstly, note that:

$$\begin{aligned}(k+(q-1)/2)^2(q+1) &= (k-1/2)^2 + k^2 q + kq^2 + \frac{q^3-q^2-q}{4} \\ &\equiv (k-1/2)^2 - \frac{q^2}{4} \pmod{2q} \\ \Rightarrow \exp\left(\frac{\pi i(k+(q-1)/2)^2(q+1)}{q}\right) &= \exp\left(\frac{\pi i(k-1/2)^2}{q}\right)\exp\left(\frac{-\pi i q}{4}\right).\end{aligned}$$

Bearing this in mind, we have:

$$\begin{aligned}\rho(T)\Theta(\tau,w) &= \sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)w\right)\rho(T)\mathbf{e}_{k+(q-1)/2 \bmod q} \\ &= \sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)w\right)\exp\left(\frac{\pi i(k+(q-1)/2)^2(q+1)}{q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q} \\ &= \exp\left(\frac{-\pi i q}{4}\right)\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)w\right)\exp\left(\frac{\pi i(k-1/2)^2}{q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q} \\ &= \exp\left(\frac{-\pi i q}{4}\right)\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i(\tau+1)}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)w\right)\mathbf{e}_{k+(q-1)/2 \bmod q} \\ &= e^{-\pi i q/4}\Theta(\tau+1,w),\end{aligned}$$

which implies the conclusion. □

Proof of Lemma 12.14

We prove the exact identity $\Theta(-1/\tau, w/\tau) = \epsilon_q^{-1}(-i)^q(-i\tau)^{1/2}e^{\pi i q w^2/\tau}\rho(S)\Theta(\tau,w)$; since $\epsilon_q \in \{\pm 1, \pm i\}$, the constant $\epsilon_q^{-1}(-i)^q$ is a root of unity, so this gives the statement. As in the proof of Lemma 11.17, the identity is a consequence of the Poisson summation formula (Fact B.8). We have:

$$\begin{aligned}\Theta(-1/\tau, w/\tau) &= \sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{-\pi i}{q\tau}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{w}{\tau}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}, \\ &= \sum_{y\in\mathbb{Z}_q}\sum_{k\in y-(q-1)/2+q\mathbb{Z}}(-1)^k \exp\left(\frac{-\pi i}{q\tau}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{w}{\tau}\right)\mathbf{e}_y \\ &= \sum_{y\in\mathbb{Z}_q}\sum_{k\in\mathbb{Z}}(-1)^{y-(q-1)/2+qk} \exp\left(\frac{-\pi i}{q\tau}\left(y-\frac{q}{2}+qk\right)^2 + 2\pi i\left(y-\frac{q}{2}+qk\right)\frac{w}{\tau}\right)\mathbf{e}_y \\ &= \sum_{y\in\mathbb{Z}_q}\sum_{k\in\mathbb{Z}}(-1)^{y-(q-1)/2} \exp\left(\frac{-\pi i}{q\tau}\left(y-\frac{q}{2}+qk\right)^2 + 2\pi i\left(y-\frac{q}{2}+qk\right)\frac{w}{\tau} + \pi i k\right)\mathbf{e}_y \text{ and} \\ \rho(S)\Theta(\tau,w) &= \frac{\epsilon_q}{\sqrt{q}}\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)w\right)\sum_{y\in\mathbb{Z}_q}\exp\left(\frac{-2\pi i(k+(q-1)/2)y}{q}\right)\mathbf{e}_y.\end{aligned}$$

Hence our goal is to show for every $y \in \mathbb{Z}_q$ that:

$$(-1)^{y-(q-1)/2}\sum_{k\in\mathbb{Z}}\exp\left(\frac{-\pi i}{q\tau}\left(y-\frac{q}{2}+qk\right)^2 + 2\pi i\left(y-\frac{q}{2}+qk\right)\frac{w}{\tau} - \pi i k\right) \tag{59}$$

$$=\frac{(-i)^q(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}}\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)w\right)\exp\left(\frac{-2\pi i(k+(q-1)/2)y}{q}\right). \tag{60}$$

To do this, note that:

$$\begin{aligned}
&\frac{-\pi i}{q\tau}\left(y-\frac{q}{2}+qk\right)^2+2\pi i\left(y-\frac{q}{2}+qk\right)\frac{w}{\tau}-\pi ik\\
=&\frac{-\pi iq}{\tau}\left(\left(k+\frac{2y-q}{2q}\right)^2-\frac{w(2y-q)}{q}-2wk+\frac{k\tau}{q}\right)\\
=&\frac{-\pi iq}{\tau}\left(k+\frac{2y-q-2qw+\tau}{2q}\right)^2-\pi iw+\frac{\pi iqw^2}{\tau}+\frac{\pi i(2y-q)}{2q}+\frac{\pi i\tau}{4q}.
\end{aligned}$$

This allows us to rewrite (59) and apply Fact B.8 to it — with $r=1$, lattice $\mathcal{L}=\mathbb{Z}$, modular variable $\tau/q\in\mathcal{H}$, and the *complex* shift $\mathbf{u}=\frac{2y-q-2qw+\tau}{2q}$, which Fact B.8 allows — using $(-i\tau/q)^{1/2}=(-i\tau)^{1/2}/\sqrt{q}$ (valid for principal branches, as $q>0$):

$$\begin{aligned}
&(-1)^{y-(q-1)/2}\exp\left(-\pi iw+\frac{\pi iqw^2}{\tau}+\frac{\pi i(2y-q)}{2q}+\frac{\pi i\tau}{4q}\right)\sum_{k\in\mathbb{Z}}\exp\left(\frac{-\pi iq}{\tau}\left(k+\frac{2y-q-2qw+\tau}{2q}\right)^2\right)\\
=&(-1)^{y-(q-1)/2}\frac{(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}}\exp\left(-\pi iw+\frac{\pi i(2y-q)}{2q}+\frac{\pi i\tau}{4q}\right)\sum_{k\in\mathbb{Z}}\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{2y-q-2qw+\tau}{2q}\right)\right)\\
=&(-1)^{y-(q-1)/2}\frac{(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}}\exp\left(-\pi iw+\frac{\pi i(2y-q)}{2q}+\frac{\pi i\tau}{4q}\right)\sum_{k\in\mathbb{Z}}\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{y}{q}-\frac{1}{2}-w+\frac{\tau}{2q}\right)\right)\\
=&-i(-1)^{y-(q-1)/2}\frac{(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}}\exp\left(-\pi iw+\frac{\pi iy}{q}+\frac{\pi i\tau}{4q}\right)\sum_{k\in\mathbb{Z}}\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{y}{q}-\frac{1}{2}-w+\frac{\tau}{2q}\right)\right).
\end{aligned} \tag{61}$$

On the other hand, working from (60), we have:

$$\begin{aligned}
&(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)w-\frac{2\pi i(k+(q-1)/2)y}{q}\right)\\
=&\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)w-\frac{2\pi i(k+(q-1)/2)y}{q}+\pi ik\right)\\
=&\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{y}{q}-\frac{1}{2}-w+\frac{\tau}{2q}\right)+\frac{\pi i\tau}{4q}-\pi iw-\frac{\pi iy(q-1)}{q}\right),
\end{aligned}$$

so (60) simplifies as follows:

$$\begin{aligned}
&\frac{(-i)^q(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}}\sum_{k\in\mathbb{Z}}\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{y}{q}-\frac{1}{2}-w+\frac{\tau}{2q}\right)+\frac{\pi i\tau}{4q}-\pi iw-\frac{\pi iy(q-1)}{q}\right)\\
=&\frac{(-i)^q(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}}\exp\left(\frac{\pi i\tau}{4q}-\pi iw+\frac{\pi iy}{q}-\pi iy\right)\sum_{k\in\mathbb{Z}}\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{y}{q}-\frac{1}{2}-w+\frac{\tau}{2q}\right)\right)\\
=&\frac{(-i)^q(-i\tau)^{1/2}e^{\pi iqw^2/\tau}(-1)^y}{\sqrt{q}}\exp\left(\frac{\pi i\tau}{4q}-\pi iw+\frac{\pi iy}{q}\right)\sum_{k\in\mathbb{Z}}\exp\left(\frac{\pi i\tau k^2}{q}-2\pi ik\left(\frac{y}{q}-\frac{1}{2}-w+\frac{\tau}{2q}\right)\right).
\end{aligned}$$

To show that this is equal to (61) (and hence (59)), it suffices to show that:

$$-i(-1)^{y-(q-1)/2}\frac{(-i\tau)^{1/2}e^{\pi iqw^2/\tau}}{\sqrt{q}} = \frac{(-i)^q(-i\tau)^{1/2}e^{\pi iqw^2/\tau}(-1)^y}{\sqrt{q}}$$
$$\Leftrightarrow -i(-1)^{-(q-1)/2} = (-i)^q,$$

which is evident as $(-1)^{-(q-1)/2} = (-1)^{(q-1)/2} = (-i)^{q-1}$. □

Proof of Lemma 12.15

All proportionality constants below are roots of unity independent of (τ, w). First, we unpack the RHS of (46), using $\mathsf{Z}^b\mathsf{X}^a\mathbf{e}_j = \zeta_q^{b(j-a)}\mathbf{e}_{j-a}$:

$$\begin{aligned}
&e^{-\pi ia^2\tau/q-2\pi iaw}\cdot \mathsf{Z}^b\mathsf{X}^a\Theta(\tau,w)\\
=&e^{-\pi ia^2\tau/q-2\pi iaw}\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)w+\frac{2\pi ib(k+(q-1)/2-a)}{q}\right)\mathbf{e}_{k+(q-1)/2-a \bmod q}\\
\propto_\zeta& e^{-\pi ia^2\tau/q-2\pi iaw}\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)w+\frac{2\pi ibk}{q}\right)\mathbf{e}_{k+(q-1)/2-a \bmod q}\\
\propto_\zeta& e^{-\pi ia^2\tau/q-2\pi iaw}\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k+a-\frac{1}{2}\right)^2+2\pi i\left(k+a-\frac{1}{2}\right)w+\frac{2\pi ibk}{q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}.
\end{aligned} \tag{62}$$

In the second step we dropped the constant $\zeta_q^{b((q-1)/2-a)}$; in the third we substituted $k \mapsto k+a$ and dropped the resulting constant $(-1)^a\zeta_q^{ab}$. Now let us unpack the LHS of (46):

$$\begin{aligned}
\Theta\left(\tau, w+\frac{a\tau+b}{q}\right) &= \sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)\left(w+\frac{a\tau+b}{q}\right)\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\\
&= \exp\left(-\frac{\pi i\tau a^2}{q}\right)\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k+a-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)\left(w+\frac{b}{q}\right)\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\\
&= \exp\left(-\frac{\pi i\tau a^2}{q}-\frac{\pi ib}{q}\right)\cdot\\
&\quad\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k+a-\frac{1}{2}\right)^2+2\pi i\left(k-\frac{1}{2}\right)w+\frac{2\pi ikb}{q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\\
&= \exp\left(-\frac{\pi i\tau a^2}{q}-\frac{\pi ib}{q}-2\pi iaw\right)\cdot\\
&\quad\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(\frac{\pi i\tau}{q}\left(k+a-\frac{1}{2}\right)^2+2\pi i\left(k+a-\frac{1}{2}\right)w+\frac{2\pi ikb}{q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q},
\end{aligned}$$

which matches (62) up to the constant $e^{-\pi ib/q}$, a root of unity, as desired. □

Proof of Lemma 12.16

All constants below are independent of (τ, w), and we work with $\propto_\zeta$ throughout. We repeatedly use the following observation: if two nowhere-vanishing holomorphic functions on the connected set $\mathcal{H}$ have the

same square, then their ratio is continuous and squares to 1, so it is a constant ± 1 — a root of unity. First we address the generators T, S:

- When $g = T$, the claim is Lemma 12.13 (here $c = 0$ and $G(T, \tau) = 1$).
- When $g = S$, we have $c = 1$ and $G(S, \tau) = \tau$, and Lemma 12.14 gives the claim with the square root $(-i\tau)^{1/2}$ in place of a branch $\tau^{1/2}$ of $G(S, \tau)^{1/2}$. These differ by a root of unity: the ratio $(-i\tau)^{1/2}/\tau^{1/2}$ is continuous on $\mathcal{H}$ and squares to the constant $-i$, so it is a constant square root of $-i$, i.e., an eighth root of unity.

Since S and T generate $\mathrm{SL}_2(\mathbb{Z})$ (Fact 10.2), it remains to show that if the claim is true for $g_1, g_2 \in \mathrm{SL}_2(\mathbb{Z})$, then it is also true for g_1^{-1} and $g_1 g_2$. First we address g_1^{-1}: the claim for g_1^{-1} at the point $(\tau', w') := g_1 * (\tau, w)$ — which ranges over all of $\mathcal{H} \times \mathbb{C}$ as (τ, w) does — follows by rearranging (47) (the inverse of a root of unity is again one):

$$\Theta(\tau, w) \propto_\zeta G(g_1, \tau)^{-1/2} \exp\left(-\frac{\pi i q c w^2}{G(g_1, \tau)}\right) \rho(g_1^{-1})\Theta(g_1 * (\tau, w)).$$

To address the $G(g_1, \tau)^{-1/2}$ term, note by Proposition 10.17 that:

$$1 = G(I, \tau) = G(g_1^{-1}, g_1 * \tau)G(g_1, \tau) \Rightarrow G(g_1, \tau)^{-1/2} = \pm\, G(g_1^{-1}, g_1 * \tau)^{1/2},$$

with a constant sign by the observation above. To address the exponential term, note that:

$$\begin{aligned}\frac{(w/G(g_1, \tau))^2}{G(g_1^{-1}, g_1 * \tau)} &= \frac{w^2}{G(g_1, \tau)^2 G(g_1^{-1}, g_1 * \tau)} \\ &= \frac{w^2}{G(g_1, \tau)} \text{ (Proposition 10.17),}\end{aligned}$$

and also that $\begin{bmatrix} a & b \\ c & d_0 \end{bmatrix}^{-1} = \begin{bmatrix} d_0 & -b \\ -c & a \end{bmatrix}$, so this accounts for the negative sign. Finally, we address $g_1 g_2$:

$$\begin{aligned}\Theta(g_1 g_2 * (\tau, w)) &\propto_\zeta G(g_1, g_2 * \tau)^{1/2} \exp\left(\frac{\pi i q c(g_1)(w/G(g_2, \tau))^2}{G(g_1, g_2 * \tau)}\right) \rho(g_1)\Theta(g_2 * (\tau, w)) \\ &\propto_\zeta G(g_1, g_2 * \tau)^{1/2} G(g_2, \tau)^{1/2} \exp\left(\frac{\pi i q c(g_1)(w/G(g_2, \tau))^2}{G(g_1, g_2 * \tau)}\right) \times \\ &\qquad \exp\left(\frac{\pi i q c(g_2) w^2}{G(g_2, \tau)}\right) \rho(g_1 g_2)\Theta(\tau, w).\end{aligned}$$

By Proposition 10.17, the two square-root factors satisfy $G(g_1, g_2 * \tau)^{1/2}\, G(g_2, \tau)^{1/2} = \pm\, G(g_1 g_2, \tau)^{1/2}$, again with a constant sign. The remaining task is to handle the exponential terms. Pulling out the factors of $\pi i q w^2$, we want to show that:

$$\begin{aligned}\frac{c(g_1)}{G(g_2, \tau)^2 G(g_1, g_2 * \tau)} + \frac{c(g_2)}{G(g_2, \tau)} &= \frac{c(g_1 g_2)}{G(g_1 g_2, \tau)} \\ \Leftrightarrow\ c(g_1) + c(g_2)G(g_1 g_2, \tau) &= c(g_1 g_2)G(g_2, \tau).\end{aligned}$$

To open this up, write:

$$g_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \qquad g_2 = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} \qquad g_1 g_2 = \begin{bmatrix} \cdot & \cdot \\ c_1 a_2 + d_1 c_2 & c_1 b_2 + d_1 d_2 \end{bmatrix}.$$

Then we have:

$$\begin{aligned} c(g_1) + c(g_2)G(g_1g_2,\tau) &= c_1 + c_2(c_1a_2\tau + d_1c_2\tau + c_1b_2 + d_1d_2) \\ &= c_2\tau(d_1c_2 + c_1a_2) + (c_1 + c_1c_2b_2 + d_1d_2c_2) \\ &= c_2\tau(d_1c_2 + c_1a_2) + (c_1(a_2d_2 - b_2c_2) + c_1c_2b_2 + d_1d_2c_2) \\ &= c_2\tau(d_1c_2 + c_1a_2) + (c_1a_2d_2 + d_1d_2c_2) \\ &= (c_1a_2 + d_1c_2)(c_2\tau + d_2) \\ &= c(g_1g_2)G(g_2,\tau), \end{aligned}$$

as desired. □

Proof of Lemma 12.19

Writing $w = z/q$, so that $F(g * (\tau,z)) = P_\beta(\Theta(g * (\tau,w)))$ and $F_g(\tau,z) = P_\beta(\rho(g)\,\Theta(\tau,w))$, we apply Lemma 12.16 and the degree-d homogeneity of P_β (a d-th power of a root of unity is again one):

$$\begin{aligned} P_\beta(\Theta(g*(\tau,w))) &\propto_\zeta P_\beta\left(G(g,\tau)^{1/2}\exp\left(\frac{\pi iqcw^2}{G(g,\tau)}\right)\rho(g)\Theta(\tau,w)\right) \\ &= G(g,\tau)^{d/2}\exp\left(\frac{\pi iqdcw^2}{G(g,\tau)}\right)P_\beta(\rho(g)\Theta(\tau,w)). \end{aligned}$$

Since $\pi iqdcw^2/G(g,\tau) = \pi idcz^2/(qG(g,\tau))$, this is the claim. □

Proof of Lemma 12.20

We have, as functions of (τ,z):

$$\begin{aligned} F_g(\tau,z+a\tau+b) &= P_\beta\left(\rho(g)\Theta\left(\tau,\frac{z+a\tau+b}{q}\right)\right) \\ &\propto_\zeta P_\beta\left(\rho(g)\,e^{-\pi ia^2\tau/q-2\pi iaw}\cdot \mathsf{Z}^b\mathsf{X}^a\Theta\left(\tau,\frac{z}{q}\right)\right) \quad \text{(Lemma 12.15, writing } w=z/q) \\ &= \exp\left(-\frac{\pi ia^2d\tau}{q} - \frac{2\pi idaz}{q}\right)P_\beta\left(\rho(g)\mathsf{Z}^b\mathsf{X}^a\Theta\left(\tau,\frac{z}{q}\right)\right) \\ &\propto_\zeta \exp\left(-\frac{\pi ia^2d\tau}{q} - \frac{2\pi idaz}{q}\right)P_\beta\left(\rho(g)\Theta\left(\tau,\frac{z}{q}\right)\right) \quad \text{(Proposition 12.9)} \\ &= \exp\left(-\frac{\pi ia^2d\tau}{q} - \frac{2\pi idaz}{q}\right)F_g(\tau,z), \end{aligned}$$

as desired. (The second and third steps use the degree-d homogeneity of P_β; a d-th power of a root of unity is again one.) □

Proof of Proposition 12.21

Since $F(\tau,z) = P_\beta(\Theta(\tau,z/q))$ and $z \mapsto z/q$ is a bijection of $\mathbb{C}$, it suffices to find a point where $P_\beta(\Theta(\tau,w)) \neq 0$. Set $(\tau,w) = (iy, iy/(2q))$ for a parameter $y > 0$. The exponent of the k-th term of Θ then becomes:

$$\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)w = -\frac{\pi y}{q}\left(\left(k-\frac{1}{2}\right)^2 + \left(k-\frac{1}{2}\right)\right) = -\frac{\pi y}{q}\left(k^2-\frac{1}{4}\right),$$

so that, normalizing by the $k = 0$ term:

$$\mathbf{v}_y := e^{-\pi y/(4q)}\,\Theta\left(iy, \frac{iy}{2q}\right) = \sum_{k\in\mathbb{Z}} (-1)^k e^{-\pi y k^2/q}\,\mathbf{e}_{k+(q-1)/2 \bmod q} = \mathbf{e}_{(q-1)/2} + E_y,$$

where the error term E_y collects the terms with $k \neq 0$. Each coordinate of E_y is bounded in absolute value by:

$$\sum_{k\neq 0} e^{-\pi y k^2/q} \leq 2\sum_{k=1}^{\infty} e^{-\pi y k/q} = \frac{2e^{-\pi y/q}}{1 - e^{-\pi y/q}} \xrightarrow{y\to\infty} 0,$$

so $\mathbf{v}_y \to \mathbf{e}_{(q-1)/2}$ as $y \to \infty$. Since P_β is a polynomial, hence continuous, $P_\beta(\mathbf{v}_y) \to P_\beta(\mathbf{e}_{(q-1)/2}) = 1$ (Proposition 12.9). So $P_\beta(\mathbf{v}_y) \neq 0$ for all large enough y, and then by the degree-d homogeneity of P_β (Proposition 12.9 again):

$$P_\beta\left(\Theta\left(iy, \frac{iy}{2q}\right)\right) = e^{\pi d y/(4q)}\,P_\beta(\mathbf{v}_y) \neq 0.$$

□

Proof of Lemma 12.24

Since $P_\beta(\mathbf{v}) = \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^d v_{x_i}$ and $\rho(g)$ is linear, the composition $P_\beta \circ \rho(g)$ is again a homogeneous polynomial of degree d: expanding $(\rho(g)\mathbf{v})_{x_i} = \sum_{r\in\mathbb{Z}_q} \rho(g)_{x_i,r}\, v_r$ and multiplying out,

$$P_\beta(\rho(g)\mathbf{v}) = \sum_{\mathbf{r}\in\mathbb{Z}_q^d} \beta_{\mathbf{r}}^{(g)} \prod_{i=1}^d v_{r_i}, \qquad \beta_{\mathbf{r}}^{(g)} := \sum_{\mathbf{x}\in\mathbb{Z}_q^d} \beta_{\mathbf{x}} \prod_{i=1}^d \rho(g)_{x_i,r_i}.$$

Now substitute $\mathbf{v} = \Theta(\tau, w)$. By definition, the r-th coordinate of $\Theta(\tau, w)$ is:

$$\Theta_r(\tau, w) = \sum_{\substack{k\in\mathbb{Z} \\ k+(q-1)/2\equiv r \pmod q}} (-1)^k \exp\left(\frac{\pi i \tau (2k-1)^2}{4q} + \pi i (2k-1) w\right),$$

using $\left(k - \frac{1}{2}\right)^2 = \frac{(2k-1)^2}{4}$ and $2\left(k - \frac{1}{2}\right) = 2k-1$. Each such series converges absolutely, locally uniformly on $\mathcal{H}\times\mathbb{C}$ (see the discussion following the definition of Θ), so the d-fold products in $P_\beta(\rho(g)\Theta(\tau, w))$ expand term by term (Tonelli's theorem for sums: on each compact set, the iterated sums of absolute values are finite, being bounded by products of the one-dimensional bounds). Writing, for a tuple $\mathbf{k} \in \mathbb{Z}^d$,

$$\mathbf{r}(\mathbf{k}) := \left(k_i + \tfrac{q-1}{2} \bmod q\right)_{i\in[d]}, \qquad \ell(\mathbf{k}) := \sum_{i=1}^d (2k_i - 1)^2, \qquad n(\mathbf{k}) := \sum_{i=1}^d (2k_i - 1),$$

we obtain the absolutely, locally uniformly convergent multiple series:

$$P_\beta(\rho(g)\Theta(\tau, w)) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \beta_{\mathbf{r}(\mathbf{k})}^{(g)}\,(-1)^{k_1+\cdots+k_d} \exp\left(\frac{\pi i \tau\, \ell(\mathbf{k})}{4q} + \pi i\, n(\mathbf{k})\, w\right). \tag{63}$$

We record three facts about these quantities:

- $\ell(\mathbf{k}) \geq d$, since each $(2k_i - 1)^2 \geq 1$;
- $\ell(\mathbf{k}) \equiv d \pmod 8$, since each odd square satisfies $(2k-1)^2 = 4k(k-1) + 1 \equiv 1 \pmod 8$ (as $k(k-1)$ is even);

- $|n(\mathbf{k})| \le \sqrt{d \cdot \ell(\mathbf{k})}$, by the Cauchy–Schwarz inequality applied to the vectors $(2k_i - 1)_i$ and $(1, \dots, 1)$.

In particular, for each fixed value ℓ of $\ell(\mathbf{k})$ there are finitely many contributing tuples (each $|2k_i - 1| \le \sqrt{\ell}$, so there are at most $(\sqrt{\ell} + 1)^d$ of them), so:

$$\Phi_{g,\ell}(w) := \sum_{\mathbf{k} \in \mathbb{Z}^d :\ \ell(\mathbf{k}) = \ell} \beta^{(g)}_{\mathbf{r}(\mathbf{k})} (-1)^{k_1 + \cdots + k_d}\ e^{\pi i\, n(\mathbf{k})\, w}$$

is a finite sum of exponentials, hence entire, and satisfies:

$$|\Phi_{g,\ell}(w)| \le B_g\, (\sqrt{\ell} + 1)^d\ e^{\pi \sqrt{d\ell} |\Im w|}, \qquad \text{where } B_g := \max_{\mathbf{r} \in \mathbb{Z}_q^d} \left|\beta^{(g)}_{\mathbf{r}}\right|. \tag{64}$$

Grouping the absolutely convergent series (63) by the value of $\ell(\mathbf{k})$ yields (48). For the asserted convergence, fix a compact subset of $\mathcal{H} \times \mathbb{C}$, on which $\Im \tau \ge y_0 > 0$ and $|\Im w| \le W$; then by (64):

$$\sum_{\ell} \left| \Phi_{g,\ell}(w)\, e^{\pi i \tau \ell/(4q)} \right| \le B_g \sum_{\ell \ge d} (\sqrt{\ell} + 1)^d\ e^{\pi \sqrt{d\ell}\, W}\ e^{-\pi \ell y_0/(4q)} < \infty,$$

since the exponent $-\pi \ell y_0/(4q)$ eventually dominates $\pi \sqrt{d\ell}\, W + d \log(\sqrt{\ell} + 1)$, which is $o(\ell)$.

For the second claim, fix $\tau \in \mathcal{H}$ and $j \ge 0$, and recall from Lemma B.6 (and its proof) that $A_{g,j}(\tau) = \frac{1}{2\pi i} \oint_{|z|=q} F_g(\tau, z)\, z^{-j-1}\, dz$. Substituting $F_g(\tau, z) = P_\beta(\rho(g)\Theta(\tau, z/q))$ and the expansion (48) — which converges uniformly on the circle $\{z : |z| = q\}$ for our fixed τ — we may integrate term by term:

$$A_{g,j}(\tau) = \sum_{\ell} e^{\pi i \tau \ell/(4q)} \cdot \frac{1}{2\pi i} \oint_{|z|=q} \Phi_{g,\ell}(z/q)\, z^{-j-1}\, dz = \sum_{\ell} e^{\pi i \tau \ell/(4q)}\, q^{-j}\, \Xi_{g;j,\ell},$$

where the last step substitutes $z = qw$ and applies Cauchy's coefficient formula to the entire function $\Phi_{g,\ell}$ on the circle $|w| = 1$. This proves (49); its absolute, locally uniform convergence in τ follows from the same majorant as above, since the same computation bounds $|q^{-j}\, \Xi_{g;j,\ell}| \le q^{-j} \max_{|w|=1} |\Phi_{g,\ell}(w)| \le B_g\, (\sqrt{\ell} + 1)^d\, e^{\pi \sqrt{d\ell}}$ by (64). □

Proof of Proposition 12.28

We have:

$$\begin{aligned}
\sum_{j=m}^{\infty} \frac{A_j(g * \tau)}{G(g,\tau)^j} z^j &= F\,(g * (\tau, z)) \\
&\propto_\zeta G(g,\tau)^{d/2} \exp\left(\frac{\pi i d c z^2}{q G(g,\tau)}\right) F_g(\tau, z) \quad \text{(Lemma 12.19)} \\
\Leftrightarrow F_g(\tau, z) &\propto_\zeta G(g,\tau)^{-d/2} \exp\left(-\frac{\pi i d c z^2}{q G(g,\tau)}\right) \left(\sum_{j=m}^{\infty} \frac{A_j(g * \tau)}{G(g,\tau)^j} z^j\right),
\end{aligned}$$

and now the conclusion follows after expanding the RHS as a power series in z (replacing the exponential term with its power series as well, and noting that it has constant term 1): the function F_g is holomorphic on $\mathcal{H} \times \mathbb{C}$ (like F, it is a composition of Θ with a polynomial and a rescaling), so Lemma B.6 applies to it, and the displayed identity (whose implicit constant does not depend on z, so applies to every Taylor coefficient) shows that its Taylor coefficients of index below m vanish while its coefficient of z^m is $\propto_\zeta G(g,\tau)^{-d/2-m} A_m(g * \tau)$. □

Proof of Lemma 12.29

We first show that A_m^2 is weakly modular of weight $d+2m$ for $\Gamma(q)$, with a one-dimensional representation of finite order. By Lemma 12.19 (and Proposition 12.18), for any $g \in \Gamma(q)$ we have:

$$\begin{aligned} F\left(g*\tau, \frac{z}{G(g,\tau)}\right) &\propto_\zeta G(g,\tau)^{d/2} \exp\left(\frac{\pi i d c z^2}{qG(g,\tau)}\right) F(\tau,z) \\ &= G(g,\tau)^{d/2}\left(\sum_{j=0}^{\infty} \frac{1}{j!}\left(\frac{\pi i d c}{qG(g,\tau)}\right)^j z^{2j}\right)\left(\sum_{j=m}^{\infty} A_j(\tau) z^j\right) \\ &= G(g,\tau)^{d/2} A_m(\tau) z^m + O(z^{m+1}). \end{aligned} \tag{65}$$

On the other hand, we have:

$$\begin{aligned} F\left(g*\tau, \frac{z}{G(g,\tau)}\right) &= \sum_{j=m}^{\infty} A_j(g*\tau)\left(\frac{z}{G(g,\tau)}\right)^j \\ &= \frac{A_m(g*\tau)}{G(g,\tau)^m} z^m + O(z^{m+1}). \end{aligned} \tag{66}$$

Equating coefficients of z^m in (65) and (66) (the constant implicit in $\propto_\zeta$ does not depend on z) implies:

$$\begin{aligned} &\frac{A_m(g*\tau)}{G(g,\tau)^m} \propto_\zeta G(g,\tau)^{d/2} A_m(\tau) \\ \Leftrightarrow\ &A_m(g*\tau) \propto_\zeta G(g,\tau)^{d/2+m} A_m(\tau). \end{aligned}$$

Squaring both sides gives $A_m^2(g*\tau) \propto_\zeta G(g,\tau)^{d+2m} A_m^2(\tau)$ for all $g \in \Gamma(q)$, now with an integer power of $G(g,\tau)$. Since A_m^2 is not identically zero (as $A_m \not\equiv 0$ by the definition of m), Lemma 10.21 shows that A_m^2 is weakly modular of weight $d+2m$ for $\Gamma(q)$, with a one-dimensional representation of finite order.

Next, we show that A_m^2 is holomorphic at every cusp $\mathfrak{c} \in \mathbb{Q}\cup\{\infty\}$, with $\operatorname{ord}_{\mathfrak{c}}(A_m^2) \geq \frac{d}{4} + 2\ell_{\mathfrak{c}}$. By Lemma 10.28, holomorphy at $\mathfrak{c}$ and $\operatorname{ord}_{\mathfrak{c}}(A_m^2)$ depend only on the class of $\mathfrak{c}$ and may be computed using the matrix $g_{\mathfrak{c}}$ fixed in Section 12.5.1; also, the cusp has width q by Proposition 12.23. Following Section 10.4.3, we therefore want to bound the exponents appearing in the expansion at ∞ of $\tau \mapsto G(g_{\mathfrak{c}},\tau)^{-d-2m}(A_m(g_{\mathfrak{c}}*\tau))^2$. By Proposition 12.28, we have $G(g_{\mathfrak{c}},\tau)^{-d-2m}(A_m(g_{\mathfrak{c}}*\tau))^2 \propto_\zeta (A_{g_{\mathfrak{c}},m}(\tau))^2$, and a nonzero constant factor does not affect the exponents in the expansion.

The function $A_{g_{\mathfrak{c}},m}$ is the m-th Taylor coefficient of $F_{g_{\mathfrak{c}}}$ (Proposition 12.28 and the uniqueness in Lemma B.6), so (49) and the minimality of $\ell^* = d + 8\ell_{\mathfrak{c}}$ in Definition 12.25 give the absolutely and locally uniformly convergent expansion:

$$A_{g_{\mathfrak{c}},m}(\tau) = \sum_{\substack{\ell \geq d+8\ell_{\mathfrak{c}} \\ \ell \equiv d \pmod 8}} q^{-m}\, \Xi_{g_{\mathfrak{c}};m,\ell}\, e^{\pi i \tau \ell/(4q)}.$$

Multiplying this series by itself (justified by absolute convergence), every term of $(A_{g_{\mathfrak{c}},m}(\tau))^2$ has the form $c\, e^{\pi i\tau\ell/(4q)} = c\, e^{2\pi i(\ell/8)\tau/q}$ for an integer $\ell \geq 2d + 16\ell_{\mathfrak{c}}$ with $\ell \equiv 2d \pmod 8$. Collecting terms by the value of ℓ (justified by absolute convergence, as in the proof of Lemma 10.30) and invoking the uniqueness in Fact B.4, the cusp expansion of A_m^2 at $\mathfrak{c}$ — taken in the local variable $e^{2\pi i\tau/q}$, as the width is q — has all its exponents at least $\frac{2d+16\ell_{\mathfrak{c}}}{8} = \frac{d}{4} + 2\ell_{\mathfrak{c}} \geq 0$. So A_m^2 is holomorphic at $\mathfrak{c}$, with $\operatorname{ord}_{\mathfrak{c}}(A_m^2) \geq \frac{d}{4} + 2\ell_{\mathfrak{c}}$. □

Proof of Lemma 12.31

It will be helpful for us to write $\begin{bmatrix} X \\ Y \end{bmatrix} := \begin{bmatrix} a & c \\ b & d_0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ i.e., $X = ax + cy$ and $Y = bx + d_0 y$. We have:

$$\begin{aligned}
B_{x,y}(g * \tau) &= \exp\left(\frac{\pi i d x^2 (g*\tau)}{4q}\right) F\left(g*\tau, \frac{x(g*\tau)+y}{2}\right) \\
&= \exp\left(\frac{\pi i d x^2 (g*\tau)}{4q}\right) F\left(g*\left(\tau, G(g,\tau)\frac{x(g*\tau)+y}{2}\right)\right) \\
&= \exp\left(\frac{\pi i d x^2 (g*\tau)}{4q}\right) F\left(g*\left(\tau, \frac{x(a\tau+b)+y(c\tau+d_0)}{2}\right)\right) \\
&\propto_\zeta \exp\left(\frac{\pi i d x^2 (g*\tau)}{4q}\right) G(g,\tau)^{d/2} \exp\left(\frac{\pi i d c}{q G(g,\tau)}\left(\frac{x(a\tau+b)+y(c\tau+d_0)}{2}\right)^2\right) \\
&\quad F_g\left(\tau, \frac{x(a\tau+b)+y(c\tau+d_0)}{2}\right) \quad \text{(Lemma 12.19)} \\
&= \exp\left(\frac{\pi i d x^2 (a\tau+b)}{4q(c\tau+d_0)}\right) G(g,\tau)^{d/2} \exp\left(\frac{\pi i d c}{4q(c\tau+d_0)}(X\tau+Y)^2\right) F_g\left(\tau, \frac{X\tau+Y}{2}\right). \qquad (67)
\end{aligned}$$

Now write $X = \tilde{x} + 2e_x$ and $Y = \tilde{y} + 2e_y$ for $e_x, e_y \in \mathbb{Z}$. Then Lemma 12.20 tells us that:

$$\begin{aligned}
F_g\left(\tau, \frac{X\tau+Y}{2}\right) &= F_g\left(\tau, \frac{\tilde{x}\tau+\tilde{y}}{2} + e_x\tau + e_y\right) \\
&\propto_\zeta \exp\left(-\frac{\pi i d}{q}(e_x^2\tau + e_x(\tilde{x}\tau+\tilde{y}))\right) F_g\left(\tau, \frac{\tilde{x}\tau+\tilde{y}}{2}\right) \\
&= \exp\left(-\frac{\pi i d}{q}(e_x^2\tau + e_x(\tilde{x}\tau+\tilde{y}))\right) \exp\left(-\frac{\pi i d \tilde{x}^2 \tau}{4q}\right) B_{g;\tilde{x},\tilde{y}}(\tau). \qquad (68)
\end{aligned}$$

Plugging this into (67) implies that $B_{x,y}(g * \tau) \propto_\zeta \Psi(\tau)\, B_{g;\tilde{x},\tilde{y}}(\tau)$, where:

$$\Psi(\tau) := G(g,\tau)^{d/2} \exp\left(\frac{\pi i d x^2 (a\tau+b)}{4q(c\tau+d_0)} + \frac{\pi i d c}{4q(c\tau+d_0)}(X\tau+Y)^2 - \frac{\pi i d}{q}(e_x^2\tau + e_x(\tilde{x}\tau+\tilde{y})) - \frac{\pi i d \tilde{x}^2\tau}{4q}\right).$$

Therefore, isolating the exponential term and pulling out a factor of $\pi i d/(4q)$, it suffices to show that the following is independent of τ and rational (the exponential is then $e^{\pi i r}$ for a rational r, a root of unity):

$$\frac{x^2(a\tau+b) + c(X\tau+Y)^2}{(c\tau+d_0)} - 4(e_x^2\tau + e_x(\tilde{x}\tau+\tilde{y})) - \tilde{x}^2\tau \qquad (69)$$

This is routine but rather tedious to check. We make use of an intermediate claim.

Claim E.5. *We have:*

$$\frac{x^2(a\tau+b) + c(X\tau+Y)^2}{(c\tau+d_0)} = X^2\tau + XY - xy.$$

Proof. We have $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} d_0 & -c \\ -b & a \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \Rightarrow x = d_0X - cY, y = -bX + aY$. Plugging these in, we want to show that:

$$\frac{(d_0X - cY)^2(a\tau+b) + c(X\tau+Y)^2}{(c\tau+d_0)} = X^2\tau + XY + (d_0X - cY)(bX - aY).$$

Indeed, we have:

$$\begin{aligned}
&(c\tau + d_0)(X^2\tau + XY + (d_0X - cY)(bX - aY)) \\
=&(c\tau + d_0)(X^2\tau + XY + b\,d_0X^2 - (ad_0 + bc)XY + acY^2) \\
=&\tau^2(cX^2) + \tau(d_0X^2 + cXY + b\,cd_0X^2 - (acd_0 + bc^2)XY + ac^2Y^2) + (d_0XY + b\,d_0^2X^2 - (ad_0^2 + b\,cd_0)XY + acd_0Y^2) \\
=&\tau^2(cX^2) + \tau(2cXY + ad_0^2X^2 - cXY\underbrace{(1 + ad_0 + bc)}_{=2ad_0} + ac^2Y^2) + (b\,d_0^2X^2 + c(1 + b\,c)Y^2 + d_0XY\underbrace{(1 - ad_0 - b\,c)}_{=-2b\,c}) \\
=&\tau^2(cX^2) + \tau(2cXY + a(d_0X - cY)^2) + (cY^2 + b(d_0X - cY)^2) \\
=&(d_0X - cY)^2(a\tau + b) + c(X\tau + Y)^2.
\end{aligned}$$

□

Plugging Claim E.5 into (69) yields an affine function in τ, so it suffices to focus on its linear coefficient, which is:

$$X^2 - 4e_x^2 - 4e_x\tilde{x} - \tilde{x}^2 = X^2 - (2e_x + \tilde{x})^2 = 0,$$

so (69) equals its constant term $XY - xy - 4e_x\tilde{y}$, an integer, and the conclusion follows.

Finally, the bijectivity claim follows from noting that $\begin{bmatrix}\tilde{x}\\ \tilde{y}\end{bmatrix} = \begin{bmatrix}a & c\\ b & d_0\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}$ mod 2, and that the matrix is by definition invertible. □

Proof of Lemma 12.34

By Lemma 10.28, holomorphy at $\mathfrak{c}$ and $\mathrm{ord}_{\mathfrak{c}}(B^2)$ depend only on the class of $\mathfrak{c}$ and may be computed using the matrix $g_{\mathfrak{c}}$ fixed in Section 12.5.1; also, the cusp has width q by Proposition 12.23. Following Section 10.4.3 (as in the proof of Lemma 12.29), we want to bound the exponents in the expansion at ∞ of $\tau \mapsto G(g_{\mathfrak{c}}, \tau)^{-3d}(B(g_{\mathfrak{c}} * \tau))^2$. Exactly as in the proof of Corollary 12.33 — but now for the single matrix $g_{\mathfrak{c}} \in \mathrm{SL}_2(\mathbb{Z})$, which need not lie in $\Gamma(q)$ — Lemma 12.31 and the bijectivity of $(x, y) \mapsto (\tilde{x}, \tilde{y})$ give:

$$G(g_{\mathfrak{c}}, \tau)^{-3d}\,(B(g_{\mathfrak{c}} * \tau))^2 \propto_{\zeta} (B_{g_{\mathfrak{c}}}(\tau))^2,$$

so it suffices to show that $(B_{g_{\mathfrak{c}}}(\tau))^2$ is a sum of terms $e^{\pi i\tau\ell/(4q)}$ with $\ell \geq 2d + 16\ell_{\mathfrak{c}} + 16H_{\mathfrak{c}}$. We bound the three factors of $B_{g_{\mathfrak{c}}}$ in turn. First, we have:

$$\begin{aligned}
B_{g_{\mathfrak{c}};0,1}(\tau) &= F_{g_{\mathfrak{c}}}\left(\tau, \frac{1}{2}\right) \\
&= P_\beta\left(\rho(g_{\mathfrak{c}})\Theta\left(\tau, \frac{1}{2q}\right)\right) = \sum_{\ell \geq d + 8\ell_{\mathfrak{c}}} \Phi_{g_{\mathfrak{c}},\ell}\left(\frac{1}{2q}\right) e^{\pi i\tau\ell/(4q)},
\end{aligned}$$

a sum of terms of the form $e^{\pi i\tau\ell/(4q)}$ for $\ell \geq d + 8\ell_{\mathfrak{c}}$, by Lemma 12.24 (at $w = \frac{1}{2q}$) and the minimality of $\ell^* = d + 8\ell_{\mathfrak{c}}$ in Definition 12.25. Next, we have:

$$\begin{aligned}
\exp\left(-\frac{\pi i d\tau}{4q}\right) B_{g_{\mathfrak{c}};1,0}(\tau) &= F_{g_{\mathfrak{c}}}\left(\tau, \frac{\tau}{2}\right) \\
&= P_\beta\left(\rho(g_{\mathfrak{c}})\Theta\left(\tau, \frac{\tau}{2q}\right)\right)
\end{aligned}$$

$$= P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{\tau}{2q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\right)$$

$$= P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}k^2 - \frac{\pi i\tau}{4q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\right)$$

$$\propto_\zeta P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}k^2 - \frac{\pi i\tau}{4q}\right)\mathbf{e}_{k \bmod q}\right) \quad \text{(Proposition 12.9)}$$

$$= \exp\left(-\frac{\pi i d\tau}{4q}\right) P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}k^2\right)\mathbf{e}_{k \bmod q}\right)$$

$$\Rightarrow B_{g_c;1,0}(\tau) \propto_\zeta P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}k^2\right)\mathbf{e}_{k \bmod q}\right).$$

(The step invoking Proposition 12.9 uses its fourth bullet — the X-invariance, up to a root of unity, of $P_\beta \circ \rho(g_c)$ — applied with $H = \mathsf{X}^{(q+1)/2}$.) Now let us consider the expansion of $P_\beta \circ \rho(g_c)$ on the RHS. The x-th coordinate of the vector in the argument is $\sum_{k\equiv x \pmod q}(-1)^k \exp\left(\frac{\pi i\tau}{4q}\cdot 4k^2\right)$, whose exponents satisfy $4k^2 \geq 4w_q(x)$, since $w_q(x) = \min_{k\equiv x \pmod q} k^2$ (Definition 5.4). Hence a monomial $X_{x_1}\dots X_{x_d}$ with nonzero coefficient in $P_\beta \circ \rho(g_c)$ expands out to terms of the form $\exp\left(\frac{\pi i\tau}{4q}\ell\right)$ for $\ell \geq 4\sum_{i\in[d]} w_q(x_i) \geq 4H_c$, where the last inequality is Definition 12.26; expanding the products term by term is justified by absolute convergence, as in the proof of Lemma 12.24. Finally, we have:

$$\exp\left(-\frac{\pi i d\tau}{4q}\right) B_{g_c;1,1}(\tau) = F_{g_c}\left(\tau, \frac{1+\tau}{2}\right)$$

$$= P_\beta\left(\rho(g_c)\Theta\left(\tau, \frac{1+\tau}{2q}\right)\right)$$

$$= P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}\left(k-\frac{1}{2}\right)^2 + 2\pi i\left(k-\frac{1}{2}\right)\frac{1+\tau}{2q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\right)$$

$$= P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}k^2 - \frac{\pi i\tau}{4q} + \frac{\pi ik}{q} - \frac{\pi i}{2q}\right)\mathbf{e}_{k+(q-1)/2 \bmod q}\right)$$

$$\propto_\zeta P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi i\tau}{q}k^2 - \frac{\pi i\tau}{4q} + \frac{\pi ik}{q} - \frac{\pi i}{2q}\right)\mathbf{e}_{k \bmod q}\right) \quad \text{(Proposition 12.9)}$$

$$\propto_\zeta \exp\left(-\frac{\pi i d\tau}{4q}\right) P_\beta \circ \rho(g_c)\left(\sum_{k\in\mathbb{Z}}(-1)^k \exp\left(\frac{\pi ik}{q}\right)\exp\left(\frac{\pi i\tau}{q}k^2\right)\mathbf{e}_{k \bmod q}\right),$$

where the last step pulls out, by degree-d homogeneity, both the τ-exponential and the constant $e^{-\pi i/(2q)}$ common to every coordinate (the latter contributes only a root of unity); and now an analogous argument (the extra k-dependent phase $e^{\pi ik/q}$ does not affect the τ-exponents) implies that this is also a sum of terms of the form $\exp\left(\frac{\pi i\tau}{4q}\ell\right)$ for $\ell \geq 4H_c$.

Multiplying the three factors together (Cauchy products of absolutely convergent series, so the minimal exponents add), we infer that $B_{g_c}(\tau)$ is a sum of terms of the form $\exp\left(\frac{\pi i\tau}{4q}\ell\right)$ for $\ell \geq (d+8\ell_c)+4H_c+4H_c = d+8\ell_c+8H_c$, and its square is a sum of terms of the same form with $\ell \geq 2d+16\ell_c+16H_c$. (Unlike in Lemma 12.29, the exponents of the individual factors here are not confined to a single residue class

modulo 8; this does not matter for the argument, which uses only the lower bound on the exponents and the uniqueness of the coefficients.) Collecting terms by the value of ℓ and invoking the uniqueness in Fact B.4 (as in the proof of Lemma 12.29), the cusp expansion of B^2 at $\mathfrak{c}$ — in the local variable $e^{2\pi i\tau/q}$, as the width is q — has all its exponents at least $\frac{2d+16\ell_{\mathfrak{c}}+16H_{\mathfrak{c}}}{8} = \frac{d}{4} + 2\ell_{\mathfrak{c}} + 2H_{\mathfrak{c}} \geq 0$. So B^2 is holomorphic at $\mathfrak{c}$, with $\mathrm{ord}_{\mathfrak{c}}(B^2) \geq \frac{d}{4} + 2\ell_{\mathfrak{c}} + 2H_{\mathfrak{c}}$. □

Proof of Lemma 12.36

Throughout, write $g := g_{\mathfrak{c}}$ and $\ell^* := d + 8\ell_{\mathfrak{c}}$, and adopt the convention $\Phi_{g,\ell} := 0$ for the indices $\ell \not\equiv d \pmod 8$ not covered by Lemma 12.24, so that sums over ℓ may range over all integers. By Lemma 12.24 (with $w = z/q$) and the minimality of ℓ^* in Definition 12.25:

$$F_g(\tau, z) = \sum_{\ell \geq \ell^*} \Phi_{g,\ell}(z/q)\, e^{\pi i\tau\ell/(4q)}, \qquad \Phi_{g,\ell}(z/q) = \sum_{\xi \in \mathbb{Z} - d/2} c_{\ell,\xi}\, e^{2\pi i\xi z/q}, \tag{70}$$

where the second display follows from the explicit formula for $\Phi_{g,\ell}$ derived in the proof of Lemma 12.24: each contributing tuple $\mathbf{k}$ has w-exponential $e^{\pi i n(\mathbf{k})w} = e^{2\pi i\xi(\mathbf{k})z/q}$ with $\xi(\mathbf{k}) := \frac{n(\mathbf{k})}{2} = \left(\sum_i k_i\right) - \frac{d}{2}$ — an element of the coset $\mathbb{Z} - \frac{d}{2}$ (half-integers when d is odd) — so that:

$$c_{\ell,\xi} := \sum_{\mathbf{k} \in \mathbb{Z}^d :\ \ell(\mathbf{k}) = \ell,\ \xi(\mathbf{k}) = \xi} \beta^{(g)}_{\mathbf{r}(\mathbf{k})}\, (-1)^{k_1 + \cdots + k_d}.$$

For each fixed ℓ the inner sum in (70) is finite: $c_{\ell,\xi} = 0$ unless $|\xi| \leq \frac{1}{2}\sqrt{d\ell}$, by the Cauchy–Schwarz bound $|n(\mathbf{k})| \leq \sqrt{d\,\ell(\mathbf{k})}$ from the proof of Lemma 12.24. We also adopt the convention $c_{\ell,\xi} := 0$ for $\ell < \ell^*$, consistent with (70).

Define $p_{\mathfrak{c}}(X) := \sum_\xi c_{\ell^*,\xi}\, X^{\xi + d/2}$ — a Laurent polynomial a priori, with integer exponents $\xi + \frac{d}{2} = \sum_i k_i$. By construction, the coefficient of $e^{\pi i\tau\ell^*/(4q)}$ in (70) is exactly $\Phi_{g,\ell^*}(z/q) = e^{-2\pi i z d/(2q)}\, p_{\mathfrak{c}}(e^{2\pi i z/q})$, so the identity (51) holds. It remains to prove that the exponents of $p_{\mathfrak{c}}$ lie in $[0, d]$ — i.e., that $c_{\ell^*,\xi} = 0$ whenever $|\xi| > d/2$ — so that $p_{\mathfrak{c}}$ is a genuine polynomial of degree at most d.

By Lemma 12.20 (with $(a, b) = (n, 0)$), for any $n \in \mathbb{Z}$ there is a nonzero constant ω_n, independent of τ and z, such that:

$$F_g(\tau, z + n\tau) = \omega_n \exp\left(-\frac{\pi i d}{q}(n^2\tau + 2nz)\right) F_g(\tau, z).$$

We expand both sides using (70); all rearrangements below are of absolutely convergent series, by the majorants from the proof of Lemma 12.24. On the left, $e^{2\pi i\xi(z+n\tau)/q} = e^{2\pi i\xi z/q}\, e^{\pi i\tau(8n\xi)/(4q)}$, so, reindexing by $\ell' := \ell + 8n\xi$:

$$F_g(\tau, z + n\tau) = \sum_{\ell,\xi} c_{\ell,\xi}\, e^{2\pi i\xi z/q}\, e^{\pi i\tau(\ell + 8n\xi)/(4q)} = \sum_{\ell',\xi} c_{\ell' - 8n\xi,\ \xi}\, e^{2\pi i\xi z/q}\, e^{\pi i\tau\ell'/(4q)}.$$

On the right, reindexing by $\ell' := \ell - 4n^2 d$ and $\xi' := \xi - nd$:

$$\omega_n e^{-\pi i d n^2\tau/q}\, e^{-2\pi i d n z/q} \sum_{\ell,\xi} c_{\ell,\xi}\, e^{2\pi i\xi z/q}\, e^{\pi i\tau\ell/(4q)} = \sum_{\ell',\xi'} \omega_n\, c_{\ell' + 4n^2 d,\ \xi' + nd}\, e^{2\pi i\xi' z/q}\, e^{\pi i\tau\ell'/(4q)}.$$

Now fix $\ell' \in \mathbb{Z}$ and compare the two sides. By the uniqueness in Fact B.4 (applied in τ with period $8q$, for each fixed z), the coefficients of $e^{\pi i\tau\ell'/(4q)}$ agree as functions of z. Both are *finite* sums of the characters $z \mapsto e^{2\pi i\xi z/q}$: on the left, $c_{\ell' - 8n\xi,\xi} \neq 0$ requires $\ell' - 8n\xi \geq \ell^*$ and $4\xi^2 \leq d(\ell' - 8n\xi)$, which has finitely many

solutions ξ; similarly on the right. Since these characters (for distinct ξ in the coset $\mathbb{Z} - \frac{d}{2}$) are linearly independent — after factoring out $e^{-\pi i d z/q}$, they become distinct integer powers of $e^{2\pi i z/q}$ — the coefficients match term by term:

$$c_{\ell - 8n\xi,\ \xi} = \omega_n \, c_{\ell + 4n^2 d,\ \xi + nd} \qquad \text{for all } n, \ell \in \mathbb{Z} \text{ and } \xi \in \mathbb{Z} - \tfrac{d}{2}. \tag{71}$$

Now suppose for the sake of contradiction that $c_{\ell^*, \xi^*} \neq 0$ for some ξ^* with $|\xi^*| > d/2$. We have two cases:

- If $\xi^* > d/2$: apply (71) with $n = 1$, $\ell = -3d + 8\ell_{\mathfrak{c}}$, and $\xi = \xi^* - d$. The right-hand side is $\omega_1 c_{\ell + 4d,\ \xi + d} = \omega_1 c_{d + 8\ell_{\mathfrak{c}},\ \xi^*} = \omega_1 c_{\ell^*, \xi^*} \neq 0$, so the left-hand side $c_{\ell - 8\xi,\ \xi}$ is nonzero. But $\ell - 8\xi = \ell - 8(\xi^* - d) = 5d + 8\ell_{\mathfrak{c}} - 8\xi^* < 5d + 8\ell_{\mathfrak{c}} - 4d = \ell^*$, contradicting the convention $c_{\ell, \xi} = 0$ for $\ell < \ell^*$ (the minimality in Definition 12.25).

- If $\xi^* < -d/2$: apply (71) with $n = -1$, $\ell = -3d + 8\ell_{\mathfrak{c}}$, and $\xi = \xi^* + d$. The right-hand side is $\omega_{-1} c_{\ell + 4d,\ \xi - d} = \omega_{-1} c_{\ell^*, \xi^*} \neq 0$, so $c_{\ell + 8\xi,\ \xi} \neq 0$. But $\ell + 8\xi = -3d + 8\ell_{\mathfrak{c}} + 8\xi^* + 8d = 5d + 8\ell_{\mathfrak{c}} + 8\xi^* < \ell^*$, again a contradiction.

So the exponents $\xi + d/2$ of $p_{\mathfrak{c}}$ all lie in $[0, d]$, completing the proof. □