Support Divergence of Mixture NPMLEs Under Finite Mixtures

Abstract

Motivated by distribution testing and empirical Bayes calibration, we ask how sparse a mixture NPMLE can be when the true mixing distribution has fixed finite support. For Gaussian location mixtures in every fixed dimension d, every maximizer has at least $c(\log n)^{d/2}/\log\log n$ atoms with high probability. For Poisson mixtures, the lower bound is $c\sqrt{\log n}/(\log\log n)^{3/2}$. Thus the NPMLE can have diverging support even under a fixed finite mixture.

Note. This draft was produced through an iterative process in which Mark Sellke repeatedly asked GPT-5.6 in OpenAI Codex to prove, refine, and write up further extensions.

Contents

1 Introduction

Mixture models describe heterogeneous populations through a probability distribution G over latent parameters. Writing $p_G(x) = \int p_\theta(x)\, dG(\theta)$, each atom of G corresponds to a component of the mixture. When the number of components is unknown, the nonparametric maximum likelihood estimator (NPMLE) maximizes the likelihood over all probability measures G, rather than fixing a support size in advance. This construction traces to Kiefer–Wolfowitz [KW56] and has closely related empirical-Bayes roots in Robbins

[Rob50]: there G is an unknown prior, and the fitted mixing distribution is used to estimate posterior quantities or Bayes rules. Yet the mixture density, the induced decision rule, and the support of the mixing distribution are different statistical objects.

We ask whether the atom count of the NPMLE adapts to a finitely supported truth. Suppose the data law is p_{G_0} for a fixed finite mixture G_0. For Gaussian location and Poisson mixtures, we show that the support of every NPMLE diverges with the sample size. Thus accurate density or empirical-Bayes estimation can coexist with a fitted atom count that does not estimate the number of latent components. Approximate, thresholded, and penalized procedures define different estimators whose atom counts may depend on stopping or tuning rules.

The question was posed explicitly for one-dimensional Gaussian location mixtures by Polyanskiy and Wu [PW20]. They proved the self-regularization bound $O_{\mathbb{P}}(\log n)$ for subgaussian data and asked in [PW20, Section 5.6] whether this order is attained even under a single Gaussian component. Lindsay's convex-geometric theory gives a maximizer with at most n atoms [Lin83], and the Lindsay–Roeder criterion makes the NPMLE unique in the one-dimensional models considered here [LR93]. Multivariate Gaussian NPMLEs can be nonunique [SGS25], so our Gaussian lower bound is uniform over all maximizers. We extend the single-Gaussian question to every fixed dimension and every fixed finite Gaussian location mixture, and prove an analogous result for finite Poisson mixtures with positive largest mean. Our companion papers determine how much variance inflation is needed to collapse the Gaussian fit to one atom in dimension one and in dimensions at least two [Ano26b, Ano26a].

Preliminaries and notation

All asymptotic notation is as $n \to \infty$, unless another limit is specified. We write $\mathbb{P}$ and $\mathbb{E}$ for probability and expectation under the distribution currently under discussion. For a deterministic positive sequence a_n, the relation $X_n = O_{\mathbb{P}}(a_n)$ means that $|X_n|/a_n$ is tight, while $X_n = o_{\mathbb{P}}(a_n)$ means that $X_n/a_n \to 0$ in probability. Deterministic $O(\cdot)$ and $o(\cdot)$ have their usual meanings. We write $a_n \sim b_n$ if $a_n/b_n \to 1$, $a_n \ll b_n$ if $a_n/b_n \to 0$, and $a_n \lesssim b_n$ if $a_n \le Cb_n$ for a constant independent of n. The reverse relations are denoted by $\gg$ and $\gtrsim$, and $a_n \asymp b_n$ means that both one-sided bounds hold. Subscripts, as in $\lesssim_\lambda$ or $\asymp_d$, indicate the parameters on which the implicit constants may depend. Constants denoted by $c, C, c_0, C_0, \dots$ may change from line to line, subject to any displayed dependencies.

A well-specified finite mixture means that the data density is p_{G_0} for a fixed, finitely supported mixing distribution G_0. We call a measure D-atomic if it is supported on at most D points. For a measure G, $\operatorname{supp}(G)$ denotes its support; for a set A, $\operatorname{conv}(A)$ denotes its convex hull; and $T_\# G$ denotes the pushforward of G by a measurable map T. We write $B(x, r)$ for a Euclidean ball, $x^+ = \max\{x, 0\}$, $\log_+ x = \max\{\log x, 0\}$, and $\mathbf{1}_A$ for the indicator of A. The operator norm is $\|\cdot\|_{\mathrm{op}}$. An L^2 norm is written $\|\cdot\|_2$ when its reference measure is clear and $\|\cdot\|_{L^2(Q)}$ when it is not. For functions on $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, ℓ^q and $\ell^q(Q)$ denote the norms with respect to counting measure and a probability measure Q, respectively. If a maximizer has infinite support, our support lower bounds hold automatically; the proofs therefore need only rule out maximizers with too few atoms.

We also fix the bracketing notation used in the likelihood bounds. If $\mathcal{F}$ is a class of square-integrable functions with reference measure Q, an η-bracket is a pair $\ell \le f \le u$, with inequalities understood Q-a.e., such that $\|u - \ell\|_{L^2(Q)} \le \eta$. The bracketing number $N_{[]}(\eta, \mathcal{F}, L^2(Q))$ is the minimum number of such brackets needed to cover $\mathcal{F}$. Near a true density p_0, we use the coordinates $\sqrt{p/p_0} - 1$. Their $L^2(p_0)$ norm is $\|\sqrt{p} - \sqrt{p_0}\|_2$, the Hellinger distance under our convention.

Gaussian location mixtures

Let $\phi_d(x) = (2\pi)^{-d/2} e^{-|x|^2/2}$, $x \in \mathbb{R}^d$. For a probability measure G on $\mathbb{R}^d$, set $p_G(x) = \int \phi_d(x-a)\, dG(a)$. Given $X_1, \dots, X_n \in \mathbb{R}^d$, an NPMLE is any maximizer of $G \mapsto \sum_{i=1}^n \log p_G(X_i)$ over probability measures on $\mathbb{R}^d$.

Our first theorem answers the support question for every fixed finite Gaussian location mixture. The constant in the lower bound may depend on the true mixture and on the dimension, but the power of $\log n$ depends only on the dimension.

Theorem 1.1. *Fix $d \geq 1$, and let G_0 be a probability measure on $\mathbb{R}^d$ with finite nonempty support. Let $X_1, \dots, X_n$ be i.i.d. with density p_{G_0}, and let $\widehat{G}_n$ be any Gaussian location NPMLE over $\mathbb{R}^d$. Then there exists a constant $c_{G_0,d} > 0$ such that*

$$\mathbb{P}\Big\{ |\operatorname{supp}(\widehat{G}_n)| \geq c_{G_0,d} \frac{(\log n)^{d/2}}{\log \log n} \Big\} \to 1.$$

For $d = 1$ and $G_0 = \delta_0$, the theorem gives $c\sqrt{\log n}/\log\log n$ fitted atoms from data drawn from a single standard Gaussian. This gives the first diverging lower bound in the setting asked about by Polyanskiy and Wu.

The proof compares the best likelihood gains available with and without a support constraint. To lower-bound the unrestricted likelihood, first suppose that $G_0 = \delta_0$. Moving small amounts of mass from zero to many separated locations produces nearly uncorrelated likelihood scores and a total gain of order $(\log n)^{d/2}$. For a general finite mixture, we perform the same perturbation at a vertex of the convex hull of its support, moving mass outward in directions for which that component dominates the other components. To upper-bound the likelihood over all D-atomic mixtures, a localized bracketing argument gives a gain of order at most $D \log\log n$. Comparing the two bounds yields the theorem.

Poisson mixtures

Poisson mixtures also acquire extra atoms under a fixed finite mixture, but the lower-bound rate is different. For $\theta \geq 0$, let $p_\theta(k) = e^{-\theta}\theta^k/k!$, $k \in \mathbb{N}_0$, with p_0 interpreted as the point mass at zero. For a probability measure G on $[0, \infty)$, write $p_G(k) = \int p_\theta(k)\, dG(\theta)$; we use the same notation for finite nonnegative measures. Given observations in $\mathbb{N}_0$, a Poisson-mixture NPMLE maximizes $\sum_i \log p_G(X_i)$ over all such probability measures G.

The Poisson theorem uses the largest true mean as the exposed component. The assumption that this largest mean is positive excludes only the degenerate law concentrated at zero.

Theorem 1.2. *Let G_0 be a probability measure on $[0, \infty)$ with finite support, and suppose its largest support point λ_* is positive. Let $X_1, \dots, X_n$ be i.i.d. with mass function p_{G_0}, and let $\widehat{G}_n^{\mathrm{Pois}}$ be any Poisson-mixture NPMLE. Then there exists $c_{G_0} > 0$ such that*

$$\mathbb{P}\Big\{ |\operatorname{supp}(\widehat{G}_n^{\mathrm{Pois}})| \geq c_{G_0} \frac{\sqrt{\log n}}{(\log\log n)^{3/2}} \Big\} \to 1.$$

To construct a likelihood gain, we move mass from the largest true mean λ_* to points $\lambda_* + a\sqrt{\lambda_*}$. After normalization, the resulting scores have the same covariance kernel as the Gaussian scores. Their untruncated fourth moments are dominated by count values far larger than any observation in the sample and would restrict the construction to $|a| \lesssim (\log n)^{1/4}$. Truncating those irrelevant tails allows $a^2 \leq \gamma \log n/\log\log n$ for every fixed $\gamma < 1$, which produces the stated rate.

In both one-dimensional models, known upper bounds are logarithmic. A fixed finite Gaussian mixture is subgaussian, so Polyanskiy and Wu's self-regularization theorem gives $|\operatorname{supp}(\widehat{G}_n)| = O_{\mathbb{P}}(\log n)$ [PW20]. For Poisson mixtures, let $M_n = \max_i X_i$. Since every observed count lies in $\{0, \dots, M_n\}$, Lindsay's theorem gives a maximizer with at most $M_n + 1$ atoms [Lin83]; uniqueness then gives the same bound for the NPMLE itself [LR93]. Under every fixed finite Poisson mixture with positive largest mean, $M_n = (1 + o_{\mathbb{P}}(1)) \log n / \log\log n$. In both models, therefore, the lower and upper bounds differ by a factor of order $\sqrt{\log n}$, apart from powers of $\log\log n$.

For Gaussian mixtures in dimensions $d \geq 2$, the gap is much larger. The only general comparison is the finite-sample bound that some maximizer has at most n atoms. No upper bound is known near the polylogarithmic lower bound $(\log n)^{d/2} / \log\log n$, and nonuniqueness prevents the bound for one selected maximizer from applying automatically to all maximizers.

Related literature and empirical Bayes motivation

In empirical Bayes applications, the NPMLE mixing distribution is used as an estimated prior; see, for example, Jiang–Zhang [JZ09], Koenker–Mizera [KM14], Dicker–Zhao [DZ16], and Soloff–Guntuboyina–Sen [SGS25]. This use raises two separate questions: how well the fitted mixture estimates the data distribution or Bayes rule, and how many atoms the maximizing mixing distribution contains. Our results address the second question under a fixed finite-mixture data law.

Previous work on NPMLE support size has emphasized computation and upper bounds. Koenker and Gu's *Minimalist g-Modeling* [KG19] documented sparse NPMLE fits; as reported by Polyanskiy and Wu [PW20], their numerical examples suggested support sizes on the $\sqrt{n}$ scale. Polyanskiy and Wu proved the much smaller upper bound $|\operatorname{supp}(\widehat{G}_n)| = O_{\mathbb{P}}(\log n)$ under subgaussian sampling and asked whether it is tight for data from a single Gaussian component. Polyanskiy and Sellke [PS25] study certified computation and support behavior for Gaussian-location NPMLEs.

A separate literature establishes statistical guarantees for mixture NPMLEs. Zhang [Zha09] and Saha–Guntuboyina [SG20] prove density-risk bounds for Gaussian mixture NPMLEs, while Doss–Wu–Yang–Zhou [DWYZ23] treat high-dimensional Gaussian location mixtures. Zhang–Volgushev [ZV26] prove adaptation to finitely supported mixing distributions in Gaussian and Poisson models. Ghosh–Guntuboyina–Mukherjee–Tran [GGMT26] obtain KL stability bounds for Gaussian-mixture NPMLEs and their approximate solutions, and Chen–Wu [CW26] prove regret–Hellinger inequalities for Gaussian empirical Bayes. These guarantees do not control the number of atoms in the optimizer. The present results show that accurate estimation can coexist with diverging fitted support under a fixed finite mixture.

Mixture likelihood-ratio statistics also have nonstandard asymptotics. Hartigan [Har85] gave an early example for normal mixtures; Azaïs–Gassiat–Mercadier [AGM09] and Jiang–Zhang [JZ19] obtained further nonstandard limits; and Zhang–Volgushev [ZV25] proved a tight-versus-divergent dichotomy for nonparametric mixture likelihood-ratio tests with bounded parameter spaces. Those works study the value of a likelihood-ratio statistic. We instead study the smallest support size among likelihood maximizers.

2 Likelihood Geometry and Entropy Tools

The proofs of Theorems 1.1 and 1.2 use a common upper bound for mixtures with a prescribed number of atoms. After constructing an unrestricted likelihood gain, we must show that no mixture with only D atoms can match it. This section supplies the two reductions used for that comparison. The first restricts the atoms of a competitor to a sample-dependent compact set without decreasing likelihood; the second uses localized bracketing entropy to bound the likelihood gain over the resulting sparse class. The model-specific applications appear in Propositions 3.4 and 4.5.

2.1 Compact Support Reduction

For Gaussian mixtures, atoms outside the convex hull of the observations can be projected onto that hull without decreasing the fitted density at any observation. This reduction depends only on the Gaussian kernel, not on the distribution that generated the data.

Lemma 2.1. *Let $C_n = \mathrm{conv}\{X_1, \ldots, X_n\} \subset \mathbb{R}^d$. For every probability measure G on $\mathbb{R}^d$, there exists a probability measure $\widetilde{G}$ supported on C_n such that $p_{\widetilde{G}}(X_i) \ge p_G(X_i)$ for every $1 \le i \le n$. Moreover, if G has at most D atoms, then so does $\widetilde{G}$.*

Proof. Let Π_{C_n} be Euclidean projection onto C_n, and let $\widetilde{G} = (\Pi_{C_n})_\# G$ be the pushforward of G. If $a \notin C_n$ and $y = \Pi_{C_n}(a)$, then $(a - y) \cdot (x - y) \le 0$ for every $x \in C_n$. Therefore

$$\begin{aligned}\log \phi_d(x - y) - \log \phi_d(x - a) &= (y - a) \cdot x - \frac{1}{2}(|y|^2 - |a|^2) \\ &= \frac{1}{2}|a - y|^2 + (a - y) \cdot (y - x) \ge 0.\end{aligned}$$

Applying this with $x = X_i$ shows that replacing an atom at a by an atom of the same mass at y does not decrease any fitted density value $p_G(X_i)$. Integrating this pointwise inequality over a gives $p_{\widetilde{G}}(X_i) \ge p_G(X_i)$ for every i. $\square$

2.2 Localized Bracketing Bound

For every candidate density p, its expected log-likelihood relative to the truth is the negative Kullback–Leibler divergence $-\mathrm{KL}(p_0, p)$. Candidates far from p_0 therefore have a larger negative mean and need not be covered as finely as candidates near p_0. Local bracketing formalizes this idea by bounding the entropy separately on each Hellinger ball. We use the following localized form of Wong–Shen [WS95, Theorem 1, the remark following it, and Lemma 7]; see also van de Geer [vdG00, Chapter 7]. The theorem controls the likelihood outside a Hellinger ball, while the truncated-log estimate in their Lemma 7 controls the innermost ball.

Lemma 2.2. *Let $X_1, \ldots, X_n$ be i.i.d. with density p_0 with respect to a common dominating measure ν, and set $P_0(dx) = p_0(x)\nu(dx)$. Let $\mathcal{P}_n$ be a class of densities p with respect to ν such that $p\, d\nu \ll P_0$; ratios p/p_0 are understood P_0-a.e. Assume the suprema below are measurable; without this assumption, interpret all probabilities as outer probabilities. For $0 < \delta \le \sqrt{2}$, define*

$$\mathcal{F}_n(\delta) = \Big\{ \sqrt{p/p_0} - 1 : p \in \mathcal{P}_n,\ \|\sqrt{p/p_0} - 1\|_{L^2(p_0)} \le \delta \Big\}.$$

Assume $A_n \to \infty$ and $A_n = o(n)$. Fix $C < \infty$, choose $K_{\mathrm{WS}} = K_{\mathrm{WS}}(C)$ sufficiently large, and set $\delta_n = (K_{\mathrm{WS}} A_n / n)^{1/2}$. If

$$J_{[],n}(\delta) := \int_0^\delta \sqrt{1 + \log N_{[]}\big(\eta, \mathcal{F}_n(\delta), L^2(p_0)\big)}\, d\eta \le C\delta\sqrt{A_n},$$

for every $\delta_n \le \delta \le \sqrt{2}$, then there is a constant $C_1 < \infty$, depending only on C, such that

$$\mathbb{P}\left\{ \sup_{p \in \mathcal{P}_n} \sum_{i=1}^n \log \frac{p(X_i)}{p_0(X_i)} \le C_1 A_n \right\} \to 1.$$

Proof. This is the standard localization of Wong–Shen's inequalities. Peel the region outside the ball of radius δ_n into dyadic Hellinger shells. On the shell with outer radius δ, apply [WS95, Theorem 1] to the Hellinger ball of radius 2δ. The assumed entropy integral is at most $2C\delta\sqrt{A_n}$. Once $K_{\mathrm{WS}}(C)$ is large enough, the theorem's critical-radius condition holds, and the probability that any density in this shell has log-likelihood ratio larger than $-cn\delta^2$ is at most $C'e^{-c'n\delta^2}$. Summing over the shells costs $O(e^{-cA_n})$.

It remains to control the ball of radius δ_n. Apply the truncated-log estimate in [WS95, Lemma 7], with the truncation level chosen as in the remark following their Theorem 1. The same entropy integral bounds the centered truncated empirical process by $C_1 A_n$, while the truncation remainder has probability $O(e^{-cA_n})$. Since $A_n \to \infty$ and $A_n = o(n)$, both applications are in their stated range. Combining the inner ball with the dyadic shells proves the lemma. $\square$

In the Gaussian application, A_n is proportional to the atomic budget times $(d+1)\log\log n$; in the Poisson application, it is proportional to the atomic budget times $\log\log n$. In both cases, the factor $\log\log n$ comes from the logarithm of the sample-dependent parameter radius.

Both entropy proofs decompose a candidate mixture into a baseline submixture near the true atoms and several additional components. The next lemma controls how their square-root density increments combine. Recording this algebra once leaves only model-specific tail estimates in the Gaussian and Poisson sections.

Lemma 2.3. *On a measure space with dominating measure ν, let p_0 be a positive density, let a be a positive subdensity, and let q be a density, all with respect to ν. For $w > 0$, set*

$$h_{a,w,q} = \frac{\sqrt{a+wq}-\sqrt{a}}{\sqrt{p_0}}, \qquad e_{a,w,q} = \|h_{a,w,q}\|_{L^2(p_0)}, \qquad J_{a,w,q} = \int \frac{w^2q^2}{a+wq}\,d\nu.$$

In this lemma, $L^2(p_0)$ abbreviates $L^2(p_0\,d\nu)$. Then $J_{a,w,q}/4 \le e_{a,w,q}^2 \le J_{a,w,q}$, and, pointwise,

$$\frac{1}{2}h_{a,w,q} \le \partial_{\log w} h_{a,w,q} \le h_{a,w,q}, \qquad \frac{1}{2} \le \partial_{\log w}\log e_{a,w,q} \le 1.$$

If $s = \sqrt{a/p_0}$, then replacing s by any nonnegative measurable function s_0 changes $h_{a,w,q}$ by at most $|s-s_0|$ pointwise.

More generally, if $p = a + \sum_{j=1}^J w_j q_j$ and $h_j = h_{a,w_j,q_j}$, then

$$\sqrt{p/p_0} = \Phi(s,h_1,\dots,h_J), \qquad \Phi(s,h_1,\dots,h_J) = \left\{ s^2 + \sum_{j=1}^J (h_j^2 + 2sh_j) \right\}^{1/2}.$$

On the nonnegative orthant, Φ is coordinatewise increasing and $\partial_{h_j}\Phi \le 1$, $\partial_s\Phi \le \sqrt{J+1}$.

Proof. The identity

$$\sqrt{a+wq}-\sqrt{a} = \frac{wq}{\sqrt{a+wq}+\sqrt{a}}$$

gives the comparison with $J_{a,w,q}$. Differentiating in $\log w$ shows that the ratio of $\partial_{\log w} h_{a,w,q}$ to $h_{a,w,q}$ is $(\sqrt{a+wq}+\sqrt{a})/(2\sqrt{a+wq}) \in [1/2,1]$; integrating against $h_{a,w,q}^2 p_0$ gives the bound for $\partial_{\log w}\log e_{a,w,q}$. The map $x \mapsto \sqrt{x^2+c}-x$ is one-Lipschitz on $[0,\infty)$, which proves the baseline comparison.

Finally, $w_j q_j/p_0 = h_j^2 + 2sh_j$, proving the composition identity. Its derivatives are

$$\partial_{h_j}\Phi = (s+h_j)/\Phi \le 1, \qquad \partial_s\Phi = (s+\sum_j h_j)/\Phi \le \sqrt{J+1},$$

the last inequality following from Cauchy–Schwarz. $\square$

3 Gaussian Support Divergence

This section proves Theorem 1.1. The proof compares the two likelihood levels established in Propositions 3.2 and 3.4: the first constructs an unrestricted gain of order $(\log n)^{d/2}$, while the second bounds the gain of every D-atomic mixture by $O(D \log\log n)$. Their comparison forces the support of the NPMLE to diverge at the stated rate.

Fix a finite mixing distribution

$$G_0 = \sum_{m=1}^{K} \pi_m \delta_{\mu_m}, \qquad \pi_m > 0,$$

on $\mathbb{R}^d$, after merging duplicate atoms. Let $p_0 = p_{G_0}$, and write likelihoods relative to p_0:

$$L_n(G; G_0) = \sum_{i=1}^{n} \log \frac{p_G(X_i)}{p_0(X_i)}.$$

The perturbation moves a small amount of mass from one true atom μ_* to many points $\mu_* + a$. We choose μ_* to be a vertex of the convex hull of the true support and restrict a to a cone pointing away from the other atoms. In the Gaussian integrals that determine the score moments, this geometry makes the μ_*-component dominate the mixture density. The scores are then comparable to those of a single Gaussian component and become nearly uncorrelated when their displacement vectors are separated.

3.1 Scores From an Exposed Component

The constructive half of the proof requires many nearly orthogonal score directions. We obtain them by perturbing an exposed atom of the true mixing distribution outward from the convex hull of the remaining atoms; the lemmas in this subsection provide the normalization, correlation decay, and moment bounds used in Proposition 3.2.

Suppose first that $K > 1$. Choose an index m_* for which μ_{m_*} is a vertex of $\operatorname{conv}\{\mu_1, \ldots, \mu_K\}$, and write $\mu_* = \mu_{m_*}$ and $\pi_* = \pi_{m_*}$. A sufficiently narrow cone inside the outward normal cone at μ_* has nonempty interior and satisfies

$$(\mu_* - \mu_m) \cdot a \geq \eta_0 |a|, \qquad a \in \mathcal{C}, \quad m : \mu_m \neq \mu_*, \tag{3.1}$$

for some $\eta_0 > 0$. We take the cone narrow enough that, for all $a, b \in \mathcal{C}$,

$$|a + b| \geq c_{\mathcal{C}}(|a| + |b|), \qquad a \cdot b \geq 0.$$

If $K = 1$, there are no competing true atoms. We set $m_* = 1$, $\mu_* = \mu_1$, $\pi_* = 1$, and choose any closed cone with nonempty interior on which $|a + b| \geq c_{\mathcal{C}}(|a| + |b|)$ and $a \cdot b \geq 0$.

For $a \in \mathcal{C}$, define

$$\Delta_a(x) = \phi_d(x - \mu_* - a) - \phi_d(x - \mu_*), \qquad g_a(x) = \frac{\Delta_a(x)}{p_0(x)},$$

$$\sigma_a^2 = \mathbb{E}_{p_0} g_a(X)^2 = \int \frac{\Delta_a(x)^2}{p_0(x)}\, dx, \qquad h_a = \frac{g_a}{\sigma_a}.$$

Then $\mathbb{E}_{p_0} h_a(X) = 0$ and $\mathbb{E}_{p_0} h_a(X)^2 = 1$.

The next lemma supplies the three estimates needed for the likelihood construction. The first identifies the normalization of each score, the second shows that separated displacement vectors have weakly correlated scores, and the third provides the moments needed for concentration. All three estimates match the corresponding one-component formulas up to constants depending on G_0.

Lemma 3.1. *There are constants $R_0 < \infty$ and $0 < c < C < \infty$, depending only on G_0, d and $\mathcal{C}$, such that, uniformly for $a, b \in \mathcal{C}$ with $|a|, |b| \geq R_0$,*

$$ce^{|a|^2} \leq \sigma_a^2 \leq Ce^{|a|^2},$$

$$\left|\mathbb{E}_{p_0} h_a(X) h_b(X)\right| \leq Ce^{-|a-b|^2/2}, \qquad a \neq b,$$

and, for each fixed integer $m \geq 3$,

$$\mathbb{E}_{p_0} |h_a(X)|^m \leq C_m e^{m(m-2)|a|^2/2}.$$

Proof. Translate so that $\mu_* = 0$. Then

$$p_0(y) = \phi_d(y) \left[\pi_* + \sum_{m \neq m_*} \pi_m \exp\{\mu_m \cdot y - |\mu_m|^2/2\}\right] = \phi_d(y) W(y).$$

By (3.1), $\mu_m \cdot a \leq -\eta_0 |a|$ for $a \in \mathcal{C}$ and $m \neq m_*$. Under the tilted density proportional to $\phi_d(y-a)^2/\phi_d(y)$, the variable Y has law $\mathcal{N}(2a, \mathrm{Id}_d)$ and

$$\frac{\phi_d(Y-a)^2}{\phi_d(Y)} = e^{|a|^2} \phi_d(Y - 2a).$$

On the event $|Y - 2a| \leq c|a|$, every competing component in $W(Y)$ is $O(e^{-c|a|})$, while the complement has probability $e^{-c|a|^2}$. Hence

$$\int \frac{\phi_d(y-a)^2}{p_0(y)}\, dy = \pi_*^{-1} e^{|a|^2} \{1 + O(e^{-c|a|})\}.$$

The unshifted-square and cross terms in Δ_a^2 are both $O(1)$, whereas the displayed integral has order $e^{|a|^2}$. This proves $\sigma_a^2 \asymp e^{|a|^2}$.

For covariance, the leading integral is

$$\int \frac{\phi_d(y-a)\phi_d(y-b)}{p_0(y)}\, dy = e^{a \cdot b}\, \mathbb{E} \frac{1}{W(Y_{a,b})}, \qquad Y_{a,b} \sim \mathcal{N}(a+b, \mathrm{Id}_d).$$

Since $a + b \in \mathcal{C}$ and $|a + b| \geq c_{\mathcal{C}}(|a| + |b|)$, the exposed-component domination used for the variance also gives

$$\mathbb{E} \frac{1}{W(Y_{a,b})} = \pi_*^{-1} \{1 + O(e^{-c(|a|+|b|)})\}.$$

Dividing by $\sigma_a \sigma_b \asymp e^{(|a|^2+|b|^2)/2}$ gives the displayed $e^{-|a-b|^2/2}$ term, with a smaller domination error. The remaining three terms in $\Delta_a \Delta_b$ contain at least one unshifted factor $\phi_d(y)$. After normalization they are bounded by $Ce^{-(|a|^2+|b|^2)/2}$, which is at most $Ce^{-|a-b|^2/2}$ because $a \cdot b \geq 0$ on the cone.

For moments, use $p_0 \geq \pi_* \phi_d(\cdot - \mu_*)$ and $|\Delta_a|^m \leq 2^{m-1}\{\phi_d(\cdot - \mu_* - a)^m + \phi_d(\cdot - \mu_*)^m\}$. The Gaussian identity

$$\int \phi_d(y-a)^m \phi_d(y)^{1-m}\, dy = \exp\{m(m-1)|a|^2/2\}$$

combined with $\sigma_a^m \geq c_m e^{m|a|^2/2}$ proves the moment bound. $\square$

3.2 Constructive Likelihood Gain

We now use the exposed scores to prove the constructive half of Theorem 1.1. A separated grid of perturbations supplies order $(\log n)^{d/2}$ nearly orthogonal directions, and optimizing their small mixing weights yields the gain stated in Proposition 3.2.

Let $L = \log n$, and fix a small constant $A_* = A_*(G_0, d) > 0$. Set

$$K_{n,d}^{G_0} = \{a \in \mathcal{C} : R_0 \le |a| \le A_*\sqrt{L}\}.$$

Choose L_d large, and let $\mathcal{A}_{n,d}$ be a maximal L_d-separated subset of $K_{n,d}^{G_0}$. Since $\mathcal{C}$ has positive solid angle,

$$J_{n,d} := |\mathcal{A}_{n,d}| \asymp_{G_0,d} (\log n)^{d/2}.$$

The following proposition gives the unrestricted half of the likelihood comparison. Its sample-dependent mixing distribution removes a small amount of mass from μ_* and redistributes it over $\mu_* + \mathcal{A}_{n,d}$. We use this distribution only as a feasible competitor; it is not asserted to be the NPMLE.

Proposition 3.2. *For suitable constants R_0, L_d, A_*, there is a random mixing distribution G_n such that*

$$\mathbb{P}\{L_n(G_n; G_0) \ge c_{G_0,d} J_{n,d}\} \to 1.$$

Proof. For $a \in \mathcal{A}_{n,d}$, let

$$d_a = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h_a(X_i), \qquad \widehat{H}_{ab} = \frac{1}{n}\sum_{i=1}^{n} h_a(X_i)h_b(X_i), \qquad H_{ab} = \mathbb{E}_{p_0} h_a(X)h_b(X).$$

Thus d_a is the empirical score in direction a, while $\widehat{H}$ and H are its empirical and population Gram matrices. By Lemma 3.1, choosing L_d large makes $\sup_a \sum_{b\neq a} |H_{ab}| < 1/5$. Gershgorin's theorem gives $\lambda_{\max}(H) \le 1.2$. The moment estimate in Lemma 3.1, with A_* small enough, gives

$$\mathbb{E}|h_a(X)|^3 \le Cn^{3A_*^2/2}, \qquad \mathbb{E}|h_a(X)|^4 \le Cn^{4A_*^2}.$$

Since $J_{n,d}$ is polylogarithmic, Chebyshev's inequality and a union bound over the $J_{n,d}^2$ entries, followed by $\|A\|_{\text{op}} \le J_{n,d} \max_{a,b} |A_{ab}|$, give

$$\|\widehat{H} - H\|_{\text{op}} = o_{\mathbb{P}}(1).$$

We fix A_* small enough that the resulting powers of n are negligible in the concentration estimates that follow; any sufficiently small constant satisfying $A_*^2 < 1/8$ suffices.

Let $S = \sum_{a\in\mathcal{A}_{n,d}} (d_a^+)^2$. We next show that a positive fraction of the score coordinates contribute. Fix a large constant B and write $\psi_B(x) = (x^+)^2 \wedge B$. The one- and two-dimensional Berry–Esseen bounds for bounded Lipschitz test functions, together with the third-moment estimate in Lemma 3.1, hold uniformly over all coordinates and pairs; their error is n^{-c} for some $c > 0$. Consequently,

$$\mathbb{E}\psi_B(d_a) = \mathbb{E}\psi_B(Z) + o(1)$$

uniformly in a, where $Z \sim \mathcal{N}(0,1)$. For a centered bivariate normal pair (Z_a, Z_b) with correlation H_{ab}, the Gaussian maximal correlation inequality gives

$$|\operatorname{Cov}(\psi_B(Z_a), \psi_B(Z_b))| \le C_B |H_{ab}|.$$

The uniform bivariate approximation and the row-sum bound for H therefore imply

$$\mathrm{Var}\left(\sum_a \psi_B(d_a)\right) = O_B(J_{n,d}) + o(J_{n,d}^2).$$

Choose B so that $\mathbb{E}\psi_B(Z) > 1/3$. Chebyshev's inequality and $\psi_B(d_a) \le (d_a^+)^2$ then show that S is at least a constant multiple of $J_{n,d}$ with high probability. The identity $\mathbb{E}\sum_a d_a^2 = J_{n,d}$ gives the matching stochastic upper bound:

$$\mathbb{P}\{S \ge cJ_{n,d}\} \to 1, \qquad S = O_{\mathbb{P}}(J_{n,d}).$$

Fix $\eta = 1/5$, put $r_a = d_a^+/(1+\eta)$, and set

$$t_a = \frac{r_a}{\sqrt{n}\,\sigma_a}.$$

Since $\inf_a \sigma_a > 0$ and $\sum_a r_a \le \sqrt{J_{n,d}S} = O_{\mathbb{P}}(J_{n,d}) = o_{\mathbb{P}}(\sqrt{n})$, we have $\sum_a t_a = o_{\mathbb{P}}(1)$. Hence $\sum_a t_a \le \pi_*$ with probability tending to one. On this event define

$$G_n = G_0 - \Big(\sum_{a\in\mathcal{A}_{n,d}} t_a\Big)\delta_{\mu_*} + \sum_{a\in\mathcal{A}_{n,d}} t_a\delta_{\mu_*+a},$$

and set $G_n = G_0$ otherwise.

On the event $\max_i |X_i| \le C_{G_0}\sqrt{L}$, which has probability tending to one, the bound $p_0 \ge \pi_*\phi_d(\cdot - \mu_*)$ gives

$$|h_a(X_i)| \le C\exp\{a\cdot(X_i-\mu_*) - |a|^2\} + C \le n^\rho$$

for all i and $a \in \mathcal{A}_{n,d}$, where $\rho < 1/2$ after choosing A_* small enough. Therefore, with

$$u_i = \frac{1}{\sqrt{n}}\sum_{a\in\mathcal{A}_{n,d}} r_a h_a(X_i),$$

we have $\max_i |u_i| = o_{\mathbb{P}}(1)$, and

$$\frac{p_{G_n}(X_i)}{p_0(X_i)} = 1 + u_i.$$

Moreover, $\sum_i u_i^2 = r^\top \widehat{H} r = O_{\mathbb{P}}(J_{n,d})$. The remainder in $\log(1+u) = u - u^2/2 + O(|u|^3)$ is therefore at most $O(\max_i |u_i| \sum_i u_i^2) = o_{\mathbb{P}}(J_{n,d})$, and hence

$$L_n(G_n; G_0) = r^\top d - \frac{1}{2} r^\top \widehat{H} r + o_{\mathbb{P}}(J_{n,d}).$$

Since $r^\top d = S/(1+\eta)$ and $r^\top \widehat{H} r \le (1.2 + o_{\mathbb{P}}(1))S/(1+\eta)^2$, the resulting coefficient of S is positive for $\eta = 1/5$. The high-probability lower bound on S proves the proposition. $\square$

3.3 Localized Upper Bound

We now bound the likelihood of every D-atomic competitor. Lemma 2.1 restricts its atoms to a ball whose radius grows like $\sqrt{\log n}$. To retain the desired $D\log\log n$ likelihood scale, the entropy must therefore depend on this radius only through its logarithm.

For $D \geq 1$, $M \geq 2$, and $0 < \delta \leq \sqrt{2}$, let $\mathcal{P}_{d,D}(M)$ be the class of Gaussian location mixtures with at most D atoms supported in $B(0, M)$, and define

$$\mathcal{F}^{G_0}_{d,D}(M, \delta) = \left\{ \sqrt{p_G/p_0} - 1 : p_G \in \mathcal{P}_{d,D}(M),\ \|\sqrt{p_G/p_0} - 1\|_{L^2(p_0)} \leq \delta \right\}.$$

The next lemma provides this dependence. The factor $\log M$, rather than a power of M, is what becomes $\log \log n$ when $M \asymp \sqrt{\log n}$.

Lemma 3.3. *For fixed G_0, fixed d, and fixed $C_0 < \infty$, there exists $C_{G_0,d,C_0} < \infty$ such that, for every $M \geq 2$, $1 \leq D \leq M^{C_0}$, and $0 < \eta \leq \delta \leq \sqrt{2}$,*

$$\log N_{[]}\big(\eta, \mathcal{F}^{G_0}_{d,D}(M, \delta), L^2(p_0)\big) \leq C_{G_0,d,C_0} D(d + 1)\big(\log M + \log(\delta/\eta)\big).$$

Proof. Write $\rho_G = \|\sqrt{p_G/p_0} - 1\|_2$. We first work in a fixed neighborhood $0 < \rho_G \leq \rho_0$. Choose disjoint bounded neighborhoods $A_1, \dots, A_K$ of the atoms of G_0, and let A_0 be their complement. The local-geometry theorem of Gassiat–van Handel [GvH14, Theorem 3.10] controls both the mass outside these neighborhoods and the perturbation formed by the remaining atoms, uniformly in the number of candidate atoms. In particular,

$$W := G(A_0) \leq C\rho_G, \qquad \left\| \sqrt{p_{G|_{A_0^c}}/p_0} - 1 \right\|_2 \leq C\rho_G. \tag{3.2}$$

The theorem directly bounds W. Its coefficient lower bound, followed by second-order Taylor expansion of the kernels on the fixed sets A_i, also gives the second inequality. The same coefficient and Taylor bounds give the pointwise envelope needed below for the innermost Hellinger ball: the normalized inner density perturbation is bounded by a fixed $L^2(p_0)$ function formed from the Gaussian kernel and its first two location derivatives on $\bigcup_i A_i$.

Fix the contribution a of the atoms in $\bigcup_i A_i$. We first construct brackets after adding one atom outside these neighborhoods; the full mixture will follow by composing at most D such brackets. Write $q_\theta = \phi_d(\cdot - \theta)$, and use the notation h_{a,w,q_θ}, e_{a,w,q_θ} from Lemma 2.3. Differentiation gives

$$\nabla_\theta h_{a,w,q_\theta} = \frac{w q_\theta(x - \theta)}{2\sqrt{p_0}\sqrt{a + w q_\theta}}.$$

Because a contains a fixed positive amount of Gaussian mass with centers in a fixed ball, the measure governing the derivative norm has a uniform exponential radial moment. More precisely, if $d\mu = w^2 q_\theta^2 (a + w q_\theta)^{-1} dx$, comparison along rays gives

$$\int e^{c|x|/(1+M)} \left(1 + \frac{|x|}{1 + M}\right)^2 d\mu \leq C \int d\mu, \qquad |\theta| \leq M. \tag{3.3}$$

To prove (3.3), put $\widetilde{q} = w q_\theta$, $v = \widetilde{q}^2/a$, and $\mu = \widetilde{q} v/(\widetilde{q} + v)$; this last expression equals the density in (3.3). If $x = t\omega$, with $|\omega| = 1$, then for $r \geq 0$,

$$\widetilde{q}((t + r)\omega) \leq e^{-r(t-M)-r^2/2}\widetilde{q}(t\omega), \qquad a((t + r)\omega) \geq e^{-r(t+B_0)-r^2/2} a(t\omega),$$

where B_0 bounds the inner centers. Hence $v((t + r)\omega) \leq e^{-r(t-2M-B_0)-r^2/2} v(t\omega)$. The harmonic-mean map $(x, y) \mapsto xy/(x + y)$ is increasing and homogeneous, so the same bound holds for μ. After polar integration, if $F(t) = t^{d-1} \int_{\mathbb{S}^{d-1}} \mu(t\omega)\, d\omega$, then

$$F(t + 1) \leq e^{-t/3} F(t), \qquad t \geq C(1 + M + d).$$

Iterating this inequality on unit shells proves (3.3). Lemma 2.3 now yields

$$\|\nabla_\theta h_{a,w,q_\theta}\|_2 \le C(1+M)e_{a,w,q_\theta}.$$

We use these derivative bounds to construct ordered pointwise brackets, which is stronger than constructing an L^2-cover. Partition the location ball into boxes of diameter $c/(1+M)$, and partition the positive weights on the $\log w$ scale. Within one location box, the ratio of any kernel to the kernel at its center is at most $\exp\{c_1(1+|x|/(1+M))\}$, where c_1 can be made smaller than the constant in (3.3). After squaring, that estimate and (3.3) bound the $L^2(p_0)$ norms of the pointwise suprema of the weight and location derivatives by Ce and $C(1+M)e$, respectively.

Counting the weight and location cells gives, uniformly over the admissible inner subdensities,

$$\log N_{[]}\left(\epsilon, \{h_{a,w,q_\theta} : |\theta| \le M,\ e_{a,w,q_\theta} \le r\}, L^2(p_0)\right) \le Cd\{\log M + \log_+(r/\epsilon)\}. \tag{3.4}$$

To see the endpoint $w = 0$, fix a preliminary location box and a point θ_0 in it. By Lemma 2.3, the logarithmic derivative of $e_{a,w,q_{\theta_0}}$ lies in $[1/2, 1]$. All weights for which this norm is at most $c\epsilon$ therefore fit into one bracket $[0, \sup_{\theta \text{ in the same box}} h_{a,w_*,q_\theta}]$ of width $C\epsilon$; above that threshold, the relevant $\log w$-interval has length $O(1 + \log(r/\epsilon))$. Meshes of size $c\epsilon/r$ in $\log w$ and $c\epsilon/[r(1+M)]$ in location complete the count. Adjusting the mesh constants gives width ϵ.

We now bracket a Hellinger shell $r/2 < \rho_G \le r$. Decompose

$$p_G = a + \sum_{j=1}^{J} w_j q_{\theta_j}, \qquad J \le D,$$

where $a = p_{G|_{A_0^c}}$. By (3.2), with $s = \sqrt{a/p_0}$, $\|s-1\|_2 \le Cr$. Positivity also gives

$$\|h_{a,w_j,q_{\theta_j}}\|_2 \le \|\sqrt{p_G} - \sqrt{a}\|_2 \le Cr.$$

Let $\mathcal{S}_D^{\text{in}}(r)$ denote the class of these inner functions:

$$\mathcal{S}_D^{\text{in}}(r) = \Big\{ \sqrt{p_H/p_0} : H \text{ is a subprobability measure}, \quad |\operatorname{supp}(H)| \le D, \\ \operatorname{supp}(H) \subset \textstyle\bigcup_i A_i, \quad \|\sqrt{p_H/p_0} - 1\|_2 \le Cr \Big\}.$$

For $a = p_H$, write $a = m\bar{a}$, where $m = H(\mathbb{R}^d)$ and $\bar{a}$ is a probability-mixture density. The reverse triangle inequality gives $|\sqrt{m} - 1| \le Cr$, and hence

$$\|\sqrt{\bar{a}/p_0} - 1\|_2 \le m^{-1/2}\|\sqrt{a/p_0} - 1\|_2 + |m^{-1/2} - 1| \le Cr.$$

Apply Gassiat–van Handel's fixed-compact local bracketing theorem to $\bar{a}$, with order $D \vee K$, and bracket the scalar $\sqrt{m}$ separately. Multiplying the nonnegative scalar and density endpoints gives

$$\log N_{[]}\left(\epsilon, \mathcal{S}_D^{\text{in}}(r), L^2(p_0)\right) \le CD(d+1)\log(CDr/\epsilon).$$

Put $\zeta = c\min\{\eta, r\}$. Bracket s at width $\zeta/(CD)$, choose an actual representative s_0 from each nonempty bracket, and apply (3.4) with baseline $a_0 = s_0^2 p_0$ to each outer component at the same width. If the inner bracket is $[L, U]$, replace L by $L_+ = \max\{L, 0\}$, put $b = U - L_+$, and widen each component bracket $[\ell_j, u_j]$ to $[(\ell_j - b)_+, u_j + b]$. Lemma 2.3 then gives genuine composed endpoints through Φ, of total width at most ζ. The logarithm of the shell bracket count is therefore at most

$$CD(d+1)\{\log M + \log D + \log_+(r/\eta)\}. \tag{3.5}$$

The shell construction leading to (3.5) covers every positive Hellinger radius. To cover the innermost Hellinger ball by a single bracket, apply the coefficient bound and the fixed Taylor envelope just described. These two bounds give a fixed $H_0 \in L^2(p_0)$ such that the inner ratio $A = a/p_0$ satisfies $|A-1| \le C\rho_G H_0$. The outer part $B = (p_G - a)/p_0$ satisfies

$$B/\rho_G \le C \sup_{|\theta|\le M} q_\theta/p_0.$$

Completing the square shows that the supremum on the right has $L^2(p_0)$-norm at most $e^{C(1+M^2)}$. Since $|\sqrt{A+B}-1| \le |A+B-1|$, there is a common envelope F_M such that

$$|\sqrt{p_G/p_0}-1| \le \rho_G F_M, \qquad \|F_M\|_2 \le e^{C(1+M^2)}.$$

Thus the ball $\rho_G \le \eta/(2\|F_M\|_2)$ lies in one bracket of width η. There are only $O\{1+M^2+\log_+(\delta/\eta)\}$ dyadic shells above it. Taking the union of their bracket collections adds at most

$$C\{\log M + \log_+(\delta/\eta)\}$$

to (3.5). Because $D \le M^{C_0}$, the term $D\log D$ is absorbed by $CD\log M$, and the shell-union term is absorbed by the asserted entropy bound.

For the remaining region $\rho_G \ge \rho_0$, no localization is needed. Pointwise derivative brackets for Gaussian kernels, combined with a simplex grid for the weights, give

$$\log N_{[]}\big(\eta^2, \{p_G : |\operatorname{supp}(G)| \le D,\ \operatorname{supp}(G) \subset B(0,M)\}, L^1\big) \le CD(d+1)\{\log M + \log(1/\eta)\}.$$

Here is the construction. A location cell I, centered at θ_I, is bracketed by $[(q_{\theta_I} - \operatorname{diam}(I)E_I)_+, q_{\theta_I} + \operatorname{diam}(I)E_I]$, where $E_I = \sup_{\theta\in I}|\nabla q_\theta|$ has uniformly bounded L^1-norm. Quantizing weights at mesh $h = c\eta^2/D$, and writing $k_j = \lfloor w_j/h \rfloor$, gives mixture endpoints $\sum_j k_j h \ell_j$ and $\sum_j (k_j+1)hu_j$. Their L^1-distance is at most $C\eta^2$, and the number of feasible integer weight vectors is at most $(C/\eta^2)^D$. Taking square roots turns these into Hellinger brackets of width η. On the present region, $\delta \ge \rho_0$, so $\log(1/\eta) \le C + \log(\delta/\eta)$. Combining the local and global regions proves the lemma. □

We now insert the entropy estimate into the localized likelihood inequality. The resulting bound applies simultaneously to all sparse competitors and will be compared with Proposition 3.2.

Proposition 3.4. *Fix $C_0 < \infty$. Let D_n be positive integers satisfying $D_n \le (\log n)^{C_0}$ and $D_n(d+1)\log\log n = o(n)$. Then*

$$\mathbb{P}\left\{\sup_{|\operatorname{supp}(G)|\le D_n} L_n(G;G_0) \le C_{G_0,d,C_0} D_n(d+1)\log\log n\right\} \to 1.$$

Proof. Since G_0 is fixed and finite, $\mathbb{P}\{\max_i |X_i| \le C_{G_0}\sqrt{\log n}\} \to 1$. On this event, Lemma 2.1 restricts all D_n-atomic competitors to $B(0, C_{G_0}\sqrt{\log n})$. Apply Lemma 3.3 with $M = C_{G_0}\sqrt{\log n}$ and growth exponent $2C_0+1$. The bracketing integral is bounded by $C\delta\sqrt{A_n}$ for $A_n = C_{G_0,d,C_0} D_n(d+1)\log\log n$, because $\int_0^1 \sqrt{1+\log(1/t)}\,dt < \infty$. Lemma 2.2 gives the desired upper bound. □

Proof of Theorem 1.1. By Proposition 3.2, the unrestricted likelihood is at least $cJ_{n,d}$ with probability tending to one, where $J_{n,d} \asymp_{G_0,d} (\log n)^{d/2}$. Proposition 3.4 gives the upper bound $CD_n(d+1)\log\log n$ for all D_n-atomic competitors. Taking

$$D_n = \left\lfloor c_{G_0,d} \frac{(\log n)^{d/2}}{\log\log n} \right\rfloor$$

with $c_{G_0,d}$ small enough, no D_n-atomic mixing distribution can maximize the likelihood on the intersection of the two high-probability events. This proves the theorem. □

3.4 Small Variance Misspecification

The one-dimensional phase-transition analysis in [Ano26b] shows that, for $\varepsilon_n = n^{-r}$, the one-atom probability is nontrivial when $0 < r < 1/4$ and vanishes when $r \geq 1/4$. The following proposition quantifies the support on the smaller-mismatch side: if ε_n is below $n^{-1/4}$ by a fixed polynomial factor, then the lower bound has the same order as under correct specification.

Proposition 3.5. *Fix $\delta > 0$. Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(0, 1-\varepsilon_n)$, where $0 \leq \varepsilon_n \leq n^{-1/4-\delta}$, and fit unit-variance Gaussian location mixtures. There is a constant $c_\delta > 0$ such that every NPMLE $\widehat{G}_n$ satisfies*

$$\mathbb{P}\left\{ |\operatorname{supp}(\widehat{G}_n)| \geq c_\delta \frac{\sqrt{\log n}}{\log\log n} \right\} \to 1.$$

Proof. The proof is the one-dimensional construction from Sections 3.2 and 3.3, with two changes: the score array is controlled by truncation rather than raw fourth moments, and the variance mismatch introduces a small deterministic damping term. We give the estimates that justify both changes.

If the proposition holds for one value of δ, it also holds for every larger value. We may therefore assume $0 < \delta < 1/4$. Write $X_i = \sigma_n Z_i$, where $\sigma_n^2 = 1-\varepsilon_n$ and $Z_1, \dots, Z_n$ are i.i.d. standard normal. Put $\gamma = \delta/2$, choose

$$\theta_0 = \frac{1}{2} - \frac{\gamma}{2}, \qquad \theta_1 = \frac{1}{2} - \frac{\gamma}{4},$$

and let u_n satisfy $n\mathbb{P}\{Z > u_n\} = 1$ for $Z \sim \mathcal{N}(0,1)$. A constant-spaced grid $\mathcal{A}_n = \{a_1, \dots, a_J\}$ in $[\theta_0 u_n, \theta_1 u_n]$ has $J \asymp_\delta \sqrt{\log n}$. For

$$q_a(z) = e^{az - a^2/2}, \qquad h_a(z) = \frac{q_a(z) - 1}{(e^{a^2} - 1)^{1/2}},$$

direct Gaussian integration gives

$$\mathbb{E} h_a(Z) h_b(Z) = \frac{e^{ab} - 1}{\sqrt{(e^{a^2}-1)(e^{b^2}-1)}} \leq 2e^{-(a-b)^2/2} \qquad (a \neq b).$$

Choosing the grid spacing large makes the off-diagonal Gram row sums smaller than $1/10$.

The earlier construction used a smaller score radius and controlled the empirical Gram matrix by raw fourth moments. Here the grid approaches $a = u_n/2$, where those moments are too large, so we replace that step by the following truncation estimates. The strict inequality $\theta_1 < 1/2$ gives $c > 0$ and $\rho_n = n^{-c}$ such that, uniformly over a, b in the grid,

$$\begin{aligned}
\mathbb{E}\Big[h_a(Z)^2 \mathbf{1}_{\{|h_a(Z)| > \rho_n \sqrt{n}\}} \Big] + n\mathbb{P}\{|h_a(Z)| > \rho_n \sqrt{n}\} &\leq \rho_n, \\
\mathbb{E}\big[|h_a(Z) h_b(Z)| \mathbf{1}_{\{|h_a(Z)h_b(Z)| > n\rho_n\}} \big] &\leq \rho_n, \\
\mathbb{E}\big[(h_a(Z) h_b(Z))^2 \mathbf{1}_{\{|h_a(Z)h_b(Z)| \leq n\rho_n\}} \big] &\leq n\rho_n.
\end{aligned}$$

To verify these estimates, note that $(e^{a^2} - 1)^{1/2} \asymp e^{a^2/2}$. If $a = \theta u_n$, the first tail event forces

$$Z \geq \left(\theta + \frac{1}{4\theta} - \frac{c}{2\theta} + o(1) \right) u_n.$$

At $c = 0$, the coefficient is uniformly larger than both 1 and 2θ on $[\theta_0, \theta_1]$. Mills' bound controls the ordinary tail, while

$$h_a(z)^2 \phi_1(z) \leq C\phi_1(z - 2a) + Ce^{-a^2} \phi_1(z)$$

controls its second moment under the tilted normal law. The product estimates then follow from Cauchy–Schwarz, after reducing c. More explicitly, repeating this tilted-tail calculation at the larger threshold $\sqrt{n\rho_n}$ first gives, for some $c_0 > 0$,

$$\sup_a \mathbb{E}\Big[h_a(Z)^2 \mathbf{1}_{\{|h_a(Z)|>\sqrt{n\rho_n}\}}\Big] \le n^{-c_0}.$$

Choose $c < c_0/2$. Since $\{|h_a h_b| > n\rho_n\}$ is contained in the union of the two events $\{|h_a| > \sqrt{n\rho_n}\}$ and $\{|h_b| > \sqrt{n\rho_n}\}$, Cauchy–Schwarz gives the product-tail bound above. The final truncated second-moment bound follows from $(h_a h_b)^2 \mathbf{1}_{\{|h_a h_b| \le n\rho_n\}} \le n\rho_n |h_a h_b|$.

Now set

$$d_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n h_{a_j}(Z_i), \qquad \widehat{H}_{jk} = \frac{1}{n} \sum_{i=1}^n h_{a_j}(Z_i) h_{a_k}(Z_i), \qquad S = \sum_{j=1}^J (d_j^+)^2.$$

Truncating the products at $n\rho_n$, applying Chebyshev's inequality to each entry, and taking a union bound gives $\|\widehat{H} - H\|_{\mathrm{op}} = o_{\mathbb{P}}(1)$. Truncating the individual scores at $\rho_n\sqrt{n}$, and repeating the bounded positive-score argument from Proposition 3.2, gives

$$S = O_{\mathbb{P}}(J), \qquad \mathbb{P}\{S \ge cJ\} \to 1.$$

Indeed, J is polylogarithmic whereas ρ_n is a negative power of n, so the entrywise union bounds, the operator-norm conversion, and the one- and two-coordinate Berry–Esseen errors all vanish. More explicitly, with $T_n = \rho_n\sqrt{n}$,

$$\big|\mathbb{E} h_a(Z) \mathbf{1}_{\{|h_a(Z)|>T_n\}}\big| \le \sqrt{\mathbb{P}\{|h_a(Z)| > T_n\} \mathbb{E}[h_a(Z)^2 \mathbf{1}_{\{|h_a(Z)|>T_n\}}]} \le \frac{\rho_n}{\sqrt{n}}.$$

Thus recentering moves a normalized score by at most ρ_n, and variance normalization changes it by $O(\rho_n)$. The probability that any of the nJ summands is truncated is at most $J\rho_n = o(1)$, so the Gaussian approximation transfers back to the original scores d_j. Finally, choose $\zeta > 0$ so that $2\theta_1(1 + \zeta - \theta_1) < 1/2$. With probability tending to one, $\max_i |Z_i| \le (1+\zeta)u_n$, and the explicit score formula then gives, for some $c_1 > 0$,

$$\max_{i,j} |h_{a_j}(Z_i)| \le n^{1/2 - c_1}$$

with probability tending to one. Thus both ingredients used in the correctly specified construction remain valid: the empirical score vector has the same Gaussian approximation, and every observed score is uniformly small enough for the likelihood Taylor expansion.

We now transfer the correctly specified construction from the standardized sample $Z_1, \ldots, Z_n$ to the observed sample. Put $\kappa_n = \varepsilon_n / \{2(1 - \varepsilon_n)\}$. A fitted atom $b_j = a_j/\sigma_n$ has density ratio

$$\frac{\phi_1(X_i - b_j)}{\phi_1(X_i)} = e^{-\kappa_n a_j^2} q_{a_j}(Z_i).$$

Thus $e^{-\kappa_n a_j^2} = 1 + O(\varepsilon_n \log n) = 1 + o(1)$ uniformly on the grid. As in Proposition 3.2, fix $\eta = 1/5$, set

$$r_j = \frac{d_j^+}{1+\eta}, \qquad t_j = \frac{r_j}{\sqrt{n}\,(e^{a_j^2} - 1)^{1/2}},$$

and move mass t_j from zero to b_j. Since $\sum_j r_j = O_{\mathbb{P}}(J)$, the total moved mass is $o_{\mathbb{P}}(1)$.

Relative to the correctly specified calculation, the only additional constant in each fitted density ratio is

$$c_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \frac{r_j\{e^{-\kappa_n a_j^2} - 1\}}{(e^{a_j^2} - 1)^{1/2}}.$$

Its contribution to the linear likelihood term satisfies

$$n|c_n| \le O_{\mathbb{P}}(J)(\log n) n^{1/2-(1/4+\delta)-\theta_0^2+o(1)} = o_{\mathbb{P}}(J),$$

because $\theta_0^2 > 1/4 - \delta$. The damping factors also change the linear and quadratic score terms by only $o_{\mathbb{P}}(J)$. The constant c_n creates the additional terms $2c_n \sum_i v_i$ and nc_n^2 in the quadratic expansion, where

$$v_i = \frac{1}{\sqrt{n}} \sum_j r_j e^{-\kappa_n a_j^2} h_{a_j}(Z_i), \qquad \sum_i v_i = O_{\mathbb{P}}(J).$$

They are $o_{\mathbb{P}}(J)$ because $n|c_n| = o_{\mathbb{P}}(J)$ and $J = o(n)$. The bound $\max_{i,j} |h_{a_j}(Z_i)| \le n^{1/2-c_1}$ makes the Taylor remainder $o_{\mathbb{P}}(J)$, exactly as in Proposition 3.2. The quadratic likelihood expansion from that proposition now applies with the damping, centering, and Taylor errors just bounded. It gives a feasible mixing distribution with likelihood gain at least $cJ \asymp_\delta \sqrt{\log n}$ with probability tending to one.

The sparse upper bound transfers without any new entropy calculation. For any fitted atom b, put $t = \sigma_n b$. Then

$$\frac{\phi_1(X_i - b)}{\phi_1(X_i)} = e^{-\kappa_n t^2} q_t(Z_i) \le q_t(Z_i).$$

After mixing, every D-atomic likelihood for $X_1, \ldots, X_n$ is therefore bounded by a D-atomic correctly specified likelihood for $Z_1, \ldots, Z_n$. Proposition 3.4 bounds the correctly specified likelihood by $CD \log\log n$ with probability tending to one. Comparing the two likelihood scales and taking $D = c_\delta \sqrt{\log n}/\log\log n$, with c_δ small enough, proves the proposition. □

The one-dimensional restriction reflects the mismatch scale on which the bound is informative, not a limitation of the score construction. The same construction gives the following higher-dimensional statement.

Remark 3.6. For every fixed $d \ge 2$, replacing the one-dimensional grid $\mathcal{A}_n$ by a constant-separated grid in the annulus

$$\{a \in \mathbb{R}^d : \theta_0 u_n \le |a| \le \theta_1 u_n\}$$

gives $J \asymp_{\delta,d} (\log n)^{d/2}$ score locations. For the multivariate Gaussian scores, the exact covariance formula implies

$$|\mathbb{E} h_a(Z) h_b(Z)| \le C \left\{ e^{-|a-b|^2/2} + e^{-(|a|^2+|b|^2)/2} \right\}.$$

Choosing the separation constant large makes the off-diagonal row sums of the first term arbitrarily small; the second has row sum at most $Je^{-\theta_0^2 u_n^2} = o(1)$. Thus the truncation, damping, and sparse-likelihood arguments are otherwise unchanged. Repeating the score and likelihood calculations on this annular grid gives

$$|\operatorname{supp}(\widehat{G}_n)| \ge c_{\delta,d} \frac{(\log n)^{d/2}}{\log\log n}$$

with probability tending to one whenever $\varepsilon_n \le n^{-1/4-\delta}$.

We state Proposition 3.5 only in one dimension because this polynomial condition is tied to the correct one-dimensional boundary. In higher dimensions it is far below the actual transition scale: the companion paper finds $\varepsilon_n \asymp 1/\log n$ for $d = 2$ and $\varepsilon_n \asymp (\log\log n)/\log n$ for $d \geq 3$ [Ano26a]. Proving quantitative support growth near those scales would require uniform control over the continuum of angular directions near the radial extremes, not merely the sub-edge Lindeberg argument used here.

4 Poisson Mixture Analogue

This section proves Theorem 1.2. As in the Gaussian case, Proposition 4.3 constructs a large unrestricted likelihood gain and Proposition 4.5 bounds every D-atomic competitor. The Poisson proof differs in its score analysis: truncation at the observed sample scale replaces the Gaussian moment estimates and permits perturbations across the full exposed tail window.

For the Poisson theorem, write

$$G_0 = \sum_{m=1}^{K} \pi_m \delta_{\lambda_m}, \qquad 0 \leq \lambda_1 < \cdots < \lambda_K = \lambda_*, \qquad \lambda_* > 0,$$

and write $p_0 = p_{G_0}$. The relative log-likelihood is

$$L_n^{\mathrm{Pois}}(G; G_0) = \sum_{i=1}^{n} \log \frac{p_G(X_i)}{p_0(X_i)}.$$

Write $\pi_* = \pi_K$. The construction perturbs the component at λ_*. This component is identifiable from the upper tail because, among the finitely many true Poisson laws, it dominates the probability of large counts.

As in the Gaussian case, a sample-dependent compactness reduction controls sparse competitors. For Poisson mixtures, the compact set is the interval up to the sample maximum.

Lemma 4.1. *Let $M_n = \max_{1\leq i\leq n} X_i$. For every probability measure G on $[0,\infty)$, there exists a probability measure $\widetilde{G}$ supported on $[0, M_n]$ such that $p_{\widetilde{G}}(X_i) \geq p_G(X_i)$ for every $1 \leq i \leq n$. Moreover, if G has at most D atoms, then so does $\widetilde{G}$.*

Proof. Let $T(\theta) = \min(\theta, M_n)$ and set $\widetilde{G} = T_\# G$. Fix an observed value $k \leq M_n$. If $M_n = 0$, then $k = 0$ and $p_\theta(0) = e^{-\theta} \leq 1 = p_{\theta=0}(0)$. Otherwise, $d \log p_\theta(k)/d\theta = k/\theta - 1$, so $\theta \mapsto p_\theta(k)$ is decreasing on $[M_n, \infty)$. Hence $p_{T(\theta)}(k) \geq p_\theta(k)$ for every $\theta \geq 0$, and integration over G proves the lemma. □

4.1 Scores Near the Largest Mean

The upper tail exposes the largest true mean λ_*. The score estimates below show that separated outward perturbations of this component are nearly orthogonal after truncation; these are the inputs to Proposition 4.3.

For $a \geq 0$, set $\theta_a = \lambda_* + a\sqrt{\lambda_*}$ and

$$\Delta_a(k) = p_{\theta_a}(k) - p_{\lambda_*}(k), \qquad g_a(k) = \frac{\Delta_a(k)}{p_0(k)},$$

$$\sigma_a^2 = \sum_{k=0}^{\infty} \frac{\Delta_a(k)^2}{p_0(k)}, \qquad h_a(k) = \frac{g_a(k)}{\sigma_a}.$$

Then $\mathbb{E}_{p_0} h_a(X) = 0$ and $\mathbb{E}_{p_0} h_a(X)^2 = 1$. Also

$$\pi_* \le \frac{p_0(k)}{p_{\lambda_*}(k)} = \pi_* + \sum_{m<K} \pi_m e^{\lambda_* - \lambda_m} \left(\frac{\lambda_m}{\lambda_*}\right)^k \le C_{G_0}, \qquad k \ge 0. \tag{4.1}$$

In (4.1), the term corresponding to $\lambda_m = 0$ at $k = 0$ is interpreted using $0^0 = 1$. Thus replacing the full denominator p_0 by the single-component denominator p_{λ_*} changes a score by at most constant factors.

The next lemma gives the covariance and concentration estimates needed to combine many such scores. A direct moment bound would be governed by counts far beyond the observed sample range. The truncated estimates discard that irrelevant part of the likelihood ratio and remain uniform up to $a^2 \asymp \log n / \log\log n$.

Lemma 4.2. *Fix $\gamma \in (0,1)$, and put $R_n = \sqrt{\gamma \log n / \log\log n}$. There are constants $A_0 = A_0(G_0)$, $c = c(G_0, \gamma) > 0$, and $C = C(G_0) < \infty$ such that, uniformly for $A_0 \le a, b \le R_n$,*

$$C^{-1} e^{a^2} \le \sigma_a^2 \le C e^{a^2}, \qquad |\mathbb{E}_{p_0} h_a(X) h_b(X)| \le C e^{-(a-b)^2/2}.$$

Moreover, with $\rho_n = e^{-c \log n / \log\log n}$, uniformly for $A_0 \le a, b \le R_n$,

$$\mathbb{E}\big[h_a(X)^2 \mathbf{1}_{\{|h_a(X)| > \rho_n \sqrt{n}\}}\big] \le \rho_n, \qquad n\mathbb{P}\{|h_a(X)| > \rho_n \sqrt{n}\} \le \rho_n,$$

$$\mathbb{E}\big[|h_a(X) h_b(X)| \mathbf{1}_{\{|h_a(X) h_b(X)| > n\rho_n\}}\big] \le \rho_n,$$

and

$$\mathbb{E}\big[(h_a(X) h_b(X))^2 \mathbf{1}_{\{|h_a(X) h_b(X)| \le n\rho_n\}}\big] \le n\rho_n.$$

Proof. Let $\tau_a = (e^{a^2} - 1)^{1/2}$ and let

$$h_{a,*}(k) = \frac{p_{\theta_a}(k)/p_{\lambda_*}(k) - 1}{\tau_a}$$

be the one-component score at λ_*. By (4.1), $\sigma_a \asymp \tau_a$ and $|h_a(k)| \asymp |h_{a,*}(k)|$, uniformly for $a \ge A_0$ and $k \ge 0$, after taking A_0 fixed and large enough. Under the pure p_{λ_*} law, the unnormalized score covariance is exactly $e^{ab} - 1$. To compare this formula with the covariance under the full mixture, write

$$W(k) = \frac{p_0(k)}{p_{\lambda_*}(k)} = \pi_* + \sum_{m<K} c_m r_m^k, \qquad 0 \le r_m < 1.$$

Thus $W(k)^{-1} = \pi_*^{-1} + O(r^k)$ for some fixed $r < 1$. If $q_a = p_{\theta_a}/p_{\lambda_*}$, then

$$\mathbb{E}_{\lambda_*}[q_a q_b] = e^{ab},$$

whereas

$$\mathbb{E}_{\lambda_*}[r^X q_a q_b] = \exp\Big\{rab - (1-r)\big((a+b)\sqrt{\lambda_*} + \lambda_*\big)\Big\}.$$

The analogous formulas with $q_a q_b$ replaced by q_a, q_b, or 1 show that

$$\sum_k \frac{\Delta_a(k)\Delta_b(k)}{p_0(k)} = \pi_*^{-1}(e^{ab} - 1) + O\Big(e^{rab - c(a+b)} + e^{-ca} + e^{-cb} + 1\Big).$$

For $a = b$, this proves $\sigma_a^2 \asymp e^{a^2}$. For arbitrary $a, b \ge A_0$, division by $\sigma_a \sigma_b \asymp e^{(a^2+b^2)/2}$ gives

$$|\mathbb{E} h_a(X) h_b(X)| \le C e^{-(a-b)^2/2};$$

the error term is smaller by the additional factor $e^{-(1-r)ab-c(a+b)}$.

We next locate the counts at which the score exceeds the truncation level. Fix C_0 so that $|h_a(k)| \le C_0|h_{a,*}(k)|$ throughout the stated range, and put $T_n = C_0^{-1}\rho_n\sqrt{n}$. Since

$$\frac{p_{\theta_a}(k)}{p_{\lambda_*}(k)} = e^{-a\sqrt{\lambda_*}}\left(1+\frac{a}{\sqrt{\lambda_*}}\right)^k,$$

the event $|h_a(X)| > \rho_n\sqrt{n}$ forces $X \ge k_a$, where

$$k_a = \left\lceil \frac{a\sqrt{\lambda_*}+\log(1+T_n\tau_a)}{\log(1+a/\sqrt{\lambda_*})} \right\rceil.$$

Equivalently,

$$k_a = \frac{\frac{1}{2}\log n + \frac{1}{2}a^2 - c\log n/\log\log n + O_{G_0}(a+1)}{\log(1+a/\sqrt{\lambda_*})} + O(1).$$

The final $O(1)$ is the rounding error. After multiplication by $\log(1+a/\sqrt{\lambda_*}) = O(\log\log n)$, it is negligible compared with the margins of order $\log n/\log\log n$ below. Uniformly for $A_0 \le a \le R_n$, and with $c > 0$ small enough, there are constants $\delta_0, c_1, c_2 > 0$, depending only on G_0 and γ, such that this threshold satisfies

$$k_a \ge (1+\delta_0)(\sqrt{\lambda_*}+a)^2, \qquad k_a \log\frac{k_a}{e\lambda_*} \ge \log n + c_1\frac{\log n}{\log\log n},$$

and

$$k_a \log\frac{k_a}{e(\sqrt{\lambda_*}+a)^2} + (\sqrt{\lambda_*}+a)^2 \ge c_2\frac{\log n}{\log\log n}.$$

To verify these inequalities uniformly in a, write $L = \log n$, $\ell = \log L$, and $y = a^2\ell/L$. On every fixed interval $y \in [\delta,\gamma]$,

$$k_a = \frac{L}{\ell}\left[1 + \frac{\log\ell - \log y + \log\lambda_* + y - 2c + o(1)}{\ell}\right].$$

Consequently, if $I_\mu(k) = k\log(k/\mu) - k + \mu$ denotes the Poisson Chernoff exponent, then, uniformly on this interval,

$$I_{\lambda_*}(k_a) = L + \{y - \log y - 1 - 2c + o(1)\}\frac{L}{\ell},$$

and, for $\mu_a = (\sqrt{\lambda_*}+a)^2$,

$$I_{\mu_a}(k_a) = \{y - \log y - 1 + o(1)\}\frac{L}{\ell}.$$

The function $y - \log y - 1$ is decreasing on $(0,1)$, so the two rate functions have a fixed positive margin on $[\delta,\gamma]$. Also $k_a/\mu_a = 1/y + o(1)$ there, uniformly, and is therefore bounded away from one.

The expansion is not uniform when y approaches zero. For that range, fix $\delta \in (0, \min\{\gamma, e^{-3}\})$, and write

$$w_a = 2\log(1+a/\sqrt{\lambda_*}) = \log(\mu_a/\lambda_*).$$

Since $a \le R_n$, the $O(a+1)$ term in the definition of k_a is $o(L/\ell)$, uniformly. Hence

$$k_a w_a \ge L + \mu_a - 3cL/\ell \tag{4.2}$$

for all sufficiently large n. If $w_a \le \ell/2$, then $\mu_a = \lambda_* e^{w_a} \le \lambda_*\sqrt{L}$, while $k_a \ge (2-o(1))L/\ell$. Thus $k_a/\mu_a \to \infty$ uniformly in this range and, eventually,

$$I_{\mu_a}(k_a) \ge k_a \ge L/\ell.$$

Suppose instead that $w_a > \ell/2$, and put $z = \mu_a \ell/L$. The assumption $y \le \delta$ gives $z \le 2\delta$ for large n, while $w_a = \ell - \log \ell + \log(z/\lambda_*) \le \ell$. If $r_a = k_a/\mu_a$, then (4.2) yields, uniformly in this range,

$$r_a \ge \frac{\ell + z - 3c}{z w_a} \ge \frac{1 - o(1)}{z}.$$

Choose a fixed $\varepsilon \in (0, 1/4)$. For large n, the right side is at least $(1-\varepsilon)/z$. Since $f(r) = r \log r - r + 1$ is increasing for $r > 1$,

$$\frac{\ell}{L} I_{\mu_a}(k_a) = z f(r_a) \ge (1-\varepsilon) \log \frac{1-\varepsilon}{z} - (1-\varepsilon) + z.$$

The expression on the right is decreasing for $0 < z < 1-\varepsilon$; at $z = 2\delta$ it is a positive constant m_δ. Consequently $I_{\mu_a}(k_a) \ge m_\delta L/\ell$, and k_a/μ_a is uniformly bounded away from one, throughout this second range. Put $\bar{m}_\delta = \min\{1, m_\delta\}$. The two w_a-ranges therefore give $I_{\mu_a}(k_a) \ge \bar{m}_\delta L/\ell$ throughout $y \le \delta$. Finally, the exact identity

$$I_{\lambda_*}(k_a) - I_{\mu_a}(k_a) = k_a w_a + \lambda_* - \mu_a$$

and (4.2) give

$$k_a \log \frac{k_a}{e\lambda_*} = I_{\lambda_*}(k_a) - \lambda_* \ge L + (\bar{m}_\delta - 3c)L/\ell.$$

Choose $c > 0$ smaller than one quarter of $\gamma - \log \gamma - 1$ and smaller than $\bar{m}_\delta/6$. The compact and small-y ranges together prove all three threshold inequalities, after decreasing δ_0, c_1, c_2.

The first rate function controls the threshold probability under p_{λ_*}, and the second controls the truncated second moment after exponential tilting. The identity

$$\left(\frac{p_{\theta_a}(k)}{p_{\lambda_*}(k)}\right)^2 p_{\lambda_*}(k) = e^{a^2} p_{(\sqrt{\lambda_*}+a)^2}(k)$$

therefore gives, by Chernoff's bound,

$$\mathbb{E}_{p_{\lambda_*}}\left[h_{a,*}(X)^2 \mathbf{1}_{\{|h_{a,*}(X)| > \rho_n \sqrt{n}\}}\right] \le e^{-c \log n/\log\log n}$$

and

$$n\mathbb{P}_{p_{\lambda_*}}\{|h_{a,*}(X)| > \rho_n \sqrt{n}\} \le e^{-c \log n/\log\log n}.$$

For the product tail, use $\{|h_a h_b| > n\rho_n\} \subset \{|h_a| > \sqrt{n\rho_n}\} \cup \{|h_b| > \sqrt{n\rho_n}\}$. Repeating the one-coordinate calculation at the larger threshold $\sqrt{n\rho_n}$ gives, for some $c_3 > 0$, a uniform second-moment tail bound $\exp\{-c_3 \log n/\log\log n\}$. Cauchy–Schwarz on the two events therefore bounds the product tail by $2\exp\{-c_3 \log n/(2\log\log n)\}$. Decreasing the constant c in the definition of ρ_n makes this at most ρ_n. Finally,

$$(h_a h_b)^2 \mathbf{1}_{\{|h_a h_b| \le n\rho_n\}} \le n\rho_n |h_a h_b|,$$

and $\mathbb{E}|h_a h_b| \le 1$. □

4.2 Constructive Likelihood Gain

The following proposition constructs the unrestricted likelihood gain. It moves a small amount of mass from λ_* to a separated grid of larger means $\lambda_* + a_j\sqrt{\lambda_*}$.

Proposition 4.3. *Fix $\gamma \in (0,1)$, and put $R_n = \sqrt{\gamma \log n / \log\log n}$. Choose Δ large enough, let $a_j = A_0 + \Delta(j-1) \le R_n$, and let J be the number of such points. Then $J \asymp_{G_0,\gamma,\Delta} \sqrt{\log n / \log\log n}$, and there is a random mixing distribution G_n such that*

$$\mathbb{P}\left\{L_n^{\mathrm{Pois}}(G_n; G_0) \ge c_{G_0,\gamma,\Delta} J\right\} \to 1.$$

Proof. For $1 \le j \le J$, write $h_j = h_{a_j}$,

$$d_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n h_j(X_i), \qquad H_{jl} = \mathbb{E} h_j(X) h_l(X), \qquad \widehat{H}_{jl} = \frac{1}{n} \sum_{i=1}^n h_j(X_i) h_l(X_i).$$

Thus d_j is the empirical score in direction a_j, while H and $\widehat{H}$ are the population and empirical Gram matrices of these scores. Choose Δ so large that the off-diagonal row sums of H are at most $1/10$; this is possible by Lemma 4.2. The truncation bounds in Lemma 4.2, followed by Chebyshev's inequality and a union bound over j, l, give

$$\|\widehat{H} - H\|_{\mathrm{op}} = o_{\mathbb{P}}(1).$$

Here are the details. Put

$$Y_i^{jl} = h_j(X_i) h_l(X_i) \mathbf{1}_{\{|h_j(X_i) h_l(X_i)| \le n\rho_n\}}.$$

The truncation lemma gives

$$|\mathbb{E} Y_i^{jl} - H_{jl}| \le \rho_n, \qquad \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i^{jl}\right) \le \rho_n.$$

Moreover, with probability tending to one, $\max_{i,j} |h_j(X_i)| \le \rho_n \sqrt{n}$, because $nJ \max_j \mathbb{P}\{|h_j(X)| > \rho_n \sqrt{n}\} \le J\rho_n \to 0$. On this event no product is truncated. Chebyshev's inequality, followed by a union bound over the J^2 pairs, therefore yields

$$\max_{j,l} |\widehat{H}_{jl} - H_{jl}| = O_{\mathbb{P}}(\rho_n^{1/4}).$$

Since $J\rho_n^{1/4} \to 0$, the operator-norm claim follows.

Let $S = \sum_{j=1}^J (d_j^+)^2$. We claim that a positive fraction of the score coordinates contribute to S. Fix a large constant B and let $\psi_B(x) = (x^+)^2 \wedge B$. To obtain a uniform Gaussian approximation, truncate each h_j at $\rho_n \sqrt{n}$, then recenter and renormalize it. If

$$A_{jl} = \{|h_j| > \rho_n \sqrt{n}\} \cup \{|h_l| > \rho_n \sqrt{n}\},$$

then the product-tail estimate gives

$$\mathbb{E}|h_j h_l| \mathbf{1}_{A_{jl}} \le \rho_n + n\rho_n \, \mathbb{P}(A_{jl}) \le \rho_n + 2\rho_n^2.$$

Also $n\mathbb{P}(A_{jl}) \le 2\rho_n$, so the original pair of normalized sums agrees with its truncated version with probability at least $1 - 2\rho_n$. Thus truncation changes each one- or two-coordinate covariance matrix by $O(\rho_n)$, and the matrices remain uniformly nonsingular. The truncated third absolute moment is at most $\rho_n \sqrt{n}\, \mathbb{E} h_j(X)^2$. The fixed-dimensional Berry–Esseen bound in Wasserstein distance therefore gives error $O(\rho_n)$ for every bounded Lipschitz function of one or two normalized coordinates. This applies to ψ_B and to the product test function

$$(x, y) \longmapsto \psi_B(x) \psi_B(y).$$

Because ρ_n decays faster than any power of J, the total approximation error over all J^2 pairs is $o(1)$. Cauchy–Schwarz also gives

$$\left|\mathbb{E}h_j(X)\mathbf{1}_{\{|h_j(X)|>\rho_n\sqrt{n}\}}\right| \le \frac{\rho_n}{\sqrt{n}}.$$

Thus recentering shifts a normalized sum by at most ρ_n, and variance normalization changes it by $O(\rho_n)$. This completes the transfer from the truncated array back to the original coordinates d_j.

For a centered Gaussian pair (Z_j, Z_l) with correlation H_{jl}, the Gaussian maximal correlation inequality gives

$$|\operatorname{Cov}(\psi_B(Z_j), \psi_B(Z_l))| \le C_B|H_{jl}|.$$

The uniform Gaussian approximation and the row-sum bound for H imply that $\sum_j \psi_B(d_j)$ has variance $O_B(J) + o(J^2)$. Choose B so that $\mathbb{E}\psi_B(Z) > 1/3$ for $Z \sim \mathcal{N}(0,1)$. Its mean is then at least $J/3 + o(J)$. Since $\sum_j \psi_B(d_j) \le S$, Chebyshev's inequality shows that $S \ge cJ$ with high probability. The upper bound follows from $\mathbb{E}\sum_j d_j^2 = J$. Thus

$$\mathbb{P}\{S \ge cJ\} \to 1, \qquad S = O_{\mathbb{P}}(J).$$

Fix $\eta = 1/5$, let $r_j = d_j^+/(1+\eta)$, and set $t_j = r_j/(\sqrt{n}\,\sigma_{a_j})$. Since $\inf_j \sigma_{a_j} > 0$ and $\sum_j r_j \le \sqrt{JS} = O_{\mathbb{P}}(J) = o_{\mathbb{P}}(\sqrt{n})$, we have $\sum_j t_j = o_{\mathbb{P}}(1)$. Hence $\sum_j t_j \le \pi_*$ with probability tending to one. On this event define

$$G_n = G_0 - \Big(\sum_{j=1}^J t_j\Big)\delta_{\lambda_*} + \sum_{j=1}^J t_j \delta_{\lambda_* + a_j\sqrt{\lambda_*}},$$

and set $G_n = G_0$ otherwise.

On the event $\sum_j t_j \le \pi_*$, for each observation,

$$\frac{p_{G_n}(X_i)}{p_0(X_i)} = 1 + u_i, \qquad u_i = \frac{1}{\sqrt{n}}\sum_{j=1}^J r_j h_j(X_i).$$

This feasibility event has probability tending to one. Intersect it with the event, also of probability tending to one, on which $\max_{i,j}|h_j(X_i)| \le \rho_n\sqrt{n}$, supplied by Lemma 4.2 and a union bound. On their intersection, $\max_i |u_i| \le \rho_n \sum_j r_j = o_{\mathbb{P}}(1)$, and Taylor expansion gives

$$L_n^{\mathrm{Pois}}(G_n; G_0) = r^\top d - \frac{1}{2} r^\top \widehat{H} r + o_{\mathbb{P}}(J).$$

Indeed, $\sum_i u_i^2 = r^\top \widehat{H} r = O_{\mathbb{P}}(J)$, so the sum of the cubic remainders is bounded by $O(\max_i |u_i| \sum_i u_i^2) = o_{\mathbb{P}}(J)$. Since $r^\top d = S/(1+\eta)$ and $r^\top \widehat{H} r \le (1.2 + o_{\mathbb{P}}(1))S/(1+\eta)^2$, the coefficient of S is positive. The lower bound on S completes the proof. $\square$

4.3 Localized Upper Bound

We now bound every D-atomic Poisson competitor. Lemma 4.1 restricts its means to the interval ending at the sample maximum, which is of order $\log n/\log\log n$. As in the Gaussian case, the required entropy estimate must depend on this parameter radius only through its logarithm.

For $D \ge 1$, $M \ge 2$, and $0 < \delta \le \sqrt{2}$, define

$$\mathcal{F}_D^{\mathrm{Pois},G_0}(M,\delta) = \left\{\sqrt{p_G/p_0} - 1 : |\operatorname{supp}(G)| \le D,\ \operatorname{supp}(G) \subset [0,M],\ \|\sqrt{p_G/p_0} - 1\|_{L^2(p_0)} \le \delta\right\}.$$

The coordinate $u = \sqrt{\theta}$ makes this logarithmic dependence available: square-root Poisson kernels have Gaussian affinity $\sum_k \sqrt{p_{u^2}(k)p_{v^2}(k)} = e^{-(u-v)^2/2}$. The next lemma combines this geometry with local identifiability around the true finite mixture.

Lemma 4.4. *For fixed G_0 and fixed $C_0 < \infty$, there exists $C_{G_0,C_0} < \infty$ such that, for every $M \geq 2$, $1 \leq D \leq M^{C_0}$, and $0 < \eta \leq \delta \leq \sqrt{2}$,*

$$\log N_{[]}\big(\eta, \mathcal{F}_D^{\mathrm{Pois},G_0}(M,\delta), L^2(p_0)\big) \leq C_{G_0,C_0} D\big(\log M + \log(\delta/\eta)\big).$$

Proof. Put $\rho_G = \|\sqrt{p_G/p_0} - 1\|_2$. We first record the local geometry needed when $\rho_G \leq \rho_0$. Choose relatively open, bounded intervals A_i around the true means λ_i, with disjoint closures, and let A_0 be their complement. We need the following local identifiability bound. Hellinger distance controls the mass outside the true-atom neighborhoods, as well as the first two local moments of the perturbation inside each neighborhood. For

$$m_i = G(A_i), \qquad b_i = \int_{A_i} (\theta - \lambda_i)\, dG(\theta), \qquad s_i = \int_{A_i} (\theta - \lambda_i)^2\, dG(\theta),$$

the Poisson analogue of the finite-mixture local-geometry bound is

$$G(A_0) + \sum_i \{|m_i - \pi_i| + |b_i| + s_i\} \leq C\rho_G. \tag{4.3}$$

We first prove the corresponding linear statement. Let μ be a finite nonnegative measure supported on A_0, and let $\eta_i, \beta_i \in \mathbb{R}$ and $\sigma_i \geq 0$. The linearized density perturbation is

$$p_\mu + \sum_i \{\eta_i p_{\lambda_i} + \beta_i \partial_\lambda p_{\lambda_i} + \tfrac{1}{2}\sigma_i \partial_\lambda^2 p_{\lambda_i}\},$$

Its ℓ^1-norm is bounded below by a fixed multiple of

$$\mu(A_0) + \sum_i (|\eta_i| + |\beta_i| + \sigma_i).$$

Suppose this lower bound fails. Normalize $\mu(A_0) + \sum_i(|\eta_i| + |\beta_i| + \sigma_i)$ to one, and choose a sequence whose linearized expression tends to zero in ℓ^1. The coefficient vectors have a convergent subsequence. The corresponding outer measures μ_n are also tight. Indeed, if v_n denotes the finite linear combination of the fixed kernel derivatives, then for every integer K,

$$\sum_{k>K} p_{\mu_n}(k) \leq \|p_{\mu_n} + v_n\|_1 + \sum_{k>K} |v_n(k)|.$$

The second term tends to zero uniformly in n as $K \to \infty$. Moreover,

$$\mu_n([R,\infty)) \inf_{\theta \geq R} \mathbb{P}_\theta\{X > K\} \leq \sum_{k>K} p_{\mu_n}(k),$$

and the infimum tends to one as $R \to \infty$, for fixed K. Thus a further subsequence satisfies $\mu_n \Rightarrow \mu$ for a finite measure μ on A_0. Passing to probability generating functions gives the Laplace transform of

$$\mu + \sum_i \{\eta_i \delta_{\lambda_i} - \beta_i \delta'_{\lambda_i} + \tfrac{1}{2}\sigma_i \delta''_{\lambda_i}\}$$

identically zero. Uniqueness of Laplace transforms first eliminates μ away from the λ_i's, and linear independence of the remaining point masses and their derivatives eliminates every coefficient. This contradicts the normalization.

To pass from linear identifiability to the nonlinear mixture bound, we use the uniform Taylor remainder

$$\|p_\theta - p_{\lambda_i} - (\theta-\lambda_i)p'_{\lambda_i} - \tfrac{1}{2}(\theta-\lambda_i)^2 p''_{\lambda_i}\|_1 \le C\ \mathrm{diam}(A_i)(\theta-\lambda_i)^2.$$

This follows from $\partial_\theta^j p_\theta(k) = \sum_{r=0}^{j}(-1)^{j-r}\binom{j}{r}p_\theta(k-r)$, with $p_\theta(m)=0$ for $m<0$, which also gives a uniform ℓ^1-bound on the third derivative. Shrinking the fixed intervals A_i therefore proves the lower bound in total variation. Since total variation is at most twice Hellinger distance, (4.3) follows.

Let $a = p_{G|_{A_0^c}}$ and $s = \sqrt{a/p_0}$. Taylor expansion in the fixed intervals and (4.3) give

$$\|s-1\|_2 \le C\rho_G, \qquad W := G(A_0) \le C\rho_G. \tag{4.4}$$

Fix this inner subdensity a. We next bracket the increment created by adding one outer component. Write $q_u = p_{u^2}$, where $0 \le u \le \sqrt{M}$, put $b = wq_u$, and set $\nu(k) = b(k)^2/\{a(k)+b(k)\}$. After decreasing ρ_0, the local-geometry bound and $\lambda_* > 0$ ensure that the inner submixture assigns a fixed positive mass to an interval $[\ell, U] \subset (0,\infty)$. The ratios of consecutive moments of the finite measure $e^{-\theta}G|_{A_0^c}(d\theta)$ are nondecreasing, and hence, using $a(k) = k!^{-1}\int e^{-\theta}\theta^k\, dG|_{A_0^c}(\theta)$,

$$\frac{a(k+1)}{a(k)} \ge \frac{c}{k+1}, \qquad \frac{q_u(k+1)}{q_u(k)} = \frac{u^2}{k+1} \quad (u>0).$$

It follows, by cross-multiplication, that, for $u>0$,

$$(k+1)\nu(k+1) \le C(1+M)^2\nu(k).$$

Put $A = C(1+M)^2$. The recursion implies, for every $t \ge 0$,

$$\frac{\sum_k e^{tk}\nu(k)}{\sum_k \nu(k)} \le \exp\{A(e^t-1)\}.$$

Indeed, the logarithmic derivative of the numerator is at most Ae^t. Taking $t = c/A$ gives uniform exponential moments on count scale $(1+M)^2$. For $u \ge 1/2$, differentiation of p_{u^2} and Lemma 2.3 give

$$\|\partial_u h_{a,w,q_u}\|_2 \le C(1+M)^2 e_{a,w,q_u}.$$

The interval $0 \le u \le 1$ is handled directly. Indeed, $a(k) \ge c\ell^k/k!$, while termwise differentiation of $p_{u^2}(k) = e^{-u^2}u^{2k}/k!$ gives

$$\sup_{0\le u\le 1}|\partial_u^j p_{u^2}(k)| \le C(k+1)/k!, \qquad j=0,1.$$

These envelopes are square summable after division by $a(k)$, and, for the mixture weights $0 < w \le 1$ considered here, the zeroth count coordinate gives $e_{a,w,q_u} \ge cw$. Hence the same derivative bound holds on $[0,1]$, with a fixed constant.

On intervals of length $c(1+M)^{-2}$ in $[1/2, \sqrt{M}]$, the exact ratio

$$\frac{p_{(u+v)^2}(k)}{p_{u^2}(k)} = e^{-2uv-v^2}(1+v/u)^{2k} \le e^{C+Ck/(1+M)^2}$$

and the exponential-moment bound for the count variable obtained above control the pointwise suprema of the parameter derivatives. On $[0, 1]$, the power-series envelopes do the same. The weight coordinate is model-free: Lemma 2.3 shows that the logarithmic derivative of the increment norm lies in $[1/2, 1]$. Repeating the parameter-box and weight-level construction in (3.4) therefore gives, uniformly in the inner subdensity,

$$\log N_{[]} \left(\epsilon, \{h_{a,w,q_u} : 0 \leq u \leq \sqrt{M},\ 0 < w \leq 1,\ e_{a,w,q_u} \leq r\}, L^2(p_0)\right) \leq C\{\log M + \log_+(r/\epsilon)\}. \tag{4.5}$$

The one-contaminant bound treats the inner subdensity as fixed. To count the full mixture class, we must also bracket the inner atoms, all of which lie in the fixed neighborhoods A_i. We verify the needed entropy bound because the Poisson family is not a translation family. Let $\mathcal{D}_D^{\text{in}}$ be the normalized nonzero Hellinger secants

$$\mathcal{D}_D^{\text{in}} = \left\{ \frac{\sqrt{p_H/p_0} - 1}{\|\sqrt{p_H/p_0} - 1\|_2} : 0 < \|\sqrt{p_H/p_0} - 1\|_2 \leq \rho_0,\ |\text{supp}(H)| \leq D, \quad \text{supp}(H) \subset \bigcup_i A_i \right\}.$$

The proof of [GvH14, Theorem 3.1] approximates precisely such normalized secants in the Banach lattice $L^2(p_0)$. As explained in their Remark 3.6, translation structure is needed for their local-geometry theorem, while the entropy argument itself applies to smooth parametric kernels. Equation (4.3) supplies the required local geometry. To verify smoothness, use the mean parameter and define, on the fixed union of the A_i's,

$$H_j(k) = \sup_\theta \frac{|\partial_\theta^j p_\theta(k)|}{p_0(k)}, \qquad 0 \leq j \leq 3.$$

The finite-difference formula for $\partial_\theta^j p_\theta$, valid also at $\theta = 0$, reduces each numerator to finitely many shifted Poisson masses. If the inner intervals lie in $[0, B]$, then, outside a fixed finite set of counts, their pointwise suprema are bounded by a polynomial in k times $p_B(k)$. Since $p_0 \geq \pi_* p_{\lambda_*}$ for some $\lambda_* > 0$, the identity

$$\sum_{k \geq 0} (1 + k)^m \frac{p_B(k)^s}{p_{\lambda_*}(k)^{s-1}} < \infty, \qquad m < \infty, \quad s \in \{2, 4\},$$

yields

$$H_0, H_1, H_2 \in \ell^4(p_0), \qquad H_3 \in \ell^2(p_0).$$

The pointwise Taylor bound and (4.3) give the analogue of [GvH14, Lemma 3.13]. Specifically, the density ratios normalized by their ℓ^1-distance from p_0 have the $\ell^4(p_0)$ envelope

$$\frac{|p_H/p_0 - 1|}{\|p_H - p_0\|_1} \leq C(H_0 + H_1 + H_2).$$

The algebraic comparison between normalized Hellinger and chi-square secants in their Lemma 3.15 is measure-free. Their Lemma 3.16 treats atoms close to a true mean by a second-order Taylor expansion and retains atoms of small weight as individual kernel terms; the displayed H_j bounds control the resulting remainder in $\ell^2(p_0)$. Finally, pad the candidate to $D \vee K$ atoms and apply the coefficient and location grids in the proof of their Theorem 3.1. The envelopes $H_0, H_1, H_2 \in \ell^4(p_0)$, together with $H_3 \in \ell^2(p_0)$, justify each Taylor remainder in counting measure, and taking the pointwise minimum and maximum over every parameter cell produces ordered brackets exactly as in their proof. Those grids have at most $(CD/\xi)^{CD}$ elements, and hence

$$\log N_{[]}\left(\xi, \mathcal{D}_D^{\text{in}}, L^2(p_0)\right) \leq CD \log(CD/\xi).$$

The general slicing argument in [GvH14, Theorem 2.2] converts this normalized-secant bound into brackets of width ϵ for the inner Hellinger ball of radius r, at entropy $CD\log(CDr/\epsilon)$. Finally, write an inner subdensity as $a = m\bar{a}$. Then $|\sqrt{m}-1| \le \|\sqrt{a/p_0}-1\|_2$ and $\|\sqrt{\bar{a}/p_0}-1\|_2 \le Cr$; bracketing $\sqrt{m}$ separately adds one scalar coordinate. This supplies the inner-submixture brackets needed for the shell construction.

We now combine an inner bracket with at most D outer-component brackets, using the composition map from Lemma 2.3. On a shell $r/2 < \rho_G \le r$, positivity and (4.4) give $\|h_{a,w_j,q_{u_j}}\|_2 \le Cr$. Put $\zeta = c\min\{\eta, r\}$, bracket the inner square root and each outer increment at width $\zeta/(CD)$, and choose an actual baseline representative from each nonempty inner bracket. Widen every component bracket by the inner bracket width and compose through Φ. Since $\partial_s\Phi \le \sqrt{D+1}$ and $\partial_{h_j}\Phi \le 1$, the resulting width is at most

$$\sqrt{D+1}\frac{\zeta}{CD} + D\left(\frac{\zeta}{CD} + 2\frac{\zeta}{CD}\right) \le \zeta$$

after increasing C. Equations (4.5) and the inner entropy bound give, for the number N_r of brackets covering this shell,

$$\log N_r \le CD\{\log M + \log D + \log_+(r/\eta)\}. \tag{4.6}$$

For the innermost ball, Taylor expansion and (4.3) give a fixed $F_0 \in \ell^2(p_0)$ such that the inner density ratio differs from one by at most $C\rho_G F_0$. The outer ratio is bounded by $C\rho_G \sup_{0\le\theta\le M} p_\theta/p_0$. Since $p_0 \ge \pi_* p_{\lambda_*}$ and $\sup_{\theta\le M} p_\theta(k) = p_{k\wedge M}(k)$, Stirling's formula for $k \le M$ and the tilted-Poisson identity for $k > M$ give a common envelope of $\ell^2(p_0)$-norm at most $e^{C(1+M^2)}$. Thus the ball $\rho_G \le \eta e^{-C(1+M^2)}$ fits in one bracket. Summing the dyadic shell collections adds only $C\{\log M + \log_+(\delta/\eta)\}$ to (4.6); $D \le M^{C_0}$ absorbs $D\log D$.

For $\rho_G \ge \rho_0$, global L^1 density brackets suffice. Use the coordinate $u = \sqrt{\theta}$, for which $\|\partial_u p_{u^2}\|_1 \le 2$. For every interval $I \subset [0, \sqrt{M}]$, the pointwise endpoints

$$\ell_I(k) = \inf_{u\in I} p_{u^2}(k), \qquad u_I(k) = \sup_{u\in I} p_{u^2}(k)$$

form a nonnegative order bracket. The fundamental theorem of calculus gives

$$\sum_k \{u_I(k) - \ell_I(k)\} \le \int_I \|\partial_u p_{u^2}\|_1\, du \le 2|I|.$$

Partitioning $[0,\sqrt{M}]$ at mesh $c\eta^2$ and quantizing the weights at mesh $c\eta^2/D$ produces mixture brackets of L^1-width at most $C\eta^2$. The number of location and weight choices is bounded by $\exp\{CD(\log M + \log(1/\eta))\}$. Thus

$$\log N_{[]}\big(\eta^2, \{p_G : |\operatorname{supp}(G)| \le D,\ \operatorname{supp}(G) \subset [0,M]\}, L^1\big) \le CD\{\log M + \log(1/\eta)\}.$$

Taking square roots and using $\delta \ge \rho_0$ completes the proof. □

We now apply the localized likelihood inequality to the entropy bound. The result controls all sparse Poisson competitors simultaneously and will be compared with Proposition 4.3.

Proposition 4.5. *Fix $C_0 < \infty$. Let D_n be positive integers satisfying $D_n \le (\log n)^{C_0}$ and $D_n \log\log n = o(n)$. Then*

$$\mathbb{P}\left\{ \sup_{|\operatorname{supp}(G)|\le D_n} L_n^{\mathrm{Pois}}(G; G_0) \le C_{G_0,C_0} D_n \log\log n \right\} \to 1.$$

Proof. Let $M_n = \max_i X_i$, and set $m_n = \lfloor 2 \log n / \log\log n \rfloor$. Since G_0 is fixed and finite, a Chernoff bound gives $\mathbb{P}\{M_n \le m_n\} \to 1$. By Lemma 4.1, on this event all D_n-atomic competitors may be restricted to $[0, m_n]$.

Let $A_n = C_{G_0,C_0} D_n \log\log n$. Since $D_n \le m_n^{C_0+1}$ eventually and $\log m_n \le \log\log n + O(1)$, Lemma 4.4, with exponent $C_0 + 1$, gives

$$\log N_{[]}\big(\eta, \mathcal{F}^{\mathrm{Pois},G_0}_{D_n}(m_n, \delta), L^2(p_0)\big) \le A_n(1 + \log(\delta/\eta)).$$

The localized bracketing integral is therefore $O(\delta\sqrt{A_n})$, and Lemma 2.2 proves the proposition. □

Proof of Theorem 1.2. Apply Proposition 4.3 with any fixed $\gamma \in (0,1)$. Since $J \asymp \sqrt{\log n / \log\log n}$, the unrestricted likelihood is at least $c\sqrt{\log n / \log\log n}$ with probability tending to one. Proposition 4.5 bounds every D_n-atomic competitor by $C D_n \log\log n$ with high probability. Taking

$$D_n = \left\lfloor c_{G_0} \frac{\sqrt{\log n}}{(\log\log n)^{3/2}} \right\rfloor$$

with c_{G_0} small enough, no D_n-atomic mixing distribution can maximize on the intersection of the two high-probability events. This proves the theorem. □

References

[AGM09] Jean-Marc Azaïs, Élisabeth Gassiat, and Cécile Mercadier. The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM Probab. Stat.*, 13:301–327, 2009.

[Ano26a] Radial phase transitions for Gaussian mixture NPMLEs in fixed and growing dimension, 2026. In preparation.

[Ano26b] Variance misspecification and the one-dimensional Gaussian mixture NPMLE, 2026. In preparation.

[CW26] Jiafeng Chen and Yihong Wu. Sharp regret-Hellinger bounds for Gaussian empirical Bayes via polynomial approximation. *arXiv preprint arXiv:2605.02070*, 2026.

[DWYZ23] Natalie Doss, Yihong Wu, Pengkun Yang, and Harrison H Zhou. Optimal estimation of high-dimensional Gaussian location mixtures. *Ann. Stat.*, 51(1):62–95, 2023.

[DZ16] Lee H Dicker and Sihai D Zhao. High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika*, 103(1):21–34, 2016.

[GGMT26] Subhroshekhar Ghosh, Adityanand Guntuboyina, Satyaki Mukherjee, and Hoang-Son Tran. Gaussian mixtures and non-parametric likelihoods through the lens of statistical mechanics. *arXiv preprint arXiv:2603.23196*, 2026.

[GvH14] Elisabeth Gassiat and Ramon van Handel. The local geometry of finite mixtures. *Trans. Amer. Math. Soc.*, 366(2):1047–1072, 2014.

[Har85] John A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 807–810. Wadsworth, Belmont, CA, 1985.

[JZ09] Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical Bayes estimation of normal means. *Ann. Stat.*, 37(4):1647–1684, 2009.

[JZ19] Wenhua Jiang and Cun-Hui Zhang. Rate of divergence of the nonparametric likelihood ratio test for Gaussian mixtures. *Bernoulli*, 25(4B), 2019.

[KG19] Roger Koenker and Jiaying Gu. Minimalist g-Modeling. *Statistical Science*, 34(2):209 – 213, 2019.

[KM14] Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.

[KW56] Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.*, pages 887–906, 1956.

[Lin83] Bruce G Lindsay. The Geometry of Mixture Likelihoods: A General Theory. *Ann. Stat.*, 11(1):86–94, 1983.

[LR93] Bruce G Lindsay and Kathryn Roeder. Uniqueness of estimation and identifiability in mixture models. *Canadian J. Stat.*, 21(2):139–147, 1993.

[PS25] Yury Polyanskiy and Mark Sellke. Nonparametric MLE for Gaussian location mixtures: Certified computation and generic behavior. *arXiv preprint arXiv:2503.20193*, 2025.

[PW20] Yury Polyanskiy and Yihong Wu. Self-Regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models. *arXiv preprint arXiv:2008.08244*, 2020.

[Rob50] Herbert Robbins. A generalization of the method of maximum likelihood: Estimating a mixing distribution (Abstract). *Annals of Mathematical Statistics*, 21(2):314–315, 1950.

[SG20] Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Ann. Stat.*, 48(2):738–762, 2020.

[SGS25] Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):1–32, 2025.

[vdG00] Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[WS95] Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Stat.*, 23(2):339–362, 1995.

[Zha09] Cun-Hui Zhang. Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica*, pages 1297–1318, 2009.

[ZV25] Yan Zhang and Stanislav Volgushev. On a surprising behavior of the likelihood ratio test in nonparametric mixture models. *arXiv preprint arXiv:2509.05610*, 2025.

[ZV26] Yan Zhang and Stanislav Volgushev. Adaptivity of the NPMLE to finitely discrete mixing distributions in Gaussian/Poisson mixtures. *arXiv preprint arXiv:2604.12087*, 2026.