THE SAMPLE COST OF ACCURATE MODEL COMPARISON IN TOTAL VARIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing between two models requires identifying which better represents a population. How much data does this decision require? For total variation (TV), a classical selector achieves a factor-3 approximation using a sample budget independent of the number of possible outcomes. We show that improving this factor brings the sample cost to the scale of distance estimation. For two known distributions on N outcomes, every fixed $c \in [1, 3)$ requires $\Omega(N/\log N)$ target samples at sufficiently small fixed additive error. An existing TV estimator supplies a matching upper bound. The lower bound holds for candidates taking just two probability values. We obtain it by extending the construction of Bousquet, Kane, and Moran: a moment-matching transformation keeps the target's preference for one candidate while making the sampled populations hard to distinguish. The transformation stays close to each prescribed starting law and matches L moments with support growing as L^2. Taking L of order $\log N$ yields the sharp sample scale.

1 INTRODUCTION

Suppose two models describe the same population, and a held-out dataset records outcomes from that population. Which model should we choose? This question arises whenever data are used to select among competing probabilistic descriptions. A natural criterion is total variation (TV): it measures the largest error in the probability a model assigns to an event. A model with small TV error predicts the frequency of every event accurately. Comparing models under this criterion therefore asks which one gives the better description of the population across its possible outcomes.

The candidates may come from different model classes or different choices of a tuning parameter. Each is an approximation to the population, and both can have positive error. The aim is to use validation data to compete with the better available candidate. This is the motivation for agnostic hypothesis selection and minimum-distance estimation (Yatracos, 1985; Devroye & Lugosi, 2001).

The task calls for a single choice. This suggests that comparison might need much less data than estimating the discrepancy of each model. A classical selection rule realizes this saving. Given two known models, it considers the event consisting of all outcomes to which the first model assigns higher probability. Estimating the frequency of this one event is enough to select a model whose TV error is at most three times the smaller error, plus an additive sampling term (Yatracos, 1985; Devroye & Lugosi, 2001). The sample budget depends on the additive precision and confidence. It is independent of the number of possible outcomes.

How does this saving change when we ask for a more accurate comparison? For example, how much data does it take to improve the approximation factor from 3 to 2.9 at a fixed additive precision? Estimating both TV distances and choosing the smaller one achieves factor 1. On an alphabet of size N, the optimal worst-case cost of this estimation route is $N/\log N$ at fixed precision (Jiao et al., 2018). These two procedures place model comparison between a budget independent of N and a budget that grows almost linearly with N. The central question is how the sample cost moves between these scales as the comparison becomes more accurate.

Prior work establishes the factor-3 boundary through hard instances for proper hypothesis selection. Bousquet et al. (2019) introduced a pair of simple candidates and target families with opposite model rankings. Aamand et al. (2026, Theorem 3.1 and Corollary 3.2) give a corrected indistinguishability argument at the square-root alphabet scale. A square-root lower bound leaves a polynomial gap

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to the cost of estimating TV. The same quantitative question appears in the statistical evaluability framework of Aiyer et al. (2026a) and is posed explicitly by Aiyer et al. (2026b, Section 1.1).

We determine the sample cost throughout the interval below 3. For every fixed $c \in [1, 3)$ and a fixed additive error in the range of Theorem 1, comparison requires $\Theta_{c,\varepsilon}(N/\log N)$ samples. The lower bound is attained by two candidates with elementary descriptions: each assigns one probability to half the alphabet and a second probability to the other half. Thus the data requirement already arises when the evaluator knows the candidates in full and their structure is simple. The known TV estimator of Jiao et al. (2018) provides the matching upper bound.

The main challenge is to make accurate comparison hard. We need two populations that favor opposite candidates and generate nearly identical sample distributions. Repeated observations provide information about how probability is spread across outcomes. For example, seeing the same outcome twice gives information about squared probabilities. Matching moments of random probability weights aligns these repeat statistics. At the same time, the weights must retain the difference in absolute errors that determines which candidate should be selected. Keeping these two properties together is the central step of our proof.

We start from a weight formulation of the construction of Bousquet et al. (2019). We then perturb each starting law by a small amount in Wasserstein distance, match its moments with those of the other law, and bound the support by $O(L^2)$ for L matching moments. Proximity to the starting laws preserves their model rankings. Moment matching controls the sample distributions through the Poisson-mixture bounds of Wu & Yang (2015). The choice $L \asymp \log N$ balances these two parts and gives the $N/\log N$ scale. Sections 3–5 develop the lower bound, and Section 6 returns to its implications for validation data and the choice of a model.

Connections to prior work. Hypothesis selection studies how approximation quality, data, and computation interact for a finite candidate family. Minimum-distance methods provide the classical guarantees (Devroye & Lugosi, 2001); subsequent work improves the organization of candidate comparisons (Mahalanabis & Štefankovič, 2008; Acharya et al., 2018; Aliakbarpour et al., 2024). Selection and aggregation also admit sharp approximation guarantees (Bousquet et al., 2019; 2021). Fixing the family to two candidates isolates the cost of comparison itself. The remaining statistical complexity comes from the number of possible outcomes and the accuracy required of the choice.

Distribution testing studies decisions about an unknown population from samples (Canonne, 2020). Tolerant testing accepts populations close to a specified reference and rejects those farther away (Parnas et al., 2006); Canonne et al. (2022) characterize how its cost changes with the two distance thresholds. Model comparison also allows the target to differ from both candidates, and its decision is determined by their relative errors. Our construction therefore needs two target families with opposite rankings. This couples the two candidate distances throughout the lower bound.

The estimation benchmark concerns properties of large discrete distributions. Estimating the full probability vector in TV has sample cost of order N at fixed precision (Han et al., 2015). Estimating a distance can use fewer samples: Valiant & Valiant (2011) construct estimators from sample multiplicities, and Jiao et al. (2018) give the minimax TV rates used here. The logarithmic gain over empirical estimation also appears in the data-amplification results of Hao & Orlitsky (2019). Theorem 1 places accurate comparison at this same estimation scale.

2 Model comparison and the main result

Write $[N] = \{1, \dots, N\}$ and let Δ_N denote the probability simplex. For $p, q \in \Delta_N$,

$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{x=1}^{N} |p_x - q_x| = \sup_{A \subseteq [N]} |p(A) - q(A)|.$$

The evaluator receives two complete probability vectors q_0, q_1 and m independent samples from a target $p \in \Delta_N$. The candidates are specified before sampling. The evaluator may use unrestricted computation and internal randomness, and outputs an index $\widehat{i} \in \{0, 1\}$. For $c \geq 1$, additive error $\varepsilon > 0$, and failure probability $\delta \in (0, 1/2)$, its guarantee is

$$\mathbb{P}_p \left\{ \mathrm{TV}(p, q_{\widehat{i}}) \leq c \min_{i \in \{0,1\}} \mathrm{TV}(p, q_i) + \varepsilon \right\} \geq 1 - \delta \tag{1}$$

for every $p, q_0, q_1 \in \Delta_N$. Let $\mathsf{C}_N(c, \varepsilon, \delta)$ be the smallest sample budget satisfying (1). This is proper hypothesis selection with two candidates and an agnostic target. The multiplicative factor c measures the allowed loss relative to the better candidate. The additive term ε sets the resolution of the comparison when that candidate has very small error. We study how the required data grow with N once these accuracy requirements are fixed.

Theorem 1 (The sample cost of accurate comparison)**.** *Fix $c \in [1, 3)$ and set $\varepsilon_c = (3-c)^2/[16(c+5)]$. For every fixed $\varepsilon \in (0, \varepsilon_c]$,*

$$\mathsf{C}_N(c, \varepsilon, 1/3) = \Theta_{c,\varepsilon}\left(\frac{N}{\log N}\right) \qquad \text{as } N \to \infty. \tag{2}$$

On even alphabets, the lower bound holds for candidates taking only two probability values. At factor 3, $\mathsf{C}_N(3, \varepsilon, \delta) = O(\varepsilon^{-2} \log(1/\delta))$, uniformly in N.

The theorem places every fixed improvement below factor 3 at the alphabet-size scale of TV estimation. Table 1 shows the change from the previous lower bound to the matching rate.

Table 1: Sample complexity for two known candidates, at fixed approximation factor, fixed additive precision in the theorem's range, and constant failure probability. The last row gives the estimation benchmark.

Task	Sample bound	Source
Factor-3 selection	$O(1)$	Classical selection
Factor-$c < 3$ selection	$\Omega(\sqrt{N})$	Aamand et al. (2026)
Factor-$c < 3$ selection	$\Theta(N/\log N)$	Theorem 1
TV distance estimation	$\Theta(N/\log N)$	Jiao et al. (2018)

2.1 How one event gives factor three

Define $A = \{x : q_0(x) > q_1(x)\}$ and let $\widehat{p}(A)$ be its empirical frequency. The set A is where q_0 puts more probability than q_1. The classical Scheffé rule compares the observed frequency with the midpoint of the two candidate probabilities: select q_0 when $\widehat{p}(A) \geq (q_0(A) + q_1(A))/2$, and select q_1 otherwise (Yatracos, 1985; Devroye & Lugosi, 2001).

The factor 3 follows from the geometry of these three distributions. Put $D = \mathrm{TV}(q_0, q_1) = q_0(A) - q_1(A)$. If the selected candidate q_b has larger target distance than the other candidate q_g, the threshold condition gives

$$D \leq 2|\widehat{p}(A) - q_g(A)| \leq 2\,\mathrm{TV}(p, q_g) + 2|\widehat{p}(A) - p(A)|.$$

The triangle inequality then implies

$$\mathrm{TV}(p, q_b) \leq \mathrm{TV}(p, q_g) + D \leq 3\,\mathrm{TV}(p, q_g) + 2|\widehat{p}(A) - p(A)|.$$

Hoeffding's inequality makes the last term at most ε with probability $1 - \delta$ for $m \geq 2\varepsilon^{-2} \log(2/\delta)$. Estimating the frequency of A therefore suffices, regardless of how many outcomes it contains.

For our two-level candidates, A is one half of the alphabet. Its frequency measures the total mass on that half. TV also adds up the coordinatewise absolute errors inside the half. Positive and negative errors can cancel in the total mass while both contribute to TV. The construction in Section 3 uses this difference to preserve nearly equal half-masses while reversing the candidates' ranking.

2.2 The matching upper bound from TV estimation

Let $\widehat{d}_i$ estimate $d_i = \mathrm{TV}(p, q_i)$ to error $\varepsilon/2$. Selecting the smaller estimate gives

$$d_{\widehat{i}} \leq \widehat{d}_{\widehat{i}} + \varepsilon/2 \leq \widehat{d}_{i^*} + \varepsilon/2 \leq d_{i^*} + \varepsilon, \qquad i^* \in \arg\min_i d_i.$$

The known-reference estimator of Jiao et al. (2018, Theorem 2) has worst-case mean squared error $O(N/(n \log n))$ from two independent Poisson samples of mean n in its large-alphabet regime. Set $n = K_\varepsilon N/\log N$. Then $\log n \asymp \log N$, and choosing K_ε large enough makes the probability that

either estimate errs by more than $\varepsilon/2$ at most $1/6$. A union bound gives this guarantee when both estimates use the same two samples.

To implement the estimator with a fixed budget $m = \lceil 4n \rceil$, draw independent $K_1, K_2 \sim \mathrm{Poi}(n)$. If $K_1 + K_2 \le m$, take consecutive blocks of these lengths from the m observations; otherwise return q_0. Coupling to an infinite IID sequence shows that this simulation agrees with the Poisson experiment except with probability at most

$$\mathbb{P}\{\mathrm{Poi}(2n) > m\} \le \exp\{-2n(2\log 2 - 1)\}.$$

For large N this is at most $1/6$, so the total failure probability is at most $1/3$. The fixed budget is $O_\varepsilon(N/\log N)$, proving the upper bound in (2).

3 A STARTING CONSTRUCTION WITH OPPOSITE RANKINGS

To prove the lower bound, we construct two random target populations, called world $+$ and world $-$. A valid selector must choose q_0 in the first world and q_1 in the second. We then show that the sample laws from these worlds are close in TV. This turns successful model comparison into successful testing between two almost indistinguishable experiments, using the two-prior form of Le Cam's method (Yu, 1997; Cai & Low, 2011).

Each world is a prior over target distributions: we first draw one target and then draw all observations independently from that target. The candidates stay fixed across both priors. Because the selection guarantee holds for every target, it also holds after averaging over either prior. A selector can therefore be used as a test by interpreting its chosen candidate as the world label. Figure 1 summarizes the construction.

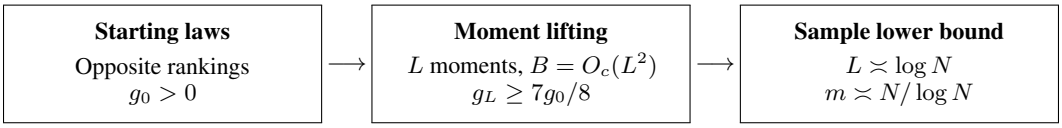

Figure 1: The proof keeps a positive ranking gap while increasing the number of matching moments. The starting laws adapt the construction of Bousquet et al. (2019); the moment lifting controls both the ranking gap and the support size. These two controls determine the sample scale.

3.1 THE INHERITED EXAMPLE IN PROBABILITY WEIGHTS

We use a mean-one weight formulation of the construction of Bousquet et al. (2019, Appendix C, Theorem 22). Their candidates exchange high and low probability levels, and their targets replace some ordinary weights by spikes and zeros. The laws below express this construction through the weight of a coordinate. The block construction of Aamand et al. (2026, Section 3) develops the same starting idea and supplies a corrected sample indistinguishability proof.

Fix $\beta \in (0, 1/2]$ and write $a = 1 - \beta$, $b = 1 + \beta$. For even N, let q_0 assign a/N to the first half of the alphabet and b/N to the second half; q_1 exchanges these values. Define random weights

$$U_0 = \begin{cases} a & \text{with probability } 1/b, \\ 2 & \text{with probability } \beta/b, \end{cases} \qquad V_0 = \begin{cases} b & \text{with probability } 1/b, \\ 0 & \text{with probability } \beta/b. \end{cases} \tag{3}$$

Each law has mean 1 and second moment $1 + \beta$. Most U_0 weights agree with the lower candidate value, while most V_0 weights agree with the higher value. The spikes at 2 and the zeros balance their means. Thus both laws have the same average weight even though their ordinary coordinates favor different candidate values.

Place U_0 on the first half and V_0 on the second. The expected half-ℓ_1 discrepancies of the raw weights divided by N are

$$r_{\mathrm{good}} = \frac{\mathbb{E}|U_0 - a| + \mathbb{E}|V_0 - b|}{4} = \frac{\beta}{2}, \tag{4}$$

$$r_{\mathrm{bad}} = \frac{\mathbb{E}|U_0 - b| + \mathbb{E}|V_0 - a|}{4} = \frac{\beta(3 - \beta)}{2(1 + \beta)}. \tag{5}$$

The factor $1/4$ combines the factor $1/2$ in TV with the half of the coordinates drawn from each law. Swapping the two weight laws swaps these discrepancies, so the preferred candidate changes between worlds. Their ratio is $(3-\beta)/(1+\beta)$, which approaches 3 as β decreases.

For the factor c in Theorem 1, choose

$$\beta = \frac{3-c}{2(c+1)}, \qquad g_0 = r_{\text{bad}} - c r_{\text{good}} = \frac{(3-c)^2}{4(c+5)} > 0. \tag{6}$$

The quantity g_0 measures how much the worse candidate exceeds the factor-c allowance. Maintaining this gap after moment matching and normalization will force the selector to identify the world.

For example, take $c = 2$. The choice in (6) gives $\beta = 1/6$, $a = 5/6$, and $b = 7/6$, with

$$r_{\text{good}} = \frac{1}{12}, \qquad r_{\text{bad}} = \frac{17}{84}, \qquad r_{\text{bad}} - 2r_{\text{good}} = \frac{1}{28}.$$

The worse candidate thus exceeds twice the better candidate's discrepancy by a fixed amount. Exchanging the weight laws exchanges which candidate has this excess error. The next step is to keep enough of this margin while matching more moments.

3.2 Repeated observations and matching moments

Repeated observations are a basic source of information in distribution testing. Collision statistics underlie uniformity tests (Paninski, 2008), and sample multiplicities play a central role in testing symmetric properties (Valiant, 2011). A paired form of our example makes the corresponding moment calculation visible. Partition the alphabet into $N/2$ pairs, each with one coordinate from each half. Draw an independent U_0 weight for each pair and assign probabilities $(U_0, 2-U_0)/N$. World $-$ exchanges the coordinates in each pair. Every pair has total probability $2/N$, so these are normalized target distributions. The law of $2-U_0$ is exactly the law of V_0.

Conditional on visiting a given pair k times, the probability of seeing its first coordinate j times is

$$\mathbb{E}\left[\binom{k}{j}(U_0/2)^j(1-U_0/2)^{k-j}\right].$$

This expression is the expectation of a degree-k polynomial in the weight. Since U_0 and V_0 match their first two moments, one or two visits produce exactly the same within-pair observations in both worlds. A third visit can reveal their difference, as Figure 2 shows. Proposition 5 gives the joint observation law and the corresponding distance calculations.



Figure 2: The seed laws adapted from Bousquet et al. (2019), at $\beta = 0.05$. (a) Their first two moments agree. (b) Assigning the laws to the two halves produces expected good and bad discrepancies with ratio 2.81; swapping the assignment reverses the ranking. (c) The paired construction has identical observation laws through two visits and different laws at three. Section 4 lifts the moment order while preserving the ranking gap.

This calculation suggests matching more moments to conceal more repeat statistics. The ranking depends on expectations of absolute-value losses, such as $|U-a|$. On a fixed bounded interval, polynomials of increasing degree approximate these losses increasingly well, so matching many

moments also brings their expectations together. We address this tension by enlarging the weight interval while keeping each law close to its starting distribution. Section 4 makes this tradeoff quantitative.

The larger weights are accommodated by sampling coordinates independently and normalizing their total mass. The paired example provides the repeat count interpretation; the independent construction supports the growing interval. Section 5 controls the normalization through a concentration bound.

4 Matching moments while preserving the ranking

Our goal is to change the moments of the two weight laws while keeping their four absolute-error expectations close to (4) and (5). Wasserstein distance supplies the required control. For laws on the real line,

$$W_1(\mu, \nu) = \inf \mathbb{E}|X - Y|,$$

where the infimum ranges over couplings with marginals μ, ν. Moving weights by average distance at most τ changes the expectation of each 1-Lipschitz loss by at most τ. In particular, this applies to the losses $|x - a|$ and $|x - b|$ that determine the model ranking.

Small transport cost is compatible with a growing support. An atom of probability w at a large location R contributes wR to the first moment and wR^k to the kth moment. A small amount of probability at large locations can therefore change higher moments substantially while its effect on absolute-error expectations remains small. The construction below uses this freedom to correct the moments of the starting pair. Its transport bound accounts for both the added probability and the rescaling needed to recover mean one.

Lemma 2 (Moment matching near a prescribed pair)**.** *Fix $0 < \ell \le h < \infty$. Suppose U_0, V_0 have mean 1 and are supported on $\{0\} \cup [\ell, h]$. For every integer $L \ge 2$ and $\tau \in (0, 1]$, there are nonnegative U, V such that*

1. *$\mathbb{E}U = \mathbb{E}V = 1$ and $\mathbb{E}U^k = \mathbb{E}V^k$ for $k = 0, \dots, L$;*
2. *$W_1(\mathcal{L}(U), \mathcal{L}(U_0)) \le \tau$ and $W_1(\mathcal{L}(V), \mathcal{L}(V_0)) \le \tau$;*
3. *both variables are supported on $[0, C_{\ell,h} L^2/\tau]$.*

The lemma preserves the individual starting laws while aligning their moments. This allows us to carry the seed's comparison gap through the transformation. Its ingredients come from polynomial methods for statistical lower bounds: size biasing and inverse weighting appear in Wu & Yang (2015, Lemma 5), and the signed polynomial-functional construction follows the approach of Carpentier & Verzelen (2017, Lemma 3). Canonne et al. (2022, Appendix C) use related duality to construct asymmetric distance gaps for tolerant testing. Here we control the transport from each prescribed law, which retains the two candidates' opposite rankings.

The connection between polynomial approximation and moment matching underlies lower bounds for nonsmooth and discrete functionals (Cai & Low, 2011; Jiao et al., 2015; Wu & Yang, 2016; 2020). For comparison, the transport guarantee preserves the expectations of both losses $|x - a|$ and $|x - b|$ under each starting law. This simultaneous control of four expectations carries the opposite rankings through the transformation.

4.1 Why a quadratic support suffices

Size biasing gives probability laws $S(dx) = x\mathcal{L}(U_0)(dx)$ and $T(dx) = x\mathcal{L}(V_0)(dx)$ on $[\ell, h]$. It shifts each positive moment down by one degree. On a larger interval $[\ell, R]$, Markov's derivative inequality (DeVore & Lorentz, 1993) bounds the difference of these moment functionals: for every polynomial q of degree at most $L - 1$,

$$|\mathbb{E}_T q - \mathbb{E}_S q| \le \frac{2(L-1)^2 W_1(S, T)}{R - \ell} \|q\|_{\infty,[\ell,R]}. \tag{7}$$

Thus an interval of length proportional to L^2/τ makes the functional small. Hahn–Banach extension and Riesz representation produce a signed measure with the required moment differences and

small total mass. Its positive and negative parts supply positive corrections to the two laws. Inverse weighting, mass at zero, and a common rescaling then recover probability laws of mean one. Lemma 6 gives the construction and proves that the total transport cost is at most τ.

4.2 THE RANKING SURVIVES THE TRANSFORMATION

Apply the lemma to the seed with $[\ell, h] = [a, 2]$ and $\tau = g_0/[4(c+1)]$. Denote the new expected good and bad discrepancies by $r^{(L)}_{\text{good}}$ and $r^{(L)}_{\text{bad}}$. Each changes by at most $\tau/2$, because it is a sum of two absolute-error expectations divided by four. Consequently,

$$g_L := r^{(L)}_{\text{bad}} - c r^{(L)}_{\text{good}} \geq g_0 - \frac{c+1}{2}\tau = \frac{7}{8} g_0. \tag{8}$$

The laws have support in $[0, B]$ with $B \leq C_c L^2$. We have therefore obtained all three properties needed by the lower bound: mean one, matching moments through degree L, and a ranking gap bounded away from zero as L grows.

A finite illustration of this transformation appears in Appendix D. At $\beta = 0.05$, increasing the moment order from 2 to 20 keeps the ratio of bad to good discrepancies above 2.80. Table 2 illustrates the change in both candidate discrepancies and their ratio as the moment order increases.

5 FROM MATCHING MOMENTS TO THE SAMPLE LOWER BOUND

We now place the constructed weights on the alphabet and analyze the observations. Matching moments controls each coordinate's count law; concentration ensures that the normalized target retains the ranking gap. Together these facts make comparison require almost-linear data.

5.1 MOMENT MATCHING CONTROLS THE OBSERVED COUNTS

For a random weight U, write $\text{MixPoi}(\lambda U)$ for the law obtained by drawing U and then a Poisson count of mean λU. The link between matching moments and Poisson-mixture indistinguishability is established by Wu & Yang (2015, Appendix B, Lemma 6). For mixing variables supported on $[0, B]$ and matching L moments, they give

$$\text{TV}\big(\text{MixPoi}(\lambda U), \text{MixPoi}(\lambda V)\big) \leq \left(\frac{e\lambda B}{2L}\right)^L.$$

A thinning representation expresses the comparison through a Poisson tail and connects it directly to the repeat-count example.

Lemma 3 (A thinning representation for matched moments)**.** *If $U, V \in [0, B]$ match moments through degree L, then*

$$\text{TV}\big(\text{MixPoi}(\lambda U), \text{MixPoi}(\lambda V)\big) \leq \mathbb{P}\{\text{Poi}(\lambda B) > L\} \leq \left(\frac{e\lambda B}{L+1}\right)^{L+1} \tag{9}$$

for $0 < \lambda B \leq L + 1$.

For intuition, draw a number of potential visits $H \sim \text{Poi}(\lambda B)$, then retain each visit with probability U/B. Conditional probabilities for $H \leq L$ are polynomials of degree at most L in the weight, so the matched moments make them agree. Only the tail $H > L$ contributes to the bound. Lemma 8 gives the full proof.

5.2 INDEPENDENT WEIGHTS GIVE NORMALIZED TARGETS WITH OPPOSITE RANKINGS

In world $+$, draw independent weights W_x from U on the first half of the alphabet and from V on the second half. In world $-$, exchange the laws. Define

$$T_N = \frac{1}{N}\sum_x W_x, \qquad p_x = \frac{W_x}{N T_N} \quad (T_N > 0), \tag{10}$$

and set $p = q_0$ when $T_N = 0$. This produces a prior on probability vectors in each world. Their mass and raw discrepancies concentrate because the weights are independent, bounded by B, and have mean one.

More explicitly, put

$$D_i = \frac{1}{2N} \sum_x |W_x - N q_i(x)|.$$

In world $+$, their expectations are $r^{(L)}_{\text{good}}$ and $r^{(L)}_{\text{bad}}$, respectively. To control their fluctuations, use mean one and $0 \le W_x \le B$ to obtain $\mathbb{E}W_x^2 \le B$. For either anchor $v \in \{a, b\}$,

$$\mathbb{E}|W_x - v|^2 = \mathbb{E}W_x^2 - 2v + v^2 \le B + b^2.$$

Independence then gives

$$\mathrm{Var}(T_N) \le B/N, \qquad \mathrm{Var}(D_i) \le (B + b^2)/(4N).$$

On $T_N > 0$, normalization changes each discrepancy by at most

$$|\,\mathrm{TV}(p, q_i) - D_i| \le \frac{1}{2} \sum_x \left| \frac{W_x}{N T_N} - \frac{W_x}{N} \right| = \frac{|T_N - 1|}{2}. \tag{11}$$

For $L \asymp \log N$, we have $B = O_c(\log^2 N)$, so the three quantities T_N, D_0, D_1 concentrate around their expectations. To see how much error the ranking can tolerate, take $t = g_0/[16(c+1)]$. With probability tending to one,

$$|T_N - 1| \le t, \qquad |D_0 - r^{(L)}_{\text{good}}| \le t, \qquad |D_1 - r^{(L)}_{\text{bad}}| \le t.$$

The total-mass bound ensures $T_N > 0$. Combining it with (11), each actual TV distance is within $3t/2$ of its corresponding expected raw discrepancy. Hence, in world $+$,

$$\mathrm{TV}(p, q_1) - c\,\mathrm{TV}(p, q_0) \ge \frac{7}{8} g_0 - \frac{3}{2}(c+1)t = \frac{25}{32} g_0 > \varepsilon,$$

since $\varepsilon \le \varepsilon_c = g_0/4$. Thus q_0 is the unique acceptable choice. The same calculation in world $-$ makes q_1 the unique acceptable choice. We next bound the information in the observations.

5.3 Choosing the moment order and sample budget

Consider independent counts $C_x \mid W_x \sim \mathrm{Poi}(2mW_x/N)$. After averaging over the weights, the two count laws $\mathsf{Q}_+, \mathsf{Q}_-$ are products of Poisson mixtures. Applying Lemma 3 coordinate by coordinate gives

$$\mathrm{TV}(\mathsf{Q}_+, \mathsf{Q}_-) \le N \left(\frac{2emB}{N(L+1)} \right)^{L+1}. \tag{12}$$

There are two costs in this expression. Combining N coordinates contributes the leading factor N. Enlarging the weight support raises the Poisson intensities through B. The matching order L supplies the power that counteracts these costs.

For a desired count-law TV bound $\zeta \in (0, 1)$, equation (12) is at most ζ when

$$m \le \frac{L+1}{2eB} N \left(\frac{\zeta}{N} \right)^{1/(L+1)}. \tag{13}$$

A fixed moment order gives the scale $N^{L/(L+1)}$. Taking $L \asymp \log N$ makes $N^{-1/(L+1)}$ a constant, while the support bound $B = O_c(L^2)$ gives $(L+1)/B = \Omega_c(1/\log N)$. These two effects yield the scale $N/\log N$.

Concretely, set $L = \lceil 2 \log N \rceil$ and $m = \lfloor k_c N / \log N \rfloor$. The support bound makes the parenthesized term in (12) at most

$$\frac{2ek_cC_cL^2}{(L+1)\log N} \le 6ek_cC_c$$

for large N, because $L \le 3 \log N$. Thus the choice $k_c = 1/(24eC_c)$ bounds this term by $1/4$ and also ensures the Poisson-tail condition $2mB/N \le L + 1$. It follows that

$$\mathrm{TV}(\mathsf{Q}_+, \mathsf{Q}_-) \le N\, 4^{-(L+1)} \le \frac{1}{4} N^{1-2\log 4} = o(1).$$

This calculation identifies the source of the logarithmic denominator. The logarithmic moment order controls the observations across all N coordinates. With this choice, the quadratic support bound gives a sample budget proportional to N/L. Meanwhile, $L^2/N \to 0$ keeps the normalization and risk fluctuations small. The same choice of L therefore satisfies both the observation and ranking requirements.

5.4 Returning to a fixed sample of observations

The factor 2 in the Poisson intensities provides enough observations to extract a fixed sample. Conditional on the weights, the total count $K = \sum_x C_x$ is Poisson with mean $2mT_N$. Uniformly order the observed symbols and take the first m when $K \ge m$; otherwise return a fixed sequence. Conditional on the weights and on $K = k \ge m$, those first m symbols have law $p^{\otimes m}$.

Lemma 9 bounds the error of this extraction, averaged over each original prior, by

$$\mathbb{P}(T_N < 3/4) + e^{-m/12} = o(1).$$

The rule uses only the observed counts and therefore defines the same transformation in both worlds. TV contracts under this transformation, so the two fixed-sample laws also have TV distance $o(1)$.

A selector satisfying (1) would identify the world with average success at least $2/3 - o(1)$: choosing the acceptable model reveals whether the target came from world $+$ or world $-$. The optimal test between two equally likely laws at TV distance $o(1)$ has success at most $1/2 + o(1)$. This contradiction proves the lower bound. Padding the even construction with one zero-probability outcome handles odd alphabets. Theorem 10 records the constants and exceptional-event bounds.

Combining this lower bound with the estimation upper bound in Section 2 proves Theorem 1.

6 Conclusion and implications for model comparison

Return to the two models and the held-out dataset in the introduction. Theorem 1 identifies the data required to choose a model with a TV guarantee for every target distribution. For each fixed $c \in [1, 3)$ and additive precision in the theorem's range, the worst-case sample cost is of order $N/\log N$. The lower bound already holds for fully known candidates with two probability values. The target's errors within each half of the alphabet determine which candidate is better, and observations must reveal enough of these errors to resolve the choice. Even a decision between two models can therefore require the same order of data as estimating their TV distances.

The multiplicative factor and additive precision specify the strength of this validation guarantee together. Our lower bound applies for $\varepsilon \le \varepsilon_c = (3-c)^2/[16(c+5)]$. For the introductory example $c = 2.9$, this gives $\varepsilon \le 1/12640 \approx 7.91 \times 10^{-5}$. The theorem fixes these accuracy requirements and then characterizes growth with the number of outcomes; the constants in the rate depend on the chosen accuracy. At factor 3, a single event suffices for a guarantee whose sample cost is independent of N. Resolving finer differences between the candidates brings the distribution of errors across outcomes into the comparison.

Choosing an existing model is also part of the task. It describes validation used to decide which fitted model to deploy. When the output can be a newly constructed distribution, aggregation achieves factor 2 with a sample bound independent of N for a fixed candidate family, precision, and confidence (Bousquet et al., 2019; 2021). These results make the permitted output a consequential part of the validation problem. For a larger candidate family, a natural next question is how its size interacts with the outcome space and the requested accuracy. The two-candidate result identifies a source of data demand already present in a single comparison: recovering which model better describes probability mass spread across the outcome space.

AI USE STATEMENT

We used an LLM research assistant to help formulate the research problem, develop proofs, and implement the numerical constructions. It also assisted with literature search and comparison, interpretation of the numerical results, and manuscript drafting. The authors are responsible for the mathematical results, references, code, figures, and text in this paper.

REFERENCES

Anders Aamand, Maryam Aliakbarpour, Justin Y. Chen, and Sandeep Silwal. How fast can you find a good hypothesis?, 2026. arXiv:2509.03734v3.

Jayadev Acharya, Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Maximum selection and sorting with adversarial comparators. *Journal of Machine Learning Research*, 19(59):1–31, 2018.

Shashaank Aiyer, Yishay Mansour, Shay Moran, and Han Shao. A theoretical framework for statistical evaluability of generative models, 2026a.

Shashaank Aiyer, Yishay Mansour, Shay Moran, Han Shao, and Tom Waknine. Scale-sensitive shattering: Learnability and evaluability at optimal scale, 2026b.

Maryam Aliakbarpour, Mark Bun, and Adam Smith. Optimal hypothesis selection in (almost) linear time. In *Advances in Neural Information Processing Systems*, volume 37, pp. 141490–141527, 2024.

Olivier Bousquet, Daniel Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Proceedings of the Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 318–341, 2019.

Olivier Bousquet, Mark Braverman, Klim Efremenko, Gillat Kol, and Shay Moran. Statistically near-optimal hypothesis selection. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pp. 909–919, 2021.

T. Tony Cai and Mark G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

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.

Clément L. Canonne, Ayush Jain, Gautam Kamath, and Jerry Li. The price of tolerance in distribution testing. In *Proceedings of the Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 573–624, 2022.

Alexandra Carpentier and Nicolas Verzelen. Adaptive estimation of the sparsity in the gaussian vector model, 2017. arXiv:1703.00167.

Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1993.

Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

Yi Hao and Alon Orlitsky. Data amplification: Instance-optimal property estimation, 2019. arXiv:1903.01432.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the L_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

Satyaki Mahalanabis and Daniel Štefankovič. Density estimation in linear time. In *Proceedings of the Conference on Learning Theory*, pp. 503–512, 2008.

Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of the IEEE Symposium on Foundations of Computer Science*, pp. 403–412, 2011.

Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6): 1927–1968, 2011.

Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen, 2015. arXiv:1504.01227.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Yihong Wu and Pengkun Yang. Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends in Communications and Information Theory*, 17(4):402–585, 2020.

Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

A Details of the starting construction

For probability measures, TV is one half of the L_1 distance. For a signed measure ν, $\|\nu\|_{\mathrm{var}}$ denotes its total absolute mass. Wasserstein distance W_1 uses absolute-value transport cost. All logarithms are natural.

A.1 Seed laws and their comparison gap

The following laws express the construction of Bousquet et al. (2019, Appendix C) in mean-one coordinate weights. We record their moments, discrepancies, and size-biased transport cost for use in the moment transformation.

Lemma 4 (Seed laws). *For $0 < \beta < 1$, put $a = 1 - \beta$ and $b = 1 + \beta$. Let*

$$\mathcal{L}(U_0) = \frac{1}{b}\delta_a + \frac{\beta}{b}\delta_2, \qquad \mathcal{L}(V_0) = \frac{1}{b}\delta_b + \frac{\beta}{b}\delta_0. \tag{14}$$

Both variables have mean one and second moment b, and V_0 has the same law as $2 - U_0$. Define

$$A_0 = \frac{\mathbb{E}|U_0 - a| + \mathbb{E}|V_0 - b|}{4}, \qquad B_0 = \frac{\mathbb{E}|U_0 - b| + \mathbb{E}|V_0 - a|}{4}.$$

Then

$$A_0 = \frac{\beta}{2}, \qquad B_0 = \frac{\beta(3-\beta)}{2(1+\beta)}. \tag{15}$$

The size-biased laws S, T of U_0, V_0 satisfy

$$S = \frac{a}{b}\delta_a + \frac{2\beta}{b}\delta_2, \qquad T = \delta_b, \qquad W_1(S,T) = \frac{4a\beta}{b} \le 4\beta. \tag{16}$$

Proof. The probabilities in each law sum to one. Their means are $(a + 2\beta)/b = 1$ and $b/b = 1$. Their second moments are $(a^2 + 4\beta)/b = b$ and $b^2/b = b$. Reflection about one maps the atoms and probabilities of the first law to those of the second. Moreover,

$$\mathbb{E}|U_0 - a| = \mathbb{E}|V_0 - b| = \beta, \qquad \mathbb{E}|U_0 - b| = \mathbb{E}|V_0 - a| = \frac{\beta(3-\beta)}{b},$$

which proves (15). Size bias gives the displayed S, T. Since T is a point mass,

$$W_1(S,T) = \frac{a}{b}(b - a) + \frac{2\beta}{b}(2 - b) = \frac{4a\beta}{b}.$$

□

A.2 The paired observation example

The two-moment seed also admits an exactly normalized construction. It illustrates the role of matching observations from repeated visits to the same part of the domain.

Proposition 5 (An exactly normalized pair construction). *Let $N = 2M$, and fix a, b, β as in Lemma 4. On every coordinate pair j, let q_0 assign probabilities $(a, b)/N$, and let q_1 assign $(b, a)/N$. Draw $W_1, \ldots, W_M$ independently with law $\mathcal{L}(U_0)$. Define random distributions p^+, p^- by assigning probabilities*

$$p^+|_j = (W_j, 2 - W_j)/N, \qquad p^-|_j = (2 - W_j, W_j)/N.$$

Let $P_+^{(m)}, P_-^{(m)}$ be the corresponding mixtures of ordered m-sample IID laws. Then

$$\mathrm{TV}(P_+^{(m)}, P_-^{(m)}) \le \binom{m}{3} M^{-2}. \tag{17}$$

If $T = \#\{j : W_j = 2\}$, then $T \sim \mathrm{Bin}(M, \beta/b)$ and

$$\mathrm{TV}(p^+, q_0) = \frac{bT}{N}, \qquad \mathrm{TV}(p^+, q_1) = \beta + \frac{(1 - 3\beta)T}{N}. \tag{18}$$

The distances are interchanged for p^-.

Proof. Every pair has mass $2/N = 1/M$ under either distribution, for every latent draw. Thus the sampled pair labels are IID uniform on $\{1, \dots, M\}$, independently of the latent variables. Given the pair labels, the probability of any ordered sequence of t within-pair labels is a degree-t polynomial in W_j. For $t \le 2$, its expectation is unchanged when W_j is replaced by $2 - W_j$, because the first two moments match. Independence across pairs therefore makes the two integrated sample laws identical on the event that no pair is visited three times. This event has the same probability in both worlds. A union bound over triples of observation indices bounds its complement by $\binom{m}{3}/M^2$, proving (17).

A pair with $W_j = a$ contributes zero to $\mathrm{TV}(p^+, q_0)$ and $2\beta/N$ to $\mathrm{TV}(p^+, q_1)$. A pair with $W_j = 2$ contributes b/N and a/N, respectively. Summing proves (18); swapping the two coordinates swaps the candidate distances. □

B Moment matching and preservation of the ranking

B.1 A moment-smoothing lemma

The construction uses size biasing and inverse weighting as in Wu & Yang (2015, Lemma 5), together with the signed-functional method of Carpentier & Verzelen (2017, Lemma 3). We track the transport from each starting law through correction and normalization.

Lemma 6 (Moment smoothing)**.** *Let $0 < \ell \le h < \infty$, and let U, V be nonnegative random variables supported on $\{0\} \cup [\ell, h]$, with $\mathbb{E}U = \mathbb{E}V = 1$. Let S, T be their size-biased probability measures:*

$$S(dx) = x\,\mathcal{L}(U)(dx), \qquad T(dx) = x\,\mathcal{L}(V)(dx),$$

and put $D = W_1(S, T)$. Fix an integer $L \ge 1$, a number $\eta > 0$, and $R \ge h$ with $R > \ell$ such that

$$\frac{2(L-1)^2 D}{R - \ell} \le \eta. \tag{19}$$

There exist nonnegative random variables U_L, V_L satisfying

$$\mathbb{E}U_L = \mathbb{E}V_L = 1, \qquad \mathbb{E}U_L^k = \mathbb{E}V_L^k \quad (0 \le k \le L),$$

whose supports are contained in

$$\left[0, R\left(1 + \frac{\eta}{2\ell}\right)\right],$$

and for which

$$W_1\big(\mathcal{L}(U_L), \mathcal{L}(U)\big) \le \eta \max\{1, \ell^{-1}\}, \qquad W_1\big(\mathcal{L}(V_L), \mathcal{L}(V)\big) \le \eta \max\{1, \ell^{-1}\}. \tag{20}$$

Proof. Let $\mathcal{P}_{L-1}$ be the space of real polynomials of degree at most $L - 1$, equipped with the supremum norm on $[\ell, R]$. Define a linear functional on this space by

$$\mathcal{F}(q) = -\int q\, d(S - T).$$

For every $q \in \mathcal{P}_{L-1}$, the coupling definition of W_1 and Markov's polynomial inequality give

$$|\mathcal{F}(q)| \le D\|q'\|_{\infty,[\ell,R]} \le \frac{2(L-1)^2 D}{R - \ell}\|q\|_{\infty,[\ell,R]} \le \eta\|q\|_{\infty,[\ell,R]}.$$

Here the factor $2/(R - \ell)$ comes from mapping $[\ell, R]$ to $[-1, 1]$. The Hahn–Banach theorem extends $\mathcal{F}$ to $C([\ell, R])$ with norm at most η. The Riesz representation theorem then provides a real signed measure ν on $[\ell, R]$ such that

$$\|\nu\|_{\mathrm{var}} \le \eta, \qquad \int x^k\, d\nu = -\int x^k\, d(S - T) \quad (0 \le k \le L - 1). \tag{21}$$

Since S, T are probability measures, $\mathcal{F}(1) = 0$. Thus the Jordan parts ν^+, ν^- have a common mass $t \le \eta/2$.

Write $\mu_U = \mathcal{L}(U)$, $\mu_V = \mathcal{L}(V)$, and define positive measures

$$\alpha_+(dx) = x^{-1}\nu^+(dx), \qquad \alpha_-(dx) = x^{-1}\nu^-(dx).$$

They are well defined because $x \geq \ell$ on the support of ν. For $1 \leq k \leq L$, equation (21) yields

$$\int x^k \, d(\mu_U + \alpha_+ - \mu_V - \alpha_-) = \int x^{k-1} \, d(S - T + \nu) = 0.$$

Each $\alpha_\pm$ has mass at most t/ℓ and first moment exactly t. Add mass at zero to whichever of $\mu_U + \alpha_+$ and $\mu_V + \alpha_-$ has smaller total mass. Denote the resulting positive measures by $\bar{\mu}_U, \bar{\mu}_V$. They have common mass Z, common first moment H, and satisfy

$$1 \leq Z \leq 1 + \frac{\eta}{2\ell}, \qquad H = 1 + t \leq 1 + \frac{\eta}{2}. \tag{22}$$

They also match every moment of order zero through L.

Let Y_U, Y_V have probability laws $\bar{\mu}_U/Z, \bar{\mu}_V/Z$, respectively, and set

$$U_L = \frac{Z}{H} Y_U, \qquad V_L = \frac{Z}{H} Y_V.$$

These variables have mean one and the required moment equalities. Their supports are bounded above by RZ/H, which is at most the claimed bound because $H \geq 1$.

It remains to bound the transport cost. The positive measure $\bar{\mu}_U - \mu_U$ has mass $Z - 1$ and first moment $H - 1$. Couple the common component μ_U/Z identically and couple the remaining components independently. Nonnegativity gives

$$W_1\big(\mathcal{L}(Y_U), \mu_U\big) \leq \frac{Z - 1 + H - 1}{Z}.$$

This bound is also valid when $Z = 1$, in which case there is no added mass. Coupling Y_U with its rescaling gives

$$W_1\big(\mathcal{L}(U_L), \mathcal{L}(Y_U)\big) \leq \left| \frac{Z}{H} - 1 \right| \mathbb{E} Y_U = \frac{|Z - H|}{Z}.$$

Consequently,

$$W_1\big(\mathcal{L}(U_L), \mu_U\big) \leq \frac{2 \max\{Z - 1, H - 1\}}{Z} \leq \eta \max\{1, \ell^{-1}\}.$$

The same argument applies to V_L. □

B.2 Applying the transformation to the seed

Corollary 7 (A separated pair with arbitrarily many matched moments)**.** *Fix $c \in [1, 3)$, and define*

$$\beta = \frac{3 - c}{2(c + 1)}, \qquad a = 1 - \beta, \qquad b = 1 + \beta, \qquad g = \frac{(3 - c)^2}{4(c + 5)}. \tag{23}$$

There is a finite constant $C_c \geq 1$ such that, for every integer $L \geq 2$, there are random variables $U_L, V_L \in [0, C_c L^2]$ with mean one and matching moments of orders zero through L. For

$$A_L = \frac{\mathbb{E}|U_L - a| + \mathbb{E}|V_L - b|}{4}, \qquad B_L = \frac{\mathbb{E}|U_L - b| + \mathbb{E}|V_L - a|}{4},$$

they satisfy

$$B_L - c A_L \geq \frac{7}{8} g. \tag{24}$$

Proof. For these choices, $0 < \beta \leq 1/2$, $1/2 \leq a < 1$, and Lemma 4 gives

$$B_0 - c A_0 = \frac{\beta}{2} \left(\frac{3 - \beta}{1 + \beta} - c \right) = g.$$

Set

$$\eta = \frac{ag}{4(c + 1)}, \qquad R = 2 + \frac{8 \beta L^2}{\eta}.$$

Apply Lemma 6 to the seed variables, using the interval $[a, 2]$. Equation (16) verifies (19). The resulting variables satisfy

$$W_1(\mathcal{L}(U_L), \mathcal{L}(U_0)) \le \eta/a, \qquad W_1(\mathcal{L}(V_L), \mathcal{L}(V_0)) \le \eta/a.$$

As absolute-value loss is 1-Lipschitz, this implies

$$|A_L - A_0| \le \frac{\eta}{2a}, \qquad |B_L - B_0| \le \frac{\eta}{2a}.$$

The gap decreases by at most $(c+1)\eta/(2a) = g/8$. For example, a valid support constant is

$$C_c = \left(1 + \frac{\eta}{2a}\right)\left(2 + \frac{8\beta}{\eta}\right).$$

□

C THE FINITE-DOMAIN SAMPLE LOWER BOUND

C.1 MIXED POISSON OBSERVATIONS

Wu & Yang (2015, Appendix B, Lemma 6) bound the TV distance between Poisson mixtures with matched moments, including the bound $(e\lambda R/(2L))^L$. The following thinning argument gives the Poisson-tail form used in the proof.

Lemma 8 (Poisson thinning and matched moments)**.** *Let $U, V \in [0, R]$ match moments of orders zero through L, where $L \ge 0$ is an integer. For $\lambda \ge 0$, let Π_U, Π_V be the mixed Poisson laws obtained by drawing a Poisson count of mean λU and λV, respectively. Then*

$$\mathrm{TV}(\Pi_U, \Pi_V) \le \mathbb{P}\{\mathrm{Pois}(\lambda R) > L\}. \tag{25}$$

If $0 < \lambda R \le L + 1$, then

$$\mathrm{TV}(\Pi_U, \Pi_V) \le \left(\frac{e\lambda R}{L+1}\right)^{L+1}. \tag{26}$$

When $\lambda R = 0$, the two mixed laws are identical.

Proof. For $R > 0$, first draw $J \sim \mathrm{Pois}(\lambda R)$, independently of X, and then draw $K \mid (J, X) \sim \mathrm{Bin}(J, X/R)$. Poisson thinning gives $K \mid X \sim \mathrm{Pois}(\lambda X)$. For fixed $J = j \le L$, the probability of each binomial count is a polynomial in X of degree at most j:

$$\mathbb{P}\{K = k \mid J = j, X\} = \binom{j}{k}(X/R)^k(1 - X/R)^{j-k}.$$

Its expectation agrees under U and V. The two joint laws of (J, K) therefore agree conditional on $J \le L$, and the mixing probability of this event is the same. Their total variation is at most $\mathbb{P}(J > L)$. Marginalizing J proves (25). For $z = \lambda R > 0$ and $s = L + 1 \ge z$, the usual exponential Markov bound for a Poisson variable gives

$$\mathbb{P}\{\mathrm{Pois}(z) \ge s\} \le e^{-z}(ez/s)^s \le (ez/s)^s.$$

The zero-mean case is immediate, and also covers $R = 0$. □

C.2 CONVERSION TO A FIXED SAMPLE BUDGET

Lemma 9 (Conversion to a fixed IID sample budget)**.** *Let $X = (X_1, \ldots, X_N)$ be any random nonnegative vector, and put $S = N^{-1}\sum_i X_i$. When $S > 0$, define $p_i = X_i/(NS)$; when $S = 0$, define p arbitrarily. For an integer $m \ge 1$, observe counts that, conditional on X, are independent with*

$$K_i \mid X \sim \mathrm{Pois}(2mX_i/N).$$

There is a Markov kernel, independent of the law of X, that maps the count vector to an ordered sample in $[N]^m$, and whose output law has TV distance at most

$$\mathbb{P}\{S < 3/4\} + e^{-m/12} \tag{27}$$

from the mixture $\mathbb{E}[p^{\otimes m}]$.

Proof. Let $K = \sum_i K_i$. If $K \geq m$, uniformly shuffle the multiset containing K_i copies of label i, and return its first m labels. If $K < m$, return any fixed element of $[N]^m$. This kernel uses only the observed counts.

Conditional on X with $S > 0$ and on the total count $K = k$, the count vector has multinomial law with parameters k, p. After uniform shuffling, its ordered labels have law $p^{\otimes k}$. Hence, for every $k \geq m$, the returned sample has law $p^{\otimes m}$. For each fixed X, the output law is therefore a mixture whose weight on $p^{\otimes m}$ is at least $\mathbb{P}(K \geq m \mid X)$. Its distance from $p^{\otimes m}$ is at most $\mathbb{P}(K < m \mid X)$. The same bound, equal to one, holds when $S = 0$. Averaging over X gives a bound of $\mathbb{P}(K < m)$.

Conditional on X, K is Poisson with mean $2mS$. On $S \geq 3/4$, this mean is at least $3m/2$. Monotonicity and the Poisson lower-tail bound imply

$$\mathbb{P}(K < m \mid X) \leq e^{-m/12}.$$

Splitting according to $S < 3/4$ proves the claim. The conditional error is averaged over the original law of X. □

C.3 The finite-domain lower bound

Theorem 10 (Fixed-factor lower bound for two known candidates)**.** *Fix $c \in [1, 3)$, $0 < \delta < 1/2$, and*

$$0 < \varepsilon \leq \frac{(3-c)^2}{16(c+5)}.$$

There are constants $\kappa > 0$ and $N_0 < \infty$, depending only on c, δ, such that for every $N \geq N_0$ there are two completely known distributions $q_0, q_1 \in \Delta_N$ for which the following holds. Any possibly randomized algorithm that observes at most m IID samples from an unknown $p \in \Delta_N$, outputs $\widehat{i} \in \{0, 1\}$, and satisfies

$$\mathbb{P}_p\left\{\mathrm{TV}(p, q_{\widehat{i}}) \leq c \min_{i \in \{0,1\}} \mathrm{TV}(p, q_i) + \varepsilon\right\} \geq 1 - \delta \quad \textit{for every } p \in \Delta_N \tag{28}$$

must use $m \geq \kappa N / \log N$. The candidates may be fixed independently of the algorithm and its sample budget.

Proof. We first treat even N. Fix a, b, g, C_c from Corollary 7. Let q_0 have probability a/N on the first $N/2$ coordinates and b/N on the second $N/2$. Let q_1 interchange these values. Since $a + b = 2$, both are probability distributions. They depend only on c, N.

Put

$$L = \lceil 2 \log N \rceil, \qquad R_L = C_c L^2, \qquad m_N = \left\lfloor \frac{N}{16 e C_c L} \right\rfloor.$$

For sufficiently large N, $L \geq 2$ and $m_N \geq 1$. Obtain $U_L, V_L \in [0, R_L]$ from Corollary 7. We construct two priors on valid target distributions. In world $+$, independently draw $X_i \sim U_L$ on the first half and $X_i \sim V_L$ on the second half. In world $-$, swap these laws. Set

$$S = N^{-1} \sum_i X_i, \qquad p_i = \frac{X_i}{NS} \quad (S > 0),$$

and take $p = q_0$ when $S = 0$. Let $P_+^{(m_N)}, P_-^{(m_N)}$ denote the corresponding mixtures of ordered m_N-sample IID laws.

Separation after normalization. Every X_i has mean one and satisfies $\mathbb{E}X_i^2 \leq R_L \mathbb{E}X_i = R_L$. Independence gives

$$\mathbb{E}S = 1, \qquad \mathrm{Var}(S) \leq R_L / N. \tag{29}$$

Define raw half-L_1 quantities

$$D_k = \frac{1}{2N} \sum_{i=1}^{N} |X_i - N q_{k,i}|, \qquad k \in \{0, 1\}.$$

In world $+$, $\mathbb{E}D_0 = A_L$, $\mathbb{E}D_1 = B_L$. Since the anchors $Nq_{k,i}$ belong to $\{a, b\}$,

$$\mathbb{E}|X_i - Nq_{k,i}|^2 \le R_L + b^2, \qquad \mathrm{Var}(D_k) \le \frac{R_L + b^2}{4N}.$$

On $S > 0$, the reverse triangle inequality gives

$$|\,\mathrm{TV}(p, q_k) - D_k| \le \frac{1}{2}\sum_i \left|\frac{X_i}{NS} - \frac{X_i}{N}\right| = \frac{|S-1|}{2}. \tag{30}$$

Let $t = g/[16(c+1)]$, which belongs to $(0, 1)$. Chebyshev's inequality and a union bound show that the event

$$|S-1| \le t, \qquad |D_0 - A_L| \le t, \qquad |D_1 - B_L| \le t$$

has probability at least $1 - \rho_N$, where

$$\rho_N = \frac{3R_L + b^2}{2Nt^2} = o(1).$$

On this event $S > 0$, and

$$\mathrm{TV}(p, q_1) - c\,\mathrm{TV}(p, q_0) \ge B_L - cA_L - \frac{3}{2}(c+1)t \ge \frac{25}{32}g > \varepsilon.$$

Thus in world $+$, outside an event of probability at most ρ_N, the only admissible output in (28) is q_0. The same proof in world $-$ makes q_1 the only admissible output with probability at least $1 - \rho_N$.

Indistinguishability at a fixed sample budget. Consider the raw Poisson count experiment

$$K_i \mid X_i \sim \mathrm{Pois}(\lambda X_i), \qquad \lambda = 2m_N/N,$$

independently over coordinates. Let Q_+, Q_- be the two integrated count laws. Because the latent weights are independent, each count law is a product of mixed Poisson laws. Lemma 8 and the product bound for TV yield

$$\mathrm{TV}(Q_+, Q_-) \le N\left(\frac{e\lambda R_L}{L+1}\right)^{L+1} \le N\,8^{-(L+1)} = o(1).$$

Indeed, the choice of m_N ensures $\lambda R_L \le L/(8e)$. By (29), $\mathbb{P}(S < 3/4) \le 16R_L/N$ in each world. Apply the same kernel from Lemma 9 to both count laws. The triangle inequality and contraction of TV under a Markov kernel give

$$\mathrm{TV}(P_+^{(m_N)}, P_-^{(m_N)}) \le N\,8^{-(L+1)} + \frac{32R_L}{N} + 2e^{-m_N/12} =: \gamma_N = o(1). \tag{31}$$

Two-prior testing bound. Suppose an algorithm satisfying (28) used at most m_N samples. It can be viewed as an algorithm on m_N observations by ignoring any unused observations. Use its output to guess the world: report $+$ for q_0 and $-$ for q_1. Under either prior, averaging the uniform guarantee and using the separation event gives success probability at least $1 - \delta - \rho_N$. Any test between the two equiprobable worlds has success probability at most $(1 + \gamma_N)/2$. Since $\delta < 1/2$ is fixed and $\rho_N, \gamma_N \to 0$, this is a contradiction for sufficiently large N. Consequently more than $m_N = \Omega_c(N/\log N)$ samples are necessary.

For odd N, use the construction on $N - 1$ coordinates and give the remaining coordinate probability zero in both candidates and every target. Reducing the constant κ, if necessary, establishes the statement for all sufficiently large N. □

D ILLUSTRATING THE MOMENT TRANSFORMATION

Table 2 illustrates how the transformation in Section 4 changes the two candidate discrepancies. We start from the laws in (3) with $\beta = 0.05$ and increase the moment order from 2 to 20. The discrepancy ratio stays close to its initial value of 2.81 as the correction aligns more moments.

Table 2: Numerical illustration at $\beta = 0.05$ and $R = 100L^2$. The good and bad discrepancies are computed from the transformed laws using the expectations in (4)–(5). Increasing the moment order leaves the candidate ranking intact.

L	Good discrepancy	Bad discrepancy	Bad/good
2	0.0250000	0.0702381	2.80952
4	0.0250007	0.0702385	2.80947
8	0.0250014	0.0702389	2.80940
12	0.0250017	0.0702391	2.80938
20	0.0250020	0.0702393	2.80935

To obtain the correction, discretize $[a, R]$ on a Chebyshev grid that includes the original positive support points $a, b, 2$. Let $\phi(x) = 2(x - a)/(R - a) - 1$ and let T_k denote the degree-k Chebyshev polynomial. At grid points z_j, solve the linear program

$$\begin{aligned} \min_{v^+, v^- \geq 0} \quad & \sum_j (v_j^+ + v_j^-), \\ \text{subject to} \quad & \sum_j (v_j^+ - v_j^-) T_k(\phi(z_j)) = d_k, \quad 0 \leq k < L, \\ & d_k = T_k(\phi(b)) - \frac{a}{b} T_k(\phi(a)) - \frac{2\beta}{b} T_k(\phi(2)). \end{aligned}$$

The right-hand side is the difference of the size-biased moments. Divide the correction weights by z_j, add mass at zero to equalize total mass, and normalize and rescale the laws as in Lemma 6.