

CLOSING THE HARMONIC GAP: EXACT IDENTITY TESTING FOR ZIPFIAN DISTRIBUTIONS

Pahan Dewasurendra
Johns Hopkins University

ABSTRACT

Instance-optimal identity testing asks how many samples are needed to test an unknown discrete distribution against a specific known null. Existing characterizations truncate a constant multiple of the testing radius. This is usually harmless, but it creates a polynomial gap for the harmonic null $q_i \propto 1/i$, the canonical obstruction highlighted in a 2024 COLT open problem. We give a constant-factor characterization for the full Zipfian class $q_i \asymp 1/(Li)$. Let $m_q(\varepsilon)$ be the last index whose strict suffix has null mass at least ε . Throughout the nondegenerate range,

$$N^*(q, \varepsilon) = \Theta\left(L\sqrt{m_q(\varepsilon)} \max\{1, (\varepsilon L)^{-2}\}\right),$$

where constants depend only on the fixed Zipf envelope. For the exact harmonic law, $L = H_k$ and $m_q(\varepsilon) = \Theta(k \exp(-\varepsilon H_k))$, so

$$N^*(q, \varepsilon) = \Theta\left(H_k \sqrt{k e^{-\varepsilon H_k}} \max\{1, (\varepsilon H_k)^{-2}\}\right).$$

Thus, for every fixed $\varepsilon \in (0, 1/3]$, the answer is $\Theta((\log k)k^{(1-\varepsilon)/2})$. The upper bound combines a coarsened instance-optimal test with one unweighted tail-collision statistic. Its key step is a water-filling inequality: if almost all of a Zipfian tail is deleted, its squared discrepancy cannot fall below the tail's collision mass. The lower bound introduces a graded-thinning prior. Tail coordinate i is retained with probability proportional to $i^{-1/3}$ and inflated when retained. This profile makes the alternatives ε -far while keeping the mixture second moment bounded up to the matching sample size. Beyond settling the harmonic example robustly, the result isolates graded thinning as the mechanism missed by constant-radius truncations.

1 INTRODUCTION

Identity testing is the simplest distribution-testing problem with a nontrivial instance geometry. A known distribution q on $[k]$ and samples from an unknown distribution p are given. The goal is to distinguish $p = q$ from $d_{TV}(p, q) > \varepsilon$. In the worst case over q , the optimal sample complexity is $\Theta(\sqrt{k}/\varepsilon^2)$ (Paninski, 2008; Diakonikolas et al., 2018). For a fixed q , the answer can be much smaller. This motivates instance-optimal, also called massively parameterized, identity testing (Valiant & Valiant, 2017; Goldreich, 2017).

The influential characterization of Valiant & Valiant (2017) uses the $\ell_{2/3}$ quasi-norm after removing the largest atom and a low-probability tail. Their lower and upper bounds remove, respectively, constant multiples of ε . Blais et al. (2017) give incomparable bounds in terms of the K -functional between ℓ_1 and ℓ_2 . Chhor & Carpentier (2022) characterize the dual local separation radius up to constants. None of these results gives a constant-factor sample complexity for every pair (q, ε) . The issue is not cosmetic. A constant change in the truncated mass can induce an arbitrarily large sample gap (Blais et al., 2017; Canonne, 2024).

The standard witness is the harmonic distribution

$$q_i^{(k)} = \frac{1}{iH_k}, \quad H_k = \sum_{j=1}^k \frac{1}{j}. \tag{1}$$

Removing tail mass ε leaves about $k \exp(-\varepsilon H_k) = k^{1-\varepsilon+o(1)}$ coordinates. Hence a constant multiplying ε changes a power of k . This example is the quantitative obstruction in the open problem of Canonne (2024).

We close this harmonic gap, and do so robustly for the Zipfian envelope $q_i \asymp 1/(Li)$. The rate depends on the exact ε -tail quantile, so it preserves the accuracy rather than a constant multiple of it. The proof also explains why existing truncation arguments lose information. The hard alternatives do not perturb the retained head by a small symmetric amount. Instead, they thin the whole relevant tail at a coordinate-dependent rate and inflate the survivors. The best retention probability is proportional to the cube root of the null mass. Within a Zipf envelope this is equivalent to $(r/i)^{1/3}$ beyond a cutoff r .

Contributions. Our results are as follows.

1. We give a constant-factor sample complexity for every fixed Zipf envelope $q_i \asymp 1/(Li)$. The exact harmonic law is a corollary with universal constants. The answer transitions at $\varepsilon \asymp 1/L$ from the classical quadratic-accuracy rate to an accuracy-dependent effective support.
2. We introduce a two-part tester. A standard identity test is applied to a coarsening that preserves a head and lumps the tail. If it accepts, any far alternative must internally redistribute almost all of the tail. One unweighted collision statistic detects exactly this case.
3. We construct a matching graded-thinning prior and analyze it directly in the fixed-sample multinomial experiment. Conditioning a single reservoir coordinate on a typical fluctuation enforces exact normalization without destroying the product second moment.
4. We prove quantitative water-filling and thinning lemmas that expose a common cube-root profile on the algorithmic and information-theoretic sides.

2 RELATED WORK

Identity and uniformity testing. Collision and chi-squared statistics yield the optimal $\sqrt{k}/\varepsilon^2$ worst-case rate (Goldreich, 2017; Canonne, 2020; 2022). Paninski (2008) proved a sharp uniformity lower bound by a random perturbation prior. High-confidence identity testing is characterized by Diakonikolas et al. (2018). We work at constant success probability and ask for the exact dependence on the known null.

Instance-optimal testing. Valiant & Valiant (2017) propose a weighted chi-squared statistic and show bounds based on a truncated $\ell_{2/3}$ quasi-norm. Their result is constant-factor sharp after allowing a constant rescaling of the testing radius. Blais et al. (2017) connect testing lower bounds to communication complexity and obtain a K -functional characterization, again with different constants at the radius. Chhor & Carpentier (2022) study local minimax separation for multinomial, binomial, and Poisson experiments. A constant-factor separation radius does not invert to a constant-factor sample complexity when the radius occurs in an exponent. The open problem of Canonne (2024) asks for a succinct functional that avoids this gap and singles out the harmonic distribution as the basic unresolved example. Our theorem resolves that example and its constant-factor Zipfian neighborhood, rather than the full arbitrary-null problem.

Lower-bound methods. Poissonization, random perturbations, and moment matching are standard tools for discrete testing (Paninski, 2008; Valiant & Valiant, 2017). Our lower prior differs from the symmetric perturbations used in prior instance-optimal bounds. It is a sparse, inhomogeneous thinning prior. Its retention profile is chosen to equalize the second-moment cost subject to deleting prescribed null mass.

Power-law nulls. Exponent-one Zipf laws are a standard model for ranked word frequencies, and discrete power-law profiles are a motivating structured family in distribution inference (Valiant & Valiant, 2021). Previous instance-optimal testing guarantees still compare different constant multiples of the radius on this family. Our exact tail-quantile rate removes that ambiguity.

Table 1: Bounds for the harmonic null at fixed ε . Polynomial factors in k are shown explicitly.

Result	Type	Harmonic dependence, suppressing ε constants
Valiant–Valiant	lower	$\Omega(k^{(1-\varepsilon)/2}/H_k)$
Blais–Canonne–Gur	lower	$\Omega(k^{1/2-\varepsilon})$
Valiant–Valiant	upper	$O(k^{(1-\varepsilon/16)/2}/H_k)$
Theorem 1	matching	$\Theta(H_k k^{(1-\varepsilon)/2})$

3 SETUP AND MAIN RESULT

For distributions p, q on $[k]$, let

$$d_{\text{TV}}(p, q) = \frac{1}{2} \sum_{i=1}^k |p_i - q_i|.$$

Let $N^*(q, \varepsilon)$ be the least integer n for which a randomized test based on n independent samples has both type-I error under q and worst-case type-II error over $\{p : d_{\text{TV}}(p, q) > \varepsilon\}$ at most $1/3$. Changing $1/3$ to another fixed constant changes all rates only by universal factors.

Fix $\kappa \geq 1$. We call a nonincreasing null q on $[k]$ κ -Zipfian if there is a scale L such that

$$\frac{1}{\kappa L i} \leq q_i \leq \frac{\kappa}{L i} \quad (i \in [k]). \quad (2)$$

Normalization implies $H_k/\kappa \leq L \leq \kappa H_k$. Define the exact tail quantile

$$m_q(\varepsilon) = \max \left(\{0\} \cup \left\{ m \in [k-1] : \sum_{i>m} q_i \geq \varepsilon \right\} \right)$$

and put $x = \varepsilon L$. The mild lower bound on $m_q(\varepsilon)$ below only excludes the endpoint where the effective tail has constant size.

Theorem 1 (Zipfian identity testing). *For every fixed $\kappa \geq 1$, there are constants $c_\kappa, C_\kappa, r_\kappa > 0$ such that every κ -Zipfian q and every $0 < \varepsilon \leq 1/3$ with $m_q(\varepsilon) \geq r_\kappa$ satisfy*

$$c_\kappa L \sqrt{m_q(\varepsilon) \max\{1, x^{-2}\}} \leq N^*(q, \varepsilon) \leq C_\kappa L \sqrt{m_q(\varepsilon) \max\{1, x^{-2}\}}. \quad (3)$$

The constants are independent of k, q , and ε .

Corollary 1 (The harmonic gap). *For $q^{(k)}$ in (1), there are universal constants in (3) with $L = H_k$ and*

$$m_q(\varepsilon) = \Theta(k e^{-\varepsilon H_k}).$$

Consequently, for every fixed $\varepsilon \in (0, 1/3]$,

$$N^*(q^{(k)}, \varepsilon) = \Theta_\varepsilon((\log k) k^{(1-\varepsilon)/2}).$$

The stronger statement (3) retains ε inside the exponent with universal constants.

The factor $\max\{1, x^{-2}\}$ describes the transition. If $x \lesssim_\kappa 1$, then $m_q(\varepsilon) = \Theta_\kappa(k)$ and (3) becomes

$$\Theta_\kappa \left(\frac{\sqrt{k}}{L \varepsilon^2} \right),$$

the usual weighted chi-squared rate. Once x is larger than a κ -dependent constant, the answer becomes $\Theta_\kappa(L \sqrt{m_q(\varepsilon)})$.

Table 1 compares the fixed-accuracy consequences. We use the total-variation convention throughout. Constant changes between total variation and ℓ_1 affect the displayed exponents in earlier bounds, so the entries state the constants as formulated by the cited results.

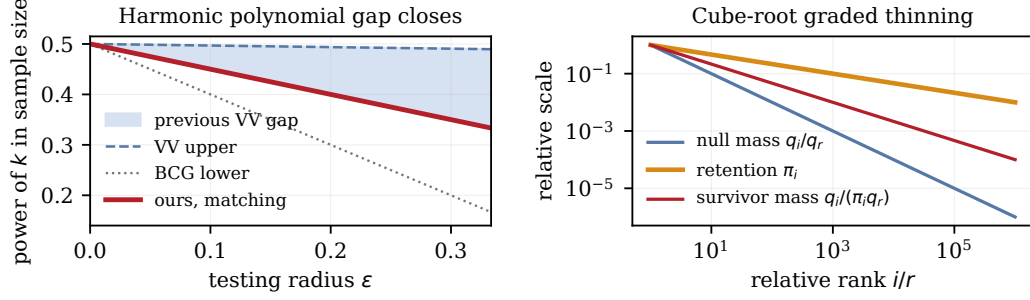


Figure 1: Left: for the harmonic null at fixed ε , previous upper and lower bounds leave a polynomial range. Our matching rate lies on the sharp lower exponent and determines the missing logarithmic factor. Right: the hard prior retains rank i with cube-root probability and inflates survivors, spreading its second-moment cost across the tail.

4 THE TWO-PART UPPER BOUND

We first give the new-regime proof. The standard small- x tester is recalled at the end. Let $m = m_q(\varepsilon)$. Choose a sufficiently large constant K_κ , then put

$$r = \lfloor m/K_\kappa \rfloor, \quad S = \{r+1, \dots, k\}. \quad (4)$$

The Zipf envelope and the definition of m imply

$$q(S) = \varepsilon + \Theta_\kappa(1/L), \quad \sum_{i \in S} q_i^2 = O_\kappa(1/(L^2 r)), \quad \|q_{[r]}\|_{2/3} = O_\kappa(\sqrt{r}/L). \quad (5)$$

We choose K_κ so that the first additive term in (5) is positive and larger than the maximum atom.

The tester splits its samples into two independent batches.

Coarsened head test. Map every $i \in S$ to one new symbol s . Apply a standard instance-optimal identity test to the resulting distribution on $[r] \cup \{s\}$ at total-variation radius $\delta = c_\kappa/L$ for a small fixed c_κ . The tail cell has mass larger than q_1 and is the largest atom. Removing it leaves a vector whose full $\ell_{2/3}$ quasi-norm is $O_\kappa(\sqrt{r}/L)$. The upper bound of Valiant & Valiant (2017) therefore uses

$$O_\kappa\left(\frac{\sqrt{r}/L}{\delta^2} + \frac{1}{\delta}\right) = O_\kappa(L\sqrt{r}) \quad (6)$$

samples.

Tail collision test. In a Poissonized batch with mean n , let N_i be the count of symbol i and compute

$$Z_S = \sum_{i \in S} ((N_i - nq_i)^2 - N_i). \quad (7)$$

Reject when Z_S exceeds a sufficiently large constant threshold. Standard de-Poissonization changes the sample size by only a universal factor.

Why are these two tests enough? Put $\Delta = p - q$ and let $D_S = \sum_{i \in S} (-\Delta_i)_+$. If the coarsened distributions are within δ , then

$$\sum_{i \leq r} |\Delta_i| + |\Delta(S)| < 2\delta.$$

Since the total negative discrepancy equals $d_{TV}(p, q)$, every ε -far alternative satisfies

$$D_S \geq \varepsilon - 2\delta. \quad (8)$$

Thus all but $O_\kappa(1/L)$ of the null tail mass is deleted somewhere. The next lemma converts this one-sided statement into collision signal.

Lemma 1 (Zipfian water filling). *Let q satisfy (2) and $S = \{r+1, \dots, k\}$. Suppose $d_i \in [0, q_i]$ obey*

$$\sum_{i \in S} d_i \geq q(S) - B/L.$$

If $k/r \geq A_{\kappa, B}$, then

$$\sum_{i \in S} d_i^2 \geq \frac{c_{\kappa, B}}{L^2 r}, \quad c_{\kappa, B} > 0. \quad (9)$$

The proof is a one-dimensional convex program. The minimizer is $d_i = \min\{q_i, \lambda\}$. The undeleted-mass budget and (2) force the last uncapped index to be at most a κ, B -dependent multiple of r . Coordinates after it are fully deleted and their squared null mass is $\Omega_\kappa(1/(L^2 r))$. Full details are in Appendix B.

Apply Lemma 1 to $d_i = (-\Delta_i)_+$. Equations (4) and (8) give, whenever the coarsened test has no signal,

$$A := \sum_{i \in S} \Delta_i^2 \geq c_\kappa / (L^2 r). \quad (10)$$

Under Poisson sampling,

$$\mathbb{E}_p Z_S = n^2 A, \quad \text{Var}_p Z_S = 2n^2 \sum_{i \in S} p_i^2 + 4n^3 \sum_{i \in S} p_i \Delta_i^2. \quad (11)$$

The inequalities

$$\sum p_i^2 \leq 2 \sum q_i^2 + 2A, \quad \sum p_i \Delta_i^2 \leq q_{r+1} A + A^{3/2} \quad (12)$$

show that $n = C_\kappa L \sqrt{r}$ makes the mean in (11) dominate its standard deviation. Under the null, $\mathbb{E}_q Z_S = 0$ and $\text{Var}_q Z_S = 2n^2 \sum_{i \in S} q_i^2 = O_\kappa(1)$ after fixing the leading constant. Chebyshev's inequality completes the new-regime upper bound.

For x below a fixed constant, apply the usual weighted identity tester directly to q . Its cost is $O_\kappa(\sqrt{k}/(L\varepsilon^2))$. This is within a κ -dependent factor of (3) on every bounded interval of x . Together with the two-part tester for larger x , it proves the upper half of Theorem 1.

5 THE GRADED-THINNING LOWER BOUND

The matching prior is easiest to describe using the same cutoff $r = \lfloor m_q(\varepsilon)/K_\kappa \rfloor$. For $i \in S$, let

$$\pi_i = (r/i)^{1/3}, \quad Y_i \sim \text{Bernoulli}(\pi_i) \quad (13)$$

independently. Set $A_i = Y_i/\pi_i - 1$ and define a provisional tail by $p_i = q_i(1 + A_i) = q_i Y_i/\pi_i$. Coordinates $2, \dots, r$ remain unchanged. Coordinate 1 is a reservoir:

$$R = \sum_{i \in S} q_i A_i, \quad p_1 = q_1 - R. \quad (14)$$

Every realization with $p_1 \geq 0$ is exactly normalized.

The deleted tail mass is

$$D = \sum_{i \in S} q_i (1 - Y_i).$$

The cube-root profile gives

$$\sum_{i \in S} q_i \pi_i \leq \frac{C_\kappa}{L}, \quad \text{Var}(D) = O_\kappa(1/(L^2 r)), \quad \text{Var}(R) = O_\kappa(1/(L^2 r)). \quad (15)$$

The choice of K_κ makes $q(S) - \varepsilon$ larger than the first term in (15) by a fixed multiple of $1/L$. Thus the mean of D exceeds ε by c_κ/L . We condition on the event

$$\mathcal{E} = \left\{ D > \varepsilon, |R| \leq \frac{B_\kappa}{L\sqrt{r}} \right\} \quad (16)$$

for a fixed sufficiently large B_κ . If r exceeds a κ -dependent constant, (15) gives $\mathbb{P}(\mathcal{E}) \geq 0.99$. Since $q_1 \geq 1/(\kappa L)$, also $p_1 \geq 0$ on $\mathcal{E}$.

Every conditioned realization is ε -far. Indeed, if $R < 0$, the deleted tail mass D is part of the total negative discrepancy. If $R \geq 0$, the positive tail discrepancy is $D + R$. In both cases

$$d_{\text{TV}}(p, q) \geq D > \varepsilon. \quad (17)$$

It remains to show that the mixture is indistinguishable. Let p, p' be independent draws from the conditioned prior and let $L_p, L_{p'}$ be their likelihood ratios against q for n multinomial samples. The standard identity

$$\mathbb{E}_q[L_p L_{p'}] = \left(1 + \sum_i \frac{(p_i - q_i)(p'_i - q_i)}{q_i}\right)^n \leq \exp\left(n \sum_i \frac{(p_i - q_i)(p'_i - q_i)}{q_i}\right) \quad (18)$$

reduces the calculation to a bilinear form. The tail contribution factorizes before conditioning. The reservoir contribution is $O_\kappa(n/(Lr))$ on $\mathcal{E}$, which is $o(1)$ at the claimed scale.

Lemma 2 (Thinning second moment). *Let $A_i = Y_i/\pi_i - 1$ with independent $Y_i \sim \text{Bernoulli}(\pi_i)$ and $\pi_i = (r/i)^{1/3}$. If A'_i is an independent copy, then for $n \leq c_\kappa L\sqrt{r}$,*

$$\mathbb{E} \exp\left(n \sum_{i \in S} q_i A_i A'_i\right) \leq \exp(C_\kappa c_\kappa^2). \quad (19)$$

Here C_κ depends only on the envelope. Moreover, c_κ can be chosen so that the right-hand side is arbitrarily close to one.

The calculation behind Lemma 2 is transparent. The centered multiplier has variance $v_i = (1 - \pi_i)/\pi_i$. A scalar Taylor bound gives

$$\log \mathbb{E} e^{nq_i A_i A'_i} \leq Cn^2 q_i^2 v_i^2,$$

uniformly because $nq_i \max\{1, v_i^2\} = O_\kappa(r^{-1/2})$. Finally,

$$\sum_{i \in S} q_i^2 v_i^2 \leq \sum_{i \in S} \frac{q_i^2}{\pi_i^2} = O_\kappa(1/(L^2 r)). \quad (20)$$

Conditioning costs at most $\mathbb{P}(\mathcal{E})^{-2}$ in the positive second-moment integral. Combining (18), (16), and Lemma 2, then choosing c_κ small and r_κ large, shows that the chi-squared divergence between the null sample law and the mixture is below a fixed small constant. Le Cam's method proves

$$N^*(q, \varepsilon) = \Omega_\kappa(L\sqrt{r}) = \Omega_\kappa(L\sqrt{m_q(\varepsilon)}). \quad (21)$$

When $x \lesssim_\kappa 1$, the classical Valiant–Valiant perturbation bound gives

$$\Omega_\kappa\left(\frac{\sqrt{k}}{L\varepsilon^2}\right) = \Omega_\kappa\left(L\sqrt{m_q(\varepsilon)} x^{-2}\right). \quad (22)$$

For x larger than a constant, (21) dominates. Equations (21) and (22) prove the lower half of Theorem 1. Appendix D supplies all constants and the conditioning argument.

6 WHY THE CUBE ROOT APPEARS

The lower prior gives a simple variational explanation. Suppose a tail coordinate is retained with probability π_i and inflated by $1/\pi_i$. Its leading mixture cost is q_i^2/π_i^2 , while its expected undeleted null mass is $q_i\pi_i$. Minimizing

$$\sum_i \frac{q_i^2}{\pi_i^2} \quad \text{subject to} \quad \sum_i q_i \pi_i \leq b \quad (23)$$

gives $\pi_i \propto q_i^{1/3}$ by the Lagrange equations. For $q_i \propto 1/i$, this is precisely (13).

There is a parallel cube-root law in the head. The weighted chi-squared tester uses $q_i^{-2/3}$, and its complexity is governed by the $\ell_{2/3}$ quasi-norm. The head and tail are therefore not unrelated proof cases. They are the small-perturbation and sparse-thinning faces of the same optimization. Constant-radius truncation captures the first face but discards the graded structure of the second.

This observation suggests a route toward the full open problem. A general functional must interpolate between symmetric perturbation cost $q_i^2 a_i^4$ for small relative changes a_i and thinning cost $q_i^2/(1 - a_i)^2$ as deletion approaches one. Proving a matching tester for arbitrary, non-Zipfian nulls remains open. Our collision argument works for the full envelope (2) because water filling changes an index cutoff only by a κ -dependent factor.

7 LIMITATIONS AND CONCLUSION

Theorem 1 resolves the canonical harmonic instance and a robust constant-factor Zipfian class, but it does not characterize every null. The stated range $\varepsilon \leq 1/3$ avoids a finite-support endpoint in which only constantly many leading coordinates remain. The techniques can cover a larger range with additional boundary cases, but that would obscure the accuracy-sensitive phenomenon.

The main conclusion is that a low-mass tail cannot always be ignored. When its mass is only $O_\kappa(1/L)$ above the testing radius, the hardest alternatives redistribute that tail through graded thinning. A coarsened head test and one collision statistic recover the missing upper bound, while the same cube-root profile yields a matching lower prior. This closes the polynomial harmonic gap throughout its natural envelope and identifies a concrete ingredient that any full instance-optimal characterization must represent.

REFERENCES

- Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *32nd Computational Complexity Conference*, volume 79 of *Leibniz International Proceedings in Informatics*, pp. 28:1–28:40, 2017. doi: 10.4230/LIPIcs.CCC.2017.28.
- Clément L. Canonne. A survey on distribution testing: Your data is big, but is it blue? *Theory of Computing Graduate Surveys*, 9:1–100, 2020. doi: 10.4086/toc.gs.2020.009.
- Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends in Communications and Information Theory*, 19(6):1032–1198, 2022. doi: 10.1561/0100000114.
- Clément L. Canonne. Open problem: Tight characterization of instance-optimal identity testing. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5312–5316, 2024. URL <https://proceedings.mlr.press/v247/canonne24a.html>.
- Julien Chhor and Alexandra Carpentier. Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and poisson families and in multinomials. *Mathematical Statistics and Learning*, 5(1–2):1–54, 2022. doi: 10.4171/MSL/32.
- Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming*, volume 107 of *Leibniz International Proceedings in Informatics*, pp. 41:1–41:14, 2018. doi: 10.4230/LIPIcs.ICALP.2018.41.
- Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. doi: 10.1017/9781108135252.
- Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. doi: 10.1109/TIT.2008.928987.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. doi: 10.1137/151002526.

Gregory Valiant and Paul Valiant. Instance optimal distribution testing and learning. In Tim Roughgarden (ed.), *Beyond the Worst-Case Analysis of Algorithms*, pp. 506–526. Cambridge University Press, 2021.

A ZIPF-ENVELOPE ESTIMATES

We collect the finite-sum estimates used in the proof. For integers $1 \leq a < b$,

$$\log \frac{b+1}{a+1} \leq \sum_{i=a+1}^b \frac{1}{i} \leq \log \frac{b}{a}, \quad (24)$$

$$\frac{1}{b+1} \leq \sum_{i=b+1}^{\infty} \frac{1}{i^2} \leq \frac{1}{b}. \quad (25)$$

More generally, for $\alpha > 1$,

$$\sum_{i=a+1}^{\infty} i^{-\alpha} \leq \int_a^{\infty} t^{-\alpha} dt = \frac{a^{1-\alpha}}{\alpha-1}. \quad (26)$$

The reverse bound with $a+1$ in place of a follows by the same integral comparison.

Let q satisfy (2), put $m = m_q(\varepsilon)$, and let $r = \lfloor m/K_\kappa \rfloor$. Maximality of m gives

$$\varepsilon \leq q(\{m+1, \dots, k\}) < \varepsilon + \frac{\kappa}{L(m+1)}.$$

Applying (24) between r and m , and choosing the fixed K_κ sufficiently large, gives constants $0 < b_\kappa < B_\kappa < \infty$ such that

$$\varepsilon + \frac{b_\kappa}{L} \leq q(\{r+1, \dots, k\}) \leq \varepsilon + \frac{B_\kappa}{L}. \quad (27)$$

Also,

$$\frac{c_\kappa}{L^2 r} \leq \sum_{i=r+1}^k q_i^2 \leq \frac{C_\kappa}{L^2 r}, \quad (28)$$

provided k/r exceeds a κ -dependent constant. Finally,

$$\begin{aligned} \sum_{i=1}^r q_i^{2/3} &= \Theta_\kappa(L^{-2/3} r^{1/3}), \\ \|q_{[r]}\|_{2/3} &= \left(\sum_{i=1}^r q_i^{2/3} \right)^{3/2} = \Theta_\kappa(\sqrt{r}/L). \end{aligned} \quad (29)$$

Here and below $\|z\|_{2/3} = (\sum_i |z_i|^{2/3})^{3/2}$.

For the exact harmonic law, (24) and maximality of m give

$$\log(k/m) = \varepsilon H_k + O(1),$$

uniformly in the nondegenerate range. Hence $m = \Theta(ke^{-\varepsilon H_k})$, proving the tail-quantile assertion in Corollary 1.

B THE WATER-FILLING LEMMA

We prove Lemma 1 in a slightly more explicit form.

Lemma 3 (Capped quadratic minimization). *Let $u_1 \geq u_2 \geq \dots \geq u_M > 0$ and $0 < a < \sum_i u_i$. The minimizer of*

$$\min \left\{ \sum_{i=1}^M d_i^2 : 0 \leq d_i \leq u_i, \sum_i d_i \geq a \right\}$$

has the form $d_i = \min\{u_i, \lambda\}$ for a unique $\lambda > 0$.

Proof. The inequality constraint is tight at a minimizer. Strict convexity gives uniqueness. The Karush–Kuhn–Tucker conditions state that every coordinate strictly below its cap has the same value λ , while a coordinate at its cap has $u_i \leq \lambda$. This is exactly the claimed form. $\square$

Proof of Lemma 1. It is enough to impose equality $\sum_{i \in S} d_i = q(S) - B/L$. Apply Lemma 3 with $u_i = q_i$. Let t be the last index for which $q_t > \lambda$. Up to rounding, the minimizer is

$$d_i = \lambda \quad (r < i \leq t), \quad d_i = q_i \quad (i > t). \quad (30)$$

The undeleted mass is at most B/L . If $t > r$, then the Zipf envelope and $\lambda < q_t \leq \kappa/(Lt)$ give

$$\begin{aligned} \frac{B}{L} &\geq \sum_{i=r+1}^{\lfloor t \rfloor} (q_i - \lambda) \\ &\geq \frac{1}{\kappa L} \log \frac{\lfloor t \rfloor + 1}{r+1} - \frac{\kappa}{L}. \end{aligned}$$

For r above a κ, B -dependent constant this implies $t \leq A'_{\kappa, B} r$. If $t \leq r$, then $d_i = q_i$ throughout S , and (28) gives the result directly. Otherwise, choose $A_{\kappa, B} \geq 2A'_{\kappa, B}$. The assumption $k/r \geq A_{\kappa, B}$ ensures that $\{t+1, \dots, 2t\}$ is available. By (30), (2), and (25),

$$\sum_{i \in S} d_i^2 \geq \sum_{i > t} q_i^2 \geq \frac{c_\kappa}{L^2(t+1)} \geq \frac{c_{\kappa, B}}{L^2 r}.$$

This proves the claim. $\square$

C DETAILS OF THE UPPER BOUND

We first record the statistic calculation.

Lemma 4 (Poisson collision moments). *Let $N_i \sim \text{Poi}(np_i)$ independently, let $\Delta_i = p_i - q_i$, and let*

$$Z_T = \sum_{i \in T} ((N_i - np_i)^2 - N_i).$$

Then

$$\mathbb{E}_p Z_T = n^2 \sum_{i \in T} \Delta_i^2, \quad (31)$$

$$\text{Var}_p Z_T = 2n^2 \sum_{i \in T} p_i^2 + 4n^3 \sum_{i \in T} p_i \Delta_i^2. \quad (32)$$

Proof. For $N \sim \text{Poi}(\lambda)$, the first four factorial moments give

$$\mathbb{E}[(N - a)^2 - N] = (\lambda - a)^2$$

and

$$\text{Var}((N - a)^2 - N) = 2\lambda^2 + 4\lambda(\lambda - a)^2.$$

Set $\lambda = np_i$ and $a = np_i$, then sum independent coordinates. $\square$

Lemma 5 (Tail collision test). *Let $S = \{r+1, \dots, k\}$ and suppose*

$$\sum_{i \in S} q_i^2 \leq \frac{C_0}{L^2 r}, \quad q_{r+1} \leq \frac{C_0}{Lr}.$$

There is a constant $C = C(C_0, c_0)$ such that, using a Poissonized sample with $n = CL\sqrt{r}$, one can distinguish $p = q$ from

$$\sum_{i \in S} (p_i - q_i)^2 \geq \frac{c_0}{L^2 r}$$

with both errors at most $1/12$.

Proof. Put $A = \sum_{i \in S} \Delta_i^2$ and $Q_2 = \sum_{i \in S} q_i^2$. Besides $\sum p_i^2 \leq 2Q_2 + 2A$, we have

$$\begin{aligned} \sum_{i \in S} p_i \Delta_i^2 &\leq \sum_{i \in S} (q_i + |\Delta_i|) \Delta_i^2 \\ &\leq q_{r+1} A + \sum_{i \in S} |\Delta_i|^3 \\ &\leq q_{r+1} A + A^{3/2}. \end{aligned}$$

The last step is monotonicity of finite-dimensional norms.

Let $A_0 = c_0/(L^2 r)$ and reject when $Z_S > n^2 A_0/2$. Under the null, Lemma 4 and $Q_2 = O(A_0)$ show that the rejection probability is $O(C^{-2})$. Under an alternative with $A \geq A_0$, the squared ratio of the standard deviation to the mean is at most

$$O\left(\frac{Q_2 + A}{n^2 A^2} + \frac{q_{r+1}}{nA} + \frac{1}{n\sqrt{A}}\right) = O(C^{-1}). \quad (33)$$

Choosing C sufficiently large and applying Chebyshev's inequality proves the lemma. $\square$

We now assemble the new-regime algorithm. Coarsen q by preserving $1, \dots, r$ and replacing S by a single cell. Call the coarsened null $\bar{q}$. By (27), the tail cell is the largest cell. The general upper bound of Valiant & Valiant (2017, Theorem 2), applied at radius $\delta = c_\kappa/L$ for a sufficiently small fixed c_κ , is at most

$$C_\kappa \max\left\{\delta^{-1}, \frac{\|\bar{q} - \max\|_{2/3}}{\delta^2}\right\} \leq C'_\kappa L\sqrt{r}, \quad (34)$$

where using the untruncated quasi-norm only enlarges the published upper bound. Equation (29) gives the last inequality.

Suppose $d_{TV}(p, q) > \varepsilon$ and $d_{TV}(\bar{p}, \bar{q}) < \delta$. Since

$$2d_{TV}(\bar{p}, \bar{q}) = \sum_{i \leq r} |p_i - q_i| + |p(S) - q(S)|,$$

the total negative discrepancy outside S is below 2δ . Therefore

$$\sum_{i \in S} (q_i - p_i)_+ \geq \varepsilon - 2\delta. \quad (35)$$

By (27), the right side of (35) is at least $q(S) - B'_\kappa/L$. Moreover, $q(\{m+1, \dots, k\}) \leq (\kappa/L) \log(k/m)$, so $k/r \geq K_\kappa e^{x/\kappa}$. For x above a κ -dependent constant, Lemma 1 with $B = B'_\kappa$ gives

$$\sum_{i \in S} (p_i - q_i)^2 \geq \sum_{i \in S} (q_i - p_i)_+^2 \geq \frac{c_\kappa}{L^2 r}. \quad (36)$$

Lemma 5 detects this alternative. Splitting samples and amplifying each component test to error $1/12$ gives total error below $1/3$.

For x below that fixed κ -dependent constant, the untruncated version of the same Valiant–Valiant upper bound uses

$$O_\kappa\left(\frac{\|q\|_{2/3}}{\varepsilon^2}\right) = O_\kappa\left(\frac{\sqrt{k}}{L\varepsilon^2}\right) = O_\kappa\left(L\sqrt{m_q(\varepsilon)} \max\{1, x^{-2}\}\right), \quad (37)$$

because (2) gives $m_q(\varepsilon) = \Theta_\kappa(k)$ whenever x stays bounded.

The argument above uses Poissonized samples only for the collision statistic. The standard comparison between multinomial and Poisson experiments gives a fixed-sample tester at constant overhead. Concretely, given $2n$ observations, draw $M \sim \text{Poi}(n)$ independently. If $M \leq 2n$, use the first M observations. Without the truncation, their cell counts are independent $\text{Poi}(np_i)$ variables. The event $M > 2n$ is exponentially unlikely, so adding it to the error budget changes only the universal constant.

D DETAILS OF THE THINNING LOWER BOUND

We prove the concentration and second-moment statements from Section 5. Recall that $S = \{r + 1, \dots, k\}$, $\pi_i = (r/i)^{1/3}$, $Y_i \sim \text{Bernoulli}(\pi_i)$, and $A_i = Y_i/\pi_i - 1$.

D.1 THE ALTERNATIVES ARE VALID AND FAR

Integral comparison gives

$$\sum_{i \in S} q_i \pi_i \leq \frac{\kappa r^{1/3}}{L} \sum_{i > r} i^{-4/3} \leq \frac{3\kappa}{L}, \quad (38)$$

$$\text{Var}(D) \leq \sum_{i \in S} q_i^2 \pi_i \leq \frac{C_\kappa}{L^2 r}, \quad (39)$$

$$\text{Var}(R) = \sum_{i \in S} q_i^2 \frac{1 - \pi_i}{\pi_i} \leq \sum_{i \in S} \frac{q_i^2}{\pi_i} \leq \frac{C_\kappa}{L^2 r}. \quad (40)$$

The exponents in the last two sums are $7/3$ and $5/3$, respectively. By (27) and (38),

$$\mathbb{E}D = q(S) - \sum_{i \in S} q_i \pi_i \geq \varepsilon + c_\kappa/L. \quad (41)$$

Fix B_κ so that $C_\kappa/B_\kappa^2 \leq 1/400$. Chebyshev's inequality yields

$$\begin{aligned} \mathbb{P}(D \leq \varepsilon) &\leq C_\kappa/r, \\ \mathbb{P}(|R| > B_\kappa/(L\sqrt{r})) &\leq 1/400. \end{aligned}$$

For a κ -dependent r_κ , the event $\mathcal{E}$ in (16) has probability $\alpha \geq 0.99$. Increasing r_κ ensures $B_\kappa/\sqrt{r} < 1/\kappa$, hence $p_1 = q_1 - R > 0$ on $\mathcal{E}$.

For completeness, let U be the positive tail discrepancy. Since $R = U - D$, if $R < 0$ then the total negative discrepancy is D . If $R \geq 0$, the reservoir contributes negative discrepancy R and the total negative discrepancy is $D + R = U$. Thus $d_{\text{TV}}(p, q) \geq D$ in both cases. Every conditioned alternative is valid and more than ε away from q .

D.2 A SCALAR CROSS-MOMENT BOUND

Lemma 6 (Centered Bernoulli multiplier). *Let Y, Y' be independent $\text{Bernoulli}(\pi)$ variables, let $A = Y/\pi - 1$, and let A' be its independent copy. Put $v = (1 - \pi)/\pi$. If*

$$|t| \max\{1, v^2\} \leq \frac{1}{2},$$

then

$$\mathbb{E}e^{tAA'} \leq \exp(Ct^2v^2) \quad (42)$$

for a universal constant C .

Proof. The variable A takes values -1 and v , has mean zero, and satisfies $\mathbb{E}A^2 = v$. Consequently, $X = AA'$ has mean zero, $\mathbb{E}X^2 = v^2$, and $|X| \leq \max\{1, v^2\}$. Taylor's theorem and the assumption give

$$e^{tX} \leq 1 + tX + Ct^2X^2.$$

Taking expectations and using $1 + z \leq e^z$ proves the result. $\square$

Proof of Lemma 2. Let $t_i = nq_i$ and $v_i = (1 - \pi_i)/\pi_i$. If $n \leq c_\kappa L\sqrt{r}$, then for every $i > r$,

$$\begin{aligned} t_i &\leq c_\kappa \kappa \sqrt{r}/i \leq c_\kappa \kappa / \sqrt{r}, \\ t_i v_i^2 &\leq t_i / \pi_i^2 \leq c_\kappa \kappa r^{-1/6} i^{-1/3} \leq c_\kappa \kappa / \sqrt{r}. \end{aligned}$$

Thus Lemma 6 applies for a sufficiently small c_κ and large r . Independence across coordinates gives

$$\begin{aligned} \log \mathbb{E} \exp \left(n \sum_{i \in S} q_i A_i A'_i \right) &\leq C n^2 \sum_{i \in S} q_i^2 v_i^2 \\ &\leq \frac{C_\kappa n^2}{L^2 r^{2/3}} \sum_{i > r} i^{-4/3} \\ &\leq \frac{C'_\kappa n^2}{L^2 r} \leq C'_\kappa c_\kappa^2. \end{aligned}$$

This proves (19). $\square$

D.3 CONDITIONING AND THE FIXED-SAMPLE EXPERIMENT

Let μ be the prior conditioned on $\mathcal{E}$ and let M_n be the resulting mixture law of n independent observations. For two valid alternatives p, p' drawn from μ , the multinomial likelihood-ratio identity is

$$\mathbb{E}_{q^{\otimes n}} L_p L_{p'} = \left(\sum_{i=1}^k \frac{p_i p'_i}{q_i} \right)^n = \left(1 + \sum_{i=1}^k \frac{\Delta_i \Delta'_i}{q_i} \right)^n. \quad (43)$$

The base in (43) is nonnegative. Using $1 + z \leq e^z$, equations (14) and (16) give

$$\begin{aligned} 1 + \chi^2(M_n, q^{\otimes n}) &\leq \mathbb{E}_{p, p' \sim \mu} \exp \left(n \sum_{i \in S} q_i A_i A'_i + n \frac{R R'}{q_1} \right) \\ &\leq \exp \left(\frac{\kappa n B_\kappa^2}{L r} \right) \alpha^{-2} \mathbb{E} \exp \left(n \sum_{i \in S} q_i A_i A'_i \right). \end{aligned} \quad (44)$$

The last expectation is under the unconditioned product prior. The factor α^{-2} follows by dropping two indicators from a nonnegative integrand.

Set $n = c_\kappa L\sqrt{r}$. Lemma 2 and (44) imply

$$1 + \chi^2(M_n, q^{\otimes n}) \leq \exp(c_\kappa \kappa B_\kappa^2 / \sqrt{r}) \alpha^{-2} \exp(C_\kappa c_\kappa^2). \quad (45)$$

Choose r_κ large and then c_κ small. The right side can be made at most 1.05. Hence

$$d_{\text{TV}}(M_n, q^{\otimes n}) \leq \sqrt{\chi^2(M_n, q^{\otimes n})/2} < 1/3.$$

No test can then have both type-I error and mixture-averaged type-II error at most $1/3$. Since every member of the mixture is ε -far, this proves $N^*(q, \varepsilon) = \Omega_\kappa(L\sqrt{r})$. Because $r = \Theta_\kappa(m_q(\varepsilon))$, this is $\Omega_\kappa(L\sqrt{m_q(\varepsilon)})$.

E THE FINE-ACCURACY REGIME AND COMPLETION OF THEOREM 1

We spell out how the standard lower bound completes the phase diagram. The lower bound of Valiant & Valiant (2017, Proposition 2), in the notation of Canonne (2024), is

$$N^*(q, \varepsilon) \geq c \max \left\{ \varepsilon^{-1}, \frac{\Phi_{\text{VV}}(q, c'\varepsilon)}{\varepsilon^2} \right\}, \quad (46)$$

for universal constants $c, c' > 0$. After enlarging r_κ , the corner-case qualification discussed by Canonne (2024) does not arise in the fine regime because $\|q\|_\infty \leq \kappa/L$.

There is a constant $x_\kappa > 0$ such that removing $O(\varepsilon)$ tail mass when $x = \varepsilon L \leq x_\kappa$ retains $\Theta_\kappa(k)$ coordinates. The Zipf envelope therefore gives

$$\Phi_{\text{VV}}(q, c'\varepsilon) = \Theta_\kappa(\sqrt{k}/L),$$

and (46) yields

$$N^*(q, \varepsilon) = \Omega_\kappa \left(\frac{\sqrt{k}}{L\varepsilon^2} \right) = \Omega_\kappa \left(L\sqrt{m_q(\varepsilon)}x^{-2} \right). \quad (47)$$

For $x \geq x_\kappa$, the factor $\max\{1, x^{-2}\}$ is bounded by a κ -dependent constant, and the graded-thinning lower bound $\Omega_\kappa(L\sqrt{m_q(\varepsilon)})$ matches it. These two bounds give the lower half of (3). Equations (36) and (37) give the matching upper half. Enlarging r_κ absorbs rounding and all finitely many cases used in the proof.