hexagon:2611.00893v1

# Simultaneous quadratic probes: reconstruction thresholds and marked Poisson limits

Brice Pouly

With mathematical and editorial assistance from ChatGPT

**Abstract**

We observe a uniformly chosen integer in $[X, 2X)$ through the four Legendre symbols of $n, n+1, 3n+1, 3n+2$ at each of $k$ randomly chosen distinct primes congruent to 3 modulo 8. Our main arithmetic estimate compares simultaneous collision probabilities with those of independent uniform words, in absolute mean at the critical scale. It includes tuples with deficient square-class rank and permits arbitrary deterministic binary labels fixed before the primes are sampled. We deduce a threshold at $k = \log_{16} X$: below it by a diverging additive amount, optimal identification tends to zero uniformly over panels. In probability over the panel, the observation becomes independent of the label below the threshold, and the proportion of origins with a unique observation tends to one above it by a diverging additive amount. If $X/16^k \to \rho$ and the positive-label proportion tends to $\alpha$, the empirical marked occupancy law converges in probability to independent Poisson laws of means $\rho\alpha$ and $\rho(1-\alpha)$. Classical occupancy calculations then give explicit identification and prediction limits. For primes in $(Y/2, Y]$ we treat $\log\log X = o(\log Y)$, $\log Y = o(\log X)$ by the quadratic large sieve, and $Y = (\log X)^A$, for each fixed $A > 2$, by a prime-character estimate under the generalized Riemann hypothesis for Dirichlet $L$-functions. The independently checked zero theorem announced by Alpöge supplies this last hypothesis. The proof combines square-class separation, the quadratic large sieve and averaged exponentiation of prime-character estimates. For the quartic Liouville label, an existing bound on ordinary averages and fixed-ratio isotopy constrain its mark density outside a set of scales of logarithmic density zero.

## 1 What four quadratic observations determine

Fix a large integer $X$. An unknown integer $n$ lies in

$$I_X = [X, 2X) \cap \mathbb{N}, \qquad |I_X| = X.$$

At a prime $p$, we observe four quadratic residues or non-residues:

$$L_1(n) = n, \quad L_2(n) = n+1, \quad L_3(n) = 3n+1, \quad L_4(n) = 3n+2. \tag{1.1}$$

If none of these values is divisible by $p$, the observation is a word in $\Gamma = \{-1, +1\}^4$, an alphabet with sixteen letters. A natural first guess is that $k$ such observations divide the $X$ candidate integers among $16^k$ possible words. This predicts a threshold near $\log_{16} X$ observations.

The letters are not independent random bits. They are quadratic characters of four values attached to the same integer, and different integers can satisfy square-class relations. The question is whether these arithmetic dependences alter the predicted threshold or the occupancy law at that threshold. We prove that they do not, for the prime ranges specified below. The proof must control rare tuples with unusually large collision probabilities, as well as typical tuples.

We also attach a deterministic binary label $f_X(n)$ to each origin. The label is fixed before the observing primes are sampled; it may depend arbitrarily on $n$ and on $X$. One example is

$$C(n) = \prod_{i=1}^{4} \lambda(L_i(n)), \qquad \lambda(m) = (-1)^{\Omega(m)}. \tag{1.2}$$

Here $\Omega(m)$ counts prime factors with multiplicity. Identifying an origin determines its label. Predicting a label can be easier: several origins with the same observed word may all have the same label. The critical law distinguishes these tasks.

### 1.1 The experiment and the two prime ranges

Let

$$\mathcal{P}_Y = \{p : Y/2 < p \le Y,\ p \equiv 3 \pmod 8\}, \qquad M = |\mathcal{P}_Y|. \tag{1.3}$$

Choose an ordered panel $\Pi = (p_1, \dots, p_k)$ uniformly without replacement from this set, independently of the unknown origin. The panel itself is known to the observer. Its order is only a convenient way of recording the observations. At a prime dividing at least one of the four forms, use a single erasure letter $*$ for the entire row. Thus

$$B_\Pi(n)_j = \begin{cases} \left(\left(\frac{L_1(n)}{p_j}\right), \dots, \left(\frac{L_4(n)}{p_j}\right)\right), & p_j \nmid \prod_i L_i(n), \\ *, & p_j \mid \prod_i L_i(n). \end{cases} \tag{1.4}$$

All decoders may use the full dictionary $\{(n, B_\Pi(n), f_X(n)) : n \in I_X\}$. No running-time restriction is imposed.

Throughout the paper one of the following assumptions holds:

(U) *Unconditional range:* $\log\log X = o(\log Y)$ and $\log Y = o(\log X)$.

(G) *Polylogarithmic range under GRH:* $Y = (\log X)^A$ for a fixed $A > 2$, and the generalized Riemann hypothesis holds for Dirichlet $L$-functions.

The two hypotheses identify the arithmetic input to each proof. For (G), the only zero statement used is that zeros in the open critical strip of Dirichlet $L$-functions lie on its critical line. The Lean development announced on 20 November 2026 by Alpöge [1] includes this statement for every modulus and every Dirichlet character, as well as its primitive-character specialization. Sutherland [29, statements 3–5] and Rubinstein [28] document independent checks of the formal development. Invoking this external theorem gives the conclusions below throughout the stated polylogarithmic range. We keep its use explicit in Lemma 4.1. For example, $Y = \exp((\log\log X)^2)$ satisfies (U). The prime number theorem in the fixed residue class gives $M \sim Y/(8\log Y)$ in either range. All logarithms without a subscript are natural. On a discrete space we use $d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2}\sum_x |\mu\{x\} - \nu\{x\}|$. Write

$$k_*(X) = \log_{16} X = \frac{\log_2 X}{4}, \qquad \alpha_X = X^{-1}|\{n \in I_X : f_X(n) = +1\}|. \tag{1.5}$$

Convergence in probability below always refers to the choice of the panel. The origin is uniform on $I_X$ when a conditional observation law or a success probability is defined.

### 1.2 Threshold and critical law

**Theorem 1.1** (Reconstruction threshold)**.** *Assume* (U) *or* (G)*, and let* $0 \le k \le M$ *be integers. For every deterministic sequence of labels* $f_X : I_X \to \{-1, +1\}$ *chosen before the panel, the following statements hold.*

(i) *If* $k - k_*(X) \to -\infty$, *then*

$$d_{\mathrm{TV}}(\mathcal{L}_\Pi(f_X(n), B_\Pi(n)), \mathcal{L}(f_X(n)) \otimes \mathcal{L}_\Pi(B_\Pi(n))) \xrightarrow{\mathbb{P}} 0. \tag{1.6}$$

*Consequently the optimal label-prediction success probability is* $\max(\alpha_X, 1-\alpha_X) + o_{\mathbb{P}}(1)$. *In addition, both the optimal identification success probability and the proportion of origins with a unique word tend to zero uniformly over all admissible panels.*

(ii) *If* $k - k_*(X) \to +\infty$, *the proportion of origins whose observed word has exactly one preimage tends to* 1 *in probability. In particular, the optimal identification and label-prediction success probabilities tend to* 1.

*Here* $\mathcal{L}_\Pi$ *means the law for a uniform origin conditional on the known panel.*

In particular the two conclusions hold at any fixed relative distance below or above $k_*$. The statement is for every label fixed in advance; it does not allow the label to be selected after inspecting the panel. The reconstruction statement concerns a uniform origin: above the threshold, the proportion of origins sharing their word with another candidate tends to zero in probability. Global injectivity would require distinct words for every pair of different origins in the catalogue, a stronger conclusion that is not asserted here.

For the Liouville label, the subcritical success probability is

$$\frac{1 + |\overline{C}_X|}{2} + o_{\mathbb{P}}(1), \qquad \overline{C}_X = \frac{1}{X} \sum_{n \in I_X} C(n). \tag{1.7}$$

No cancellation of this dyadic mean is assumed. Section 9.1 bounds it on a set of scales of logarithmic density one.

For a word $w \in \Gamma^k$, let

$$N_{\Pi,+}(w) = |\{n \in I_X : B_\Pi(n) = w,\ f_X(n) = +1\}|,$$
$$N_{\Pi,-}(w) = |\{n \in I_X : B_\Pi(n) = w,\ f_X(n) = -1\}|.$$

Words with an erasure are not in $\Gamma^k$. The probability measure

$$\nu_{\Pi,X} = \frac{1}{16^k} \sum_{w \in \Gamma^k} \delta_{(N_{\Pi,+}(w), N_{\Pi,-}(w))} \tag{1.8}$$

describes the occupancies of a word drawn uniformly from all non-erased words, including the unoccupied words.

**Theorem 1.2** (Marked critical law)**.** *Assume* (U) *or* (G). *Along any sequence on which*

$$\rho_X = \frac{X}{16^k} \longrightarrow \rho \in (0, \infty), \qquad \alpha_X \longrightarrow \alpha \in [0, 1], \tag{1.9}$$

*the empirical measure* (1.8) *converges in probability to the law of two independent random variables*

$$Z_+ \sim \mathrm{Pois}(\rho\alpha), \qquad Z_- \sim \mathrm{Pois}(\rho(1-\alpha)).$$

*The convergence holds in total variation on* $\mathbb{N}_0^2$. *For a uniform origin, let* $\mathrm{Uniq}_{\Pi,X}$ *be the probability that its word is unique,* $\mathrm{Id}_{\Pi,X}$ *the optimal probability of identifying the origin, and* $\mathrm{Pred}_{\Pi,X}$ *the optimal probability of predicting its label. Then*

$$\mathrm{Uniq}_{\Pi,X} \xrightarrow{\mathbb{P}} e^{-\rho}, \tag{1.10}$$

$$\mathrm{Id}_{\Pi,X} \xrightarrow{\mathbb{P}} \frac{1 - e^{-\rho}}{\rho}, \tag{1.11}$$

$$\mathrm{Pred}_{\Pi,X} \xrightarrow{\mathbb{P}} \frac{1}{2} + \frac{1}{2\rho} \mathbb{E}|Z_+ - Z_-|. \tag{1.12}$$

The choice of a uniform word in (1.8) and that of a uniform origin in (1.10)–(1.12) give different laws. An origin visits a word in proportion to its occupancy. This is why unique reconstruction and optimal identification have different limits. At $\rho = 1$ they are approximately 0.367879 and 0.632121, respectively. If the label is balanced, the optimal prediction limit is approximately 0.836835. We derive these numbers in Section 8. The integer-valued number of probes makes subsequences natural in (1.9).

### 1.3 Relation to earlier work

Prescribed quadratic-character signatures at several small primes have long been studied. Lehmer, Lehmer and Shanks [18] investigated integers with prescribed quadratic characters; pseudosquares provide a particularly well-studied signature. Pomerance and Shparlinski [20] obtained distribution results for pseudosquares and pseudopowers. Hanson, Vaughan and Zhang [11] studied least integers realizing prescribed Legendre-symbol vectors and their relation to binary quadratic forms. These works concern simultaneous observations at several primes, but do not study the empirical marked occupancy of a fixed dyadic catalogue under a randomly sampled panel.

A second relevant line obtains Poisson statistics by combining local conditions through the Chinese remainder theorem. Kurlberg and Rudnick [17] treated spacings between quadratic residues; Granville and Kurlberg [10] developed general criteria for Poisson spacing statistics; and Kurlberg [16] studied value sets of polynomials, including exceptional configurations in local correlations. These are antecedents for both the Poisson mechanism and the need to control failures of generic independence. The statistic and scale here require an additional comparison. At criticality, the panel modulus $Q_\Pi = \prod_{j=1}^{k} p_j$ satisfies

$$\frac{\log Q_\Pi}{\log X} \sim \frac{\log Y}{\log 16} \longrightarrow \infty,$$

so our fixed catalogue has length $Q_\Pi^{o(1)}$. Averaging over a complete period modulo $Q_\Pi$ therefore does not describe this catalogue directly.

Nor does a uniform local factorization hold for every nonzero shift. For example, let $\Omega_p^+$ consist of residues at which all four symbols equal $+1$. Character expansion and Weil's bound give $|\Omega_p^+|/p = 2^{-4} + O(p^{-1/2})$. The conditions for $n$ and $n+1$ involve seven distinct affine factors, since $L_2(n) = L_1(n+1)$. Hence

$$\frac{1}{p}\#\{n \bmod p : n, n+1 \in \Omega_p^+\} = 2^{-7} + O(p^{-1/2}),$$

whereas the product of the two densities has main term $2^{-8}$. Thus the uniform local-correlation hypothesis in [10, Theorem 1] cannot be invoked directly for these four-form sets. Our collision argument treats the relevant arithmetic exceptions with their increased collision weights.

The limiting occupancy calculation is classical; see Barbour, Holst and Janson [2, Chapter 6, Section 6.2], and Boucheron, Gamboa and Léonard [6, Proposition 2.3] for an empirical occupancy law. The language of random maps on fixed inputs also connects the problem to universal hashing [7]. Pairwise universality alone does not imply a critical Poisson occupancy law: over $\mathbb{F}_q$, the maps $x \mapsto ax + b$ with independent uniform $a, b$ have independent uniform outputs at any two distinct inputs, yet with probability $1 - 1/q$ they permute the full catalogue $\mathbb{F}_q$. As $q \to \infty$, its empirical occupancy law converges in probability to $\delta_1$, rather than Pois(1). Higher simultaneous collisions are essential here. The subcritical independence from a fixed label is also related to privacy amplification [5] and the hashing lemma of Håstad, Impagliazzo, Levin and Luby [12, Lemma 4.8]. Those general principles do not themselves establish the required collision bounds for these Legendre panels.

The uniform square-class fibre bounds used below have direct antecedents in de la Bretèche, Wang and Xu [8, Lemmas 3.2–3.3], with the earlier inputs of Vaughan–Wooley and Evertse–Silverman specified alongside Lemma 3.2. We retain a proof adapted to the four forms. Pylaiev's recent preprint [26, Theorem 1.5] gives fixed-order estimates for square products with squarefree polynomial values. Its squarefree-value condition does not apply directly to our raw product $D(n) = n(n+1)(3n+1)(3n+2)$, which is always divisible by 4.

Other Poisson laws involving quadratic non-residues have a different source of randomness. Basak, Nath and Zaharescu [4, Theorem 1] study chains of small non-residues modulo a single large prime; Basak, Berndt and Zaharescu [3] study related distributions and partitions. The random multiplicative models in [21, 25] instead randomize signs at primes and count rare patterns. Here the candidate integers and their labels are fixed, and the panel is random. In particular, the marked limit for (1.2) preserves the actual label density and does not establish $\alpha = 1/2$ or a new cancellation estimate for $C$. In the prime-factor observation experiment of Pouly [23], the origin has harmonic weight and supplies the sampled prime factors. Here the origin has uniform weight and the panel is drawn independently of it. These different sampling laws lead to different decision limits.

The arithmetic contribution is the absolute averaged comparison of simultaneous collisions in Proposition 2.2. It is at the critical weighted scale, covers every fixed partition of origins into collision groups, and permits arbitrary prescribed labels. The treatment of deficient square-class rank in Proposition 5.1 and the averaged exponentiation in Lemma 6.1 supply the two main steps beyond the generic one-prime estimate. Neither the alphabet count $16^k$ nor the Poisson and decision formulas are claimed as new probabilistic mechanisms. The analytic inputs are Heath-Brown's quadratic large sieve [13, Corollary 2], the GRH prime-character estimate in Montgomery–Vaughan [19, Theorem 13.7], and Weil's finite-field character bound [31].

Section 2 isolates the collision estimate; Sections 3–6 prove it. Section 7 passes to the empirical marked law, and Section 8 derives the decisions and the full additive threshold.

## 2 The collision estimate behind the limit

It is convenient to work temporarily with binary observations even at primes dividing a form. Replace each zero Legendre symbol by $+1$, separately in the four columns, and write $\widetilde{B}_\Pi$ for the resulting word in $\Gamma^k$. This is an auxiliary convention. We first bound its effect and the effect of sampling with replacement.

**Lemma 2.1** (Two reductions)**.** *For $k = O(\log X)$, uniformly in every distinct panel, the number of origins at which $B_\Pi$ and $\widetilde{B}_\Pi$ differ is*

$$R_\Pi \ll k(X/Y + 1) = o(X). \tag{2.1}$$

*An ordered independent sample of $k$ primes from $\mathcal{P}_Y$ has a repetition with probability $O(k^2/M) = o(1)$.*

*Proof.* For $p > 3$, each affine form has one zero modulo $p$. A union bound over the four forms and the panel gives $R_\Pi \ll \sum_{j\le k}(X/p_j + 1)$. Both prime ranges imply $k/Y = o(1)$ and $k/X = o(1)$. The repetition bound is the union bound over pairs of sampled positions; $M \asymp Y/\log Y$ and $(\log X)^2/M = o(1)$. Conditional on having no repetition, an independent ordered sample has exactly the required panel law. □

Until Section 7, primes are sampled with replacement and the auxiliary binary convention is used. Fix $v \ge 1$ and a partition $\mathcal{G}$ of $\{1, \dots, v\}$ into $b \ge 1$ nonempty groups. For an ordered tuple $\boldsymbol{n} = (n_1, \dots, n_v)$, let $q_{\mathcal{G}}(\boldsymbol{n})$ be the probability that a single random prime makes the

four-bit observations agree within each group. No equality or inequality is imposed between different groups. Independence of the sampled primes makes the corresponding panel collision probability $q_{\mathcal{G}}(\boldsymbol{n})^k$.

If all four bits at all origins were independent and fair, this single-prime probability would be

$$q_0 = 16^{-(v-b)}. \tag{2.2}$$

The following averaged comparison is the central estimate. It is stronger than an unweighted assertion that most tuples have no square-class relation.

**Proposition 2.2** (Averaged simultaneous collisions)**.** *Under* (U) *or* (G)*, for every fixed partition $\mathcal{G}$ as above and every sequence $k = \log_{16} X + O(1)$,*

$$\sum_{\substack{n_1,\dots,n_v \in I_X \\ \text{pairwise distinct}}} \left| q_{\mathcal{G}}(\boldsymbol{n})^k - 16^{-k(v-b)} \right| = o(X^b). \tag{2.3}$$

*All constants and errors may depend on the fixed partition, a fixed bound for $|k - \log_{16} X|$, and, in* (G)*, on $A$. In* (U) *the error is understood along the chosen admissible function $Y(X)$. No error depends on the labels.*

For two origins in one group, this compares pair collisions to $16^{-k}$. Higher occupancies require the same comparison for more origins. Concentration of the empirical occupancy law requires two groups. Keeping the absolute value in (2.3) permits arbitrary restrictions on the labels afterwards.

The scale of the right side matters. There are about $X^v$ tuples, and at the critical scale a typical one has collision probability of order $X^{-(v-b)}$. Their total contribution is therefore of order $X^b$. An exceptional family of size $o(X^v)$ is not automatically negligible at this scale: its members may collide much more often. We control this weighted exceptional contribution in Section 5.

## 3 Square classes on a dyadic interval

For a positive integer $m$, write $\mathrm{sf}(m)$ for the unique squarefree positive integer such that $m/\mathrm{sf}(m)$ is a square. Equivalently, $\mathrm{sf}(m)$ represents the class of $m$ in $\mathbb{Q}_{>0}^{\times}/(\mathbb{Q}_{>0}^{\times})^2$. We need both an injectivity fact and a bound for less restrictive fibres.

**Lemma 3.1** (Separation of two kernels)**.** *Each of the maps*

$$n \longmapsto \mathrm{sf}(n(n+1)), \qquad n \longmapsto \mathrm{sf}(D(n)), \qquad D(n) = \prod_{i=1}^{4} L_i(n),$$

*is injective on $I_X$.*

*Proof.* Suppose $n(n+1) = da_n^2$. The class $d = 1$ is impossible for $n \geq 1$, since $n^2 < n(n+1) < (n+1)^2$. The number

$$\beta_n = 2n + 1 + 2a_n\sqrt{d}$$

is a norm-one unit in $\mathbb{Z}[\sqrt{d}]$. If $m > n$ has the same class, $\beta_m/\beta_n = u + v\sqrt{d} > 1$ is another integral norm-one unit. Its rational part $u$ is odd and strictly greater than 1, so $\beta_m/\beta_n \geq 3 + 2\sqrt{2}$. On the other hand,

$$\frac{\beta_n}{n} = \left(1 + \sqrt{1 + \frac{1}{n}}\right)^2$$

decreases with $n$. Thus $\beta_m/\beta_n < m/n < 2$ for two origins in the same dyadic interval, a contradiction.

For the second map, put $x_n = 9n(n+1)+1$. A direct identity gives

$$x_n^2 - 9D(n) = 1. \tag{3.1}$$

If $D(n) = da_n^2$, use the units $x_n + 3a_n\sqrt{d}$. The case $d = 1$ would give $(x_n - 3a_n)(x_n + 3a_n) = 1$, impossible for $n \geq 1$. For two origins $m > n$ with the same class, the rational part of the quotient of their units is an integer $u \equiv 1 \pmod 9$, and $u > 1$. Hence $u \geq 10$. The irrational part is positive, so multiplying back gives $x_m > 10x_n$. But $m < 2n$ implies $x_m < 4x_n$. □

The first assertion already implies that prescribing all four individual square classes specifies at most one origin in $I_X$. For the second map, Pouly [24, Lemma 3.1] proves a stronger separation: if $m > n$ and $\mathrm{sf}(D(m)) = \mathrm{sf}(D(n))$, then $m > 6n$. Its proof uses the quotient of the associated units modulo 18. The quartic assertion also explains why the squarefree kernel of the full product is itself a rich observable; it is not the same observation as a fixed number of its quadratic signs.

**Lemma 3.2** (Uniform kernel fibres)**.** *For every nonempty $S \subseteq \{1,2,3,4\}$, set $F_S(n) = \prod_{i\in S} L_i(n)$. Uniformly over squarefree $d$,*

$$|\{n \in I_X : \mathrm{sf}(F_S(n)) = d\}| \begin{cases} \ll X^{1/2}, & |S| = 1, \\ \leq X^{o(1)}, & |S| \geq 2. \end{cases} \tag{3.2}$$

*In the second line, for every $\varepsilon > 0$ there is a constant depending on $\varepsilon$ and the fixed forms, but not on $d$, which bounds the fibre by that constant times $X^{\varepsilon}$.*

For $|S| \geq 2$, these estimates are special cases of known uniform bounds for square classes of polynomial values. In the notation of de la Bretèche–Wang–Xu [8, Lemmas 3.2–3.3], take $P = F_S$, whose roots are distinct. For $|S| = 2$, their Lemma 3.2 with $N = 2X$ gives $O_{F_S,\varepsilon}(X^{\varepsilon})$ uniformly in $d$. For $|S| \geq 3$, their Lemma 3.3 bounds the number of integral points on $F_S(x) = dy^2$ by $2^{O_{F_S}(1+\omega(d))}$. A nonempty fibre has $d \ll X^4$, and then $\omega(d) \ll \log X/\log\log X$, so this is $X^{o(1)}$, uniformly in $d$. The quadratic estimate is obtained there from Vaughan–Wooley [30, Lemma 3.5]; the higher-degree estimate follows from Evertse–Silverman [9, Theorem 1(b)]. We retain a proof adapted to the four affine forms, making explicit the fixed norm equations and the uniformity in the prescribed classes.

*Proof.* The first bound follows by writing $L_i(n) = du^2$. For the second, any prime greater than 3 divides at most one of the four values at a fixed origin: the determinants of two of the affine forms are nonzero and have no prime factor greater than 3. A prescribed class $d$ therefore allows at most

$$2^8\, 4^{\omega(d)} = X^{o(1)}$$

assignments of the individual classes $d_i = \mathrm{sf}(L_i(n))$, $i \in S$. The factor $2^8$ bounds the parity choices at 2 and 3. For a nonempty fibre, $d \ll X^4$, which makes this estimate uniform.

Fix one such assignment and two distinct indices $i, j \in S$. Write $L_i = a_i n + b_i$, $L_j = a_j n + b_j$, and $L_i = d_i u^2$, $L_j = d_j v^2$. Then

$$(a_i L_j + a_j L_i)^2 - 4a_i a_j d_i d_j (uv)^2 = (a_i b_j - a_j b_i)^2 =: c_{ij} > 0. \tag{3.3}$$

Here $c_{ij} \in \{1, 4, 9\}$ is fixed. If $a_i a_j d_i d_j$ is a square, integer factorization of (3.3) gives $O(1)$ possibilities. Otherwise consider the algebraic integer

$$z = a_i L_j + a_j L_i + 2uv\sqrt{a_i a_j d_i d_j}$$

in the associated real quadratic field. It has norm $c_{ij}$, so its principal ideal divides $(c_{ij})$. The number of such ideal divisors is bounded independently of the field: each of the finitely many rational primes dividing $c_{ij}$ has at most two prime ideals above it, with bounded exponents. For a fixed principal ideal its generators differ by units. The two real embeddings relevant here lie between $c/X$ and $CX$. Every real quadratic unit whose absolute value exceeds 1 has that absolute value at least $(1+\sqrt{5})/2$: its trace is an integer, and its norm is $+1$ or $-1$. There are consequently $O(\log(2X))$ possible generators. The rational part of $z$ is an affine function of $n$ with nonzero slope, so each generator specifies at most one origin. Multiplying by the number of class assignments proves the claim. □

# 4 Prime-character estimates in the two ranges

Put $H = \log X$ and, for squarefree $d \geq 1$, define

$$b_Y(d) = \frac{1}{M} \sum_{p \in \mathcal{P}_Y} \left(\frac{d}{p}\right). \tag{4.1}$$

The classes 1 and 2 must be retained separately: $b_Y(1) = 1$ and $b_Y(2) = -1$. In (G), all other classes of polynomial size have small averages. In (U), the large sieve leaves a set of exceptional classes of subpolynomial cardinality.

**Lemma 4.1** (Uniform estimate under GRH)**.** *In range* (G)*, for every fixed $C > 0$, uniformly for squarefree $d \leq X^C$ with $d \notin \{1, 2\}$,*

$$|b_Y(d)| \ll_C \frac{H \log Y}{\sqrt{Y}}. \tag{4.2}$$

*In particular one may fix $s \in (0, 1)$, depending on $A$, such that the right side is $O_C(H^{-s})$.*

*Proof.* The symbol $(d/p)$ is the value at $p$ of the real primitive character with fundamental discriminant $d$ or $4d$. Expand the condition $p \equiv 3 \pmod 8$ in the four characters modulo 8. None of the resulting products is principal unless $d = 1$ or 2. Their moduli are $O(X^C)$. Under Dirichlet GRH, [19, Theorem 13.7, equation (13.21)] gives, for a nonprincipal character modulo $q$,

$$\sum_{p \leq y} \chi(p) \ll \sqrt{y}\, \log(qy).$$

The statement includes imprimitive characters. Apply it at $Y$ and $Y/2$ to each product character and divide by $M \asymp Y/\log Y$. Since $\log Y = o(H)$, this proves (4.2). Any fixed $0 < s < \min(1, A/2 - 1)$ is admissible after increasing $X$. □

**Lemma 4.2** (Few exceptional classes without GRH)**.** *In range* (U)*, for every fixed $C > 0$, the set*

$$\mathcal{E}_C = \{d \leq X^C : d \text{ squarefree},\ |b_Y(d)| > H^{-2}\} \cup \{1, 2\} \tag{4.3}$$

*has cardinality $X^{o(1)}$.*

*Proof.* We use the version of the quadratic large sieve that permits non-squarefree coefficients. If $\mathcal{S}(Q)$ is the set of real primitive characters of conductor at most $Q$, then Heath-Brown [13, Corollary 2, p. 238] proves

$$\sum_{\chi \in \mathcal{S}(Q)} \left| \sum_{m \leq N} a_m \chi(m) \right|^2 \ll_\varepsilon (QN)^\varepsilon (Q + N) \sum_{\substack{m_1, m_2 \leq N \\ m_1 m_2 \text{ a square}}} |a_{m_1} a_{m_2}|. \tag{4.4}$$

The form of the last sum is essential because prime factors will repeat.

Set $t = \lceil CH/\log Y \rceil$, and expand $(\sum_{p\in\mathcal{P}_Y} \chi(p))^t$ as a Dirichlet polynomial of length $N = \lfloor Y^t \rfloor$. Its coefficient at $m$ counts ordered $t$-tuples of primes with product $m$. The square-product sum in (4.4) counts $2t$ prime positions in which every prime occurs an even number of times. Pairing equal occurrences bounds it by $(2t-1)!!\, M^t \le (2t)^t M^t$. Different positive squarefree $d$ induce different real primitive characters of conductor at most $4X^C$. Markov's inequality and (4.4) yield

$$|\mathcal{E}_C| \ll_{\varepsilon,C} (X^C Y^t)^{\varepsilon}(X^C + Y^t)\left(\frac{2tH^4}{M}\right)^t + 2. \tag{4.5}$$

Now $X^C \le Y^t \le X^C Y$ and $M \asymp Y/\log Y$. The factor $Y^t$ in $X^C + Y^t$ is cancelled by the corresponding factor in $M^t$. The remaining logarithm, apart from the $\varepsilon$-factor, is at most

$$O_C\big(t\log(2tH^4\log Y)\big) = o(H).$$

Indeed, $\log H = o(\log Y)$ and $\log Y = o(H)$. Thus (4.5) is at most $X^{2C\varepsilon+o(1)} + 2$. Taking $\varepsilon > 0$ arbitrarily small proves the result; no estimate with an $X$-dependent $\varepsilon$ is used. □

For a fixed number $v$ of origins we take $C > 4v$. In (G) set $\mathcal{E} = \{1, 2\}$ and use the exponent $s$ of Lemma 4.1; in (U) set $\mathcal{E} = \mathcal{E}_C$ and $s = 2$. In both cases

$$|\mathcal{E}| = X^{o(1)}, \qquad |b_Y(d)| \ll_v H^{-s} \quad (d \notin \mathcal{E},\ d \ll_v X^{4v}). \tag{4.6}$$

The estimate $H^{-s}$ tends to zero, but under (G) it need not be $o(H^{-1})$. Section 6 addresses exactly this loss when a collision probability is raised to $k \asymp H$.

## 5 Controlling tuples with square-class relations

Fix the partition $\mathcal{G}$ from Proposition 2.2 and put $h = v - b$. For each group and each of the four forms, equality of its bits is imposed by comparing all its origins to one chosen origin. The products of these $4h$ comparisons form a vector space $V$ over $\mathbb{F}_2$ of dimension $4h$. More explicitly, a vector $\xi \in V$ selects some of the $4v$ form values so that, in each group, each form index is selected an even number of times. Its evaluated square class is

$$d_{\boldsymbol{n}}(\xi) = \mathrm{sf}\left(\prod_{(j,i)\ \text{selected by}\ \xi} L_i(n_j)\right).$$

All such classes are $\ll_v X^{4v}$.

Away from primes dividing any form value, Fourier expansion of the equality indicators on the sign cube gives

$$q_{\mathcal{G}}(\boldsymbol{n}) = 2^{-4h}\sum_{\xi\in V} b_Y(d_{\boldsymbol{n}}(\xi)) + O_v(H/Y). \tag{5.1}$$

There are $O_v(H/\log Y)$ excluded primes in $\mathcal{P}_Y$, since their product divides a fixed product of polynomial size in $X$. Dividing by $M \asymp Y/\log Y$ justifies the error, including the auxiliary treatment of zero symbols.

Let $U = U(\boldsymbol{n}) \subseteq V$ be the subspace spanned by the nonzero vectors $\xi$ with $d_{\boldsymbol{n}}(\xi) \in \mathcal{E}$. Put

$$r = 4h - \dim U.$$

For vectors outside $U$, (4.6) applies. For vectors in $U$, use $|b_Y| \le 1$. Consequently, at $k = \log_{16} X + O(1)$,

$$q_{\mathcal{G}}(\boldsymbol{n}) \le 2^{-r} + O_v(H^{-s} + H/Y), \qquad q_{\mathcal{G}}(\boldsymbol{n})^k \le X^{-r/4+o(1)}. \tag{5.2}$$

The second estimate is uniform over tuples for fixed $v$. Its subpolynomial factor is permitted in the exceptional count below. It would be insufficient by itself for the main contribution.

**Proposition 5.1** (Negligible weighted exceptional tuples)**.** *For fixed $v, \mathcal{G}$ and $k = \log_{16} X + O(1)$,*

$$\sum_{\substack{\boldsymbol{n} \in I_X^v \text{ pairwise distinct} \\ U(\boldsymbol{n}) \neq 0}} q_{\mathcal{G}}(\boldsymbol{n})^k = o(X^b). \tag{5.3}$$

*The number of tuples in this sum is also $o(X^v)$.*

*Proof.* There are only finitely many possible formal subspaces $U$ and bases, with a bound depending on $v$. Choose a basis of $U$ from its exceptional generators and prescribe the class in $\mathcal{E}$ of every basis vector. There are $|\mathcal{E}|^{O_v(1)} = X^{o(1)}$ such assignments. Fix one assignment for now. Multiplication in the group of square classes then specifies the evaluated class of every vector in $U$. These prescribed classes have size $X^{O_v(1)}$.

Let $\mathcal{S}$ be the exact stratum with this $U$ and these prescribed basis classes. We estimate its size by counting tuples satisfying the prescribed relations. This larger constrained set may include tuples with additional relations; the probability bound (5.2) is used only on $\mathcal{S}$.

**Successive choices of origins.** Choose one initial origin in each group. Order the remaining origins after the initial origin of their own group. Let $V_j$ be the comparison space supported on the origins exposed so far, and let $U_j = U \cap V_j$. An initial origin contributes no comparisons. A noninitial origin increases $\dim V_j$ by four, and increases $\dim U_j$ by a number $c_j \in \{0, 1, 2, 3, 4\}$. Projection to its four new coordinates identifies $U_j/U_{j-1}$ with a $c_j$-dimensional subspace of $\mathbb{F}_2^4$. The newly imposed relations therefore prescribe $c_j$ independent products of the four new forms after the previous origins have been fixed.

Set $a_j = 4 - c_j$. Summing over noninitial origins gives $\sum_j a_j = r$. An initial origin has at most $X$ choices. For a noninitial one the number of choices is at most $X^{e(a_j)+o(1)}$, where

| $a$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $e(a)$ | 0 | 0 | 0 | 1/2 | 1. |

(5.4)

One relation invokes Lemma 3.2. Any two independent vectors in $\mathbb{F}_2^4$ have a nonzero combination involving at least two coordinates, which gives the $X^{o(1)}$ bound when $c_j \ge 2$. When $c_j = 4$, all four individual classes are prescribed and Lemma 3.1 gives at most one origin. The useful comparison is

$$e(a) - a/4 \le 0, \qquad e(a) - a/4 \le -1/4 \quad \text{for } a = 1, 2, 3. \tag{5.5}$$

Combining the successive choices with (5.2) gives the following exponent for the total weighted contribution of the fixed stratum:

$$\begin{aligned} \sum_{\boldsymbol{n} \in \mathcal{S}} q_{\mathcal{G}}(\boldsymbol{n})^k &\le X^{\,b + \sum_j e(a_j) - r/4 + o(1)} \\ &= X^{\,b + \sum_j (e(a_j) - a_j/4) + o(1)}. \end{aligned} \tag{5.6}$$

Here $j$ runs over the noninitial origins. The factor $X^b$ comes from the initial origin in each group, and $\sum_j a_j = r$ accounts for the full collision-probability cost. Thus any increment

with $a_j \in \{1,2,3\}$ makes the contribution at most $X^{b-1/4+o(1)}$. It remains to save a power when all increments are zero or four.

We may change the initial origins and their exposure order in the remaining cases. The parity definition of $V$, the exact subspace $U$, and the prescribed evaluation of its basis do not depend on these choices. Only the filtration $U_j = U \cap V_j$ and its individual increments change. Their sum still gives $\sum_j a_j = 4(v-b) - \dim U = r$, so both (5.2) and (5.6) remain valid. Below, an exposure order is chosen from the indices of a specified irregularity. There are only finitely many such choices for fixed $v$, and a finite union of the resulting bounds suffices.

**Every exceptional tuple has an arithmetic irregularity.** Fix $\eta = 1/(100(v+1))$. Write $L_i(n) = d_i(n)s_i(n)^2$ with $d_i(n)$ squarefree; the *square part* in the following argument is the root $s_i(n)$, not $s_i(n)^2$. Call a tuple irregular if at least one of the following holds:

(i) Some $s_i(n_j) > X^\eta$.

(ii) Some $L_i(n_j)$ has gcd greater than $X^\eta$ with one of the prescribed exceptional basis classes.

(iii) Two form values at distinct origins have gcd greater than $X^\eta$.

Every tuple in the fixed nonzero stratum is irregular. To see this, suppose otherwise and choose a nonzero basis relation. Each selected individual class has size $\gg X^{1-2\eta}$. Its gcd with each other selected class is at most $X^\eta$, or is bounded by a fixed constant when the forms are at the same origin. Removing all common prime divisors from one selected class leaves a factor of size at least

$$c_v X^{1-2\eta-(4v-1)\eta}. \tag{5.7}$$

That factor occurs in exactly one selected squarefree class, so it divides the evaluated class of the basis relation. It also divides the original form value. Since $1-(4v+2)\eta > 0$, this contradicts the absence of the second irregularity.

**Large square parts and prescribed divisors.** For the first irregularity, make the affected origin initial in its group. Its possible values are bounded by

$$\sum_{X^\eta < s \ll \sqrt{X}} O(X/s^2+1) \ll X^{1-\eta} + X^{1/2}. \tag{5.8}$$

For the second, again make the affected origin initial and sum over divisors $g > X^\eta$ of the relevant fixed class. That class has $X^{o(1)}$ divisors, and each divisibility condition allows $O(X/g+1)$ origins. There are at most $X^{1-\eta+o(1)}$ possibilities. In both cases the remaining successive choices, combined with (5.5) and (5.2), contribute at most $X^{b-\eta+o(1)}$.

**A large gcd between two origins.** Discard the first irregularity, so all square parts are at most $X^\eta$. For two unrestricted origins the number of pairs with a specified pair of forms having gcd greater than $X^\eta$ is at most

$$\sum_{X^\eta < g \ll X} O((X/g+1)^2) \ll X^{2-\eta} + X\log X. \tag{5.9}$$

If the origins are in different groups, choose them both as initial origins and use this saving. If they are in the same group, expose the pair first, as its initial origin and first addition, before any other noninitial origin. If that addition has $a = 4$, (5.9) replaces $X^2$ possibilities for the pair. If $1 \le a \le 3$, the strict inequality in (5.5) suffices.

It remains to treat the zero-rank addition $a = 0$. The whole four-dimensional comparison space of this first pair lies in $U$. Hence the four classes

$$h_i = \mathrm{sf}(L_i(m)L_i(n))$$

are fixed by the prescribed basis classes. Write $L_i(n) = d_i s_i^2$, $L_i(m) = d_i' t_i^2$ and $g_i = \gcd(d_i, h_i)$. Then

$$d_i' = \frac{d_i h_i}{g_i^2}, \qquad \frac{L_i(m)}{L_i(n)} = \frac{h_i t_i^2}{g_i^2 s_i^2}. \tag{5.10}$$

Each $g_i$ is a divisor of the fixed integer $h_i$, so it has $X^{o(1)}$ possibilities. For $i = 1, 2$, the inequalities $s_i, t_i \le X^\eta$ imply that the pair of ratios in (5.10) has at most $X^{4\eta+o(1)}$ possible values. The two ratios

$$R_1 = m/n, \qquad R_2 = (m+1)/(n+1)$$

are distinct because $m \neq n$, and they specify the ordered pair uniquely:

$$n = \frac{R_2 - 1}{R_1 - R_2}, \qquad m = R_1 n.$$

Thus $X^{4\eta+o(1)}$ replaces the generic factor $X$ for an initial origin and its zero-rank addition. This is a power saving.

The intermediate increments save at least $1/4$ in the exponent of (5.6). Each of the first two irregularities saves $\eta$ in an initial-origin count. A large gcd between two free origins gives the same saving by (5.9); for a zero-rank addition, the saving is $1 - 4\eta$. Since $\eta < 1/4$ and $1 - 4\eta > \eta$, every case is bounded by $X^{b-\eta+o(1)}$ for a fixed assignment. Summing over the finitely many indices and formal bases, and the $X^{o(1)}$ assignments of basis classes, still gives

$$\sum_{\substack{\boldsymbol{n} \in I_X^v \text{ pairwise distinct} \\ U(\boldsymbol{n}) \neq 0}} q_{\mathcal{G}}(\boldsymbol{n})^k \le X^{b-\eta+o(1)} = o(X^b),$$

which proves (5.3). For the unweighted assertion, the same union of irregularities, estimated by (5.8), the divisor bound, and (5.9), has $X^{v-\eta+o(1)} = o(X^v)$ tuples. □

The last case is the reason that generic square-class rank alone does not finish the argument. A rank-zero addition pays no probability factor. Two affine ratios, together with small square parts, instead limit how many such pairs can occur.

# 6 Raising a one-prime estimate to many probes

For tuples with $U = 0$, (5.1) gives $q = q_0 + O_v(H^{-s} + H/Y)$. In (U), $s = 2$ and $k = O(H)$, so

$$q^k = q_0^k(1 + o(1)) \tag{6.1}$$

uniformly on these tuples. In (G), $s$ can be smaller than 1; multiplying the pointwise error by $k$ need not give a vanishing quantity. We now prove the mean estimate that replaces this unsuccessful uniform exponentiation.

**Lemma 6.1** (Averaged exponentiation under GRH)**.** *In range* (G)*, for every fixed partition and* $k = \log_{16} X + O(1)$*,*

$$\frac{1}{X^v} \sum_{\substack{\boldsymbol{n} \in I_X^v \\ U(\boldsymbol{n}) = 0}} \left| \frac{q_{\mathcal{G}}(\boldsymbol{n})^k}{q_0^k} - 1 \right| = o(1). \tag{6.2}$$

*Repeated origins are allowed in this sum.*

*Proof.* If $h = 0$, there are no comparisons and $q = q_0 = 1$. Assume $h > 0$. For a fixed prime $p$, let $I_p(\boldsymbol{n})$ be the indicator of all within-group comparisons, with the auxiliary convention at zero. Let $m_p$ be its mean when the $v$ origins are uniform and independent modulo $p$.

For one uniform residue $n$ mod $p$, every four-bit pattern has probability

$$\frac{1}{16} + O(p^{-1/2}). \tag{6.3}$$

Indeed, expand its indicator into the sixteen products of the four characters. Every nonempty product is the character of a polynomial with distinct simple roots modulo $p > 3$ and degree at most four. Weil's character-sum bound [31, p. 206] gives $O(\sqrt{p})$ for its complete sum. Changing the values at the four zero residues adds $O(1/p)$ to the probability. Independence of the $v$ uniform residues therefore gives

$$m_p = q_0 + O_v(p^{-1/2}), \qquad \overline{m} := M^{-1} \sum_{p \in \mathcal{P}_Y} m_p = q_0 + O_v(Y^{-1/2}). \tag{6.4}$$

For independent uniform origins in the actual dyadic interval, put

$$\Delta = M^{-1} \sum_{p \in \mathcal{P}_Y} (I_p - m_p).$$

Then $q = \overline{m} + \Delta$. In a comparison model, let the $v$ residues at every prime be uniform and independent, also independently across the primes. In that model the summands defining $\Delta$ are independent, centered and bounded. Hoeffding's inequality [14, Theorem 2], integrated against the moment weight, gives $\mathbb{E}|\Delta|^\ell \le (C\ell/M)^{\ell/2}$.

Write $\mathbb{E}_X$ for expectation over $v$ independent uniform origins in $I_X$, and $\mathbb{E}_{\mathrm{CRT}}$ for expectation in the comparison model. For an ordered list $\boldsymbol{p} = (p_1, \dots, p_{2j})$ of primes, with repetitions allowed, put

$$G_{\boldsymbol{p}} = \prod_{\ell=1}^{2j} (I_{p_\ell} - m_{p_\ell}), \qquad Q_{\boldsymbol{p}} = \prod_{p \in \{p_1, \dots, p_{2j}\}} p.$$

The product defining $Q_{\boldsymbol{p}}$ uses each distinct prime once. The function $G_{\boldsymbol{p}}$ has absolute value at most one and depends only on the $v$ residues modulo $Q_{\boldsymbol{p}}$. For a single uniform origin in an interval of $X$ consecutive integers, the total variation distance of its residue law modulo $Q$ from uniformity is at most $Q/(2X)$. Tensorizing over the $v$ origins and using the Chinese remainder theorem therefore gives

$$|\mathbb{E}_X G_{\boldsymbol{p}} - \mathbb{E}_{\mathrm{CRT}} G_{\boldsymbol{p}}| \le \frac{v Q_{\boldsymbol{p}}}{X} \le \frac{v Y^{2j}}{X}.$$

Consequently, for every integer $j \ge 1$,

$$\begin{aligned} \mathbb{E}_X \Delta^{2j} &= M^{-2j} \sum_{\boldsymbol{p} \in \mathcal{P}_Y^{2j}} \mathbb{E}_X G_{\boldsymbol{p}} \\ &= \mathbb{E}_{\mathrm{CRT}} \Delta^{2j} + O_v(Y^{2j}/X). \end{aligned} \tag{6.5}$$

The implicit constant is independent of $j$: there are $M^{2j}$ terms, each with coefficient $M^{-2j}$. Combining (6.5) with the comparison-model moment bound and Cauchy–Schwarz yields

$$\mathbb{E}_X |\Delta|^j \le (C_v j/M)^{j/2} + O_v(Y^j/\sqrt{X}), \tag{6.6}$$

again with constants independent of $j$. These are unconditional moments over the origins; restricting their nonnegative integrands to $U = 0$ can only decrease them.

Choose the exponent $s \in (0,1)$ from Lemma 4.1. On $U = 0$, equations (5.1) and (6.4) give $|\Delta| \ll_v H^{-s}$. Thus, for a fixed constant $C'_v > 0$,

$$T := k|\Delta|/q_0 \le C'_v H^{1-s}.$$

Choose $K_v \ge 8e\max(1, C'_v)$ and put $J = \lceil K_v H^{1-s} \rceil$. Using $(J+1)! \ge ((J+1)/e)^{J+1}$, we obtain uniformly on $U = 0$

$$\begin{aligned} 0 \le e^T - \sum_{j=0}^{J} \frac{T^j}{j!} &\le e^T \frac{T^{J+1}}{(J+1)!} \\ &\le \exp(C'_v H^{1-s}) \left(\frac{eC'_v}{K_v}\right)^{J+1} \\ &\le \exp\Big(-(K_v \log 8 - C'_v) H^{1-s}\Big) = o(1). \end{aligned}$$

The coefficient $K_v \log 8 - C'_v$ is positive. At the same time, because $s > 0$ is fixed and $\log Y = A \log H$,

$$J \log Y = o(H), \qquad J \log H = o(H).$$

We may therefore average the truncated series on $U = 0$ and bound its terms by the unconditional moments (6.6). The nonconstant terms from the first part of that bound sum to at most

$$\sum_{j=1}^{J} \left(\frac{C_v k}{\sqrt{jM}}\right)^j = o(1), \tag{6.7}$$

because $k/\sqrt{M} \to 0$ for $A > 2$. The completion errors sum to

$$X^{-1/2} \exp(O_v(J \log(kY + 2))) = X^{-1/2+o(1)}.$$

It follows that $\mathbb{E}[\mathbf{1}_{U=0}(e^{k|\Delta|/q_0} - 1)] = o(1)$.

Finally $k|\overline{m} - q_0| = O_v(H/\sqrt{Y}) = o(1)$. For $z \ge -1$ and integer $k \ge 1$, $|(1+z)^k - 1| \le e^{k|z|} - 1$. Apply this with $z = (q - q_0)/q_0$ and use the preceding exponential estimate. This proves (6.2). □

*Proof of Proposition 2.2.* On $U = 0$, use (6.1) in (U) or Lemma 6.1 in (G). Since $X^v q_0^k \asymp X^b$, the contribution to (2.3) is $o(X^b)$. For $U \ne 0$, the sum of $q^k$ is $o(X^b)$ by Proposition 5.1. The sum of the comparison term $q_0^k$ over the same tuples is also $o(X^b)$, because there are $o(X^v)$ such tuples. □

# 7 From collisions to a marked Poisson law

We first prove the empirical limit for the auxiliary binary words and independent sampled primes. Write $B = 16^k$ for the number of words in this section; the letter $M$ still denotes the number of available primes. Put $X_+ = X\alpha_X$ and $X_- = X - X_+$. For nonnegative integers $s, t$, define the empirical factorial moment

$$F_{s,t}(\Pi) = \frac{1}{B} \sum_{w \in \Gamma^k} (N_{\Pi,+}(w))_s (N_{\Pi,-}(w))_t,$$

where $(u)_j = u(u-1)\cdots(u-j+1)$ for $j \ge 1$ and $(u)_0 = 1$. In this section the counts initially refer to $B_\Pi$.

**Lemma 7.1** (Empirical factorial moments)**.** *Under* (1.9)*, for every fixed* $s, t \geq 0$*,*

$$F_{s,t}(\Pi) \longrightarrow (\rho\alpha)^s\big(\rho(1-\alpha)\big)^t \quad \text{in } L^2 \tag{7.1}$$

*for independent sampled primes and the auxiliary binary convention.*

*Proof.* The cases of total degree at most one are deterministic:

$$F_{0,0} = 1, \qquad F_{1,0} = X_+/B, \qquad F_{0,1} = X_-/B.$$

They have the claimed $L^2$ limits, including when $\alpha = 0$ or $\alpha = 1$. We may therefore assume $v = s + t \geq 2$. For a uniform word chosen independently of the panel, the product of falling factorials counts ordered tuples of distinct origins, $s$ positive and $t$ negative, all producing that word. Conditional on their words coinciding, a uniform word equals the common word with probability $B^{-1}$. Proposition 2.2, for one group, therefore gives

$$\begin{aligned}\mathbb{E}F_{s,t} &= B^{-1} \sum_{\substack{n_1,\dots,n_v \text{ distinct} \\ \text{prescribed labels}}} q_{\{1,\dots,v\}}(\boldsymbol{n})^k \\ &= B^{-v}(X_+)_s(X_-)_t + o(1) \longrightarrow (\rho\alpha)^s\big(\rho(1-\alpha)\big)^t. \end{aligned} \tag{7.2}$$

The error remains valid for arbitrary prescribed labels because the unlabelled error was summed in absolute value.

For the second moment of $F_{s,t}$, choose two uniform words independently. The two ordered lists of origins form two collision groups, with coefficient $B^{-2}$. If the lists are disjoint, Proposition 2.2 for two groups gives the square of the limit in (7.2), with error $o(1)$. More explicitly, the main count is $B^{-2v}(X_+)_{2s}(X_-)_{2t}$. If the lists overlap, their common origin forces the two words to agree. The remaining distinct origins then belong to one connected collision group. For each of the finitely many overlap patterns, Proposition 2.2 bounds the sum of collision probabilities by $O_v(X)$. The contribution is thus $O_v(X/B^2) = o(1)$. This proves (7.1). □

*Proof of the empirical-law part of Theorem 1.2.* From any sequence of $X$ we may select a subsequence along which all the convergences in Lemma 7.1 hold almost surely, by a diagonal extraction over the countably many fixed pairs $(s, t)$. For each resulting realization, the first moments give tightness of the empirical measures on $\mathbb{N}_0^2$. Higher moments give the uniform integrability needed to identify all factorial moments of any weak limit. Those moments are the moments of the indicated pair of independent Poisson variables. This law is moment-determinate, since its exponential moments are finite in a neighbourhood of zero. Every subsequential limit is consequently the same Poisson product law.

On the discrete space $\mathbb{N}_0^2$, weak convergence to a probability measure implies total variation convergence: first restrict to a finite box having arbitrarily large limiting mass, and then sum the pointwise differences within that box. The subsequence argument proves total variation convergence in probability for the original sequence. The same reasoning, using the second moments, permits integration of functions bounded in absolute value by a constant times $1 + x + y$.

Lemma 2.1 now transfers convergence in probability to the distinct-prime panel. This conditioning step is not an assertion about the expectations of arbitrary unbounded statistics. Finally remove all origins having an actual erasure from their auxiliary binary bins. There are $R_\Pi = o(X)$ such origins, uniformly in the panel. At most $R_\Pi$ bins change, so the total variation distance between the two empirical occupancy measures is at most

$$R_\Pi/B = o(1)$$

at critical load. This proves the claimed limit for the actual non-erased words. □

# 8 Identification, prediction and the full threshold

## 8.1 The three decision quantities

For any fixed panel and its full observation alphabet, let $N(w) = N_+(w) + N_-(w)$ include the bins with erasures. The uniform prior on origins gives the exact identities

$$\mathrm{Uniq}_{\Pi,X} = \frac{1}{X}\sum_w \mathbf{1}_{\{N(w)=1\}}, \tag{8.1}$$

$$\mathrm{Id}_{\Pi,X} = \frac{1}{X}\sum_w \mathbf{1}_{\{N(w)>0\}}, \tag{8.2}$$

$$\mathrm{Pred}_{\Pi,X} = \frac{1}{X}\sum_w \max(N_+(w), N_-(w)). \tag{8.3}$$

For identification, a decoder can choose one origin from each nonempty bin, and can correctly identify no more than one origin in that bin. Randomizing the choice cannot improve its average success. For label prediction, choosing a majority label in every bin attains (8.3).

In the auxiliary model the sums are over $B$ words, so each right side is $B/X$ times an empirical mean. The law of $Z = Z_+ + Z_-$ is $\mathrm{Pois}(\rho)$. Applying Section 7 to the first two bounded functions gives

$$\rho^{-1}\mathbb{P}(Z = 1) = e^{-\rho}, \qquad \rho^{-1}\mathbb{P}(Z > 0) = \frac{1 - e^{-\rho}}{\rho}.$$

The function $\max(x, y)$ has linear growth. Its empirical mean converges by uniform integrability; for example the contribution on $x + y > K$ is at most the empirical second moment of $x + y$ divided by $K$. Since $\max(x, y) = (x + y + |x - y|)/2$, this gives (1.12).

Changing the word assigned to one origin changes the number of singleton bins by at most two. It changes each optimal success probability in (8.2)–(8.3) by at most $1/X$: for any fixed decoder only that origin's success indicator can change, and one then optimizes. Thus the $R_\Pi = o(X)$ erasures change all three limits by $o(1)$. This completes the proof of Theorem 1.2.

## 8.2 A balanced label at critical load

If $\alpha = 1/2$, the two Poisson means are $\rho/2$. For $j \geq 0$ define the modified Bessel functions by the convergent series

$$I_j(\rho) = \sum_{\ell \geq 0} \frac{(\rho/2)^{2\ell+j}}{\ell!(\ell + j)!}.$$

Directly from the two Poisson mass functions, $\mathbb{P}(Z_+ - Z_- = j) = e^{-\rho} I_j(\rho)$. The difference $Z_+ - Z_-$ has the Skellam distribution [27]; the equal-mean case was studied earlier by Irwin [15]. We derive the required absolute moment directly. Writing $D = Z_+ - Z_-$, the Poisson shift identity gives

$$\mathbb{E}[D\mathbf{1}_{\{D\geq 1\}}] = \frac{\rho}{2}(\mathbb{P}(D \geq 0) - \mathbb{P}(D \geq 2)).$$

By symmetry, $\mathbb{E}|D| = \rho(\mathbb{P}(D = 0) + \mathbb{P}(D = 1))$. Consequently the balanced prediction limit is

$$\frac{1}{2} + \frac{e^{-\rho}}{2}(I_0(\rho) + I_1(\rho)). \tag{8.4}$$

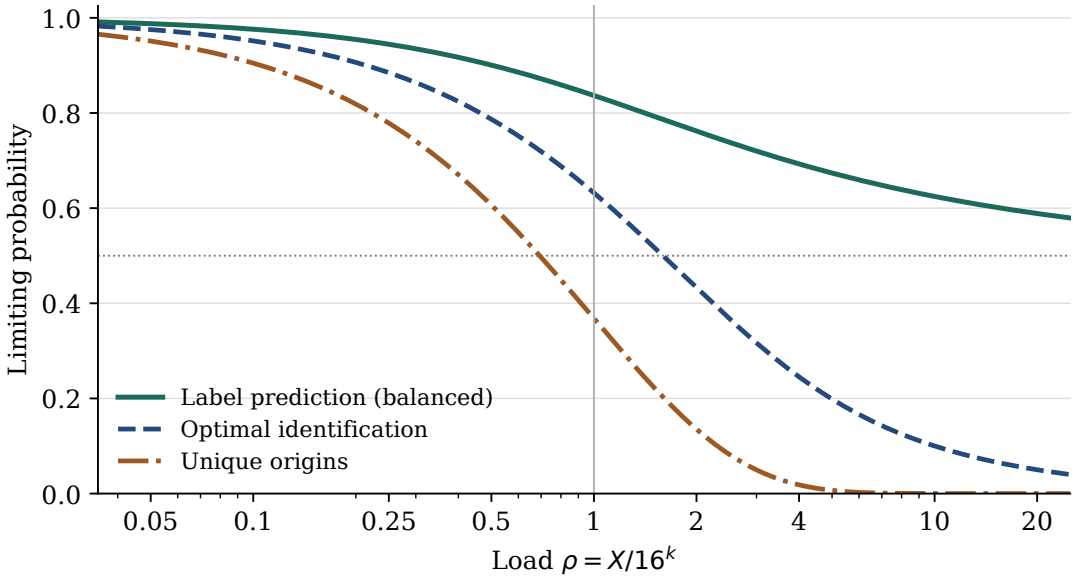


Figure 1: The three critical limits as functions of the load $\rho = X/16^k$. Label prediction is shown for $\alpha = 1/2$. These are the formulas of Theorem 1.2, not fitted curves or estimates from simulated Liouville data.

At load one the three limits are

| | |
|---|---|
| unique origins: | $0.3678794412\ldots$, |
| optimal identification: | $0.6321205588\ldots$, |
| optimal balanced-label prediction: | $0.8368350115\ldots$. |

The last value is a consequence of balanced marks, not an assertion that the Liouville labels have limiting density $1/2$.

### 8.3 Leaving the critical window

We finish by proving the full additive threshold. Couple all distinct-prime panels as prefixes of a uniform permutation of $\mathcal{P}_Y$. Appending a row refines the partition into observed words. The fraction of nonunique origins decreases under refinement.

*Proof of Theorem 1.1, upper side.* Fix $D > 0$ and consider the prefix of length $\lceil k_*(X) + D \rceil$. Its load lies in the compact interval $[16^{-D-1}, 16^{-D}]$. Along every subsequence one may extract a convergent load subsequence. Applying Theorem 1.2 with constant labels, since uniqueness is independent of the labels, bounds the limiting fraction of nonunique origins by $1 - \exp(-16^{-D})$. For any $k - k_*(X) \to +\infty$, this prefix is eventually contained in the $k$-row panel. Monotonicity gives the same bound for that panel. Letting $D \to \infty$ proves that the fraction of nonunique origins tends to zero in probability. All origins with unique words can be identified and labelled correctly. □

For the lower side, write $\mathcal{D}_{\Pi,X}$ for the distance in (1.6). For each fixed panel a direct calculation of the two binary label probabilities gives

$$\mathcal{D}_{\Pi,X} = \frac{1}{X} \sum_{w} |(1 - \alpha_X) N_+(w) - \alpha_X N_-(w)|. \tag{8.5}$$

This distance is nondecreasing when a row is appended: the shorter observation is a deterministic function of the longer one, and total variation contracts under that map.

*Proof of Theorem 1.1, lower side.* For identification, a counting bound already settles the subcritical range. Every fixed panel has at most $16^k$ non-erased words, and at most $R_\Pi$ occupied words with an erasure, since each of the latter contains at least one erased origin. Equations (8.1)–(8.2) give, for every panel,

$$0 \le \mathrm{Uniq}_{\Pi,X} \le \mathrm{Id}_{\Pi,X} = \frac{|B_\Pi(I_X)|}{X} \le \frac{16^k + R_\Pi}{X}. \tag{8.6}$$

If $k - k_*(X) \to -\infty$, then $k = O(\log X)$, and Lemma 2.1 yields

$$\sup_\Pi \mathrm{Id}_{\Pi,X} \le 16^{\,k-k_*(X)} + O\Big(\frac{k}{Y} + \frac{k}{X}\Big) \longrightarrow 0.$$

This proves the subcritical identification and uniqueness statements uniformly over the panels. The label-independence assertion requires the critical law.

At a fixed critical load, with $\alpha_X \to \alpha$, the empirical law and its first-moment integrability give

$$\mathcal{D}_{\Pi,X} \xrightarrow{\mathbb{P}} \frac{1}{\rho}\mathbb{E}|(1-\alpha)Z_+ - \alpha Z_-|.$$

The random variable inside the absolute value has mean zero and variance $\rho\alpha(1-\alpha)$. Cauchy–Schwarz bounds this limit by

$$\sqrt{\frac{\alpha(1-\alpha)}{\rho}} \le \frac{1}{2\sqrt{\rho}}. \tag{8.7}$$

The actual erasures change (8.5) by $o(1)$, either directly or by coupling the joint laws and their word marginals.

Fix $D > 0$ and use the prefix of length $\lfloor k_*(X) - D \rfloor$, which is nonnegative for large $X$. Its load is in $[16^D, 16^{D+1}]$. From any subsequence one may extract convergent loads and convergent label proportions. The bound (8.7) then gives an upper limit $16^{-D/2}/2$ in probability. If $k - k_*(X) \to -\infty$, the $k$-row observation is eventually a shorter prefix. Contraction of total variation gives the same bound for it. Let $D \to \infty$ to prove (1.6).

Under the product law in (1.6), no decoder improves on the constant majority-label rule, whose success is $\max(\alpha_X, 1-\alpha_X)$. The success of any fixed rule changes by at most the total variation distance. Optimizing over rules gives the prediction assertion. □

# 9 A finite illustration and the scope of the results

## 9.1 The quartic Liouville label on dyadic intervals

Write $M_C(t) = t^{-1}\sum_{n\le t} C(n)$ and set

$$B = \frac{196567331}{402653184}, \qquad a_- = \frac{1-B}{2}, \qquad a_+ = \frac{1+B}{2}.$$

There is a set $E \subset \mathbb{N}$ with $\sum_{X\le T,\, X\in E} X^{-1} = o(\log T)$ such that

$$\limsup_{\substack{X\to\infty \\ X\notin E}} |\overline{C}_X| \le B. \tag{9.1}$$

Indeed, the ordinary-average corollary on fixed progressions in Pouly [22], with $q = 1$ and shift zero, bounds the limsup of $|M_C(X)|$ by $B$ outside a set of logarithmic density zero. The fixed-ratio isotopy proposition there, applied to the same four forms at ratio $1/2$, gives

$M_C(2X) - M_C(X) = o(1)$ outside another such set. Take their union for $E$. Replacing the two endpoints in the dyadic sum costs at most $2/X$, so

$$2\alpha_X - 1 = \overline{C}_X = 2M_C(2X) - M_C(X) + O(X^{-1}) = M_C(X) + o(1) \qquad (X \to \infty,\ X \notin E).$$

This establishes (9.1) for the uniform dyadic catalogue and fixes the permitted scales before the panel is drawn.

On these scales all limit points of $\alpha_X$ lie in $[a_-, a_+]$. In the subcritical regime (1.7) therefore gives $\mathrm{Pred}_{\Pi,X} \le a_+ + o_{\mathbb{P}}(1)$, with $a_+ \simeq 0.74409$. At critical load $\rho > 0$, pass to a subsequence on which $\alpha_X \to \alpha$. Each of the two limiting mark intensities in Theorem 1.2 then lies in $[\rho a_-, \rho a_+]$. More precisely, let $F_\rho(u)$ be the right-hand side of (1.12) with $\alpha = u$. Conditioning on the total $m = Z_+ + Z_- \sim \mathrm{Pois}(\rho)$ gives

$$F_\rho(u) = \frac{1}{2} + \frac{e^{-\rho}}{2\rho} \sum_{m=1}^{\infty} \frac{\rho^m}{m!} \mathbb{E}|2\ \mathrm{Bin}(m,u) - m|.$$

For fixed $m$ the expectation is a symmetric polynomial in $u$ whose Bernstein coefficients $|2j - m|$, $0 \le j \le m$, form a convex sequence. Its second derivative is $m(m-1)$ times a Bernstein average of their second differences when $m \ge 2$. Hence $F_\rho$ is convex and symmetric, and the admissible limits satisfy

$$F_\rho(1/2) \le F_\rho(\alpha) \le F_\rho(a_+).$$

At unit load the limiting success is thus between 0.8368350 and 0.8766688 on these subsequences. The bounds follow by retaining $m \le 18$ in the series. Since the expectation is at most $m$, the omitted part at $\rho = 1$ is at most $19/(36 \cdot 18!)$; the supplement implements the rational interval calculation. Attainment of either extremal prediction value would require additional information about the limit points of $\alpha_X$.

### 9.2 Finite-panel measurements

Table 1 illustrates the distinction between the three decision quantities. For each displayed value of $X$, we set $Y = \lfloor (\log X)^4 \rfloor$ and draw the indicated number of distinct-prime panels from (1.3). The labels are the exact Liouville products $C(n)$. All entries are averages of exact finite-panel counts; the common erasures are retained.

| $X$ | $k$ | Panels | $\rho_X$ | $\mathrm{Id}_{\Pi,X}$ | $\mathrm{Uniq}_{\Pi,X}$ | $\mathrm{Pred}_{\Pi,X}$ |
|---|---|---|---|---|---|---|
| $2^{14}$ | 4 | 16 | 0.25 | 0.884903 | 0.779179 | 0.944519 |
| $2^{15}$ | 4 | 16 | 0.50 | 0.788101 | 0.608316 | 0.901075 |
| $2^{16}$ | 4 | 16 | 1.00 | 0.632864 | 0.368642 | 0.837325 |
| $2^{17}$ | 4 | 16 | 2.00 | 0.433181 | 0.136527 | 0.761930 |
| $2^{18}$ | 4 | 16 | 4.00 | 0.246308 | 0.019331 | 0.693147 |
| $2^{20}$ | 5 | 8 | 1.00 | 0.632652 | 0.368776 | 0.837010 |

Table 1: Finite-panel averages for the Liouville label. The table is an illustration of the observables, with no asymptotic error bound inferred from these samples.

**Code and data availability.** The PDF on Zenodo includes the embedded attachment `quadratic_probes_data.zip`. Its files are also in the source archive available from the same record. The package contains `verify_probes.py` and its reference output `verification_results.json`, recording the prime populations, all 88 sampled panels, exact counts, label proportions and sampling variability. The calculation uses seed 20261121, Python 3.12.14 and NumPy 2.3.5. From the extracted directory run

```
python verify_probes.py --output results.json
python compare_results.py verification_results.json results.json
python critical_bounds.py
```

The comparator requires exact equality for prime populations, sampled panels, counts, parameters and checks. It allows up to eight units in the last place for summaries and Poisson benchmarks, and reports platform and timing information separately. An independent rerun reproduced every panel and all discrete results. The calculation compares residue-table symbols with Euler's criterion and a prime-power Liouville sieve with independent trial factorization. It also checks the Pell identity (3.1) and both kernel injections on a finite range. These checks support reproducibility of the illustration; the asymptotic conclusions follow from the preceding proofs.

At $X = 2^{20}$ the measured dyadic label bias is $\overline{C}_X = -0.000789642333984375$, and the average erased-origin fraction is approximately 0.000729. The small finite bias explains why the balanced comparison is informative at this scale. It gives no conclusion about the limiting bias.

The proof identifies the arithmetic features needed for this observation problem: uniform fibres for products of forms, separation when all four classes are prescribed, and control of zero-rank additions. A different family of forms would require these properties to be checked again. Counting the output letters alone does not establish a reconstruction threshold. The large-sieve proof uses $\log\log X = o(\log Y)$ in the growing-moment large-sieve estimate; it does not cover every fixed polylogarithmic range treated by the Dirichlet zero theorem.

The reconstruction statements are about the information in the observations. An exhaustive dictionary search is allowed throughout, and no algorithm polynomial in $\log X$ is supplied. For the label $C$, the marked law retains the actual positive-label proportion. Section 9.1 uses separate arithmetic information to bound that proportion on the indicated scales.

## Statement on AI assistance

ChatGPT provided mathematical and editorial assistance, including development and expansion of the arguments, targeted mathematical checks, reference checks, and preparation of the source files and computations. The numerical checks described above are finite calculations. The proof presented here is a written proof; no formal verification is claimed.

## References


[1] L. Alpöge, announcement of a Lean proof of the Riemann hypothesis and its Dirichlet generalization, 20 November 2026. https://alpo.ge/remark.pdf.

[2] A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford Studies in Probability 2, Clarendon Press, Oxford, 1992. Chapter 6, Section 6.2. doi: 10.1093/oso/9780198522355.001.0001.

[3] D. Basak, B. C. Berndt and A. Zaharescu, *Distributions associated with small quadratic non-residues and their partitions*, Canadian Journal of Mathematics, First View (2026), 1–41, published online 2 June 2026. doi: 10.4153/S0008414X26102296.

[4] D. Basak, K. Nath and A. Zaharescu, *Poisson Behavior for Chains of Small Quadratic Non-Residues*, International Mathematics Research Notices **2024** (2024), no. 4, 3356–3390; published online 24 July 2023. Theorem 1. doi: 10.1093/imrn/rnad168.

[5] C. H. Bennett, G. Brassard, C. Crépeau and U. M. Maurer, *Generalized Privacy Amplification*, IEEE Transactions on Information Theory **41** (1995), no. 6, 1915–1923. doi: 10.1109/18.476316.

[6] S. Boucheron, F. Gamboa and C. Léonard, *Bins and balls: Large deviations of the empirical occupancy process*, The Annals of Applied Probability **12** (2002), no. 2, 607–636. Proposition 2.3. doi: 10.1214/aoap/1026915618.

[7] J. L. Carter and M. N. Wegman, *Universal Classes of Hash Functions*, Journal of Computer and System Sciences **18** (1979), no. 2, 143–154. doi: 10.1016/0022-0000(79)90044-8.

[8] R. de la Bretèche, V. Y. Wang and M. W. Xu, *Random Multiplicative Functions and Making Squares from Polynomial Values*, preprint, version 1, 7 July 2026. arXiv:2607.06398v1. Lemmas 3.2–3.3. doi: 10.48550/arXiv.2607.06398.

[9] J.-H. Evertse and J. H. Silverman, *Uniform bounds for the number of solutions to* $Y^n = f(X)$, Mathematical Proceedings of the Cambridge Philosophical Society **100** (1986), no. 2, 237–248. Theorem 1(b), p. 238. doi: 10.1017/S0305004100066068.

[10] A. Granville and P. Kurlberg, *Poisson statistics via the Chinese Remainder Theorem*, Advances in Mathematics **218** (2008), no. 6, 2013–2042. doi: 10.1016/j.aim.2008.04.001.

[11] B. Hanson, R. C. Vaughan and R. Zhang, *The least number with prescribed Legendre symbols and representation by binary quadratic forms of small discriminant*, Journal of Number Theory **179** (2017), 3–16. doi: 10.1016/j.jnt.2017.03.004.

[12] J. Håstad, R. Impagliazzo, L. A. Levin and M. Luby, *A Pseudorandom Generator from any One-way Function*, SIAM Journal on Computing **28** (1999), no. 4, 1364–1396. Lemma 4.8. doi: 10.1137/S0097539793244708.

[13] D. R. Heath-Brown, *A mean value estimate for real character sums*, Acta Arithmetica **72** (1995), no. 3, 235–275. Corollary 2, p. 238. doi: 10.4064/aa-72-3-235-275.

[14] W. Hoeffding, *Probability Inequalities for Sums of Bounded Random Variables*, Journal of the American Statistical Association **58** (1963), no. 301, 13–30. Theorem 2, p. 16, equation (2.6). doi: 10.1080/01621459.1963.10500830.

[15] J. O. Irwin, *The Frequency Distribution of the Difference between Two Independent Variates Following the Same Poisson Distribution*, Journal of the Royal Statistical Society **100** (1937), no. 3, 415–416. doi: 10.1111/j.2397-2335.1937.tb04518.x.

[16] P. Kurlberg, *Poisson spacing statistics for value sets of polynomials*, International Journal of Number Theory **5** (2009), no. 3, 489–513. doi: 10.1142/S1793042109002237.

[17] P. Kurlberg and Z. Rudnick, *The distribution of spacings between quadratic residues*, Duke Mathematical Journal **100** (1999), no. 2, 211–242. Theorems 1–2. doi: 10.1215/S0012-7094-99-10008-1.

[18] D. H. Lehmer, E. Lehmer and D. Shanks, *Integer Sequences Having Prescribed Quadratic Character*, Mathematics of Computation **24** (1970), no. 110, 433–451. doi: 10.1090/S0025-5718-1970-0271006-X.

[19] H. L. Montgomery and R. C. Vaughan, *Multiplicative Number Theory I: Classical Theory*, Cambridge Studies in Advanced Mathematics 97, Cambridge University Press, 2007. Theorem 13.7, p. 425, equations (13.19)–(13.21). doi: 10.1017/CBO9780511618314.

[20] C. Pomerance and I. E. Shparlinski, *On Pseudosquares and Pseudopowers*, in *Combinatorial Number Theory: Proceedings of the Integers Conference 2007*, Walter de Gruyter, Berlin, 2009, 171–184. Theorem 1. doi: 10.1515/9783110208504.171.

[21] B. Pouly, *Long runs and rare patterns of a random completely multiplicative function*, preprint, version 3, 2026. doi: 10.5281/zenodo.22872154.

[22] B. Pouly, *An explicit bound below one half for a quartic Liouville correlation*, 2026. Proposition on isotopy for even affine Liouville products and corollary on ordinary averages on each fixed progression. doi: 10.5281/zenodo.24443184.

[23] B. Pouly, *Reconstruction and sign prediction from sampled prime factors*, 2026. doi: 10.5281/zenodo.24443266.

[24] B. Pouly, *Small roots of split quadratic congruences and fundamental units in a quartic family*, 2026. doi: 10.5281/zenodo.24443299.

[25] B. Pouly, *Sharp Poisson approximation and second-order correction for long multiplicative runs*, preprint, 2026. hexagon:2611.00447v1. doi: 10.5281/zenodo.24220727.

[26] I. Pylaiev, *Random Chowla and Goldbach Conjectures*, preprint, 2026, version 1, 10 November 2026. Theorem 1.5. arXiv:2611.14579v1.

[27] J. G. Skellam, *The Frequency Distribution of the Difference between Two Poisson Variates Belonging to Different Populations*, Journal of the Royal Statistical Society **109** (1946), no. 3, 296. doi: 10.1111/j.2397-2335.1946.tb04670.x.

[28] M. Rubinstein, *An independent check of the Lean proof of RH and GRH*, verification report and logs, accessed 22 November 2026. https://mrubinst.github.io/grh-lean-verification/.

[29] A. V. Sutherland, *Verification files for the Lean proof of the Riemann Hypothesis*, statements 3–5 and verification logs, 22 November 2026. https://math.mit.edu/~drew/grh/.

[30] R. C. Vaughan and T. D. Wooley, *Further improvements in Waring's problem*, Acta Mathematica **174** (1995), no. 2, 147–240. Lemma 3.5, p. 162. doi: 10.1007/BF02392467.

[31] A. Weil, *On Some Exponential Sums*, Proceedings of the National Academy of Sciences of the United States of America **34** (1948), no. 5, 204–207, especially p. 206. doi: 10.1073/pnas.34.5.204.