Learning Finite Gaussian Location Mixtures via Gated Moments and Posterior Covers

Zhijie Zheng[1] Chengzhi Xie[2] Yucong He[3]

[1]Peking University
[2]National University of Singapore
[3]Beijing Institute of Technology

Abstract

In an exact-real model, we give an improper distribution learner for mixtures of at most k Gaussians in $\mathbb{R}^d$ with arbitrary means and weights and known positive-definite common covariance. With probability at least $1-\delta$ over training from independent samples, it returns a sampling program whose output law is within ε total variation of the target. The sample complexity, number of training operations, real-register description size, and deterministic number of operations per output are fixed polynomials in d, k, ε^{-1} and $\log(1/\delta)$. Operation counts use exact real arithmetic, elementary functions, and ideal Gaussian randomness. The bounds are uniform over coincident means, unbounded mean locations, and arbitrarily small positive weights. For identity covariance, Diakonikolas and Kane obtained quasipolynomial dependence on the number of components for unrestricted means (2020), and polynomial complexity under a bounded-radius assumption (2025). Our result extends polynomial distribution learning to unrestricted mean configurations.

The main step is a localization procedure whose error depends on omitted probability mass, even when the omitted means are arbitrarily distant. The Kullback–Leibler divergence from the normalized retained submixture to the fitted density is bounded by a universal constant times this mass plus a prescribed accuracy. Robust posterior mean estimation then refines the cover. This control permits bounded-radius moment estimates to be composed across logarithmically many noise scales at polynomial cost. Independent Gaussian views and gated, recursively compressed moments supply the local density representation.

For identity covariance and a known lower bound on positive weights, we also give a separate learner with polynomial bit cost from fixed-precision quantized observations. Its guarantee compares the discrete output to the target after an explicit Gaussian convolution.

Contents

1 Introduction and main result **1**
1.1 From local moment estimates to a global learner . . . 3

2 Three components of the learner **3**
2.1 An explicit schedule without a given radius bound . . . 5
2.2 Proof of the statistical guarantee . . . 5
2.3 Polynomial cost on every outcome . . . 6

3 Independent views and finite full-jet gates **6**
3.1 Conditional independence and repeated evaluation of gate histories . . . 6
3.2 The analytic jet bound . . . 7

4 Recursive compression of accepted moments **9**

5 A finite square-root density dictionary **11**
5.1 Projection error on the original covered atoms . . . 11
5.2 Hermite approximation and a trace-one comparator . . . 11
5.3 Finite Hermite evaluation . . . 12
5.4 Radial clipping and deterministic bounds . . . 13

6 Directional likelihood fitting with a finite optimizer **13**
6.1 Finite concentration events . . . 14
6.2 Penalized fit and directional stationarity . . . 14
6.3 The directional KL estimate . . . 15
6.4 A deterministic polynomial optimization cap . . . 16

7 Global and posterior rejection with deterministic caps **16**

8 A posterior cover with dimension-free mass loss **17**
8.1 Posterior localization including remote atoms . . . 17
8.2 Conditional KL and finite-covariance regularity . . . 18
8.3 An explicit weighted robust mean filter . . . 18
8.4 The finite list and its covered mass . . . 20

9 Computational guarantees and finite-bit learning **20**

A Finite-bit learning with a minimum weight and Gaussian smoothing **21**
A.1 Analytic reconstruction from a given cover . . . 21
A.2 Bounded-radius posterior refinement . . . 27
A.3 Rational linear algebra and stable gates . . . 28
A.4 Rational windows and posterior evaluation . . . 29
A.5 A quantitative precision and bit budget . . . 30
A.6 Discrete rejection and Gaussian smoothing . . . 34
A.7 Removing the radius assumption with quantized pilot samples . . . 35

1 Introduction and main result

A Gaussian mixture can have nearly indistinguishable components and, at the same time, groups of components separated by arbitrarily large distances. Distribution learning remains meaningful across both extremes: it seeks an approximation to the combined distribution, so coincident components can be represented together and components of sufficiently small total weight can be omitted at a controlled cost in total variation. The computational question is whether this flexibility permits learning in time polynomial in the dimension, the number of components, and the inverse accuracy, uniformly over the geometry and mixing weights.

Diakonikolas and Kane [DK20, Theorem 30 of the full version] gave a total-variation density learner for identity-covariance Gaussian mixtures with arbitrary weights and unrestricted means, with quasipolynomial dependence on the number of components. Their later implicit-moment framework [DK25] gives polynomial complexity when the means lie in a ball of radius $O(\sqrt{\log k})$. The unrestricted-mean algorithm outputs a Gaussian mixture with additional components. The bounded-radius algorithm constructs a sampler from an approximate density-ratio evaluator. Extending the latter type of guarantee to arbitrary mean configurations requires local estimates that remain reliable in the presence of mass outside the region being modeled.

In the exact-real model specified below, we obtain an improper sampling program with fixed-polynomial complexity for arbitrary means and weights. This result improves the component-count dependence of [DK20] and extends the polynomial distribution-learning guarantee of [DK25] to unrestricted mean configurations. The exact-real resource bounds are uniform in the radius of the means and the smallest positive weight. The essential advance is to control localization error by omitted probability mass throughout the algorithm. This makes bounded-radius moment methods applicable to the full distribution even when distant, low-weight components are present.

Let $\nu = \sum_{i=1}^{s} w_i \delta_{\mu_i}$, where $s \le k$, $w_i > 0$, $\sum_i w_i = 1$, and $\mu_i \in \mathbb{R}^d$. We observe independent samples from

$$P = \nu * \mathcal{N}(0, I_d). \tag{1}$$

The means may be arbitrary, including coincident or nearly coincident pairs. The output is a sampling program whose law may lie outside the class of k-component Gaussian mixtures.

Theorem 1.1 (Exact-real distribution learning)**.** *Given $d, k \ge 1$, $0 < \varepsilon, \delta < 1/4$, and independent samples from (1), there is an algorithm using*

$$\mathrm{poly}\big(d, k, \varepsilon^{-1}, \log(1/\delta)\big)$$

samples and operations that returns a finite randomized sampling program. Its real-register description size and deterministic operation count per invocation satisfy a bound of the same form. With probability at least $1 - \delta$ over training, the distribution of a single output $\widehat{P}$ satisfies $\mathrm{TV}(P, \widehat{P}) \le \varepsilon$. The exponents in these polynomial bounds are universal.

The model counts real additions, subtractions, multiplications, divisions, comparisons, square roots, exponentials, and logarithms as elementary operations and supplies independent ideal standard Gaussian draws. Integers and prescribed ceilings can be implemented by comparisons and doubling. Uniform randomness is available, for example, as $\exp(-(G_1^2 + G_2^2)/2)$ with independent standard normals. Register size is the number of real coefficients and program instructions; binary encoding length is a separate measure. Every loop has a prescribed finite cap, including on unsuccessful training outcomes. The final cap can introduce an atom at a fixed fallback point; its mass contributes to the total variation error.

For a known positive-definite common covariance, whitening reduces the problem to (1), and the inverse transformation maps the learned sampler back to the original coordinates. Theorem 1.1 therefore applies with this covariance supplied and measures resources in exact-real operations, including when the means have arbitrary magnitude.

Separately, given a lower bound α on all positive weights, Appendix A gives a finite-bit learner from observations quantized to $b = O((\log A)^2 + \log(1/\delta))$ fractional bits per coordinate, with polynomial bit cost in $A = 2 + d + k + \alpha^{-1} + \varepsilon^{-1} + \log(1/\delta)$ and the actual input encoding length. With probability at least $1 - \delta$ over training, convolving its discrete output law with $\mathcal{N}(0, \tau^2 I_d)$, $\tau = \varepsilon/(100d)$, gives a distribution within ε total variation of P. The quantized observations and supplied weight bound lead to a separate reconstruction algorithm for identity covariance.

Relation to prior work. Polynomial sample bounds are already available for much broader classes of Gaussian mixtures. Acharya, Jafarpour, Orlitsky, and Suresh [AJOS14] gave a learner for spherical mixtures with polynomial sample complexity and running time polynomial in the dimension for fixed k and accuracy; its dependence on k is exponential. Sample compression yields sample bounds optimal up to logarithmic factors for general Gaussian mixtures, with exponential running time [ABH+20]. For fixed k and a fixed bound on the means, He and Cheng [HC26] give a proper estimator attaining the optimal Hellinger rate in polynomial arithmetic time, with the polynomial degree depending on k. For known common covariance, Theorem 1.1 gives an improper sampler whose polynomial resource bounds have universal exponents, including in k, in the stated exact-real model.

Recovering the individual means has a separate statistical constraint. Regev and Vijayaraghavan [RV17] showed that separation below order $\sqrt{\log k}$ can require superpolynomially many samples for parameter estimation, even for uniform mixtures of identity-covariance Gaussians. Total-variation learning instead permits nearby components to be represented together, as in the improper hypothesis class used here.

Implicit high-order moment representations were introduced in the clustering work of Liu and Li [LL22] and developed into a general estimation framework by Diakonikolas and Kane [DK25]. These representations make high-order moment information accessible without storing the full tensors. Our recursive tensor compression and Hermite evaluation build on this approach. We combine them with localization guarantees that control the contribution of every component accepted by a gate, so the local estimates can be used to learn the full mixture. Gatmiry, Kelner, and Lee [GKL25] obtain a diffusion-based quasipolynomial learner under a minimum-weight assumption, with complexity depending on a global mean-radius bound. Their method uses warm starts and denoising across noise levels.

Gaussian posterior means also have the classical score representation known as Tweedie's formula; see Efron [Efron11]. Our posterior refinement uses samples from a fitted conditional distribution and a covariance bound to obtain a cover whose omitted mass is controlled by the density-fitting error.

The robust estimation argument uses the high-probability regular-subset theorem of Diakonikolas, Kane, and Pensia [DKP21, Theorem 1.4]. We combine its regular-subset guarantee with a finite weighted filter to estimate posterior means. Gaussian concentration and hypercontractivity are standard inputs [BLM13, OD14]; the gate, compression, approximation, fitting, and cover arguments are proved below. The convex optimization step uses a Frank–Wolfe method with an approximate linear oracle; its elementary recurrence is included, with [Jaggi13] as background.

1.1 From local moment estimates to a global learner

Local moment estimation and total-variation approximation respond differently to distant components: a component of small weight may contribute little to total variation while dominating high-order moments. The learner must therefore localize the moment estimates and control the mass lost in doing so. It alternates between fitting a density near a partial cover of the means and using that density to refine the cover at a lower noise level. The central invariant bounds the omitted latent mass. Each refinement increases this mass by at most a universal constant factor plus a prescribed additive error, independently of the distances to omitted means. This uniform coefficient is what allows repeated local learning to retain polynomial cost.

Moments of the accepted measure. Constructing the local density requires moments that remain valid after localization. Filtering an observation and reusing it as an unbiased moment sample generally fails: a standard Gaussian becomes half-normal after conditioning on being nonnegative. We form conditionally independent views and gate only one view. The resulting moments describe the actual accepted latent measure, including any remote components that the gate accepts. Recursive compression of these moments yields a positive semidefinite square-root density dictionary.

A linear cost for omitted mass. In fitting this dictionary, a bound based on the score range incurs a factor such as $b \log(A/\kappa)$ for omitted mass b. Repeating that loss over scales need not yield a polynomial algorithm. Approximate stationarity in the direction of a comparison density instead controls a nonnegative likelihood ratio averaged over the whole mixture. Dropping the omitted part can only decrease this average; renormalizing the retained mass therefore costs only a factor $1/(1-b)$. Taking logarithms and accounting for approximation error gives $D_{\mathrm{KL}}(F\|q) \le -\log(1-b) + a$, where F is the normalized retained submixture. Thus the coefficient of the omitted mass is universal.

Posterior refinement with dimension-free mass loss. To obtain new centers, we add independent Gaussian noise with the current noise covariance to an observation and sample from the fitted posterior of the original observation. For the clean channel, twice the posterior mean minus the noisy query is the posterior mean of the latent location; Gaussian localization puts this point near the generating mean on most queries. By the KL chain rule, the fitted posterior is within a fixed small total variation distance of the clean one except on a set of queries whose probability is bounded by a constant times $D_{\mathrm{KL}}(F\|q)$. A posterior first-moment estimate based only on a Euclidean second moment can lose a factor $\sqrt{d}$. We bound the clean posterior covariance in operator norm, use a robust mean filter, and apply the same affine transformation to its estimate. This gives cover radius $O(\sigma\sqrt{\log(k/\beta)})$ and omitted mass $O(\beta + D_{\mathrm{KL}}(F\|q))$, with a dimension-independent constant in the latter. Together, the fitting and refinement bounds allow logarithmically many noise levels while keeping all resource bounds polynomial. Section 2 makes this composition explicit and proves the main theorem from the three component guarantees. The subsequent sections establish those guarantees; Appendix A develops the finite-bit result.

2 Three components of the learner

Write $P_\sigma = \nu * \mathcal{N}(0, \sigma^2 I_d)$ for $\sigma \ge 1$. Samples are obtained from fresh $X \sim P$ by returning $X + \sqrt{\sigma^2 - 1}\, G$. The three results below form one refinement step. The dictionary represents the normalized covered submixture; the fit learns from the full observation distribution; and the posterior cover translates the fitting error into a bound on the latent mass omitted at the next scale.

Theorem 2.1 (Partial-cover dictionary). *Suppose $m \geq 1$ given centers c_j cover at least $1-b$ of the latent mass within radius r, where $b \leq 1/2$. For $0 < \theta < 1/4$, a finite sample-based procedure constructs*

$$Q = \frac{1}{m}\sum_{j=1}^{m} \mathcal{N}(c_j, \sigma^2 I_d), \qquad z : \mathbb{R}^d \longrightarrow \mathbb{R}^N,$$

with $z_0 \equiv 1$ and a known deterministic bound $\|z(x)\| \leq L_z$. With any prescribed success probability, there exists $S \succeq 0$, $\operatorname{tr} S \leq 1$, such that

$$\left\| z^T S z - F/Q \right\|_{L^2(Q)} \leq \theta, \tag{2}$$

where F is the normalized submixture of all covered atoms. The matrix S and the assignment of atoms to centers are used only in the analysis and are neither inputs nor outputs of the algorithm. The feature dimension, norm bound, samples, operation count, and stored circuits are polynomial in $d, k, m, \theta^{-1}, \exp(O((r/\sigma)^2))$, and logarithmic inverse failure probability.

Theorem 2.2 (Fitting with linear omitted-mass loss). *Let Q, z be fixed before fresh score samples are drawn from $P' = (1-b)F + bH$, with $b \leq 1/2$ and arbitrary H. Suppose (2) holds. For $0 < a \leq 1$, set*

$$\kappa = \frac{a}{64}, \quad \theta \leq \frac{a\sqrt{\kappa}}{64}, \quad \gamma = \frac{a}{1024}, \quad r_0 = \frac{a}{256}, \quad u = \epsilon_R = \frac{a}{16}. \tag{3}$$

A finite procedure returns $C \succeq 0$, $\operatorname{tr} C \leq 4$, and the numerator

$$f(x) = \kappa + z(x)^T C z(x), \qquad \kappa \leq f(x) \leq A_f := \kappa + 4L_z^2,$$

such that, with the prescribed confidence, $q = Qf/Z$, $Z = \mathbb{E}_Q f$, satisfies

$$D_{\mathrm{KL}}(F\|q) \leq -\log(1-b) + a \leq 2b + a. \tag{4}$$

Its costs are polynomial in N, L_z, a^{-1} and logarithmic inverse failure probability. No exact evaluation of Z is needed. The bounds on C and f hold on every outcome.

Theorem 2.3 (Posterior mass cover). *Let $F = \nu' * \mathcal{N}(0, \sigma^2 I_d)$, where ν' has at most k atoms, and suppose $D_{\mathrm{KL}}(F\|q) \leq e$. Fix $0 < \beta, \delta_{\mathrm{cov}} < 1$. There are universal constants $C_{\mathrm{cov}}, C_{\mathrm{rad}} \geq 1$ with the following property. Given fresh observations from $(1-b)F + bH$, $b \leq 1/2$, and independent samples from the q-posterior for the channel $Y = X + \sigma G$, the procedure in Section 8 returns*

$$N_{\mathrm{cov}} = \left\lceil \frac{4k}{\beta} \log \frac{2k}{\delta_{\mathrm{cov}}} \right\rceil \tag{5}$$

centers whose radius-$C_{\mathrm{rad}}\sigma\sqrt{\log(2k/\beta)}$ balls omit at most $3\beta + C_{\mathrm{cov}}e$ of ν'-mass, with probability at least $1 - \delta_{\mathrm{cov}}$. The number of posterior samples and the arithmetic cost are polynomial in $d, k, \beta^{-1}, \log(1/\delta_{\mathrm{cov}})$ and the posterior sampling cost. A sufficiently small fixed total-variation error in each independent posterior draw is allowed, by changing universal constants.

The radius and mass guarantees concern the retained latent measure. If its original mass is $1-b$, the total omitted original mass after this refinement is at most

$$b + (1-b)(3\beta + C_{\mathrm{cov}}e). \tag{6}$$

All points produced by the procedure are retained in the center list, whose size may exceed k.

2.1 An explicit schedule without a given radius bound

Fix the universal constants of Theorem 2.3. Define

$$A_* = 1 + 2C_{\mathrm{cov}}, \qquad H = \log(1/\delta), \qquad J = \left\lceil \log_4(32(d+H+2)) \right\rceil, \tag{7}$$

$$\beta = \frac{\varepsilon^2}{1000(J+1)A_*^{J+1}}, \qquad L_* = \log(2k/\beta), \qquad \sigma_j = 2^{J-j} \quad (0 \le j \le J). \tag{8}$$

All these quantities are fixed before data are processed. Draw

$$N_0 = \lceil (k/\beta)\log(16k/\delta) \rceil$$

fresh original observations and retain them all as initial centers. Their cover radius is bounded by

$$R_0 = \sqrt{d} + \sqrt{2\log(16N_0/\delta)}. \tag{9}$$

Every atom of weight at least β/k is observed except with probability $\delta/16$, by a union bound. The other atoms have total mass at most β. Gaussian norm concentration and a second union bound show that every one of the N_0 observations lies within R_0 of its generating mean except with probability $\delta/16$. Hence the initial cover omits at most β mass with probability at least $1 - \delta/8$.

The schedule ensures that this initial cover remains useful even when the required confidence is very high. Indeed,

$$\log(16N_0/\delta) \le 2\log 16 + 2L_* + 2H, \qquad R_0^2 \le 16(d + H + L_*).$$

Since $4^J \ge 32(d+H+2)$ and $L_* \ge 1$, this gives

$$R_0/\sigma_0 \le \sqrt{L_*}. \tag{10}$$

Algorithm. Set $r_0 = R_0$. At each stage $j = 0, \dots, J$, apply Theorem 2.1 at noise σ_j, radius r_j, and tolerance $\theta = \beta\sqrt{\beta/64}/64$. Fix the resulting feature circuits. Apply Theorem 2.2 with $a = \beta$, using independent artificial Q_j samples and independent fresh observations from P_{σ_j}. Store $Q_j, f_j, \kappa_j, A_{f,j}$. If $j < J$, use Theorem 2.3 with fresh P_{σ_j} observations and the capped posterior sampler below; retain its entire center list and set

$$r_{j+1} = C_{\mathrm{rad}}\sigma_j\sqrt{L_*}. \tag{11}$$

Assign failure probability $\delta/[4(J+1)]$ for each complete dictionary-and-fit call and each cover call. The initial cover and these calls have total failure probability less than δ.

2.2 Proof of the statistical guarantee

Let b_j be the original latent mass outside the stage-j cover. Each stage may cover a different set of atoms. On the events where the stated guarantees hold, $b_0 \le \beta$. At a stage with $b_j \le 1/2$, the normalized covered submixture F_j satisfies

$$D_{\mathrm{KL}}(F_j \| q_j) \le 2b_j + \beta.$$

By (6),

$$b_{j+1} \le b_j + (1 - b_j)\{3\beta + C_{\mathrm{cov}}(2b_j + \beta)\} \le A_* b_j + (3 + C_{\mathrm{cov}})\beta.$$

Since $(3 + C_{\mathrm{cov}})/(A_* - 1) \le 2$, induction gives

$$b_j \le 3A_*^j \beta \le \varepsilon^2/1000 \qquad (0 \le j \le J). \tag{12}$$

This verifies the hypothesis $b_j \le 1/2$ at every stage. The induction conditions only on previously completed training; all subsequent observation batches are fresh. Conditional guarantees and a sequential union bound therefore justify the adaptive sequence of stages.

At stage J, $\sigma_J = 1$. Pinsker's inequality yields

$$\mathrm{TV}(P, q_J) \le b_J + \sqrt{(2b_J + \beta)/2}.$$

Generate $X \sim Q_J$ and accept with probability $f_J(X)/A_{f,J}$. Cap proposals at

$$M_{\mathrm{out}} = \left\lceil \frac{A_{f,J}}{\kappa_J} \log \frac{4}{\varepsilon} \right\rceil, \tag{13}$$

returning 0 on timeout. Each proposal succeeds with probability at least $\kappa_J/A_{f,J}$ on every training outcome. The capped law differs from q_J by at most $\varepsilon/4$. Therefore

$$\mathrm{TV}(P, \widehat{P}) \le \varepsilon^2/1000 + \sqrt{3\varepsilon^2/2000} + \varepsilon/4 < \varepsilon.$$

The prescribed error bound applies to the law of a single output; joint accuracy for several outputs can be obtained by allocating the error across invocations.

2.3 Polynomial cost on every outcome

Put $P_* = 2 + d + k + \varepsilon^{-1} + \log(1/\delta)$. We have $J = O(\log P_*)$, $A_*^{J+1} \le P_*^{O(1)}$, $\beta^{-1} \le P_*^{O(1)}$, and $L_* = O(\log P_*)$. Initial and refined lists have polynomial length; each refined list has the size prescribed by (5), independently of the previous list. By (10) and (11),

$$r_j/\sigma_j \le \max\{1, 2C_{\mathrm{rad}}\}\sqrt{L_*}.$$

Consequently all dictionary calls have fixed-polynomial cost, as do fitting and posterior refinement. In particular, high confidence enters the local Hermite degree through $\log\log(1/\delta)$ or $\log(d + \log(1/\delta))$, preserving polynomial dependence on $\log(1/\delta)$.

All moment batches use a fixed number of unconditioned trials, each multiplied by its acceptance indicator. Empirical entries are clipped before finite Jacobi rotations. Every feature block has a deterministic radial cap and an unchanged constant coordinate. Optimization stays in a fixed PSD trace ball. Posterior and final rejection loops are capped; the robust filter performs at most as many updates as there are input points. The procedures below explicitly handle zero denominators and degenerate matrices. These facts give deterministic polynomial costs on every outcome, including those on which the statistical guarantees fail. Sections 3–8 establish the three component guarantees used above.

3 Independent views and finite full-jet gates

3.1 Conditional independence and repeated evaluation of gate histories

The gate and the moment query must share a latent component while using independent observation noise. For a fresh observation $X = \mu + G$ and an independent $Z \sim \mathcal{N}(0, I_d)$, put

$$A_0 = X + Z, \qquad B_0 = X - Z.$$

Conditional on μ, these are independent $\mathcal{N}(\mu, 2I_d)$ variables, because their joint law is Gaussian and their cross-covariance is zero. A stored history is a randomized measurable indicator $H(A_0, R)$ whose auxiliary randomness R is independent of B_0 conditional on μ. Thus

$$\mathbb{E}[\mathbf{1}_H \psi(B_0)] = \sum_i w_i a_H(\mu_i) \mathbb{E}_{\mathcal{N}(\mu_i, 2I_d)} \psi, \quad a_H(\mu) = \mathbb{P}(H \mid \mu). \tag{14}$$

Evaluate all gates in a stored history on one fresh shared observation A_0, retaining the stored normalizations and drawing fresh auxiliary randomness. Independent observations for different historical gates would generally define a different acceptance function.

For scale $\sigma \geq 1$, independently add noise of covariance $2(\sigma^2 - 1)I_d$ to each side. At center c and normalizing radius S, the two sides have latent mean $u = (\mu - c)/S$ and covariance vI_d, where $v = 2\sigma^2/S^2$. Any artificial replica coordinates in a query receive independent copies on the two sides. Conditional on the latent atom, the entire gate history remains independent of the query side. All learned coefficients are fixed before new original samples are drawn.

3.2 The analytic jet bound

The gate needs to suppress large conditional second moments while retaining nearby latent atoms. Summing the squared derivatives of a polynomial gives a positive quantity with controlled logarithmic variation, which will supply both bounds. Let $F : \mathbb{R}^n \to \mathbb{R}^r$ be a degree-at-most-q polynomial map, supplied by a division-free arithmetic circuit. Set

$$J_F(x) = \sum_{h=0}^{q} \frac{v^h}{h!} \left\| D^h F(x) \right\|_F^2, \qquad M(u) = \mathbb{E} \left\| F(u + \sqrt{v} G) \right\|^2.$$

When F is nonzero, $J_F > 0$ everywhere: a highest nonzero derivative is a nonzero constant. The evaluation rule accepts a zero polynomial automatically.

Lemma 3.1 (Jet moments)**.** *For a nonzero F,*

$$M(u) \leq \mathbb{E} J_F(u + \sqrt{v} G) \leq 2^q M(u), \qquad \mathbb{E} J_F(u + \sqrt{v} G)^{-2} \leq e^{12q} / M(u)^2. \tag{15}$$

Proof. Write $F = \sum_{\ell=0}^{q} F_\ell$ in orthogonal Hermite chaoses under $\mathcal{N}(u, vI)$. Differentiation of normalized Hermite polynomials gives

$$\frac{v^h}{h!} \mathbb{E} \left\| D^h F \right\|^2 = \sum_{\ell \geq h} \binom{\ell}{h} \mathbb{E} \left\| F_\ell \right\|^2.$$

Summing over h proves the first assertion. To prove the second, let $A_h = \sqrt{v^h/h!}\, D^h F$. For every unit e,

$$\left\| \partial_e A_h \right\|_F \leq \sqrt{(h+1)/v}\, \left\| A_{h+1} \right\|_F.$$

Cauchy–Schwarz therefore gives $|\partial_e J_F| \leq 2\sqrt{q/v}\, J_F$, hence $\|\nabla \log J_F\| \leq 2\sqrt{q/v}$. The Gaussian exponential concentration inequality [BLM13], applied to $L = \log J_F(u + \sqrt{v} G)$, gives $\mathbb{E} e^{t(L - \mathbb{E}L)} \leq e^{2qt^2}$. In particular $M \leq \mathbb{E} J_F \leq e^{\mathbb{E}L + 2q}$ and $\mathbb{E} e^{-2L} \leq e^{-2\mathbb{E}L + 8q} \leq e^{12q}/M^2$. Global Lipschitz continuity ensures that these expectations are finite. The constant-polynomial case is immediate. □

We implement the gate through random contractions of this derivative collection. For independent $g_\ell \sim \mathcal{N}(0, I_n)$ and $\gamma_\ell \sim \mathcal{N}(0, 1)$, define

$$W(x) = \sum_{h=0}^{q} \frac{v^{h/2}}{\sqrt{h!}} D^h F(x)[g_1, \ldots, g_h] \prod_{\ell=h+1}^{q} \gamma_\ell. \tag{16}$$

This is a contraction of an order-q tensor T_x against independent blocks (γ_ℓ, g_ℓ). The derivative orders occupy disjoint support blocks, so $\|T_x\|_F^2 = J_F(x)$ and $\mathbb{E}\,\|W(x)\|^2 = J_F(x)$.

The number of sketches is chosen to match the inverse second moment in Lemma 3.1. The argument below gives a fourth-root bound for the probability that one sketch is small. For a fixed gate input, requiring eight independent sketches to be small raises this bound to a square, which can then be averaged using (15).

Lemma 3.2 (A finite gate)**.** *Take eight independent sketches* (16) *and accept exactly when every squared norm is at most $T > 0$. For a Gaussian gate input $A \sim \mathcal{N}(u, vI_n)$, its acceptance probability $\alpha(u)$ satisfies, with $K_q = 3^{8q}e^{12q}$,*

$$\alpha(u) \le \min\{1, K_q T^2/M(u)^2\}, \qquad \alpha(u)M(u) \le \sqrt{K_q}T, \tag{17}$$

$$1 - \alpha(u) \le 8 \cdot 2^q M(u)/T. \tag{18}$$

When $M = 0$, interpret the gate as accepting with probability one.

Proof. For a nonzero matrix A and $0 < s < 1$, singular-value decomposition and convexity of $t^{-s/2}$ give

$$\mathbb{E}\,\|AG\|^{-s} \le c_s\,\|A\|_F^{-s}, \qquad c_s = \mathbb{E}|G_1|^{-s}.$$

Indeed, normalize the squared singular values λ_j to sum to one and use $(\sum_j \lambda_j G_j^2)^{-s/2} \le \sum_j \lambda_j |G_j|^{-s}$. To apply this inequality to the sketch, fix x and let $T^{(j)}$ be T_x after contracting its first j slots against the independent Gaussian blocks (γ_ℓ, g_ℓ). Thus $T^{(0)} = T_x$ and $T^{(q)} = W(x)$. Conditional on the first $j-1$ blocks, flatten $T^{(j-1)}$ into a matrix whose columns correspond to the next slot and whose rows contain all remaining indices, including the output index. Matrix Frobenius norm is unchanged by this flattening, so

$$\mathbb{E}\big[\big\|T^{(j)}\big\|_F^{-s} \mid T^{(j-1)}\big] \le c_s \big\|T^{(j-1)}\big\|_F^{-s}.$$

A nonzero matrix maps a Gaussian vector to a nonzero vector almost surely; induction therefore justifies these conditional bounds for every nonzero T_x. Iterating and using $\|T_x\|_F^2 = J_F(x)$ gives

$$\mathbb{E}\,\|W\|^{-s} \le c_s^q\,\|T_x\|_F^{-s} = c_s^q J_F(x)^{-s/2}.$$

For $q = 0$ the contraction is deterministic and the same formula holds. Since $c_{1/2} < 3$,

$$\mathbb{P}(\|W\|^2 \le T) \le 3^q (T/J_F)^{1/4}.$$

Independence of eight sketches gives $g(x) \le \min\{1, 3^{8q}T^2/J_F(x)^2\}$. Markov's inequality and a union bound give $1 - g(x) \le 8J_F(x)/T$. Average these two inequalities over A and use Lemma 3.1. Finally $\min(M, K_q T^2/M) \le \sqrt{K_q}T$. □

Finite implementation. Evaluate the circuit for F in $\mathbb{R}[t_1, \dots, t_q]/(t_1^2, \dots, t_q^2)$ at $x + \sum_\ell t_\ell g_\ell$. The coefficient of $t_1 \cdots t_h$ is exactly $D^h F(x)[g_1, \dots, g_h]$, with no factorial. Multiplication enumerates disjoint subset pairs and costs $O(3^q)$ scalar operations. A size-S_F circuit therefore implements each sketch in $O((S_F + n + r + q)\,\mathrm{poly}(q)3^q)$ operations. The stored representation consists of the circuit and its subset coefficients, leaving the n^q tensor implicit. Zero polynomials produce zero sketches automatically.

Corollary 3.3 (The actual accepted moment)**.** *Append this gate to an arbitrary fixed history that can be evaluated on fresh samples, and let $a_1(u)$ be the resulting acceptance probability. On the independent query side,*

$$\mathbb{E}[\mathbf{1}_{H_1} \|F(B)\|^2] = \sum_i w_i a_1(u_i) M(u_i) \le \sqrt{K_q} T. \tag{19}$$

Proof. On the gate side, the history and the new gate may be dependent, but $a_1(u) \le \alpha(u)$. Apply conditional independence and (17). This bounds the unnormalized accepted moment directly. □

4 Recursive compression of accepted moments

Fix a center and normalization with $vD \le C_0$, where C_0 is a fixed universal constant. Let $0 < \tau < 1$. The construction below produces co-isometries Φ_h with at most k rows and nested acceptance functions a_h for $1 \le h \le D$. Put $\Phi_0 = 1$,

$$s_h(u) = \left\|\Phi_h u^{\otimes h}\right\|^2, \quad s_0 = 1, \quad e_h(u) = \|u\|^2 s_{h-1}(u) - s_h(u) \ge 0.$$

Proposition 4.1 (Compression with a local retention guarantee)**.** *For any $\xi > 0$, the finite construction satisfies, simultaneously with the prescribed confidence,*

$$a_D(u) \ge a_0(u) - \tau \qquad \text{for every latent atom with } \|u\| \le 1, \tag{20}$$

$$\sum_i w_i a_D(u_i) e_h(u_i) \le 2k\xi \qquad (1 \le h \le D). \tag{21}$$

Its costs are polynomial in $d, k, \tau^{-1}, \xi^{-1}, \exp(O(D))$, logarithmic inverse failure probability, and the evaluation cost of the prior history.

Proof. At each degree, the gate must retain nearby atoms while making the next matrix estimate reliable. These are separate requirements: retention is checked atom by atom, whereas matrix estimation uses the full accepted measure, including any remote atoms that pass the gate.

After Φ_{h-1} is fixed, form $R_h = \Phi_{h-1} \otimes I_d$, a co-isometry with $n_h \le kd$ rows T_a. Set $y_h(u) = R_h u^{\otimes h}$. With $p_T(x) = \langle T, x^{\otimes h} \rangle$, use the real matrix-valued polynomial

$$F_{h,ac}(b, z') = \Re\{p_{T_a}(b + iz') p_{T_c}(b + iz')\}.$$

Its degree is $2h$. For independent $b \sim \mathcal{N}(u, vI_d)$, $z' \sim \mathcal{N}(0, vI_d)$, cancellation of holomorphic circular-Gaussian moments gives

$$\mathbb{E}F_h(b, z') = y_h(u) y_h(u)^T. \tag{22}$$

The product polynomials are evaluated by recursive contraction circuits without constructing the full tensors.

For any order-q tensor T of Frobenius norm at most one, expansion about u, the circular Wick identity, and Minkowski give

$$\|p_T(u + \sqrt{v}(G + iZ))\|_{L^2} \le \sum_{r=0}^{q} \binom{q}{r} \|u\|^{q-r} (2v)^{r/2} \sqrt{r!} \le (\|u\| + \sqrt{2vq})^q. \tag{23}$$

Symmetrizing a tensor only decreases its norm; contractions with u are bounded by its norm times the relevant power of $\|u\|$. These observations justify the estimate for arbitrary, possibly nonsymmetric, rows. Applied to each product row with $q = 2h$, it yields, on $\|u\| \le 1$,

$$M_h(u) := \mathbb{E}\,\|F_h\|_F^2 \le (kd)^2(1 + 2\sqrt{C_0})^{4D}, \qquad \mathbb{E}J_{F_h} \le (kd)^2 A_0^D,$$

where $A_0 = 4(1 + 2\sqrt{C_0})^4$. Append the gate of Lemma 3.2 with the common threshold

$$T = 8D(kd)^2 A_0^D/\tau. \tag{24}$$

At that stage, the acceptance probability of each atom within the prescribed radius decreases by at most τ/D. A union bound on the original gate-side marginals gives (20), regardless of dependencies among historical gates.

The gate is now fixed. Use fresh unconditioned trials to estimate $\mathbf{1}_{H_h} F_h$ on the query side. Its population mean is the PSD rank-at-most-k matrix

$$M_h^{\mathrm{pop}} = \sum_i w_i a_h(u_i) y_h(u_i) y_h(u_i)^T. \tag{25}$$

Every entry has second moment at most $V_h = \sqrt{K_{2h}}T$ by Corollary 3.3. Clip each entry to $[-B_h, B_h]$, where

$$B_h = \max\{1, 4n_h V_h/\xi\}.$$

The clipping bias is at most $V_h/B_h \le \xi/(4n_h)$. A fixed batch of $O(n_h^2 B_h^2 \xi^{-2} \log(2Dn_h^2/\delta'))$ trials makes the total error in every entry at most $\xi/(2n_h)$, by Hoeffding's inequality and a union bound. The symmetric empirical matrix is therefore within $\xi/2$ of (25) in operator norm.

Apply maximal-pivot Jacobi rotations to this empirical matrix. Each rotation decreases the squared off-diagonal Frobenius norm by a factor at most $1 - 2/[n_h(n_h - 1)]$. Since entries were clipped, a prescribed cap of $O(n_h^2 \log(2 + n_h B_h/\xi))$ rotations suffices to make the off-diagonal norm at most $\xi/2$. The one-dimensional or already diagonal cases need no rotations. The accumulated basis is exactly orthogonal. Deleting the residual off-diagonal entries gives a matrix within ξ of M_h^{pop}. Select its largest $\min(k, n_h)$ diagonalized eigenvalues, with eigenvector rows U_h, and put

$$\Phi_h = U_h(\Phi_{h-1} \otimes I_d).$$

This is an exact co-isometry on every outcome. The min–max principle and the rank bound imply that the compression of M_h^{pop} to the orthogonal complement of U_h has operator norm at most 2ξ. It is PSD of rank at most k, hence trace at most $2k\xi$. That trace is $\sum_i w_i a_h(u_i) e_h(u_i)$. Since $a_D \le a_h$ pointwise, (21) follows. Use independent batches and distribute the failure probability among the D stages. All displayed magnitudes are $\exp(O(D))$ times polynomial factors, establishing the cost. □

Control of remote accepted mass. Compression also identifies how much mass can remain far from the chosen center. Write $L = \|u\|^2 > 1$. The recursion gives

$$s_D(u)/L^D = 1 - \sum_{h=1}^{D} e_h(u)/L^h.$$

For $L \ge L_0 > 1$, the accepted mass on which the sum exceeds $1/2$ is at most $4k\xi/(L_0 - 1)$. On the complement, $s_D \ge L^D/2$. Jensen's inequality applied to (22) gives $M_D(u) \ge \left\| y_D(u) y_D(u)^T \right\|_F^2 \ge s_D(u)^2$. Therefore the accepted mass outside $\|u\|^2 < L_0$ is at most

$$\frac{4k\xi}{L_0 - 1} + \frac{16K_{2D}T^2}{L_0^{4D}}. \tag{26}$$

For fixed C_0, choosing a sufficiently large universal L_0, then $D = O(\log((kd+1)/(\tau\beta)))$ and $\xi = O(\beta/k)$, makes this at most β at polynomial cost. The condition $vD = O(1)$ keeps the moment bounds exponential in the degree, permitting logarithmic degree at polynomial cost. With $v = 1$, the corresponding bounds can grow factorially in the degree.

5 A finite square-root density dictionary

We prove Theorem 2.1. The construction has three sources of approximation error: projecting and truncating the Hermite expansion, replacing its features by finite evaluations, and clipping those features to a deterministic bound. A single positive semidefinite comparator controls all three errors; component assignments enter only its analysis. Define

$$t^2 = \max\{(r/\sigma)^2, \log(16m/\theta), 1\}, \quad S = \sigma t, \quad D = \lceil 64t^2 \rceil. \tag{27}$$

For each center, start a fresh history and apply Proposition 4.1 with $\tau = 1/4$. The view variance is $v = 2/t^2$, so $vD \le 130$. The proposition's constant-C_0 formulation covers this choice as well as the case $vD \le 1$. All degree-dependent costs are $\exp(O(D))$.

5.1 Projection error on the original covered atoms

Assign each covered atom to one center within radius r. These disjoint sets I_j are used only in the analysis. Let $s = \sum_{i \text{ covered}} w_i = 1 - b \ge 1/2$ and $u_i = (\mu_i - c_j)/S$ for $i \in I_j$. Put $\Pi_{j,h} = \Phi_{j,h}^T \Phi_{j,h}$. For

$$r_h(u) = \left\| (I - \Pi_{j,h}) u^{\otimes h} \right\|^2,$$

the orthogonal projection identity gives $r_h = \|u\|^2 r_{h-1} + e_h$. Hence $r_h \le \sum_{\ell \le h} e_\ell$ for $\|u\| \le 1$. The acceptance probability for each atom within the prescribed radius is at least $1 - \tau$, so (21) implies

$$\sum_{i \in I_j} w_i \left\| (I - \Pi_{j,h}) u_i^{\otimes h} \right\|^2 \le \frac{2kh\xi}{1-\tau}. \tag{28}$$

The estimate controls the weighted sum directly. Summing over centers introduces only a factor m.

5.2 Hermite approximation and a trace-one comparator

Let $g_j = \mathcal{N}(c_j, \sigma^2 I_d)$, $z = (x - c_j)/\sigma$, and $a_i = (\mu_i - c_j)/\sigma = t u_i$. The square root of a component's density ratio is

$$f_i(z) = \exp(a_i \cdot z/2 - \|a_i\|^2/4) = e^{-\|a_i\|^2/8} \sum_{h=0}^{\infty} \frac{(t/2)^h}{\sqrt{h!}} \langle u_i^{\otimes h}, H_h(z) \rangle.$$

Here the normalized Hermite tensor satisfies $\|\langle T, H_h \rangle\|_2 \le \|T\|_F$ and $\|\langle T, H_h \rangle\|_4 \le 3^{h/2} \|T\|_F$. Project the terms through degree D and denote the resulting polynomial by p_i. Its feature coefficients are

$$b_{i,h} = e^{-\|a_i\|^2/8} \frac{(t/2)^h}{\sqrt{h!}} \Phi_{j,h} u_i^{\otimes h}, \qquad \|b_i\|^2 \le e^{-\|a_i\|^2/4} \sum_{h=0}^{D} \frac{(\|a_i\|^2/4)^h}{h!} \le 1.$$

The coefficient bound follows directly from orthogonality in the ambient tensor space. It will give the positive semidefinite comparator a trace bound independent of the component locations within the cover.

Set $B = \sqrt{D+1}\, e^{3t^2/8}$ and $C_* = e^{t^2/4} + B$. The factorial sum obeys

$$\sum_{h=0}^{D} \frac{(\sqrt{3}t/2)^h}{\sqrt{h!}} \le \sqrt{D+1} \left(\sum_{h=0}^{D} \frac{(3t^2/4)^h}{h!} \right)^{1/2} \le B.$$

Beyond $D \ge 3et^2$, consecutive terms have ratio at most $1/2$, and $h! \ge (h/e)^h$ bounds the tail by 2^{-D}. Weighted Minkowski, (28), and hypercontractivity give

$$E_{\mathrm{lat}} := \left(\sum_{i \text{ covered}} w_i \left\| f_i - p_i \right\|_4^2 \right)^{1/2} \le 2^{-D} + B\sqrt{\frac{2kmD\xi}{1-\tau}}. \tag{29}$$

Also $\|f_i\|_4 = e^{\|a_i\|^2/4} \le e^{t^2/4}$ and $\|p_i\|_4 \le B$, hence

$$\left\| f_i^2 - p_i^2 \right\|_2 \le C_* \left\| f_i - p_i \right\|_4 .$$

The fourth-moment approximation to each square root thus gives the second-moment control needed for the density dictionary.

Define the exact features

$$z_{j,h,\ell}(x) = \sqrt{g_j(x)/Q(x)}\, \langle (\Phi_{j,h})_\ell, H_h((x-c_j)/\sigma) \rangle,$$

and the block-diagonal PSD matrix with blocks $S_j = \sum_{i \in I_j} (w_i/s) b_i b_i^T$. It has trace at most one. Since $g_j/Q \le m$, a triangle inequality and weighted Cauchy–Schwarz give

$$\left\| F/Q - z^T S z \right\|_{L^2(Q)} \le \sqrt{m/s}\, C_* E_{\mathrm{lat}} \le \sqrt{2m}\, C_* E_{\mathrm{lat}}. \tag{30}$$

The feature count is at most $N_{\mathrm{raw}} = m(1+kD)$.

5.3 Finite Hermite evaluation

For an order-q row tensor T of norm at most one, define

$$H_T(x) = \frac{\mathbb{E}_Z \Re p_T(x + iZ)}{\sqrt{q!}}.$$

The circular Wick identity gives $\mathbb{E}_{X,Z} |p_T(X+iZ)|^2 = 2^q q! \left\| \mathrm{sym}\, T \right\|_F^2 \le 2^q q!$. With $n = \lceil 8 \cdot 2^q / \xi_H^2 \rceil$, the block polynomial

$$V(x) = \frac{1}{n} \sum_{a=1}^{n} \frac{\Re p_T(x + iZ_a)}{\sqrt{q!}}$$

satisfies $\mathbb{P}(\|V - H_T\|_2 > \xi_H) \le 1/8$. Draw an odd $b_H = O(\log(N_{\mathrm{raw}}/\delta'))$ independent blocks and define $\widehat{H}_T(x)$ as their pointwise median. A Chernoff bound and a union bound imply that, simultaneously for every fixed feature, at least three quarters of its blocks have L^2 error at most ξ_H.

On this event, fix x. At least half of all block values lie on the side of the median away from $H_T(x)$; their errors are at least the median's error. Since at most one quarter of the blocks are bad, at least $b_H/4$ good blocks have this property. Applying the same counting argument with zero in place of $H_T(x)$ also bounds the median's magnitude. Thus for every $p \ge 1$,

$$|\widehat{H}_T - H_T|^p \le \frac{4}{b_H} \sum_{\ell \text{ good}} |V_\ell - H_T|^p, \qquad |\widehat{H}_T|^p \le \frac{4}{b_H} \sum_{\ell \text{ good}} |V_\ell|^p.$$

Applying hypercontractivity to each polynomial block and then using the preceding pointwise inequalities gives

$$\left\|\widehat{H}_T - H_T\right\|_4 \le 4^{1/4} 3^{q/2} \xi_H, \qquad \mathbb{E}|\widehat{H}_T|^8 \le 4 \cdot 7^{4q}(1+\xi_H)^8. \tag{31}$$

The median is a finite arithmetic-and-comparison circuit. Each evaluation of p_T uses the stored recursive circuits to contract the implicit tensor, whose ambient representation has d^q coordinates.

Choose $\xi_H \le 1$ so the first error in (31) is at most ρ. For $n_{\mathrm{loc}} = 1 + kD$, $\|p_i - \widehat{p}_i\|_4 \le \sqrt{n_{\mathrm{loc}}}\rho$. The same comparator matrix consequently satisfies

$$\left\|F/Q - z_{\mathrm{raw}}^T S z_{\mathrm{raw}}\right\|_2 \le \sqrt{2m}(C_* + \sqrt{n_{\mathrm{loc}}}\rho)(E_{\mathrm{lat}} + \sqrt{n_{\mathrm{loc}}}\rho). \tag{32}$$

For example, it suffices to take

$$\xi \le \frac{(1-\tau)\theta^2}{1024 k m^2 D B^2 (C_*+1)^2}, \qquad \rho \le \frac{\theta}{16\sqrt{2m}(C_*+1)\sqrt{n_{\mathrm{loc}}}}.$$

The choice (27) also gives $2^{-D} \le \theta/[16\sqrt{2m}(C_*+1)]$. To check this explicitly, use $C_*+1 \le 3\sqrt{D+1}e^{3t^2/8}$, $D \le 65t^2$, and $e^{-t^2} \le \theta/(16m)$; the remaining exponential decay dominates the polynomial factor by a universal margin. Equation (32) is then at most $3\theta/16$.

5.4 Radial clipping and deterministic bounds

Since $g_j/Q \le m$, (31) yields the known eighth-moment bound

$$\mathbb{E}_Q \|z_{\mathrm{raw}}\|^8 \le H_8 := 4 N_{\mathrm{raw}}^4 m^3 7^{4D} (1+\xi_H)^8.$$

Radially clip the raw feature block at radius $L_{\mathrm{clip}} \ge 1$, where $L_{\mathrm{clip}}^2 \ge 4\sqrt{H_8}/\theta$. For a trace-one PSD S,

$$\mathbb{E}_Q |z_{\mathrm{raw}}^T S z_{\mathrm{raw}} - z_{\mathrm{clip}}^T S z_{\mathrm{clip}}|^2 \le \mathbb{E}_Q[\|z_{\mathrm{raw}}\|^4 \mathbf{1}_{\|z_{\mathrm{raw}}\| > L_{\mathrm{clip}}}] \le H_8 / L_{\mathrm{clip}}^4.$$

Append the constant *after* clipping: $z = (1, z_{\mathrm{clip}})$. Pad S with a zero row and column. Its trace is unchanged, $z_0 = 1$ exactly, and $\|z\| \le \sqrt{1 + L_{\mathrm{clip}}^2}$ on every outcome. The total L^2 error is less than θ.

All inverse tolerances, moment bounds, and circuit sizes are polynomial in $d, k, m, \theta^{-1}, \exp(O(t^2))$. Gaussian replicas are fixed after compression and before fitting. This proves Theorem 2.1 for arbitrary positive weights, using the unnormalized accepted moments throughout.

6 Directional likelihood fitting with a finite optimizer

This section proves Theorem 2.2. The optimizer will control a likelihood ratio in a direction determined by the dictionary comparator. Nonnegativity allows us to discard the contaminating contribution when bounding the contribution of the retained submixture. A penalty controls the fitted numerator's normalization error, while optimization over a compact PSD trace ball preserves the matrix constraint.

Let $g = dF/dQ$, $s = z^T S z$, $\operatorname{tr} S \le 1$, and $\|s - g\|_2 \le \theta$. Write $L = L_z$ and define

$$\mathcal{K} = \{C \succeq 0 : \operatorname{tr} C \le 4\}, \quad f_C = \kappa + z^T C z, \quad A_f = \kappa + 4L^2, \quad G = \mathbb{E}_Q z z^T.$$

An independent batch of artificial samples estimates G by the empirical Gram matrix $\widehat{G}$. Thus $\widehat{G} \succeq 0$, $\widehat{G}_{00} = 1$, and $\left\|\widehat{G}\right\|_{\mathrm{op}} \le L^2$ on every outcome.

6.1 Finite concentration events

Entrywise application of Hoeffding's inequality and $\|M\|_{\mathrm{op}} \le N \max_{ij} |M_{ij}|$ imply

$$\left\|\widehat{G} - G\right\|_{\mathrm{op}} \le \gamma \tag{33}$$

with failure probability at most δ_G using $O(N^2 L^4 \gamma^{-2} \log(2N^2/\delta_G))$ artificial samples.

For a fresh score batch P_n from $P' = (1-b)F + bH$, we also require

$$\sup_{C,D \in \mathcal{K}} \left| P_n \frac{f_D}{f_C} - P' \frac{f_D}{f_C} \right| \le \epsilon_R. \tag{34}$$

The ratios lie in $[\kappa/A_f, A_f/\kappa]$. Their Lipschitz constants with respect to D and C, measured in Frobenius norm, are respectively L^2/κ and $A_f L^2/\kappa^2$. A pair of nets of radius $\epsilon_R \kappa^2/(16 A_f L^2)$ has log cardinality at most

$$2p \log\left(1 + \frac{256 A_f L^2}{\epsilon_R \kappa^2}\right), \qquad p = N(N+1)/2.$$

Hoeffding's inequality and the net approximation prove (34) if, for example,

$$n \ge \frac{8(A_f/\kappa)^2}{\epsilon_R^2} \left[2p \log\left(1 + \frac{256 A_f L^2}{\epsilon_R \kappa^2}\right) + \log \frac{2}{\delta_R} \right]. \tag{35}$$

The nets are used to prove uniform concentration. Uniformity over the entire fixed trace ball permits C, D to depend on both empirical samples. The fixed median and clipping circuits satisfy the same boundedness and parameter-Lipschitz estimates.

6.2 Penalized fit and directional stationarity

Choose

$$\lambda \ge \frac{2[\log(A_f/\kappa) + 1]}{r_0^2}, \qquad \mathcal{F}(C) = -P_n \log f_C + \frac{\lambda}{2} r(C)^2, \qquad r(C) = \kappa + \mathrm{tr}(\widehat{G}C) - 1.$$

Choosing equality in this bound gives an explicit parameter value. The Hessian norm of $\mathcal{F}$ on $\mathcal{K}$ is bounded by

$$H_F = L^4(\kappa^{-2} + \lambda). \tag{36}$$

The domain has Frobenius diameter at most 8. Produce $\widehat{C} \in \mathcal{K}$ with

$$\mathcal{F}(\widehat{C}) \le \min_{C \in \mathcal{K}} \mathcal{F}(C) + e_{\mathrm{obj}}, \qquad e_{\mathrm{obj}} \le \min\{1, u^2/(128 H_F)\}. \tag{37}$$

For every $D \in \mathcal{K}$ this implies

$$\langle \nabla \mathcal{F}(\widehat{C}), D - \widehat{C} \rangle \ge -u. \tag{38}$$

Indeed, if the directional derivative were less than $-u$, a step of length $u/(64 H_F) \le 1$ would improve the objective by more than $u^2/(128 H_F)$, using (36). The hypotheses ensure $u \le 1$ and $H_F \ge 1$.

Let $a_0 = \mathbb{E}_Q s$, $\widehat{a} = \mathrm{tr}(S\widehat{G})$. On (33),

$$|a_0 - 1| \le \theta, \qquad |\widehat{a} - a_0| \le \gamma.$$

If $\theta + \gamma \le 1/4$, then $3/4 \le \widehat{a} \le 5/4$ and

$$D = \frac{(1-\kappa)S}{\widehat{a}} \quad \text{satisfies} \quad D \in \mathcal{K}, \quad \operatorname{tr} D \le 2, \quad r(D) = 0. \tag{39}$$

Applying (38) in this direction gives the exact identity and inequality

$$1 - P_n(f_D/f_{\widehat{C}}) - \lambda r(\widehat{C})^2 \ge -u, \qquad P_n(f_D/f_{\widehat{C}}) \le 1 + u.$$

Exact empirical normalization of the comparator makes the penalty term nonpositive in the preceding inequality. This converts approximate stationarity into an upper bound on the empirical likelihood ratio. The algorithm needs only the fitted matrix. Combining with (34),

$$P'(f_D/f_{\widehat{C}}) \le 1 + u + \epsilon_R. \tag{40}$$

Objective comparison with D also yields

$$\frac{\lambda}{2} r(\widehat{C})^2 \le P_n \log f_{\widehat{C}} - P_n \log f_D + e_{\mathrm{obj}} \le \log(A_f/\kappa) + 1.$$

Therefore $|r(\widehat{C})| \le r_0$. Writing $Z_C = \mathbb{E}_Q f_C$, we obtain

$$|Z_{\widehat{C}} - 1| \le r_0 + 4\gamma =: \zeta, \qquad |Z_D - 1| \le 2\gamma \le \zeta. \tag{41}$$

6.3 The directional KL estimate

Let $q = Qf_{\widehat{C}}/Z_{\widehat{C}}$, $h = q/Q$, and $q_D = Qf_D/Z_D$. Equation (39) gives

$$q_D/Q = \frac{s+\rho}{a_0+\rho}, \qquad \rho = \frac{\kappa\widehat{a}}{1-\kappa} \le \frac{5}{2}\kappa.$$

With $R_Z = (1+\zeta)/(1-\zeta)$, (40) implies

$$P'(q_D/q) \le R_Z(1 + u + \epsilon_R).$$

The comparison with the retained submixture uses two nonnegative quantities: the floor in the comparator and the contribution of the omitted submixture. More explicitly,

$$\mathbb{E}_Q \frac{gs}{h} \le (a_0+\rho)\mathbb{E}_F \frac{q_D}{q} \le \frac{a_0+\rho}{1-b}\,\mathbb{E}_{P'} \frac{q_D}{q}.$$

The first inequality drops the nonnegative term containing ρ, and the second uses $P' \ge (1-b)F$. Substituting $a_0 \le 1+\theta$ and the bound on ρ gives

$$\mathbb{E}_Q \frac{gs}{h} \le M := \frac{(1+\theta+5\kappa/2)R_Z(1+u+\epsilon_R)}{1-b}. \tag{42}$$

In particular $M \ge 1$. Put

$$X = \mathbb{E}_Q(g^2/h), \qquad e_0 = \theta\sqrt{(1+\zeta)/\kappa}.$$

The preceding bound involves the approximate comparator. The next estimate replaces it by the true density ratio; the positive floor controls the denominator during this replacement. The lower bound $h \ge \kappa/(1+\zeta)$ and $g \in L^2(Q)$ ensure $X < \infty$. Weighted Cauchy–Schwarz gives

$$\left|\mathbb{E}_Q \frac{g(g-s)}{h}\right| \le \sqrt{X}\sqrt{\mathbb{E}_Q \frac{(g-s)^2}{h}} \le e_0\sqrt{X}.$$

Consequently $X \le M + e_0\sqrt{X}$, whence $\sqrt{X} \le \sqrt{M} + e_0$. Since $M \ge 1$, this implies

$$\log X \le \log M + 2\log(1 + e_0/\sqrt{M}) \le \log M + 2e_0.$$

Thus the comparator error contributes additively after taking logarithms. Jensen's inequality gives

$$\begin{aligned} D_{\mathrm{KL}}(F\|q) = \mathbb{E}_F \log(g/h) &\le \log X \\ &\le -\log(1-b) + \theta + \frac{5}{2}\kappa + \log R_Z + u + \epsilon_R + 2\theta\sqrt{(1+\zeta)/\kappa}. \end{aligned} \tag{43}$$

The positive floor and weighted Cauchy–Schwarz control the L^2 approximation error uniformly over $\|g\|_2$.

Under (3), $\zeta = a/128$. For $0 \le \zeta \le 1/128$, $\log((1+\zeta)/(1-\zeta)) \le 3\zeta$. Substituting all terms in (43) bounds their sum after $-\log(1-b)$ by less than a. This proves (4) and gives the linear omitted-mass cost required by the scale recurrence, in place of the score-range cost $b\log(A_f/\kappa)$.

6.4 A deterministic polynomial optimization cap

The following implementation establishes (37). Start with $C_0 = 0$. At iteration t, compute the exact gradient and choose $V_t \in \mathcal{K}$ approximately minimizing $\langle \nabla\mathcal{F}(C_t), V\rangle$. A true minimizer is zero or $4vv^T$ for a minimum-eigenvalue unit vector. Finite Jacobi rotations give a unit vector with Rayleigh quotient within η_{eig} of the minimum: make the off-diagonal Frobenius residual at most $\eta_{\mathrm{eig}}/2$, take a smallest diagonal entry, and use Weyl's inequality. Select the better of its rank-one vertex and zero. This gives a linear oracle with additive error $4\eta_{\mathrm{eig}}$. All gradient entries have a known polynomial bound from $f \ge \kappa$, $\|z\| \le L$, and $|r(C)| \le 1 + A_f$, so the Jacobi rotation count is prescribed in advance.

Use $C_{t+1} = (1-\eta_t)C_t + \eta_t V_t$, $\eta_t = 2/(t+2)$. Every iterate is exactly PSD with trace at most four, including on outcomes where the statistical guarantees fail. Smoothness gives, for $\Delta_t = \mathcal{F}(C_t) - \min\mathcal{F}$,

$$\Delta_{t+1} \le (1-\eta_t)\Delta_t + 4\eta_t\eta_{\mathrm{eig}} + 32H_F\eta_t^2.$$

The first step has $\eta_0 = 1$; induction then yields $\Delta_t \le 128H_F/(t+2) + 4\eta_{\mathrm{eig}}$. Taking $\eta_{\mathrm{eig}} = e_{\mathrm{obj}}/32$ and $t \ge \lceil 1024H_F/e_{\mathrm{obj}}\rceil$ proves (37), with a strict accuracy margin. This standard conditional-gradient recurrence [Jaggi13] applies to repeated eigenvalues and singular feature-Gram matrices. Its cost is a fixed polynomial in the quantities stated in Theorem 2.2.

7 Global and posterior rejection with deterministic caps

Suppose $q = Qf/Z$, where $\kappa \le f \le A_f$ and Q is a known Gaussian reference mixture. A proposal $X \sim Q$ accepted with probability $f(X)/A_f$ has law q. Its success probability is $Z/A_f \ge \kappa/A_f$ on every outcome. After M proposals, timeout has probability at most $\exp(-M\kappa/A_f)$. Conditional on acceptance, the output has law q. Returning a fixed fallback on timeout keeps the overall law within the stated total variation distance of q.

For a channel $Y = X + \mathcal{N}(0, tI_d)$ and reference

$$Q = \sum_j \pi_j \mathcal{N}(c_j, s^2 I_d),$$

the reference posterior at any y is an explicit mixture. Its component weights are proportional to $\pi_j\mathcal{N}(y; c_j, (s^2+t)I_d)$; component means and covariance are

$$\frac{tc_j + s^2 y}{s^2 + t}, \qquad \frac{s^2 t}{s^2 + t} I_d.$$

Accepting a reference-posterior draw with probability $f(X)/A_f$ gives the exact q-posterior. The unknown normalizing constant Z cancels. The per-trial success probability is at least κ/A_f uniformly for every query y.

In the cover step, $s^2 = t = \sigma^2$. The weights are proportional to $\exp(-\|y - c_j\|^2/(4\sigma^2))$, the means are $(y+c_j)/2$, and the covariance is $\sigma^2 I_d/2$. For a fixed sufficiently small $\rho > 0$, independently cap each requested posterior draw at

$$M_{\text{post}} = \lceil (A_f/\kappa) \log(1/\rho) \rceil . \tag{44}$$

Return zero on timeout. For every fixed y, these draws are independent and have a common law $\widetilde{Q}_y$ satisfying $\mathrm{TV}(\widetilde{Q}_y, Q_y) \le \rho$. This fixed error is included in the robust estimator's contamination budget for each posterior draw. Gaussian density weights are positive in the exact-real model, regardless of the magnitude of y or the centers. The uniform acceptance bound gives the claimed exact-real operation count. The finite-bit construction in Appendix A uses its own numerical error analysis and sampling procedure.

8 A posterior cover with dimension-free mass loss

We prove Theorem 2.3. Gaussian geometry will control the clean posterior covariance, while the KL chain rule will identify queries for which the fitted posterior is close enough to use the robust filter. The final counting argument converts these query guarantees into covered latent mass. Let $U \sim \nu' = \sum_j w_j \delta_{\mu_j}$ and consider the clean joint law

$$X = U + \sigma G_1, \qquad Y = X + \sigma G_2,$$

with independent standard Gaussians. Let P_y be the clean conditional law of X given $Y = y$, with mean m_y. Gaussian conditioning gives

$$X \mid (Y = y, U = \mu_i) \sim \mathcal{N}((y + \mu_i)/2, \sigma^2 I_d/2), \quad 2m_y - y = \mathbb{E}[U \mid Y = y], \tag{45}$$

and $\mathrm{Cov}(P_y) = \sigma^2 I_d/2 + \mathrm{Cov}(U \mid Y = y)/4$.

8.1 Posterior localization including remote atoms

Put $L = \log(2k/\beta)$ and call an atom heavy when $w_j \ge \beta/(2k)$. The remaining atoms have total mass at most $\beta/2$. Conditional on a heavy generating atom j, write $Y = \mu_j + \sqrt{2}\sigma G$. For each distinct μ_i set $v_{ij} = (\mu_i - \mu_j)/\|\mu_i - \mu_j\|$. With conditional probability at least $1 - \beta/2$, all $\langle G, v_{ij}\rangle \le \sqrt{2L}$. Let E_{geom} be the event that the generating atom is heavy and these geometric inequalities hold; its failure probability under the clean law is at most β.

On this event Bayes' rule, using only the generating component in the posterior denominator, gives for $D_{ij} = \|\mu_i - \mu_j\|$

$$\pi_i(y) := \mathbb{P}(U = \mu_i \mid Y = y) \le \frac{w_i}{w_j} \exp\left(-\frac{D_{ij}^2}{4\sigma^2} + \frac{D_{ij}\langle G, v_{ij}\rangle}{\sqrt{2}\sigma} \right) \le \exp\left(3L - \frac{D_{ij}^2}{8\sigma^2} \right).$$

The last bound follows by completing a square; coincident means contribute zero squared distance. Therefore, for every $v \ge 0$,

$$\mathbb{P}(\|U - \mu_j\|^2 > v \mid Y = y) \le \min\{1, k \exp(3L - v/(8\sigma^2))\}.$$

Integrating the tail at the splitting point $8\sigma^2(3L + \log k)$ yields

$$\mathbb{E}[\|U - \mu_j\|^2 \mid Y = y] \le 8\sigma^2(3L + \log k + 1) \le 40\sigma^2 L. \tag{46}$$

Thus, on E_{geom},

$$\|2m_y - y - \mu_j\| \le \sqrt{40}\sigma\sqrt{L}, \qquad \mathrm{Cov}(P_y) \preceq 16\sigma^2 L I_d. \tag{47}$$

The integrated tail controls the second-moment contribution of distant means, yielding the operator covariance bound required by the robust filter.

8.2 Conditional KL and finite-covariance regularity

Let Q_y be the q-posterior for the same channel. The joint KL chain rule gives

$$D_{\mathrm{KL}}(F\|q) = D_{\mathrm{KL}}(F_Y\|q_Y) + \mathbb{E}_{Y\sim F_Y} D_{\mathrm{KL}}(P_Y\|Q_Y). \tag{48}$$

Choose a sufficiently small universal $\epsilon_0 > 0$ and set $\kappa_0 = 2\epsilon_0^2$. By Markov's and Pinsker's inequalities,

$$\mathbb{P}_{F_Y}(D_{\mathrm{KL}}(P_Y\|Q_Y) > \kappa_0) \le e/\kappa_0, \qquad D_{\mathrm{KL}}(P_y\|Q_y) \le \kappa_0 \Longrightarrow \mathrm{TV}(P_y, Q_y) \le \epsilon_0.$$

With the independent capped sampler of Section 7, choose $\rho \le \epsilon_0$, so $\mathrm{TV}(P_y, \widetilde{Q}_y) \le 2\epsilon_0$ for a query satisfying the KL bound.

We use the following consequence of the regular-subset theorem cited below. For a fixed sufficiently small $\alpha > 0$, there are universal C, K such that if $\mathrm{Cov}(D) \preceq B^2 I_d$ and

$$n \ge C[d\log(2d) + \log(1/\tau)], \tag{49}$$

then an i.i.d. sample from D contains, with probability at least $1 - \tau/2$, a subset G of at least $(1 - \alpha/2)n$ indices for which

$$\frac{1}{n}\sum_{i\in G}(x_i - m)(x_i - m)^T \preceq KB^2 I_d. \tag{50}$$

This follows from [DKP21, Theorem 1.4 and Definition 1.2] by fixing its contamination parameter to a small universal constant and increasing C. Its stability bound controls the second moment about the population mean. The zero-covariance case holds directly. We combine this regular-subset guarantee with the explicit filter below.

For a fixed y satisfying the KL bound, maximally couple independent pairs of clean P_y draws and actual $\widetilde{Q}_y$ draws. Choose ϵ_0 so $2\epsilon_0 \le \alpha/8$. A Chernoff bound ensures at most $\alpha n/2$ pairs differ, except with probability e^{-cn}.

Delete these differing indices from the regular subset. The remaining observed subset has size at least $(1 - \alpha)n$ and still obeys (50), because deletion decreases an unnormalized PSD sum. The deletion bound holds for the surviving subset even when its observations are dependent. Increase C in (49) to absorb both failures.

8.3 An explicit weighted robust mean filter

Fix $\alpha = 1/100$. Suppose observed vectors $x_1, \ldots, x_n$ contain an unknown set G_0 with $|G_0| \ge (1-\alpha)n$ and satisfying (50) for an unknown mean m and known scale B. Initialize $w_i = 1/n$. Repeat at most n times:

1. Compute $W = \sum_i w_i$. If $W < 1/2$, return zero. Otherwise form
$$a = \sum_i w_i x_i / W, \qquad M = \sum_i w_i (x_i - a)(x_i - a)^T / W.$$

2. If $M = 0$, return a. Otherwise use the finite spectral procedure below to find a unit v with $\lambda_{\max}(M)/2 \le q := v^T M v \le \lambda_{\max}(M)$.

3. If $q \le 100KB^2$, return a. Otherwise set $g_i = [v \cdot (x_i - a)]^2$ and $g_{\max} = \max_{w_i > 0} g_i$. If $g_{\max} = 0$, return a; otherwise replace each positive w_i by $w_i(1 - g_i/g_{\max})$.

If the cap is reached without a return, return zero. Every nonterminal update eliminates at least one positive weight, so the algorithm terminates within its deterministic cap on every input.

Lemma 8.1 (Filter guarantee)**.** *Under the stated condition on the subset, the filter returns a with $\|a - m\| < 3\sqrt{K}B$.*

Proof. Let $G = \sum_{i \in G_0} w_i$, $H = \sum_{i \notin G_0} w_i$. We prove inductively that total removed good weight is at most total removed bad weight. Then $G \ge 1 - 2\alpha$, $H \le \alpha$, $W \ge 1 - 2\alpha$, and $w_i \le 1/n$. The weighted good second moment about m is bounded by $KB^2 I_d$.

Put $\lambda = \lambda_{\max}(M)$, $\Delta = a - m$, and $u = \sum_{i \in G_0} w_i (x_i - m)$. Directional Cauchy–Schwarz gives $\|u\| \le \sqrt{GK}B \le \sqrt{K}B$. For the contribution of the other indices, apply Cauchy–Schwarz in each unit direction and use the definition of M:
$$\left\| \sum_{i \notin G_0} w_i (x_i - a) \right\| \le \sqrt{HW\lambda} \le \sqrt{\alpha\lambda}.$$

Here $H \le \alpha$ and $W \le 1$, since the weights only decrease. Weighted centering therefore gives
$$G\Delta = u + \sum_{i \notin G_0} w_i (x_i - a), \qquad \|\Delta\| \le \frac{\sqrt{K}B + \sqrt{\alpha\lambda}}{G}.$$

The good filtering score is bounded by
$$\begin{aligned} S_{\text{good}} = \sum_{i \in G_0} w_i [v \cdot (x_i - a)]^2 &\le 2KB^2 + 2G\,\|\Delta\|^2 \\ &\le (2 + 4/G)KB^2 + (4\alpha/G)\lambda \le 7KB^2 + 5\alpha\lambda. \end{aligned}$$

The total score is $S = Wq \ge 0.49\lambda$. If an update occurs, $\lambda \ge q > 100KB^2$, so $S_{\text{good}} < 0.12\lambda < S/2$. Removed good and bad weights are respectively $S_{\text{good}}/g_{\max}$ and $(S - S_{\text{good}})/g_{\max}$. The update therefore preserves the invariant.

When the algorithm terminates by the covariance criterion, $\lambda \le 2q \le 200KB^2$, and therefore
$$\|a - m\| \le \frac{(1 + \sqrt{2})\sqrt{K}B}{0.98} < 3\sqrt{K}B.$$

The $M = 0$ case obeys the same bound. The low-mass safeguard cannot occur. Since positive good weight remains, the process must satisfy a return criterion before all n weights can be eliminated. □

Deterministic spectral step. For nonzero PSD M in dimension d, let $r = \lceil \log(2d)/(2\log(4/3)) \rceil$. Compute $y_j = M^r e_j$ for every coordinate vector, discard zero vectors, and choose a largest Rayleigh quotient $y_j^T M y_j / \|y_j\|^2$. A top unit eigenvector has squared overlap at least $1/d$ with one coordinate. For that coordinate, the squared norm of the powered vector restricted to eigenvalues at least $3\lambda/4$ is at least λ^{2r}/d, while the remainder is at most $(3/4)^{2r}\lambda^{2r} \le \lambda^{2r}/(2d)$. Its quotient is at least $\lambda/2$. Normalize the best vector by a square root. This uses $O(d^3 \log(2d))$ operations without a spectral gap or a magnitude-dependent loop. The entire filter costs $O(n^2 d^2 + nd^3 \log(2d))$ operations. The bound counts exact-real operations, including on large numerical values.

8.4 The finite list and its covered mass

For each of N_{cov} independent observations X from $(1-b)F + bH$, add fresh σG to obtain y. Draw $n = C[d\log(2d) + \log(4N_{\mathrm{cov}}/\delta_{\mathrm{cov}})]$ independent capped posterior samples. Apply the filter with the known $B = 4\sigma\sqrt{L}$ and return the center $c_y = 2a - y$.

Let E be the clean joint event that E_{geom} holds and conditional KL is at most κ_0. Its failure probability is at most $e_* := \beta + e/\kappa_0$. Put $g_j = \mathbb{P}(E \mid U = \mu_j)$ and

$$I = \{j : g_j \ge 1/2,\ w_j \ge \beta/k\}.$$

Since $\sum_j w_j(1-g_j) \le e_*$, atoms with $g_j < 1/2$ have total mass at most $2e_*$. The additional atoms of weight below β/k have mass at most β. Thus

$$\sum_{j \notin I} w_j \le 3\beta + 2e/\kappa_0. \tag{51}$$

Each fresh contaminated observation has probability at least $(1-b)w_j g_j \ge \beta/(4k)$ of being clean, generating $j \in I$, and satisfying E. The number of observations in (5) therefore ensures that this event occurs for every $j \in I$, except with probability $\delta_{\mathrm{cov}}/2$.

For each such good query, (47) and Lemma 8.1 give, except with probability $\delta_{\mathrm{cov}}/(4N_{\mathrm{cov}})$,

$$\|c_y - \mu_j\| \le (24\sqrt{K} + \sqrt{40})\sigma\sqrt{L}.$$

A union bound over the queries completes the proof, taking $C_{\mathrm{cov}} \ge 2/\kappa_0$ and $C_{\mathrm{rad}} \ge 24\sqrt{K} + \sqrt{40}$. Queries outside the good event, including queries generated by H, may return arbitrary centers, which are also retained. The deterministic caps apply to every query. The coupling argument already accounts for the error of the capped posterior draws.

9 Computational guarantees and finite-bit learning

Theorem 1.1 is an exact-real theorem. Appendix A separately proves a fixed-precision observation result under a known lower bound on positive weights, with bit complexity charged to the actual input encoding length and error measured after Gaussian smoothing. Its proof uses a distinct signed-reconstruction algorithm. In the finite-bit model, the precision is fixed before sampling and each oracle call supplies a fresh quantized observation. The weight lower bound enters both its reconstruction and its complexity bounds.

Several local ingredients admit direct finite rational implementations. For example, replacing the hard gate by the acceptance probability

$$\prod_{r=1}^{8} \frac{T}{T + \|W_r\|^2}$$

preserves Lemma 3.2: use $(1+x)^{-1} \le x^{-1/4}$ for the upper acceptance bound and $1-\prod_r(1+x_r)^{-1} \le \sum_r x_r$ for retention. All denominators are at least T. Rational stereographic plane rotations also preserve orthogonality exactly. Extending the unrestricted-weight learner to finite-bit computation requires a joint error analysis for quantized observations, finite normal sources, branch decisions, coefficient growth, and the final continuous output mechanism.

The output convention matters for total variation: a purely finite-bit output law is supported on a countable set, whereas (1) is absolutely continuous, so the former has total variation distance one from P. Appendix A therefore evaluates its discrete output after an explicit Gaussian convolution. The main theorem provides a sampling program with universal polynomial resource bounds. Reducing the polynomial exponents and supporting normalized density evaluation are natural goals for further work.

A Finite-bit learning with a minimum weight and Gaussian smoothing

This appendix develops a finite-bit reconstruction algorithm under a minimum-weight assumption, with error measured after an explicit Gaussian convolution of its output.

Theorem A.1 (Fixed-precision observations and smoothed TV)**.** *There are universal integers $C, c \ge 1$ with the following property. Suppose*

$$P = \sum_{j=1}^{k_0} w_j \mathcal{N}(\mu_j, I_d), \qquad k_0 \le k, \qquad w_j \ge \alpha > 0,$$

where the means are arbitrary finite vectors. The parameters $\alpha, \varepsilon, \delta$ are rational, $0 < \alpha \le 1$, and $0 < \varepsilon, \delta < 1/2$. Put

$$A = 2 + d + k + \alpha^{-1} + \varepsilon^{-1} + \log(1/\delta), \qquad \overline{A} = 2 + d + k + \lceil \alpha^{-1} \rceil + \lceil \varepsilon^{-1} \rceil + \lceil \log_2(1/\delta) \rceil,$$

and set $b = C[(\lceil \log_2 \overline{A} \rceil)^2 + \lceil \log_2(1/\delta) \rceil + 1]$ and $\tau = \varepsilon/(100d)$. An oracle returns independent $Q_b(X)$, $X \sim P$, where $Q_b(x) = 2^{-b}\lfloor 2^b x \rfloor$ coordinatewise. From at most A^c observations, an algorithm produces a finite rational randomized program S. If L is the total encoding length of parameters and observations actually read, its training bit cost, description size, per-output bit cost, and per-output random-bit count are $\mathrm{poly}(A, L)$*. With probability at least $1 - \delta$, the discrete output law Q of S satisfies*

$$\mathrm{TV}(Q * \mathcal{N}(0, \tau^2 I_d), P) \le \varepsilon.$$

No new target observations are required to generate outputs. Equivalently, after the rational output of S, one ideal Gaussian-vector draw gives a continuous output with the displayed TV guarantee. This final draw is an explicit additional source of randomness and is not a finite-bit operation.

All finite computations below include explicit checks and deterministic caps; if a check fails, the algorithm returns the program that always outputs zero. The bit cost of the integer parts of remote coordinates is included in L. We first describe and prove an analytic bounded-radius procedure, then construct its finite-bit implementation, and finally remove the need for a radius bound as input.

A.1 Analytic reconstruction from a given cover

In this subsection, assume that centers $c_1, \dots, c_M$ cover *every* positive-weight mean within radius B. Let $\phi_{c,a}$ denote the density of $\mathcal{N}(c, a^2 I_d)$ and write $\phi_c = \phi_{c,1}$.

For fresh $X \sim P$ and $W \sim \mathcal{N}(0, I_d)$, put $U = X + W$, $V = X - W$. Conditional on a latent mean, they are independent with distribution $\mathcal{N}(\mu_j, 2I_d)$. Let $g(U)$ be a nearest-center gate and define the positive measures

$$q_{ij} = \mathbb{P}(g(U) = i \mid j), \qquad \nu_i = \sum_j w_j q_{ij} \delta_{\mu_j - c_i}.$$

They have at most k atoms and mass at most one, and $p = \sum_i \int \phi_{c_i+z}\, d\nu_i(z)$. For $r = \|\mu_j - c_i\|$, selection by the nearest-center gate over a center covering μ_j implies a one-dimensional $\mathcal{N}(0,2)$ fluctuation of at least $(r-B)/2$. Therefore, with $A_B = e^{B^2/16}$,

$$q_{ij} \le e^{-(r-B)_+^2/16} \le A_B e^{-r^2/32}, \qquad \int F(\|z\|^2)\, d\nu_i(z) \le A_B \sup_{y \ge 0} e^{-y/32} F(y) \tag{52}$$

for nonnegative F.

Set $D = d + 1$ and $v(z) = (z/\sqrt{T}, 1)$. With an independent standard Z', the implicit order-t statistic

$$S_{it} = \mathbf{1}_{g(U)=i} \Re(a^{\otimes t}), \qquad a = ((V - c_i)/\sqrt{T} + i\sqrt{2/T} Z', 1)$$

has expectation $M_{it} = \int v(z)^{\otimes t} d\nu_i(z)$. For every tensor C with $\|C\|_F \le 1$ and $T \ge 64t$,

$$\mathbb{E}\langle C, S_{it} \rangle^2 \le A_B e^{4t^2/T}. \tag{53}$$

Indeed, conditional on the atom and gate, $a = v + G$ with circular complex spatial noise of covariance $4I_d/T$. Distinct holomorphic degrees are orthogonal, so Wick pairing bounds its squared contraction by

$$\sum_{\ell=0}^{t} \binom{t}{\ell}^2 \ell! (4/T)^\ell \|v\|^{2(t-\ell)} \le (\|v\|^2 + 4t/T)^t \le e^{4t^2/T + t\|z\|^2/T}.$$

The bound then follows from (52). If ν_i^T restricts to $\|z\|^2 \le T$ and M_{it}^T is its augmented moment, then for $T \ge 32t$,

$$\left\| M_{it} - M_{it}^T \right\|_F \le A_B 2^{t/2} e^{-T/32}. \tag{54}$$

This follows because $e^{-y/32}(1 + y/T)^{t/2}$ decreases for $y \ge T$. Both formulas use the independence of the query observation from the gate, conditional on the latent atom.

Biased positive-moment compression. Let a positive k-atomic measure of mass at most one be supported on $\|v\| \le \sqrt{2}$. Suppose the available order-2ℓ statistics have mean within Frobenius error β_0 of $M_{2\ell} = \int v^{\otimes \ell} (v^{\otimes \ell})^T d\nu$ and directional second moments at most V_0. For $\ell \le r$, fix

$$\zeta = \eta^2/(128\sqrt{k}\, 4^r), \qquad \beta_0 \le \zeta/4.$$

There are sequential co-isometries Φ_ℓ with at most k rows and PSD cores C_ℓ such that

$$A_\ell = \Phi_\ell^T C_\ell \Phi_\ell \succeq 0, \quad \operatorname{rank} A_\ell \le k, \qquad \|A_\ell - M_{2\ell}\|_F \le \eta, \tag{55}$$

using $\mathrm{poly}(D, k, r, V_0, \eta^{-1}, 2^r, \log(1/\delta'))$ operations and samples.

For the implicit implementation, we account for the fixed bias in each matrix tolerance and use fresh samples to control the additional estimation error. The resulting projection loss is then propagated across degrees before the final small core is estimated.

Start at $\Phi_0 = 1$. For $Q_\ell = \Phi_{\ell-1} \otimes I_D$, estimate $K_\ell = Q_\ell M_{2\ell} Q_\ell^T$ to error ζ, using a fresh entrywise median-of-means batch; its dimension is at most kD. Each entry is a unit-Frobenius contraction. Block means of $O(V_0/u^2)$ samples and a median of $O(\log(1/\gamma))$ blocks give entry error u with failure probability γ. Taking $u = \zeta/(4kD)$ and a union bound gives the claimed matrix accuracy including bias.

Select top-k singular rows V_ℓ and set $\Phi_\ell = V_\ell Q_\ell$. Write S for the orthogonal projector onto the selected singular subspace; at this stage $S = V_\ell^T V_\ell$. If K is PSD of rank at most k and $\left\|\widetilde{K} - K\right\|_F \le \zeta$, optimal rank approximation gives $\|(I-S)K\|_F \le 2\zeta$, hence $\operatorname{tr}((I-S)K) \le 2\sqrt{k}\zeta$. Consequently, with $E_\ell = \int \left\|(I - \Phi_\ell^T \Phi_\ell) v^{\otimes \ell}\right\|^2 d\nu$, the residuals from the preceding projection and the new rank-k projection are orthogonal, giving

$$E_\ell = \int \|v\|^2 \left\|(I - \Phi_{\ell-1}^T \Phi_{\ell-1}) v^{\otimes(\ell-1)}\right\|^2 d\nu + \operatorname{tr}((I-S)K_\ell).$$

The first term is at most $2E_{\ell-1}$ because $\|v\| \le \sqrt{2}$, and the second is at most $2\sqrt{k}\zeta$ by the rank-approximation bound. Hence

$$E_\ell \le 2E_{\ell-1} + 2\sqrt{k}\zeta, \quad E_0 = 0,$$

which yields $E_\ell \le 2\sqrt{k}\zeta(2^\ell - 1)$. For an orthogonal projector P, the identity obtained by adding and subtracting $(Px)x^T$ gives

$$\left\|xx^T - (Px)(Px)^T\right\|_F \le 2\|x\| \, \|(I-P)x\|.$$

Apply this inequality to each atom and then use Cauchy–Schwarz, the mass bound, and $\|v\| \le \sqrt{2}$. Substituting the chosen ζ and using $\ell \le r$ gives $\left\|M_{2\ell} - \Phi_\ell^T \Phi_\ell M_{2\ell} \Phi_\ell^T \Phi_\ell\right\|_F \le 2\sqrt{2^\ell E_\ell} \le \eta/4$. This explains the factor 4^r in the precision allocation. Estimate the final small core from another fresh block, then project it onto the PSD cone. Nonexpansiveness of that projection proves (55).

Write $V_{\ell,j}$ for the block of V_ℓ corresponding to last coordinate j. For complex vectors $a_1, \dots, a_\ell$, the recursion

$$b_0 = 1, \qquad b_q = \sum_{j=1}^{D} (a_q)_j V_{q,j} b_{q-1}$$

computes $\Phi_q(a_1 \otimes \cdots \otimes a_q)$ in $O(qDk^2)$ operations. The moment query uses the ordinary transpose, giving $\mathbf{1}_g \Re(bb^T)$ rather than bb^*. This recursion evaluates every compressed statistic through the stored frames and small cores.

Dimension-free partial traces. We convert the compressed matrix approximation to mixed moments through partial traces. Controlling these maps in nuclear norm gives an error bound independent of the ambient dimension. Let $D = d+1$, let $P : \mathbb{R}^D \to \mathbb{R}^d$ select the spatial coordinates, and let e_D be the last basis vector. For an operator A on $(\mathbb{R}^D)^{\otimes \ell}$ and $0 \le b \le \ell$, put $r = \ell - b$ and define the restricted partial trace

$$\mathcal{T}_b(A) = \sum_{j_1, \dots, j_b \le d} (I_{D^r} \otimes e_{\boldsymbol{j}}^T) A (I_{D^r} \otimes e_{\boldsymbol{j}}), \qquad e_{\boldsymbol{j}} = e_{j_1} \otimes \cdots \otimes e_{j_b}.$$

The empty trace is the identity. This map is positive and, for $A \succeq 0$,

$$\operatorname{tr} \mathcal{T}_b(A) = \operatorname{tr}\left[(I_{D^r} \otimes (P^T P)^{\otimes b}) A\right] \le \operatorname{tr} A.$$

Writing a Hermitian E as $E_+ - E_-$ and using the triangle inequality consequently gives $\|\mathcal{T}_b(E)\|_* \le \operatorname{tr} E_+ + \operatorname{tr} E_- = \|E\|_*$, where $\|\cdot\|_*$ denotes nuclear norm.

For $v = (z/\sqrt{T}, 1)$, the identity

$$\mathcal{T}_b(v^{\otimes \ell}(v^{\otimes \ell})^T) = \left\|z/\sqrt{T}\right\|^{2b} v^{\otimes r}(v^{\otimes r})^T$$

identifies the desired mixed moment

$$R_{itb} = \int \left\|z/\sqrt{T}\right\|^{2b} (z/\sqrt{T})^{\otimes t} d\nu_i^T(z).$$

For even $t = 2r$, take $\ell = r + b$ and apply $L = P^{\otimes r}$ on both sides of $\mathcal{T}_b(A_\ell)$. For odd $t = 2r - 1$, take the same $\ell = r + b$ and apply

$$L = P^{\otimes r}, \qquad R = P^{\otimes (r-1)} \otimes e_D^T$$

on the left and right, respectively. The resulting $L\mathcal{T}_b(A_\ell)R^T$ is a $d^r \times d^{r-1}$ matrix, viewed as an order-t tensor. The final coordinate is fixed before contraction and contributes the factor $v_D = 1$. For example, $t = 1, b = 1$ uses order four: tracing one spatial pair and pinning the remaining right coordinate produces $\int \left\|z/\sqrt{T}\right\|^2 (z/\sqrt{T}) d\nu_i^T(z)$.

All left and right maps have operator norm at most one. For $E = A_\ell - M_{2\ell}$, (55) and $\operatorname{rank} E \le 2k$ therefore imply

$$\left\|L\mathcal{T}_b(E)R^T\right\|_F \le \left\|L\mathcal{T}_b(E)R^T\right\|_* \le \|E\|_* \le \sqrt{2k}\, \|E\|_F \le \sqrt{2k}\eta.$$

This proves the claimed error using order $t + 2b$ when t is even and $t + 2b + 1$ otherwise. For $t = b = 0$, estimate the gate indicator separately and clip to $[0, 1]$; its truncation bias is bounded by (54) at order zero.

These maps can also be implemented directly on the small cores. Starting from $C^{(0)} = C_\ell$, compute

$$C^{(j+1)} = \sum_{a \le d} V_{\ell-j,a}^T C^{(j)} V_{\ell-j,a} \qquad (0 \le j < b).$$

An induction using $\Phi_h = V_h(\Phi_{h-1} \otimes I_D)$ shows that $\mathcal{T}_j(A_\ell) = \Phi_{\ell-j}^T C^{(j)} \Phi_{\ell-j}$. Each update costs $O(dk^3)$ operations. For the remaining bilinear form, use the preceding sequential contractions with spatial input vectors embedded by P^T, and with one final input e_D in the odd-order right factor. Derivatives insert a coordinate vector in each surviving position, adding only a polynomial factor in the order. The nuclear-norm bound controls the statistical error, and the small-core recursion implements the partial trace.

A bounded radial polynomial. For $0 < a \le 1/4$, $0 < \rho < 1/4$, there is a degree-$s \le C_w a^{-1/2} \log(1/\rho)$ polynomial p, with $C_w = 7000$ sufficient, satisfying

$$0 \le p \le 1 \ (0 \le u \le 1), \quad p \ge 1 - \rho \ (u \le a), \quad p \le \rho \ (u \ge 2a),$$
$$\sum_b |a_b| \le 4(s+1)^2 64^s \quad \text{if } p^2 = \sum_b a_b u^b.$$

Approximate $a/(a+u)$ by its Chebyshev series after $x = 2u - 1$. With $A_0 = 1 + 2a$ and $r_0 = A_0 - \sqrt{A_0^2 - 1}$, this series is

$$\frac{2a}{\sqrt{A_0^2 - 1}} \left(1 + 2\sum_{j \ge 1} (-r_0)^j T_j(x)\right).$$

Its tail after $n_0 = \lceil \log(200)/\sqrt{a} \rceil$ terms is at most $2e^{-n_0\sqrt{a}} \le 1/100$. Rescale the truncated polynomial Q to $q = (Q+1/100)/(1+2/100)$. On $[0,1]$ it lies in $[0,1]$, on $[0,a]$ it is at least $25/51$, and on $[2a,1]$ it is at most $53/153$. Take the binomial upper-tail polynomial with $n = \lceil 1000\log(1/\rho)+100 \rceil$ trials and threshold $\lceil 5n/12 \rceil$. Hoeffding's inequality gives the stated bounds on the pass and stop intervals. A polynomial bounded by one on $[0,1]$ has shifted-Chebyshev coefficients of magnitude at most two; the monomial coefficient sum of $T_j(2u-1)$ is at most 8^j by its recurrence. Squaring gives the stated coefficient bound.

Compatible parameters and signed reconstruction. The radial window and the moment oracle use different truncation scales. The window restricts the mass used for density reconstruction, while the larger moment cutoff makes the oracle's bias small enough to meet the compression tolerance. We choose their ratio to account for the growth of the reconstruction coefficients.

Given $0 < \xi < 1/2$, set

$$L_0 = 128(B^2 + \log(10^4 kM/\xi) + 1), \quad m = \lceil 64L_0 \rceil, \quad T = HL_0, \quad \rho = e^{-128L_0 \log(eH)}.$$

Choose the universal constant H sufficiently large to satisfy the condition below. Use the radial polynomial with $a = 1/H$, degree $s \le S_H L_0$, $S_H = 128C_w\sqrt{H}\log(eH) + 1$. Let N_m be the least even integer at least $m + 4s + 2$, $r = N_m/2$, and $D_H = 70 + 4S_H$, so $N_m \le D_H L_0$. Choose H so

$$H/32 \ge 1000(D_H + \log(eH) + 1). \tag{56}$$

This is possible since $D_H = O(\sqrt{H}\log H) = o(H)$. Put

$$W_0 = \max_{t \le m} T^{t/2}/\sqrt{t!}, \qquad \eta = \frac{\xi}{10^4 M(m+1)\sqrt{2k}W_0 A_{\mathrm{coef}}}, \quad A_{\mathrm{coef}} = \sum_b |a_b|,$$

and use ζ from (55). The factorial bound gives $\log W_0 \le 33L_0\log(eH)$. Together with the coefficient bound it yields

$$\log(1/\eta) \le (10D_H + 40\log(eH) + 20)L_0, \quad \log(4/\zeta) \le (30D_H + 100\log(eH) + 100)L_0.$$

These inequalities verify the simultaneous moment requirements. Indeed, $B^2/16 \le L_0$ and $N_m \le D_H L_0$ give

$$\log(A_B 2^{N_m/2} e^{-T/32}) \le (1 + D_H - H/32)L_0 \le -(30D_H + 100\log(eH) + 100)L_0 \le \log(\zeta/4).$$

The middle inequality follows directly from (56). The same condition gives $H \ge 64D_H$, hence $T \ge 64N_m$. Thus the bias (54) through every required order is at most $\zeta/4$, and the variance bound (53) applies to all those orders. Its common upper bound $A_B e^{4N_m^2/T}$ and all inverse tolerances are $e^{O_H(L_0)}$. In particular, H is a universal constant, fixed before the input.

Define the positive compact measure $\widetilde{\nu}_i = p(\|z\|^2/T)^2 \mathbf{1}_{\|z\|^2 \le 2L_0}\nu_i$. Its discarded mass is at most $2\rho + A_B e^{-L_0/32}$. Its normalized Hermite coefficient is $F_{it} = \int z^{\otimes t} d\widetilde{\nu}_i/\sqrt{t!}$. The Hermite tail after m has L^2 norm at most $\sqrt{2}(e/32)^{32L_0}$, by $t! \ge (t/e)^t$. The accessible traced coefficient

$$F'_{it} = \frac{T^{t/2}}{\sqrt{t!}} \sum_b a_b R_{itb}$$

differs from F_{it} by at most $\rho^2 W_0$, because the remaining region $2L_0 < \|z\|^2 \le T$ is in the stop interval. Its learned version $\widehat{F}_{it}$ has further error at most $\sqrt{2k}\eta W_0 A_{\mathrm{coef}}$. It follows that

$$h(x) = \sum_i \phi_{c_i}(x) \sum_{t=0}^{m} \langle \widehat{F}_{it}, H_t(x - c_i)\rangle, \qquad \|h - p\|_1 \le \xi. \tag{57}$$

The table below bounds the four sources of error for one center. Summing over centers contributes a factor M to the global L^1 error.

Source of error	Bound	Allocation
Discarded positive mass	$2\rho + A_B e^{-L_0/32}$	$\xi/(50M)$
Hermite tail	$\sqrt{2}(e/32)^{32L_0}$	$\xi/(100M)$
Radial stop interval	$(m+1)\rho^2 W_0$	$\xi/(100M)$
Learned coefficients	$(m+1)\sqrt{2k}\eta W_0 A_{\text{coef}}$	$\xi/(10^4 M)$

Each bound is at most the allocated error. To verify the three analytic rows, note that $L_0 \geq 128$, $m+1 \leq 66L_0$, and

$$e^{-L_0/128} \leq \frac{\xi}{10^4 kM}, \quad A_B e^{-L_0/32} \leq e^{-L_0/128}, \quad \sqrt{2}(e/32)^{32L_0} \leq e^{-L_0/128}.$$

Also $2\rho \leq 2e^{-L_0/128}$, while $\log(m+1) \leq L_0$ and the bound on W_0 give

$$(m+1)\rho^2 W_0 \leq \exp\bigl(-222L_0 \log(eH)\bigr) \leq e^{-L_0/128}.$$

The final row equals its allocation by the definition of η. The total error from discarded mass and local L^2 approximation is therefore bounded by

$$M\left[2\rho + A_B e^{-L_0/32} + \sqrt{2}(e/32)^{32L_0} + (m+1)\rho^2 W_0 + (m+1)\sqrt{2k}\eta W_0 A_{\text{coef}}\right] < \xi.$$

The compression tolerance already includes the moment truncation bias, bounded above by $\zeta/4$. Also $\sum_{t\leq m} \left\|\widehat{F}_{it}\right\|_F \leq (m+1)e^{L_0} + 1 =: Q_0$. Since $L_0 = O(B^2 + \log(kM/\xi) + 1)$ and H is universal, all training and representation costs are $\text{poly}(d, k, M, e^{B^2}, \xi^{-1}, \log(1/\delta'))$.

Randomized evaluation and sampling. For each order-t tensor, use $\Re\langle C, (z+iG)^{\otimes t}\rangle/\sqrt{t!}$ as an unbiased evaluator of its Hermite contraction. Circular Wick pairing gives integrated second moment at most $2^t \|C\|_F^2$, and for its gradient at most $t2^{t-1}\|C\|_F^2$. Averaging n independent replicas of each complete local sum gives total integrated L^1 error at most $M2^{m/2}Q_0/\sqrt{n}$. This is polynomially small for polynomial n. Fresh evaluator randomness at each query is permitted.

Let $q = M^{-1}\sum_i \phi_{c_i}$ and $K_0 = Me^{B^2}$. Assigning each true component to a covering center and completing a square gives $\int p^2/q \leq K_0$. With training and evaluation error at most $\varepsilon/100$ each, propose from q and accept with probability $\text{clip}(\widehat{h}(x)/q(x), 0, C_{\text{cap}})/C_{\text{cap}}$, where $C_{\text{cap}} = 100K_0/\varepsilon$. Put $g(x) = \mathbb{E}_{\text{eval}}\,\text{clip}(\widehat{h}(x)/q(x), 0, C_{\text{cap}})$. Since clipping is one-Lipschitz,

$$\int q|g - p/q| \leq 2\varepsilon/100 + K_0/C_{\text{cap}} = 3\varepsilon/100.$$

Thus $Z = \int qg \geq 1/2$ and $\text{TV}(qg/Z, p) \leq 3\varepsilon/100$. A cap of $\lceil 2C_{\text{cap}} \log(100/\varepsilon)\rceil$ proposals, with a fresh q fallback, adds at most $\varepsilon/100$. This establishes the sampling guarantee under a given cover, with a deterministic cap on the number of proposals. The accepted density is obtained by averaging the evaluator randomness *after* clipping.

A.2 Bounded-radius posterior refinement

Assume $\|\mu_j\| \le R$, all given centers also lie in this ball, and the cover radius at noise $a \ge 1$ is Ba. Scaling (57) gives a signed h_a with

$$\int |h_a - p_a| \le \xi, \qquad \int \|x\| \, |h_a - p_a| \le (R + a\sqrt{d})\xi. \tag{58}$$

To obtain the weighted estimate, apply the preceding error decomposition before summing its contributions. Each discarded nonnegative component measure has first moment at most $R + a\sqrt{d}$ times its mass, because its Gaussian mean lies in the radius-R ball. For a local residual ψ with reference density $\phi_{c,a}$, Cauchy–Schwarz gives

$$\int \|x\| \, \phi_{c,a}(x)|\psi(x)| \, dx \le \sqrt{\|c\|^2 + a^2 d} \, \|\psi\|_{L^2(\phi_{c,a})} \le (R + a\sqrt{d}) \, \|\psi\|_{L^2(\phi_{c,a})} .$$

Thus the discarded mass and each local L^2 error acquire the same factor, and their previously allocated sum proves the second estimate in (58).

Set $\sigma = 2a$ and smooth by variance $3a^2$. A degree-t Gaussian-weighted Hermite term is multiplied by $(a/\sigma)^t = 2^{-t}$ after smoothing; Gaussian conditioning or the Hermite generating function proves this identity. Let $m(y) = \mathbb{E}[\mu_J \mid Y = y]$, $n(y) = p_\sigma(y)m(y) = yp_\sigma(y) + \sigma^2 \nabla p_\sigma(y)$. Put $h_\sigma = h_a * \phi_{0,\sqrt{3}a}$ and $b(y) = yh_\sigma(y) + \sigma^2 \nabla h_\sigma(y)$. If $\Delta = h_a - p_a$, then

$$b(y) - n(y) = \int [x - (y - x)/3]\Delta(x)\phi_{0,\sqrt{3}a}(y - x) \, dx.$$

Tonelli and (58) give $\|h_\sigma - p_\sigma\|_1 \le \xi$ and $\int \|b - n\| \le (R + 2a\sqrt{d})\xi$. A local term contributes $\phi_{c,\sigma}(y)[cf(z) + \sigma \nabla f(z)]$, $z = (y - c)/\sigma$; the term involving the unbounded query vector cancels. The evaluator variance above supplies random estimates a_0, b_0 with

$$\int \mathbb{E}|a_0 - p_\sigma| \le 2\xi, \qquad \int \mathbb{E} \, \|b_0 - n\| \le 3S_0\xi, \quad S_0 = 1 + R + \sigma\lceil\sqrt{d}\rceil,$$

at polynomial cost in the local parameters.

Let $q = M^{-1} \sum_i \phi_{c_i,\sigma}$ and use reference ratios $A_0 = a_0/q$, $B_0 = b_0/q$. For $\lambda > 0$ define

$$\widetilde{m}(y) = \mathrm{Proj}_{B(0,R)} \frac{B_0(y)}{\max(A_0(y), \lambda)}.$$

For arbitrary scalar a_0 and vector b_0, $p > 0$, $n = pm$, $\|m\| \le R$, the projected ratio with positive denominator satisfies

$$p \, \|\widehat{m} - m\|^2 \le 4R \, \|b_0 - n\| + 8R^2|a_0 - p|.$$

For $a_0 \ge p/2$, use nonexpansiveness and $\|\widehat{m} - m\| \le 2R$; for $a_0 < p/2$, use $\|\widehat{m} - m\| \le 2R$ and $|a_0 - p| \ge p/2$. Replacing a_0 by $\max(a_0, \lambda q)$ adds at most λq to its absolute error. Therefore

$$\mathbb{E} \, \|\widetilde{m}(Y) - m(Y)\|^2 \le 28S_0^2\xi + 8R^2\lambda < \alpha\sigma^2 \tag{59}$$

for $\xi = \lambda = \alpha\sigma^2/(1000S_0^2)$.

To localize each component, write $\gamma_j = \phi_{\mu_j,\sigma}$ and posterior weights $\pi_\ell = w_\ell\gamma_\ell/p_\sigma$. The inequalities $p_\sigma \ge w_j\gamma_j, w_\ell\gamma_\ell$ give

$$\mathbb{E}_{Y \sim \gamma_j} \pi_\ell(Y) \le \sqrt{w_\ell/w_j} e^{-\|\mu_\ell - \mu_j\|^2/(8\sigma^2)}.$$

Sum over remote ℓ, use $\sum_\ell \sqrt{w_\ell} \le \sqrt{k}$, and integrate the resulting exponential squared-distance tail. Conditional Jensen's inequality yields

$$\mathbb{E}_{\gamma_j} \|m(Y) - \mu_j\|^2 \le 8\sigma^2(1 + \tfrac{1}{2}\log(k/w_j)).$$

Since $w_j \ge \alpha$, (59) gives componentwise estimator error at most σ^2, and thus $\mathbb{E}_{\gamma_j,\mathrm{eval}} \|\widetilde{m}(Y) - \mu_j\|^2 \le 18\sigma^2 \log(ek/\alpha)$.

Choose an integer $B_0 = \lceil \sqrt{320(1 + \lceil \log_2(k/\alpha) \rceil)} \rceil$ and $B = 4B_0$. The ideal estimator is within $(B_0/2)\sigma$ of the generating mean with probability at least $3/4$. A numerical center at distance at most one from this estimator is within $B_0\sigma$, since $\sigma \ge 2$ and $B_0 \ge 2$. Consequently $O(\alpha^{-1}\log(k/\gamma))$ independent candidates cover every mean with failure γ.

In the analytic comparison experiment, a fresh sample from P_a is $X + \sqrt{a^2 - 1}\, G$, and a fresh posterior query from P_{2a} is $X + \sqrt{4a^2 - 1}\, G$, with independent $X \sim P$ and $G \sim \mathcal{N}(0, I_d)$. The former is divided by a before applying the unit-noise reconstruction procedure. Both expressions are valid because $a \ge 1$; the finite implementation uses the quantized X and certified approximations of the displayed square roots, whose errors are accounted for in the numerical analysis below. Both sample types are constructed directly from target observations and Gaussian randomness.

Start at $a_0 = \max(1, R)$ with center zero; at each step use $a_{\ell+1} = \max(1, a_\ell/2)$, learn, smooth to $\sigma = 2a_\ell$, and retain every candidate. The radius is at most $B_0\sigma \le 4B_0 a_{\ell+1} = B a_{\ell+1}$. At unit noise, apply the sampler based on the given cover. There are $O(1 + \log(1 + R))$ rounds; all costs are polynomial in $d, k, 1 + R, \alpha^{-1}, \varepsilon^{-1}, \log(1/\delta)$. Every call uses fresh data. The cover invariant and the inverse-polynomial tolerances are established before numerical precision is chosen.

A.3 Rational linear algebra and stable gates

The spectral algorithm used below preserves exact rational orthogonality, even at repeated eigenvalues. Given rational symmetric B with $\|B\|_F \le H_0$, $H_0 \ge 1$, and tolerance $\theta > 0$, maximal-pivot Jacobi can be implemented with rational plane rotations to produce rational orthogonal U and rational diagonal D with $\left\|B - UDU^T\right\|_F \le \theta$. Its bit cost is polynomial in the input encoding length, matrix dimension n, and $\log(2 + H_0/\theta)$.

The proof first controls the residual of each rotation, then bounds the growth of rational denominators. If the off-diagonal norm exceeds θ, a maximal pivot b has $|b| \ge \theta/n$. For its 2×2 block with diagonal entries a, d, put $h = (a - d)/2$ and choose

$$u = \begin{cases} -b/(|h| + \sqrt{h^2 + b^2}), & h \ge 0, \\ b/(|h| + \sqrt{h^2 + b^2}), & h < 0. \end{cases}$$

It has $|u| \le 1$ and solves $2hu + b(1 - u^2) = 0$. The stereographic half-angle $t_* = u/(1 + \sqrt{1 + u^2})$ gives an annihilating rotation. Approximate it by a dyadic t with $|t - t_*| \le \theta/(1000nH_0)$ and use the exactly rational rotation

$$c(t) = \frac{1 - t^2}{1 + t^2}, \qquad s(t) = \frac{2t}{1 + t^2}.$$

The first denominator in u is at least $|b| \ge \theta/n$, and the other is at least two. Certified rational square-root bisection therefore attains the stated precision in polynomial bit time without an eigengap. Bounded derivatives of c, s make the new pivot at most $|b|/2$. All other squared off-diagonal entries rotate in pairs and preserve their sum, so the squared off-diagonal norm contracts by at least $1 - 1/[n(n-1)]$. Stop once its exact rational square is at most θ^2. A fixed $O(n^2 \log(2 + H_0/\theta))$ cap suffices; $n = 1$ and diagonal matrices require no rotations.

If t has p fractional bits, each rotation has a common denominator with $O(p)$ bits. Store U using their product and the transformed matrix using the common denominator of the input multiplied by the square of this product. Over polynomially many rotations, denominator lengths grow additively, and numerators are bounded by this length plus $O(\log H_0 + \log n)$. This proves the claimed bit bound. A common denominator for the input itself has at most the sum of the input denominators' bit lengths.

Select the largest absolute entries of D to obtain a rational rank-k frame; the rank-approximation proof leading to (55) now has error $\zeta + \theta$. For a final core, replace negative diagonal entries by zero and conjugate by U, giving an exactly rational PSD matrix within the same additive accuracy. Use spectral tolerances at most a fixed small fraction of ζ, and tighten statistical tolerances accordingly. Before each new Jacobi call, round its dense input to a prescribed dyadic grid and clip entries to a known box containing the population target, such as $[-2^r - 1, 2^r + 1]$. Entrywise clipping cannot increase Frobenius error relative to that target. This rounding prevents denominator lengths from accumulating across successive spectral inputs. Each subsequent fresh-sample argument applies to the rational frame actually selected, including at repeated eigenvalues.

For finite-bit gates, use

$$g(U, \Theta) = \arg\min_i \{\|U - c_i\|^2 + \Theta_i\}, \qquad \Theta_i \overset{\text{iid}}{\sim} \text{Unif}[0, 1].$$

Draw new offsets for every statistic. If a covering center is within B and $r = \|\mu - c_i\| > B + 1$, selection by the gate implies a one-dimensional $\mathcal{N}(0, 2)$ fluctuation of at least $(r - B - 1)/2$: the difference of squared distances is decreased by at most one. Thus (52)–(54) remain true with $B + 1$ in place of B. The offsets and gate side remain independent of the query side given the atom.

If every score is perturbed by at most e, a change in the selected center implies that two original scores differ by at most $2e$. Conditional on the centers and U, each $\Theta_i - \Theta_j$ has density at most one. A union bound gives a disagreement probability at most $4M^2e$. The random offsets give this bound for arbitrary centers, including coincident ones. The implemented offsets are finite-bit approximations coupled to these continuous offsets.

A.4 Rational windows and posterior evaluation

The schedules below use rational thresholds that can be compared exactly. Replace L_0 by the integer

$$L_1 = 128[(B + 1)^2 + 1 + \lceil \log_2(10^4 kM/\xi) \rceil],$$

where the cover bound B is an integer upper bound. Choose a sufficiently large universal integer H, put $t_0 = \lceil \sqrt{HL_1} \rceil$, $T = t_0^2$, and $a = L_1/T$. Then $HL_1 \le T \le H'L_1$, where $H' = H + 2\lceil \sqrt{H} \rceil + 1$. The compatibility proof above continues to hold after increasing the fixed H: every unfavorable coefficient is $O(\sqrt{H} \log H + \log H + 1)$ whereas the favorable bias coefficient is $H/32$.

To construct a rational radial window, truncate the Chebyshev series at the universal degree $6(\lceil \sqrt{H} \rceil + 1)$, whose error is at most $1/200$. Approximate its monomial coefficients by rationals with total absolute error at most $1/200$. Rescale as before to obtain a rational $q \in [0, 1]$ with the same strict margins on the pass and stop intervals. Use $\rho = 2^{-v}$, $v = C_\rho L_1$, where $C_\rho = 256(1 + \lceil \log_2 H \rceil)$, and binomial amplification degree $1000v + 100$. All its coefficients are rational; its exact bound $0 \le p \le 1$ proves the same coefficient-sum estimate. Universal constants are fixed before any input.

Encode the factorial normalization through the rational tensors

$$D_{it} = \frac{t_0^t}{t!} \sum_b a_b \widehat{R}_{itb}, \qquad f_i(z) = \sum_{t=0}^{m} \mathbb{E}_G \Re \langle D_{it}, (z + iG)^{\otimes t} \rangle.$$

The corresponding normalized Hermite coefficient is $\widehat{F}_{it} = \sqrt{t!}D_{it}$, so all preceding analytic bounds concern exactly this stored function. Use the conservative rational envelopes

$$W_1 = (4H')^{33L_1}, \quad A_1 = 4(s+1)^2 64^s, \quad \eta = \frac{\xi}{10^7 M(m+1)(2k)W_1 A_1}, \quad z_0 = \frac{\eta^2}{512k4^r},$$

with $m = 64L_1$ and $2r$ the least even integer at least $m + 4s + 2$. They leave constant-factor slack for all matrix, coefficient, and evaluation errors. Their magnitudes and reciprocal tolerances remain $e^{O_H(L_1)}$.

Reference ratios use softmax weights

$$\omega_i(y) = \frac{e^{-\|y-c_i\|^2/(2\sigma^2)}}{\sum_j e^{-\|y-c_j\|^2/(2\sigma^2)}},$$

with factors $M\omega_i$. Subtract the maximum logit before evaluation. The denominator is at least one in exact arithmetic. For a rational tolerance $0 < e < 1$, compute the integer $B_{\exp} = \lceil \log_2(10M/e) \rceil$ by rational comparisons with powers of two. Omit terms whose shifted logit is below $-B_{\exp}$ and approximate the remaining exponentials to $e/(10M)$. Each omitted term is at most $\exp(-B_{\exp}) \le 2^{-B_{\exp}} \le e/(10M)$, so omitted total mass and normalization error are $O(e)$. Exponentials on the interval $[-B_{\exp}, 0]$ can be evaluated by range reduction and the Taylor series with certified factorial remainder in polynomial bit time. Softmax is two-Lipschitz from logit sup norm to weight ℓ^1 norm, by its derivative $\mathrm{diag}(\omega) - \omega\omega^T$. Maxima, minima, and clipping are continuous, Lipschitz operations even where their active branches change.

The posterior denominator in reference units is at least the known inverse-polynomial λ from (59). A rational approximation inside the radius-R ball is obtained as follows. Approximate the preprojection vector by rational $\widehat{z}$ within $\nu/4$, and compute a rational upper bound u on its norm with $\|\widehat{z}\| \le u \le \|\widehat{z}\| + \nu/4$. Return $\widehat{z}$ if $u \le R$, and $(R/u)\widehat{z}$ otherwise. The output is exactly in the ball and within $\nu/2$ of the ideal projection. Division occurs only for $u > R \ge 1$ in the local routines used below. This meets the accuracy requirement for the numerical centers in the cover argument.

A.5 A quantitative precision and bit budget

The numerical argument treats three operations separately. Smooth evaluations are compared by explicit perturbation bounds, gate decisions by coupling, and spectral choices by their approximation residuals. Concentration then applies directly to the rational frames stored by the algorithm, including when small matrix perturbations change the selected eigenspaces. Let

$$A_R = 2 + d + k + (1 + R) + \alpha^{-1} + \varepsilon^{-1} + \log(1/\delta).$$

In the bounded-radius routine, all finite loop counts, array dimensions, required inverse analytic tolerances, coefficient magnitude bounds, rejection caps, and observation scales have fixed-polynomial bounds in A_R. In particular, every polynomial contraction order, window degree, and Hermite degree is at most $C_0 \log A_R$. The number of scales is also $O(\log A_R)$. These statements follow from the explicit analytic schedules, using $(B+1)^2 = O(\log(ek/\alpha))$ and the polynomial list sizes. Choose a deterministic integer $N = A_R^{O(1)}$, rounded upward by a computable rule, that dominates their sums and products by fixed constants. Elementary-function evaluation loops, whose lengths depend on bit precision, are counted separately after the precision is chosen.

Let $t_* \le C_1 \log A_R$ bound the number of factors in each product term and the degree of a single fixed-frame scalar or vector evaluation expression between successive roundings of dense arrays, plus

one. This bound applies to each individual evaluation. Polynomially many such evaluations and sums contribute additional fixed powers of N. Increase N so $N \geq \max\{2, t_*\}$, and take $t_* \geq 2$. We may further enlarge N by a fixed power to bound each coordinate dimension, batch count, fixed coefficient magnitude, observation scale, and inverse denominator by N. We also require

$$N \geq \max\left\{\frac{100kD}{z_0}, \frac{100}{\xi}, \frac{100S_0^2}{\alpha\sigma^2}, \frac{100}{\tau\varepsilon}\right\}$$

for every call, using the smallest scheduled tolerances and the largest scheduled scales. All these quantities were fixed by the analytic and rational schedules and are polynomial in A_R. Concretely, one may first choose $t_* = \lceil C_1 \log A_R \rceil$ and then $N = \lceil C_2 A_R^{C_3} \rceil$ with sufficiently large fixed constants. The source-box bounds below depend on N through $\log N$ and on confidence through $\log(1/\delta) \leq A_R$, so they are dominated by the same fixed-power choice of N. Numerical precision is chosen only after these bounds. Constants here and below are universal fixed constants of the described construction.

Finite Gaussian sources. For rational $0 < h, \beta < 1/2$, a bounded-bit draw $\widehat{Z}$ admits a coupling to $Z \sim \mathcal{N}(0,1)$ with $\mathbb{P}(|Z - \widehat{Z}| > h) \leq \beta$ and $|\widehat{Z}| \leq G + 1$, where $G = \lceil \sqrt{2\lceil \log_2(4/\beta) \rceil} \rceil$. To see this, couple a uniform U to the midpoint of a p-bit dyadic cell. On $[-G-1, G+1]$ the normal density is at least the rational $\ell = 2^{-4(G+1)^2}$. Choose $2^{-p} \leq h\ell/100$. Certified CDF bisection with absolute CDF error $h\ell/100$ finds the quantile within h on $|Z| \leq G$; if an intermediate CDF interval contains the target, return that point with its certified inverse-error bound. Endpoints are clipped. Gaussian tails bound the exceptional probability. The integral of the series for $e^{-x^2/2}$, with $O(G^2 + q)$ terms and $q + O(G^2)$ guard bits, gives CDF accuracy 2^{-q} by the factorial remainder bound. Its normalization constant is computable to the same accuracy by rational square roots and, for example, the convergent Machin arctangent formula for π. Hence the bit cost is polynomial in $\log(1/h) + \log(1/\beta)$ and input encoding lengths. Independent calls have independent couplings.

Reserve in advance all original-noise and artificial-normal coordinates allowed by the deterministic caps, including unused coordinates. There are at most N after enlarging it by a fixed power. With $G = \lceil \sqrt{2\lceil \log_2(100N/\delta) \rceil} \rceil$, all exact coordinates lie in this box except with probability $\delta/50$. Couple artificial sources with total failure at most $\delta/50$. The local routine guards incoming centered observations at a rational upper bound for $R + \sqrt{d}(G+1) + 1$ and returns the fallback if this bound is exceeded. On the coupled good event, sufficiently fine quantization passes these guards. Every later local query lies in a box of polynomial radius in A_R.

Structured evaluation bounds. A sequential order-r contraction expands into at most $N^{O(r)}$ products of $O(r)$ factors, although it is evaluated by its polynomial-size recurrence. The factors consist of input coordinates and entries of the stored frames and cores. On the source box, input coordinates have polynomial magnitude, and frame entries have magnitude at most one. A traced core remains PSD and its trace cannot increase, because $\sum_{j \leq d} V_{\ell,j} V_{\ell,j}^T \preceq I$; its initial trace is at most $\sqrt{k}\, k(2^r + 1) + 1$. Thus all core entries have polynomial magnitude on every outcome. Window coefficients and factorial-normalized factors have the bounds in Section A.4. A derivative of order at most two multiplies coefficient sums by at most the square of the contraction order. Consequently, by fixed-factor enlargement of t_* and fixed-power enlargement of N, every polynomial value or derivative used in a fixed-frame evaluation has product length at most t_* and coefficient absolute sum at most N^{2t_*}. These enlargements preserve $t_* = O(\log A_R)$ and $N = A_R^{O(1)}$.

Lemma A.2 (Perturbation of a fixed-frame evaluation)**.** *Let $N \geq \max\{2, t\}$, $t \geq 2$, and*

$$F(x) = \sum_{\ell} c_\ell \prod_{j=1}^{r_\ell} x_{\ell j}, \qquad r_\ell \leq t, \qquad \sum_{\ell} |c_\ell| \leq N^{2t}.$$

If $|x_{\ell j}| \leq N$ and $|\widehat{x}_{\ell j} - x_{\ell j}| \leq h \leq 1$, then

$$|F(x)| \leq N^{3t}, \qquad |F(\widehat{x}) - F(x)| \leq N^{4t} h.$$

The factors need not be independent or distinct.

Proof. The value bound follows term by term. For one product, telescoping gives

$$\left| \prod_{j=1}^{r} \widehat{x}_j - \prod_{j=1}^{r} x_j \right| \leq \sum_{j=1}^{r} |\widehat{x}_j - x_j| \prod_{a<j} |\widehat{x}_a| \prod_{a>j} |x_a| \leq r h (N+1)^{r-1}.$$

Multiply by the coefficient absolute sum. Since $r \leq t \leq N$ and $N + 1 \leq 2N \leq N^2$, the result is at most $N^{2t} N (N^2)^{t-1} h \leq N^{4t} h$. □

Apply the lemma with $t = t_*$ and primitive tolerance $h = 2^{-q}$. Additions, scaling, and the approximated square roots in the source transformations give factor error at most $N^2 h$. The precision chosen below makes this at most one. The lemma therefore gives polynomial-evaluation error at most $N^{6t_*} h$, including vector dimensions and the fixed finite sums after another fixed enlargement of t_*. Fixed rational coefficients are evaluated exactly; if an elementary approximation occurs as a factor, its error is included in h.

The remaining operations have explicit stability bounds. If logits differ in sup norm by e, the softmax weights differ in ℓ^1 norm by at most $2e$. If $A, \widehat{A}$ are scalar numerators and $B, \widehat{B}$ are vector numerators, then, with $d = \max(A, \lambda)$ and $\widehat{d} = \max(\widehat{A}, \lambda)$,

$$\left\| \frac{B}{d} - \frac{\widehat{B}}{\widehat{d}} \right\| \leq \lambda^{-1} \left\| B - \widehat{B} \right\| + \lambda^{-2} \left\| \widehat{B} \right\| \, |A - \widehat{A}|.$$

Here $\lambda^{-1} \leq N$ and clipping the denominator is one-Lipschitz. Ball projection is nonexpansive; its rational approximation is assigned a separate error. In each comparison, the rational centers, frames, cores, and coefficient tensors are fixed and identical in the two evaluations; only source coordinates and elementary-function evaluations are perturbed. Combining these estimates gives the following error bounds, with errors measured in absolute value for scalars and Euclidean norm for vectors:

Quantity	Relevant bound	Error bound
Moment entries, scores, and polynomial evaluations	Lemma A.2	$N^{6t_*} h$
Softmax weights	ℓ^1 Lipschitz constant 2	$N^{8t_*} h$
Density and gradient numerators	$M \leq N$; local size $\leq N^{4t_*}$	$N^{14t_*} h$
Floored posterior ratio	$\lambda^{-1} \leq N$, $\left\|\widehat{B}\right\| \leq N^{5t_*}$	$N^{24t_*} h$

For example, a reference numerator is $M \sum_i \omega_i f_i$. Its perturbation is bounded by $M \max_i |\widehat{f}_i - f_i| + M \max_i |f_i| \, \|\widehat{\omega} - \omega\|_1$. The same inequality applies to vector numerators. Their bounds in

the table, followed by the displayed ratio inequality, give the last two rows. Certified exponential evaluation adds at most a polynomial multiple of h to the weight error; it is absorbed in the second row. Polynomially many additions, norm conversions, and replicas add fixed powers of N, and medians are one-Lipschitz in their scalar input values. Thus

$$K = N^{100t_*}, \qquad E_q = K2^{-q} \tag{60}$$

bounds the aggregate absolute perturbation in the fixed-frame statistics, reference logits, polynomial values and derivatives, floored ratios, and gate scores. The exponent leaves slack for the separately certified softmax and projection evaluations. Spectral frames are controlled by their residual guarantees and discrete sampling decisions by coupling, as detailed below. The bound in (60) follows from the product length of each evaluation. Polynomial contractions and sums are evaluated exactly on rational input approximations using common denominators, and dense arrays are rounded at the prescribed stages.

For gates, the preceding disagreement bound and a union bound give at most $4N^3E_q$ total discrepancy on the source box. Choose

$$q \geq \lceil \log_2(200/\delta) \rceil + (100t_* + 10)\lceil \log_2 N \rceil. \tag{61}$$

Then $E_q \leq \delta/(200N^{10})$. The matrix consequences can be checked before any spectral choice: a compressed matrix has at most kD rows, so entry error E_q gives Frobenius error at most $kDE_q \leq z_0/100$. The rational Jacobi residual and the statistical matrix error have separate allocations in Section A.3. The resulting low-rank approximation therefore satisfies the compression bound with its prescribed slack. The restricted partial traces are evaluated exactly on the rational cores, with error controlled by the dimension-independent nuclear-norm bound above. The coefficient sum W_1A_1 then propagates that error through the explicitly chosen η to the allocated density error.

For a fresh posterior query on the coupled source box, (60) bounds the displacement from the exact comparison evaluation with the same stored coefficients. Taking the certified projection error at most $1/4$ gives a total center displacement below one, as allowed in Section A.2. The choices $N \geq 100/\xi$ and $N \geq 100S_0^2/(\alpha\sigma^2)$ also leave a fixed-factor margin in the scalar and vector accuracy allocations preceding (59). Gate decisions are treated separately: $4N^3E_q \leq \delta/50$. These implications are deterministic on the coupling and gate-agreement event. Fresh-sample concentration is applied before conditioning on that event. Since $t_* = O(\log A_R)$ and $\log N = O(\log A_R)$, the precision is

$$q = O((\log A_R)^2 + \log(1/\delta)).$$

Statistical transfer and bit lengths. For each new batch, define exact comparison observations using the unrounded underlying i.i.d. sample, exact fresh auxiliary randomness, and the *actually stored rational* centers and frames from previous batches. Conditional on the earlier finite transcript, the comparison statistics obey (53)–(54); concentration applies before conditioning on source-box or gate-agreement events. On their intersection with those separately charged events, implemented statistics differ by at most (60). Median-of-means is one-Lipschitz in its block values. The allocated numerical errors therefore preserve (55), (57), and (59) with constant slack. Every candidate remains exactly in the prescribed ball.

Each new moment matrix is rounded to a fixed dyadic mesh finer than $z_0/(100kD)$ before rational Jacobi. All frames consequently have polynomial encoding length, independently of the denominators used in earlier computations. A sequential chain has $O(\log A_R)$ such frames; a trace step is linear in the old core. Exact polynomial contractions have polynomially many products of these bounded-length rationals and polynomially many additions. Common denominators across a

polynomial-size list have at most the sum of the individual bit lengths, hence remain polynomial. Polynomial coefficient construction, finite-replica averaging, and the rational projection likewise have polynomial bit length. Centers are rounded to a sufficiently fine fixed dyadic grid and, if needed, projected inward by the certified construction; the resulting subunit change is included in the cover-radius slack. These rounding operations ensure polynomial encoding lengths across all scales.

The failure probabilities for the source bounds, source coupling, gate agreement, and concentration estimates are assigned separate constant fractions of δ, with all analytic accuracy targets tightened by a fixed factor. The sum can be made at most $\delta/2$. All statistical calls use fresh samples; their guarantees hold conditional on the preceding transcript, before conditioning on subsequent coupling events. The operation caps and guards prove the resource bound on every outcome. Thus the bounded-radius learner is implementable from q-bit observations with $\mathrm{poly}(A_R, L)$ training cost and a rational stored representation satisfying its analytic guarantees. The next subsection converts this rational representation into a discrete sampler with the smoothed total variation guarantee.

A.6 Discrete rejection and Gaussian smoothing

Fix a successful rational representation. Tighten the analytic training and evaluation error budgets so its ideal continuous capped rejection sampler has law H with $\mathrm{TV}(H, P) \le \varepsilon/4$. Its proposal and evaluator counts are fixed polynomials. Couple all their normals to finite-bit draws with compact-box failure at most $\varepsilon/100$. Couple a uniform center choice by $\lfloor MU_p \rfloor$, where U_p is a p-bit midpoint; discrepancy from uniform is at most $M2^{-p}$. The acceptance probability is a clipped reference-normalized evaluator, so its absolute error obeys (60) on this box. A fresh independent uniform draw for acceptance changes the decision with probability at most that error plus its grid width, conditional on agreement at all previous steps. A union bound over the fixed proposal cap makes all categorical and acceptance choices agree except with probability $\varepsilon/20$. This requires only $O((\log A_R)^2 + \log(1/\varepsilon))$ fractional bits after increasing constants. The same treatment applies to the fallback branch.

Choose coordinate accuracy so the matched output vectors differ by at most $\kappa = \tau\varepsilon/20$. This adds $O(\log(d/(\tau\varepsilon))) = O(\log A_R)$ bits. The resulting finite-bit rational program has law Q and admits a coupling $Z \sim Q$, $X \sim H$ with $\mathbb{P}(\|Z - X\| > \kappa) \le \varepsilon/20$.

For any such coupling with exceptional probability β,

$$\mathrm{TV}(\mathcal{L}(Z) * \mathcal{N}(0, \tau^2 I), \mathcal{L}(X) * \mathcal{N}(0, \tau^2 I)) \le \beta + \kappa/(2\tau),$$

by Pinsker for equal-covariance Gaussians and averaging over the coupling. Moreover,

$$\mathrm{TV}(P * \mathcal{N}(0, \tau^2 I), P) \le \tau^2 \sqrt{d/8},$$

because the componentwise KL is $\frac{d}{2}[\tau^2 - \log(1 + \tau^2)] \le d\tau^4/4$. Contraction under convolution and the triangle inequality give

$$\mathrm{TV}(Q * \mathcal{N}(0, \tau^2 I), P) \le \varepsilon/4 + \varepsilon/20 + \varepsilon/40 + \tau^2 \sqrt{d/8} < \varepsilon/2.$$

The remaining error budget is reserved for estimating group weights and sampling a group. The Gaussian perturbation is applied only once, to the final output. Training is analyzed separately through the exact comparison observations and the quantization bounds established above.

A.7 Removing the radius assumption with quantized pilot samples

Quantized pilot samples reduce the arbitrary-radius problem to bounded-radius groups, with R a fixed polynomial in the global A. Choose a universal integer H_* below and set $L_* = H_*\overline{A}$, $R_{\mathrm{loc}} = 7kL_*$. Draw $n_0 = \lceil \alpha^{-1} \log(8k/\delta) \rceil$ pilots, using a computable rational upper bound for this count. Form the graph on their actual quantized values with edges between points at distance at most $4L_*$. Compare rational squared distances exactly. Abort if it has more than k components; otherwise select one quantized anchor x_s from each component.

Every component of positive weight is represented among the pilot samples except with probability $\delta/8$. On the event that each original Gaussian noise in a predetermined raw-sample prefix has norm at most $L_*/2$, and $\sqrt{d}\,2^{-b} \le 1 \le L_*/2$, each quantized observation is within L_* of its latent mean. Pilots from the same latent component form a clique. Thus graph components partition whole represented latent components. An edge connects means at distance at most $6L_*$, so a simple path on at most k latent labels bounds every group mean's distance to its anchor by $[6(k-1)+1]L_* \le R_{\mathrm{loc}}$.

A later quantized observation from group s is within $2L_*$ of some pilot of its own component. Every pilot outside group s is more than $4L_*$ from that pilot, hence more than $2L_*$ from the observation. Consequently nearest-pilot routing is exactly the latent-group routing on this event. Including equality in the edge threshold makes the intergroup inequality strict under exact rational comparisons.

Let n_* be a deterministic polynomial sample upper bound for the finite-bit bounded-radius learner at R_{loc}, group accuracy $\varepsilon/3$, and group confidence $\delta/(8k)$. Reserve k separate group blocks of

$$N_{\mathrm{grp}} = \left\lceil 8\alpha^{-1}[n_* + \log(8k/\delta)] \right\rceil$$

raw observations and one weight block of $n_{\mathrm{wt}} = \lceil 18k^2\varepsilon^{-2} \log(16k/\delta) \rceil$. Each group uses only its dedicated block, taking its first n_* routed observations and aborting if there are too few. Under ideal latent routing, the group's weight $p_s \ge \alpha$, its normalized component weights are at least α, and a Chernoff bound shows that the probability of too few acceptances is at most $\delta/(8k)$. Extend each block by an independent infinite stream only in the analysis. Before conditioning on the bounded-noise event, the first n_* latent-group acceptances are i.i.d. from the conditional group mixture.

Subtract the actual quantized anchor exactly. Its lattice membership implies

$$Q_b(X) - x_s = Q_b(X - x_s).$$

Conditioned on the pilots, the ideal centered observations therefore have exactly the form of quantized input required by the local routine, with means in the radius-R_{loc} ball. Couple actual and ideal execution on the same finite prefix and charge routing disagreements separately. The local theorem thus applies to the unconditioned ideal stream; the bounded-noise event is used only to couple its routing to the implemented procedure.

The prefix length is $N_{max} = n_0 + kN_{\mathrm{grp}} + n_{\mathrm{wt}}$. There exist fixed $C_*, p_*, D_* \ge 1$ independent of H_* and the input such that

$$N_{max} \le C_* H_*^{p_*} \overline{A}^{D_*}.$$

To see this, fix the universal polynomial sample bound of the already constructed bounded-radius algorithm before substituting $R_{\mathrm{loc}} = 7kH_*\overline{A}$. The substitution introduces a fixed power of H_* and changes the fixed exponent of $\overline{A}$. Gaussian exponential integration and Markov's inequality give, for sufficiently large H_*,

$$\mathbb{P}(\|G\| > L_*/2) \le e^{-H_*^2\overline{A}^2/32}.$$

Choose H_* once so $H_*^2/32 \geq \log(100C_*) + p_* \log H_* + D_* + 3$. Since $\log(1/\delta) \leq \overline{A}$ and $\log \overline{A} \leq \overline{A}$, the union bound over the prefix is then at most $\delta/20$. The radius is therefore fixed independently of the runtime on a particular input.

The fresh weight block supplies rational empirical frequencies $\widehat{p}_s$ summing to one. Hoeffding's inequality and ideal routing bound their total ℓ^1 error by, for example, $\varepsilon/5$ with the allocated confidence, after increasing the sample count by a constant factor. A final program chooses a group by bounded-bit inversion of these rational frequencies, invokes its rational sampler, and adds its anchor exactly. With p random bits, group-choice discrepancy is at most $2k2^{-p}$, which is assigned $\varepsilon/20$. All groups use the same global $\tau = \varepsilon/(100d)$ in Section A.6; its argument works for any specified inverse-polynomial τ. Per-group accuracy can therefore be tightened without changing the final smoothing variance. Convexity yields total smoothed-TV error at most $\varepsilon/3 + \varepsilon/10 + \varepsilon/20 < \varepsilon$, with sufficient margin for the preceding constant-factor error allocations.

Allocate the training failure budget across pilot representation, finite-prefix noise, ideal group counts, group learning and numerical events, and weight estimation. Their prescribed bounds sum to at most δ after fixed-factor tightening. Since $A_{R_{\mathrm{loc}}} \leq A^{O(1)}$ and $\log(k/\delta) = O(\log A + \log(1/\delta))$, the local precision (61) is bounded by

$$C[(\lceil \log_2 \overline{A} \rceil)^2 + \lceil \log_2(1/\delta) \rceil + 1]$$

for a sufficiently large universal C. Read all observations at this one precision, rounding finer observations down if necessary. Exact subtraction before local guards prevents remote mean magnitudes from entering fractional precision. The cost of reading, comparing, and adding these large coordinates is charged to their actual encoding length L. All other quantities have polynomial bit length in A; every execution, including an unsuccessful one, obeys a deterministic cap. This proves Theorem A.1.

AI methodology

OpenAI Codex assisted with consolidating the theoretical arguments, revising the English exposition, developing proof details and checking their consistency, LaTeX preparation, and verification of selected primary references. This assistance affected the introduction, proof exposition, and bibliography. The authors retain responsibility for all claims and the final manuscript.

References

[ABH+20] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of Gaussian mixtures via compression schemes. *Journal of the ACM*, 67(6), Article 32, 1–42, 2020. doi:10.1145/3417994.

[AJOS14] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Near-optimal-sample estimators for spherical Gaussian mixtures. *Advances in Neural Information Processing Systems*, 27, 2014. https://proceedings.neurips.cc/paper_files/paper/2014/hash/9d4f24b6cc1070a2e533a90309863878-Abstract.html.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi:10.1093/acprof:oso/9780199535255.001.0001.

[DK20] Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 184–195, 2020. doi:10.1109/FOCS46700.2020.00026. Full version: arXiv:2012.07774v1, 2020.

[DK25] Ilias Diakonikolas and Daniel M. Kane. Implicit high-order moment tensor estimation and learning latent variable models. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1228–1247, 2025. doi:10.1109/FOCS63196.2025.00064. Full version: arXiv:2411.15669v2, 2025.

[DKP21] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020. Revised version: arXiv:2007.15618v2, 2021.

[Efron11] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi:10.1198/jasa.2011.tm11181.

[GKL25] Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of Gaussians using diffusion models. In *Proceedings of the Thirty Eighth Conference on Learning Theory*, PMLR 291:2403–2456, 2025. https://proceedings.mlr.press/v291/gatmiry25b.html. Full version: arXiv:2404.18869v2, 2025.

[HC26] Hengzhi He and Guang Cheng. Sharp proper estimation of fixed-component Gaussian location mixtures in polynomial time. arXiv:2608.12701v1, 2026.

[Jaggi13] Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, PMLR 28(1):427–435, 2013. https://proceedings.mlr.press/v28/jaggi13.html.

[LL22] Allen Liu and Jerry Li. Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1248–1261, 2022. doi:10.1145/3519935.3520012. Full version: arXiv:2112.00706v1.

[OD14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi:10.1017/CBO9781139814782. Author's updated text: arXiv:2105.10386, 2021.

[RV17] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated Gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96, 2017. doi:10.1109/FOCS.2017.17.