
Gaussian Sharp Mean Estimation under Finite Covariance in Fixed Dimension

Anonymous Author
Anonymous Institution

Abstract

We study estimation of the mean of an arbitrary Borel probability distribution on $\mathbb{R}^d$, assuming only that its covariance Σ is finite. For every fixed dimension d , we construct a Borel measurable estimator that may depend on the target confidence level. Its Euclidean error matches the Gaussian sample-mean benchmark up to a relative factor $1 + o(1)$ in the high-confidence regime. Writing $L = \log(1/\delta)$ and letting $\mathcal{Q}_\delta(\Sigma)$ denote the upper δ -quantile of $\|Z\|$ for $Z \sim \mathcal{N}(0, \Sigma)$, we prove

$$\mathbb{P}\left\{\|\hat{\mu} - \mu\| > \left[1 + O_d\left(\frac{\log L}{L} + \frac{L}{n}\right)\right] \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}}\right\} \leq \delta$$

whenever L is sufficiently large and L/n is sufficiently small. The guarantee holds uniformly over all distributions with finite covariance, including those whose covariance Σ is singular or arbitrarily ill-conditioned, and the estimator requires no knowledge of Σ . For Gaussian observations, $\mathcal{Q}_\delta(\Sigma)/\sqrt{n}$ is the exact confidence radius of the sample mean, and a lower bound based on two Gaussian location models shows that this benchmark is optimal to first order. Our estimator is computable in polynomial time for every fixed d , using $n^{O(d^5)}$ arithmetic operations and comparisons.

1 INTRODUCTION

Estimating the mean of a distribution from independent observations is one of the fundamental problems in statistics. Suppose that n independent observations are drawn from a distribution with finite covariance matrix Σ . The sample mean has covariance Σ/n , but its tails can be much heavier than in the Gaussian case, because a few extreme observations may dominate the average. Robust mean estimators overcome this difficulty and achieve deviation bounds with optimal dependence on sample size, confidence level, and covariance, up to universal constants, without assumptions beyond finite covariance [25, 27]. This leads to a sharper question: *under the sole assumption of finite covariance, can a robust mean estimator match the leading constant in the Gaussian confidence radius, or is a strictly larger constant unavoidable in the presence of heavy tails?*

Let $X_1, \dots, X_n$ be independent observations from a distribution on $\mathbb{R}^d$ with mean μ and covariance matrix $\Sigma \succeq 0$, and let $\delta \in (0, 1)$ be the target failure probability. We take the Gaussian confidence radius associated with Σ as our benchmark and define

$$\mathcal{Q}_\delta(\Sigma) := \inf \{r \geq 0 : \mathbb{P}_{Z \sim \mathcal{N}(0, \Sigma)}(\|Z\| > r) \leq \delta\}.$$

If the observations are Gaussian, then the sample mean $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ satisfies $\sqrt{n}(\bar{X}_n - \mu) \sim \mathcal{N}(0, \Sigma)$, and hence has Euclidean confidence radius $\mathcal{Q}_\delta(\Sigma)/\sqrt{n}$. We ask whether, without knowing Σ , one can construct a robust estimator that attains this Gaussian benchmark to first relative order uniformly over distributions with finite covariance.

We study this question in an asymptotic regime where the dimension d is fixed, while $L := \log(1/\delta) \rightarrow \infty$ and $L/n \rightarrow 0$. Thus the confidence level tends to one, while the logarithmic confidence parameter remains small relative to the sample size. The uniformity over distributions is essential: both the underlying distribution and its covariance matrix may vary with n and δ . Consequently, pointwise asymptotic normality for a fixed distribution does not establish the required

uniform guarantee. In dimensions $d > 1$, a further difficulty arises because sharp control of every projection need not yield a single Euclidean estimate at the same radius.

Suppose that the data determine a set $\mathcal{A} \subset \mathbb{R}^d$ of plausible values of the mean and let $r > 0$ denote the desired estimation error. Sharp control of the projections in every direction may imply that

$$\text{diam}(\mathcal{A}) := \sup_{\mu_1, \mu_2 \in \mathcal{A}} \|\mu_1 - \mu_2\| \leq 2r.$$

This only bounds the distance between pairs of plausible means. To construct an estimator with error at most r , one must instead find a point $\hat{\mu}$ whose distance from every element of $\mathcal{A}$ is at most r , that is, $\sup_{\mu' \in \mathcal{A}} \|\hat{\mu} - \mu'\| \leq r$. The diameter bound alone does not guarantee the existence of such a point. Indeed, if $\mathcal{A}$ consists of the vertices of a regular simplex with edge length $2r$, then

$$\inf_{x \in \mathbb{R}^d} \sup_{\mu' \in \mathcal{A}} \|x - \mu'\| = r \sqrt{\frac{2d}{d+1}} > r.$$

Thus, even if the directional information is sharp enough to control the distance between every two plausible means at the Gaussian scale, converting these constraints into a single Euclidean estimate can introduce a fixed geometric loss. This obstruction is closely related to the geometry studied by Gupta, Hopkins and Price [13]. Our construction avoids this loss by retaining the joint constraints on the proposed means and covariance bounds, rather than reducing them to a diameter bound, until the final center is selected.

1.1 Main results

For distributions with finite covariance, the optimal deviation scale is already known up to universal constants. In particular, robust mean estimators achieve an error of order $\sqrt{\text{tr}(\Sigma)/n} + \sqrt{\|\Sigma\|_{\text{op}} L/n}$ with probability at least $1 - \delta$ [25]. In our regime, d is fixed and $L \rightarrow \infty$. Since $\text{tr}(\Sigma) \leq d \|\Sigma\|_{\text{op}}$, the first term is at most $\sqrt{d/L}$ times the second. Since $\sqrt{d/L} = o(1)$, the leading scale is $\sqrt{\|\Sigma\|_{\text{op}} L/n}$.

The Gaussian benchmark identifies the correct leading coefficient on this scale, uniformly over all nonzero $\Sigma \succeq 0$,

$$\mathcal{Q}_\delta(\Sigma) = \sqrt{2\|\Sigma\|_{\text{op}} L} \left[1 + O_d\left(\frac{\log L}{L}\right) \right].$$

Thus the remaining question is whether the known rate can be attained without any fixed multiplicative loss in its leading constant. Equivalently, in the high-confidence regime the target coefficient in front

of $\sqrt{\|\Sigma\|_{\text{op}} L/n}$ is $\sqrt{2}$. Such a constant matters even asymptotically, because its square enters the leading sample size required to achieve a prescribed accuracy. The following theorem attains this coefficient in every fixed dimension under finite covariance using an estimator determined by the sample and confidence level.

Theorem 1 (Gaussian sharp mean estimation). *Fix an integer $d \geq 1$. There exist constants $C_d < \infty$, $e \leq L_d < \infty$, and $c_d > 0$, depending only on d , such that the following holds. For every $n \geq 1$ and $0 < \delta < 1$, put $L = \log(1/\delta)$. If $L \geq L_d$ and $L/n \leq c_d$, then there exists a Borel measurable estimator $\hat{\mu}_{n,\delta}^{(d)}$ such that, for every Borel probability distribution on $\mathbb{R}^d$ with mean μ and finite covariance $\Sigma \succeq 0$, and for n independent observations drawn from that distribution,*

$$\mathbb{P} \left\{ \left\| \hat{\mu}_{n,\delta}^{(d)} - \mu \right\| > \left[1 + C_d \left(\frac{\log L}{L} + \frac{L}{n} \right) \right] \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}} \right\} \leq \delta.$$

It follows that, whenever d is fixed, $L \rightarrow \infty$, and $L/n \rightarrow 0$, the confidence radius is $(1 + o(1))\mathcal{Q}_\delta(\Sigma)/\sqrt{n}$ uniformly over the entire class of distributions with finite covariance. This includes sequences of covariance matrices approaching singularity or becoming increasingly ill-conditioned.

Theorem 1 characterizes the first-order radius in the high-confidence regime. The second-order contribution of the spectrum of Σ and the exact Gaussian radius at fixed δ remain open. Our second result shows that the same estimator admits a polynomial-time implementation for fixed d .

Theorem 2 (Computational complexity). *Fix $d \geq 1$. In the regime of Theorem 1, there exists an algorithm that computes $\hat{\mu}_{n,\delta}^{(d)}$ with arithmetic complexity $n^{O(d^5)}$.*

Algorithm 1 achieves this bound. The computational model underlying this arithmetic complexity bound is specified in Appendix M.

From a statistical perspective, the guarantee is sharp to first order. Proposition G.1 shows that even for a Gaussian location family with a fixed covariance matrix Σ , any estimator whose confidence guarantee holds uniformly over $\mu \in \mathbb{R}^d$ must have radius at least $\sqrt{\|\Sigma\|_{\text{op}}/n} \Phi^{-1}(1 - \delta)$, where Φ denotes the standard normal distribution function. Since $\Phi^{-1}(1 - \delta) = \sqrt{2L}(1 + o(1))$ and $\mathcal{Q}_\delta(\Sigma) = \sqrt{2\|\Sigma\|_{\text{op}} L}(1 + o(1))$ for fixed d , this lower bound is $(1 - o(1))\mathcal{Q}_\delta(\Sigma)/\sqrt{n}$. Together with the upper bound $(1 + o(1))\mathcal{Q}_\delta(\Sigma)/\sqrt{n}$ from Theorem 1, this shows that the Gaussian bench-

mark is attained with the optimal leading constant in the stated regime.

1.2 Related work

The literature on robust mean estimation under finite covariance includes median of means methods [34], geometric aggregation [16, 29], influence function estimators [5], and PAC-Bayesian methods [6]. Devroye, Lerasle, Lugosi and Oliveira [11] distinguish estimators designed for a prescribed confidence level from those required to work simultaneously over a range of confidence levels; our estimator is allowed to depend on δ .

For vectors with finite covariance, Lugosi and Mendelson [25] established the optimal deviation scale up to universal constants, with extensions to general norms in [26]. Subsequent work developed polynomial-time procedures at the same statistical scale, including Hopkins [15], Cherapanamjeri, Flammarion and Bartlett [7], and Lei, Luh, Venkat and Zhang [24]. These results determine the correct order of the error and resolve much of the computational problem. The present work addresses the finer question of the leading confidence constant, showing that the Gaussian benchmark can be attained to first order.

In dimension one, Catoni [5] obtains the Gaussian leading constant when an upper bound on the variance is known, while Lee and Valiant [21] remove this prior knowledge under only finite variance. In higher dimensions, an additional geometric obstruction arises: sharp scalar estimates in every direction do not automatically yield a single Euclidean center with the same radius. Related refinements address robustness, distribution-dependent guarantees, self-normalization, and sharper deviations under additional structure [10, 23, 30–32].

Lee and Valiant [22] obtain a complementary sharp result when the effective dimension is large relative to the confidence parameter. Writing $d_{\text{eff}} = \text{tr}(\Sigma)/\|\Sigma\|_{\text{op}}$, their sample complexity is optimal up to a factor $1 + o(1)$ when $d_{\text{eff}}/\log^2(1/\delta) \rightarrow \infty$. Since $d_{\text{eff}} \leq d$, this regime excludes fixed d as $\delta \rightarrow 0$. Their result and Theorem 1 therefore address different asymptotic balances between dimension and confidence.

The closest comparison from the geometric viewpoint is Gupta, Hopkins and Price [13]. Under a known bound $\Sigma \preceq \sigma^2 I_d$, their Theorem 1 gives radius $(1 - \tau)J_d\sigma\sqrt{2L/n}$ for $d \geq 2$ and $n \geq CL \geq C^2d$, where $J_d = \sqrt{2d/(d+1)}$ and $\tau, C > 0$ are universal constants. Thus, their result improves the Jung coefficient by a constant factor. However, the resulting co-

efficient is independent of n and δ , and therefore does not converge to one as the sample size grows and the confidence level increases while the dimension remains constant. Theorem 1 addresses this regime without assuming prior knowledge of the covariance. The constants in our result may depend on d , and we make no claim when the dimension grows with the sample size. Gupta, Hopkins, and Price also establish a Jung-type lower bound for adversarial contamination in the population limit, but their result is established under a different statistical model.

Additional related work and the detailed comparison are given in Appendix A.

To the best of our knowledge, this is the first multivariate mean estimator for fixed $d \geq 2$ that matches the Gaussian confidence radius up to a $1 + o(1)$ factor under only a finite covariance assumption, as $\log(1/\delta) \rightarrow \infty$ with $\log(1/\delta) = o(n)$, without requiring prior knowledge of the covariance.

1.3 Construction outline

We say that a pair (m, V) with $m \in \mathbb{R}^d$ and $V \succeq 0$ is accepted if the quadratic likelihood $\mathcal{L}_n(m, V)$ defined in (1) satisfies $\mathcal{L}_n(m, V) \leq T$, where $T = L + O_d(\log L)$. To each accepted pair, we attach a ball of radius approximately $\sqrt{2\|V\|_{\text{op}}T/n}$. Lemma 2.4 ensures that (μ, Σ) is accepted with probability at least $1 - \delta$. It therefore remains to show that the balls associated with all accepted pairs have a common point, even when their scales differ or vanish.

Lemma 2.1 associates each accepted pair with a distribution having the proposed mean and covariance bounded by the proposed matrix, while Corollary 2.3 shows that the multivariate Chernoff divergence of the distributions associated with any finite family of accepted pairs is at most T/n . If the corresponding balls failed to intersect, a common Gaussian convolution can be chosen so that their squared normalized common radius lies in the range of Lemma 3.2. That lemma then gives a lower bound on the Chernoff divergence that is strictly larger than T/n , yielding a contradiction. Helly’s theorem gives the intersection property for every finite family of accepted pairs. The pair formed by the empirical mean and empirical covariance is always accepted, and its associated ball is compact; compactness therefore extends the finite intersection property to the entire family. The estimator is the minimum-norm point of this intersection. Whenever (μ, Σ) is accepted, the intersection is contained in its associated ball, which yields the Gaussian radius.

Sections 2–3 establish the likelihood, information, and geometric bounds used above. Section 4 defines the

intersection and the resulting estimator and proves Theorem 1. Section 5 gives a polynomial characterization of rejection and proves Theorem 2. The auxiliary proofs and detailed algorithmic analysis are deferred to the appendices.

2 LIKELIHOOD AND CHERNOFF DIVERGENCE BOUNDS

A candidate is a pair (m, V) with $m \in \mathbb{R}^d$ and $V \succeq 0$, interpreted as a proposed mean and a covariance upper bound. We first define a statistic for testing candidates and then show how the resulting relative entropy representations yield a multivariate Chernoff divergence bound for any finite family of accepted candidates. Throughout, vector norms are Euclidean, matrix norms are operator norms, and logarithms are natural.

2.1 Quadratic likelihood and relative entropy

Write $\hat{P}_n = n^{-1} \sum_i \delta_{X_i}$ and let $\mathcal{M}(m, V)$ be the Borel probability measures with mean m and covariance bounded by $V \succeq 0$. Every such distribution is supported on $m + \text{Range}(V)$: the variance in each null direction of V is zero. If any observation is outside this affine space, set $\mathcal{L}_n(m, V) = +\infty$.

Otherwise, write $V = U\Lambda U^\top$ on its range, where $U^\top U = I_r$, $r = \text{rank}(V)$, and $\Lambda \succ 0$. Define $y_i = (\Lambda^{-1/2} U^\top (X_i - m))$ and set

$$\mathcal{L}_n(m, V) = \sup_{\substack{A \succeq 0 \\ \text{tr } A = 1}} \sum_{i=1}^n \log(y_i^\top A y_i). \quad (1)$$

The matrix A has size $(r+1) \times (r+1)$, and we use the convention $\log 0 = -\infty$. If $V = 0$ and all observations equal m , take $y_i = 1$, so the statistic is zero. For a threshold $T \geq 0$, a candidate (m, V) is accepted if $\mathcal{L}_n(m, V) \leq T$. The constant coordinate includes affine as well as quadratic terms, while the trace constraint normalizes their expectations under the proposed moments.

Lemma 2.1 (Relative entropy representation). *If all observations lie in $m + \text{Range}(V)$, then*

$$\frac{\mathcal{L}_n(m, V)}{n} = \min_{P \in \mathcal{M}(m, V)} D_{\text{KL}}(\hat{P}_n \| P). \quad (2)$$

The minimum is attained, including when V is singular.

The relative entropy is defined by $D_{\text{KL}}(Q \| P) = \int \log(dQ/dP) dQ$ if $Q \ll P$, and is infinite otherwise. The minimization over the second argument is a reverse information projection problem; see, for example, [19]. The proof in Appendix B also shows why

the representing distribution may place residual mass outside the observations. Thus, if a candidate (m, V) is accepted at threshold T , Lemma 2.1 supplies a distribution $P \in \mathcal{M}(m, V)$ such that $D_{\text{KL}}(\hat{P}_n \| P) \leq T/n$.

2.2 Multivariate Chernoff divergence

For $P_1, \dots, P_N$, choose a common dominating probability measure ν and write $p_i = dP_i/d\nu$. For weights $\beta \in \Delta_N := \{\beta_i \geq 0 : \sum_i \beta_i = 1\}$, let $I_\beta := \{i \in \{1, \dots, N\} : \beta_i > 0\}$ denote the support of β , and define

$$\mathcal{W}_\beta = \int \prod_{i \in I_\beta} p_i^{\beta_i} d\nu, \\ C^*(P_1, \dots, P_N) = \sup_{\beta \in \Delta_N} (-\log \mathcal{W}_\beta).$$

This multivariate Chernoff divergence [33] reduces to the usual Chernoff information for two distributions [9]. Zero weights omit the corresponding factors and permit comparison of any subfamily.

Lemma 2.2 (Relative entropy upper bound for Chernoff divergence). *If $D_{\text{KL}}(Q \| P_i) \leq s < \infty$ for every i , then*

$$C^*(P_1, \dots, P_N) \leq s.$$

Corollary 2.3 (Chernoff divergence bound for accepted candidates). *If $\mathcal{L}_n(m_i, V_i) \leq T$ for $1 \leq i \leq N$, there are $P_i \in \mathcal{M}(m_i, V_i)$ with*

$$D_{\text{KL}}(\hat{P}_n \| P_i) \leq T/n, \quad C^*(P_1, \dots, P_N) \leq T/n.$$

The proofs are in Appendix B.3. For each fixed sample, the corollary applies deterministically to every finite family of accepted candidates. Accordingly, the probability estimate in the main proof concerns the acceptance of (μ, Σ) .

2.3 Calibration under finite covariance

For a fixed $A \succeq 0$ with $\text{tr } A = 1$ and independent Y_i satisfying $\mathbb{E}[Y_i Y_i^\top] = I_k$, the product $\prod_i Y_i^\top A Y_i$ has expectation one. The statistic instead chooses A after seeing the sample. The following lemma controls the cost of that optimization.

Lemma 2.4 (Quadratic likelihood maximal inequality). *Fix $d \geq 1$. There is a positive integer a_d , depending only on d , such that for independent $Y_1, \dots, Y_n \in \mathbb{R}^k$, $1 \leq k \leq d+1$, with $\mathbb{E}[Y_i Y_i^\top] = I_k$, and every $u \geq e$,*

$$\mathbb{P} \left\{ \sup_{\substack{A \succeq 0 \\ \text{tr } A = 1}} \sum_{i=1}^n \log(Y_i^\top A Y_i) > u + a_d \log u \right\} \leq e^{-u}.$$

The proof is given in Appendix C. Taking $u = L$ gives the acceptance threshold $T_d = L + a_d \log L$ under only finite covariance. Its $O_d(\log L)$ correction is independent of n , which allows L to diverge arbitrarily slowly relative to the sample size. Section 4.1 applies the lemma to the pair (μ, Σ) , including singular covariance matrices.

3 A GEOMETRIC LOWER BOUND FOR MULTIVARIATE CHERNOFF DIVERGENCE

For $1 \leq i \leq N$, let P_i be a distribution with mean μ_i and covariance matrix Σ_i , and set $\lambda_i := \|\Sigma_i\|_{\text{op}}$. Suppose that $\lambda_i > 0$ for every i . Define the squared normalized common radius

$$R = \min_{z \in \mathbb{R}^d} \max_i \frac{\|\mu_i - z\|^2}{\lambda_i}.$$

The minimum exists by continuity and coercivity. Equivalently, $\sqrt{R}$ is the smallest $\rho \geq 0$ such that the balls $\overline{B}(\mu_i, \rho\sqrt{\lambda_i})$, $1 \leq i \leq N$, have a common point. The proof first treats distributions that satisfy the required boundary and balance conditions at a minimizing center, then reduces every finite family to a subfamily with these properties.

Lemma 3.1 (Chernoff bound under boundary and balance conditions). *Let $N \geq 2$ and $d \geq 1$. Suppose $R > 0$ and that there exist a minimizing center z^* and weights $\gamma_1, \dots, \gamma_N > 0$, with $\sum_{i=1}^N \gamma_i = 1$, such that*

$$\|\mu_i - z^*\|^2 = R\lambda_i \quad (1 \leq i \leq N), \quad \sum_{i=1}^N \gamma_i(\mu_i - z^*) = 0.$$

If $R \leq \frac{1}{1024N^2}$, then

$$C^*(P_1, \dots, P_N) \geq \frac{R}{2} - 7000N^3R^2.$$

The bound is uniform over all ratios of covariance scales.

The boundary equalities identify the tight constraints, while the positive weights make the weighted displacements from the minimizing center sum to zero. The proof is in Appendix D. Together with the reduction in Corollary D.1, it gives the following form used by the estimator.

Lemma 3.2 (Multivariate Chernoff lower bound). *Let $P_1, \dots, P_N$ be Borel probability measures on $\mathbb{R}^d$ with finite second moments, means μ_i , covariance matrices Σ_i , and positive covariance scales $\lambda_i = \|\Sigma_i\|_{\text{op}}$. Define*

$$R = \min_{z \in \mathbb{R}^d} \max_{1 \leq i \leq N} \frac{\|\mu_i - z\|^2}{\lambda_i}.$$

Put $b_d = 7000(d+1)^3$. If $0 < R \leq \frac{1}{1024(d+1)^2}$, then

$$C^*(P_1, \dots, P_N) \geq \frac{R}{2} - b_d R^2.$$

The remainder is uniform over all ratios of covariance scales.

This bound follows from Lemma 3.1 and the reduction in Corollary D.1; the proofs are in Appendix D. Together with Corollary 2.3, it yields the key comparison: the distributions representing any finite family of accepted candidates have Chernoff divergence at most T/n , whereas a sufficiently large squared normalized common radius within the admissible range of Lemma 3.2 yields a lower bound exceeding this threshold. In Section 4.1, a common Gaussian convolution is applied to these representing distributions; its variance is chosen so that every covariance scale is positive and the squared normalized common radius lies within that admissible range.

4 THE ESTIMATOR AND STATISTICAL GUARANTEE

We first specify the geometric selector. Let $\kappa_d(s) = 2s(1 + 16b_d s)$, where $b_d = 7000(d+1)^3$. For an acceptance threshold T and a parameter $s \geq T/n$, define

$$\mathcal{C}(T, s) = \bigcap_{\mathcal{L}_n(m, V) \leq T} \overline{B}\left(m, \sqrt{\kappa_d(s)} \|V\|_{\text{op}}\right).$$

Whenever this set is nonempty, let $\hat{\mu}(T, s)$ denote its unique point of minimum norm.

For Theorem 1, let T_δ and s_δ be the deterministic thresholds constructed in Appendix L.4. There is a constant $\epsilon_d < \infty$, depending only on d , such that, with $T_d = L + a_d \log L$,

$$T_d \leq T_\delta \leq T_d + \epsilon_d, \quad T_\delta \leq ns_\delta \leq T_\delta + 1. \quad (3)$$

In particular, $s_\delta \geq T_\delta/n$ and $ns_\delta = L + O_d(\log L)$. We define $\hat{\mu}_{n,\delta}^{(d)} = \hat{\mu}(T_\delta, s_\delta)$.

Lemma 4.1 (Measurability of the geometric selector). *Fix d , n , and δ in the regime of Theorem 1. The map $(X_1, \dots, X_n) \mapsto \hat{\mu}(T_\delta, s_\delta)$ is Borel measurable.*

Proof. The set of constant samples is closed, and on this set the selector equals the common observation. For nonconstant samples, Appendix L.3 gives a first-order formula over the reals, with the sample coordinates as parameters, that defines $\mathcal{C}(T, s)$. By quantifier elimination [3], the graph of this correspondence is semialgebraic. Imposing the additional condition that $\|z\|^2$ be minimal over the corresponding fiber and

eliminating the auxiliary variables yields the graph of the unique minimum-norm selector. This graph is semialgebraic, and hence the selector is Borel measurable. $\square$

4.1 Proof of Theorem 1

The preceding sections supply the inputs to the construction. We prove that the intersection defining the estimator is nonempty for every sample and that its selected point has the required statistical guarantee. The likelihood and Chernoff bounds are used as stated; their proofs are in the appendices.

Proof. Fix d , n , and δ in the stated regime. All constants below depend only on d .

1. Fix the threshold and the radius. Set $T = T_\delta$ and $s = s_\delta$. By (3),

$$T \geq T_d, \quad T/n \leq s, \quad ns = L + O_d(\log L). \quad (4)$$

Enlarge L_d so that $ns \leq 2L$ whenever $L \geq L_d$, and decrease c_d so that $L/n \leq c_d$ implies

$$0 < s \leq s_0 := \min \left\{ \frac{1}{64b_d}, \frac{1}{4096(d+1)^2} \right\}. \quad (5)$$

For example, it suffices to take $c_d \leq s_0/2$ after fixing L_d . These are restrictions depending only on d .

Write $\kappa = \kappa_d(s)$ and abbreviate the associated ball by $B(m, V) = \overline{B}\left(m, \sqrt{\kappa \|V\|_{\text{op}}}\right)$. Two consequences of (5) will drive the argument. First, $\kappa = 2s(1 + 16b_d s) \leq \frac{5}{2}s < 1/[1024(d+1)^2]$. Second, the inflated radius makes the geometric lower bound strictly larger than s , the upper bound on the Chernoff divergence:

$$\begin{aligned} \frac{\kappa}{2} - b_d \kappa^2 - s &= b_d s^2 [16 - 4(1 + 16b_d s)^2] \\ &\geq \frac{39}{4} b_d s^2 > 0. \end{aligned} \quad (6)$$

The bound $16b_d s \leq 1/4$ follows from (5). The leading radius coefficient $2s$ comes from the coefficient $1/2$ in Lemma 3.2; its inflation supplies the strict gap needed for the contradiction.

2. Acceptance of (μ, Σ) . Let $r = \text{rank}(\Sigma)$. For every null direction a of Σ , $\mathbb{E}[(a^\top (X_i - \mu))^2] = a^\top \Sigma a = 0$. Applying this to a finite basis of the null space gives $X_i - \mu \in \text{Range}(\Sigma)$ almost surely. If $r > 0$, standardizing on this range and adjoining the constant coordinate produces independent vectors $Y_i \in \mathbb{R}^{r+1}$ with $\mathbb{E}[Y_i Y_i^\top] = I_{r+1}$. Lemma 2.4, applied with $u = L$, yields

$$\mathbb{P}\{\mathcal{L}_n(\mu, \Sigma) > T\} \leq \mathbb{P}\{\mathcal{L}_n(\mu, \Sigma) > T_d\} \leq \delta. \quad (7)$$

If $\Sigma = 0$, all observations equal μ almost surely and the statistic is zero, so the same conclusion holds. Let $\mathcal{E} = \{\mathcal{L}_n(\mu, \Sigma) \leq T\}$. Then $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. The geometric argument below is deterministic for each realized sample; the event $\mathcal{E}$ enters in Step 5, where $B(\mu, \Sigma)$ becomes one of the balls defining the intersection.

3. Intersections of at most $d+1$ balls. Fix the sample and a family of $N \leq d+1$ accepted candidates (m_i, V_i) . A single ball is nonempty, so assume $N \geq 2$. By Lemma 2.1, there are representing distributions $P_i \in \mathcal{M}(m_i, V_i)$ satisfying $D_{\text{KL}}(\hat{P}_n \| P_i) \leq T/n$. Corollary 2.3 therefore gives

$$C^*(P_1, \dots, P_N) \leq T/n \leq s. \quad (8)$$

Put $v_i = \|V_i\|_{\text{op}}$ and suppose, toward a contradiction, that the corresponding balls have empty intersection. Define

$$R_V = \inf_z \max_i \frac{\|m_i - z\|^2}{v_i}.$$

For $v_i = 0$, the ratio is zero at $z = m_i$ and infinite elsewhere. Its sublevel sets are precisely finite intersections of closed balls, and each finite sublevel set is compact. Thus the infimum is attained whenever it is finite. Empty intersection at level κ implies $R_V > \kappa$; equality cannot occur through an unattained infimum.

Let Σ_i denote the covariance matrix of P_i . The corresponding covariance scales satisfy $\lambda_i := \|\Sigma_i\|_{\text{op}} \leq v_i$. Consequently their squared common radius is at least R_V . Some λ_i may vanish, and this radius may exceed the local range of Lemma 3.2. Apply the same Gaussian convolution to every representing law and define

$$\begin{aligned} P_i^{(\tau)} &= P_i * \mathcal{N}(0, \tau I_d), \\ R(\tau) &= \inf_z \max_i \frac{\|m_i - z\|^2}{\lambda_i + \tau}. \end{aligned}$$

For $\tau > 0$, the means remain m_i and the covariance scales are $\lambda_i + \tau > 0$. Lemma F.1 gives continuity of $R(\tau)$ on $(0, \infty)$, its right limit $R(0) \geq R_V > \kappa$, and its limit zero at infinity. The intermediate value theorem therefore supplies $\tau > 0$ with $R(\tau) = \kappa$.

All hypotheses of Lemma 3.2 now hold, by the first inequality established in Step 1. Its lower bound and (6) imply

$$C^*(P_1^{(\tau)}, \dots, P_N^{(\tau)}) \geq \frac{\kappa}{2} - b_d \kappa^2 > s.$$

On the other hand, data processing under the common convolution (Lemma B.1) and (8) give

$$C^*(P_1^{(\tau)}, \dots, P_N^{(\tau)}) \leq C^*(P_1, \dots, P_N) \leq s.$$

This contradiction proves the required intersections. The estimator itself is determined by the balls associated with accepted candidates; the representing distributions and the common Gaussian convolution establish that every finite subfamily of these balls has nonempty intersection.

4. The full intersection and the selector. Helly's theorem [2, 14], applied to the balls associated with any finite family of accepted candidates, shows that these balls have a common intersection. To extend this conclusion to all accepted candidates, we use a compact member of the family. Define the empirical moments

$$\bar{X}_n = \frac{1}{n} \sum_i X_i, \\ \hat{\Sigma}_n = \frac{1}{n} \sum_i (X_i - \bar{X}_n)(X_i - \bar{X}_n)^\top.$$

The empirical measure belongs to $\mathcal{M}(\bar{X}_n, \hat{\Sigma}_n)$, so Lemma 2.1 gives $\mathcal{L}_n(\bar{X}_n, \hat{\Sigma}_n) = 0$. This pair is accepted for every sample. Its ball $B_0 = B(\bar{X}_n, \hat{\Sigma}_n)$ is compact, even when it is a singleton.

The closed subsets $B_0 \cap B(m, V)$ indexed by accepted candidates have the finite intersection property by Helly's theorem. Compactness of B_0 implies that their full intersection $\mathcal{C}(T, s)$ is nonempty. Since $\mathcal{C}(T, s)$ is closed, convex, and contained in B_0 , it is compact. We therefore define the selector associated with (T, s) by $\hat{\mu}(T, s) := \arg \min_{z \in \mathcal{C}(T, s)} \|z\|^2$. The minimizer exists by compactness and is unique by strict convexity. In particular, when all observations equal x , the ball B_0 is $\{x\}$ and the estimator equals x .

5. The selected point has the required error bound. On $\mathcal{E}$, the ball $B(\mu, \Sigma)$ is one of the balls defining $\mathcal{C}(T, s)$. Hence, with $\lambda = \|\Sigma\|_{\text{op}}$,

$$\|\hat{\mu}(T, s) - \mu\| \leq \sqrt{\kappa_d(s)\lambda}. \quad (9)$$

If $\lambda > 0$, factor the right-hand side as

$$\sqrt{\kappa_d(s)\lambda} = \sqrt{\frac{2\lambda L}{n}} \sqrt{\frac{ns}{L}} \sqrt{1 + 16b_d s}.$$

By (4), $ns/L = 1 + O_d(\log L/L)$ and $s \leq 2L/n$. The elementary inequality $\sqrt{1+x} \leq 1+x/2$ therefore bounds the last two factors by $1 + C_d(\log L/L + L/n)$ after increasing C_d to absorb their product term. A scalar projection of a Gaussian vector onto a top eigenvector of Σ gives the comparison

$$\sqrt{2\lambda L} \leq \left(1 + C \frac{\log L}{L}\right) \mathcal{Q}_\delta(\Sigma);$$

the scalar tail calculation is in Appendix F.3. Combining these bounds with (9) proves the desired radius on $\mathcal{E}$. Equation (7) gives its probability. If $\Sigma = 0$, the ball $B(\mu, \Sigma)$ is $\{\mu\}$ and the error is zero almost surely. Lemma 4.1 gives Borel measurability. Taking $\hat{\mu}_{n,\delta}^{(d)} = \hat{\mu}(T, s)$ completes the proof. $\square$

5 POLYNOMIAL-TIME COMPUTATION

Theorem 1 defines the estimator geometrically. We now show that, for fixed d , the same selector can be computed in polynomial time. For the thresholds constructed in Appendix L.4, write $T_\delta = \log c$, set $s = s_\delta$, and let $\zeta = \kappa_d(s)$. Appendix M gives the arithmetic procedure for computing these quantities and its cost. We characterize the intersection by a quantified polynomial formula, apply quantifier elimination, and then compute its unique point of minimum norm using standard algorithms in real algebraic geometry.

5.1 Polynomial description of the intersection

The intersection may be restricted to isotropic candidates. For $v = \|V\|_{\text{op}}$, we have $V \preceq vI_d$, and Lemma 2.1 gives $\mathcal{L}_n(m, vI_d) \leq \mathcal{L}_n(m, V)$. Moreover, (m, V) and (m, vI_d) define the same ball. Since isotropic candidates already belong to the original family, the restriction leaves the intersection unchanged; see Lemma L.1. If all observations equal x , the accepted candidate $(x, 0)$ forces the intersection to be $\{x\}$. Otherwise, it suffices to consider scalar candidates (m, v) with $v > 0$.

For the nonconstant sample, put $w_i = (1, X_i^\top)^\top$ and

$$S(m, v) = \begin{pmatrix} 1 & m^\top \\ m & mm^\top + vI_d \end{pmatrix}.$$

The likelihood representation (Lemma L.2) is

$$\exp\{\mathcal{L}_n(m, vI_d)\} = \max_{\substack{H \succ 0 \\ \text{tr } H = 1}} \frac{\prod_i w_i^\top H w_i}{[\text{tr}(HS(m, v))]^n}.$$

For $v > 0$, $S(m, v)$ is positive definite, so the denominator is positive. Clearing this denominator yields an equivalent polynomial test; Appendix L gives the derivation.

Put $p_d = (d+1)(d+2)/2$. For $\theta \in \mathbb{R}^{p_d}$, let L_θ be the lower triangular matrix of size $d+1$ whose entries are the coordinates of θ , ordered row by row. Set $H_\theta = L_\theta L_\theta^\top$, so that $H_\theta \succeq 0$ and $\text{tr } H_\theta = \|\theta\|^2$. Define

$$P_c(m, v, \theta) = \prod_{i=1}^n w_i^\top H_\theta w_i - c[\text{tr}(H_\theta S(m, v))]^n. \quad (10)$$

Every positive semidefinite matrix of trace one has this representation with $\|\theta\| = 1$. Lemma L.2 therefore implies that, for $v > 0$, rejection is equivalent to the existence of θ satisfying $\|\theta\|^2 = 1$ and $P_c(m, v, \theta) > 0$.

A point z belongs to the intersection if and only if every isotropic candidate (m, vI_d) with $v > 0$ whose associated ball excludes z is rejected. By the rejection criterion above,

$$\begin{aligned} z \in \mathcal{C}(\log c, s) &\iff \forall (m, v) \in \mathbb{R}^d \times \mathbb{R} : \\ &[v > 0 \wedge \|z - m\|^2 > \zeta v] \\ \implies \exists \theta \in \mathbb{R}^{p_d} : &\|\theta\|^2 = 1 \wedge P_c(m, v, \theta) > 0. \end{aligned} \quad (11)$$

Thus (m, v) parameterizes the isotropic candidate (m, vI_d) , while θ witnesses its rejection whenever $v > 0$ and the associated ball excludes z . For fixed d , the numbers of free and quantified variables are independent of n ; only the degrees of the defining polynomials grow with n . Hence standard algorithms in real algebraic geometry apply in polynomial time as n grows.

Algorithm 1: Mean estimation in $\mathbb{R}^d$

Input: Observations $X_1, \dots, X_n \in \mathbb{R}^d$ and confidence level $\delta \in (0, 1)$

Output: Estimate $\hat{\mu}_{n,\delta}^{(d)} \in \mathbb{R}^d$

- 1 **if** $X_1 = \dots = X_n$ **then return** X_1 ;
 - 2 Compute c, s, ζ by the threshold construction in Appendix L.4;
 - 3 **if** $s > s_0$ **then return** $n^{-1} \sum_{i=1}^n X_i$;
 - 4 Form the polynomial P_c in (10);
 - 5 Apply quantifier elimination to the condition in (11) to obtain a quantifier-free formula defining $\mathcal{C}(\log c, s)$;
 - 6 Using this formula, compute the unique point $z^* \in \mathcal{C}(\log c, s)$ minimizing $\|z\|^2$ by the procedure described below;
 - 7 **return** z^* ;
-

The minimizer in Step 6 is well defined. For the parameters above, Section 4.1 shows that $\mathcal{C}(\log c, s)$ is nonempty, compact, and convex whenever the sample is nonconstant and $s \leq s_0$. Hence $\|z\|^2$ attains its minimum on $\mathcal{C}(\log c, s)$, and strict convexity makes the minimizer unique. To compute this minimizer, let $\varphi(z)$ be the quantifier-free formula obtained in Step 5 and introduce a scalar ρ . Eliminating z from $\exists z [\varphi(z) \wedge \|z\|^2 \leq \rho]$ yields a univariate semialgebraic condition in ρ . Its smallest feasible value ρ^* can be determined by sign determination and root ordering, after which semialgebraic sampling of $\{z : \varphi(z) \wedge \|z\|^2 = \rho^*\}$ recovers the unique minimizer. Appendix M gives the construction and complexity bounds.

5.2 Proof of Theorem 2

Proof. When the sample is constant, the algorithm returns X_1 ; when $s > s_0$, it returns $\bar{X}_n$. Otherwise, Lemma L.1 and (11) imply that Step 5 produces a quantifier-free description of $\mathcal{C}(\log c, s)$. Section 4.1 shows that this set is nonempty, compact, and convex, so Step 6 computes its unique point of minimum norm, namely $\hat{\mu}(\log c, s) = \hat{\mu}_{n,\delta}^{(d)}$.

Formula (11) has d free variables, quantified blocks of sizes $d+1$ and p_d , and four polynomial comparisons of degree at most $4n$. Let N_{QE} and N_{sel} count the operations and comparisons for quantifier elimination and the subsequent selection. The standard bounds give [3, 40–42]

$$\begin{aligned} N_{\text{QE}} &\leq n^{O(d(d+1)p_d)} = n^{O(d^4)}, \\ N_{\text{sel}} &\leq n^{O(d^2(d+1)p_d)} = n^{O(d^5)}. \end{aligned}$$

Including threshold computation and polynomial formation, the total complexity is $(n + \ell_\delta)^{O(d^5)}$ (Appendix M). In the regime of Theorem 1, $\ell_\delta = O(L + 1) = O(n)$, yielding $n^{O(d^5)}$. $\square$

6 CONCLUSION

We have shown that, in every fixed dimension d , a single estimator can match the Gaussian sample-mean confidence radius $\mathcal{Q}_\delta(\Sigma)/\sqrt{n}$ up to a multiplicative factor $1 + o(1)$, uniformly over all distributions with finite covariance and without prior knowledge of Σ . Our construction relies on three components: a quadratic likelihood statistic whose relative entropy representation links accepted candidates to a common empirical reference measure, a multivariate Chernoff divergence lower bound with sharp leading coefficient $1/2$ and a remainder uniform over all ratios of covariance scales, and an intersection argument via Helly’s theorem that converts the joint constraints associated with accepted candidates into a single Euclidean estimate without the geometric loss incurred by diameter-based reductions. A matching lower bound from two Gaussian location models confirms that the leading constant is optimal in the stated regime.

Several directions remain open. Our guarantee captures the leading covariance dependence through $\|\Sigma\|_{\text{op}}$ but not the finer contribution of the remaining spectrum, and Gaussian sharp accuracy is established only in the asymptotic regime $L \rightarrow \infty$ rather than at a fixed confidence level δ . Achieving Gaussian sharp accuracy at fixed δ would require non-asymptotic likelihood calibration and remainder control that preserves the Gaussian leading constant, while identifying the next-order spectral dependence remains open.

AI use statement

In this work, we used generative AI tools to improve sentence structure, clarity, and grammar; to flag potential mathematical issues for review; and to assist with literature search. We did not use generative AI to generate the research results or derive the main mathematical arguments. All mathematical suggestions were independently checked by the authors, and literature-related claims were verified against the original sources before incorporation. We take responsibility for the final content of this work.

References

- [1] Shubhada Agrawal, Sandeep Juneja, and Peter W. Glynn. Optimal δ -correct best-arm selection for heavy-tailed distributions. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 61–110. PMLR, 2020. URL <https://proceedings.mlr.press/v117/agrawal20a.html>.
- [2] Alexander Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002. doi: 10.1090/gsm/054.
- [3] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer, second edition, 2006. doi: 10.1007/3-540-33099-2.
- [4] Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007.
- [5] Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454.
- [6] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-bayesian bounds for matrices, vectors, and linear least squares regression, 2017. URL <https://arxiv.org/abs/1712.02747>.
- [7] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-gaussian rates. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806. PMLR, 2019. URL <https://proceedings.mlr.press/v99/cherapanamjeri19b.html>.
- [8] Yeshwanth Cherapanamjeri, Nilesch Tripuraneni, Peter L. Bartlett, and Michael I. Jordan. Optimal mean estimation without a variance. In *Proceedings of the Thirty-Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 356–357. PMLR, 2022. URL <https://proceedings.mlr.press/v178/cherapanamjeri22a.html>.
- [9] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952. doi: 10.1214/aoms/1177729330.
- [10] Trung Dang, Jasper C. H. Lee, Maoyuan Song, and Paul Valiant. Optimality in mean estimation: Beyond worst-case, beyond sub-gaussian, and beyond $1 + \alpha$ moments. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0cddb777d3441326544e21b67f41bdc8-Abstract-Conference.html.
- [11] Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. doi: 10.1214/16-AOS1440.
- [12] Shivam Gupta, Jasper C. H. Lee, and Eric Price. Finite-sample symmetric mean estimation with fisher information rate. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4777–4830. PMLR, 2023. URL <https://proceedings.mlr.press/v195/gupta23a.html>.
- [13] Shivam Gupta, Samuel Hopkins, and Eric Price. Beyond Catoni: Sharper rates for heavy-tailed and robust mean estimation. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2232–2269. PMLR, 2024. URL <https://proceedings.mlr.press/v247/gupta24a.html>.
- [14] Eduard Helly. Über mengen konvexer körper mit gemeinschaftlichen punkten. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32: 175–176, 1923.
- [15] Samuel B. Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020. doi: 10.1214/19-AOS1843.
- [16] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. URL <https://www.jmlr.org/papers/v17/14-273.html>.

- [17] Fritz John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays Presented to R. Courant on His 60th Birthday*, pages 187–204. Interscience, New York, 1948.
- [18] Fumio Kubo and Tsuyoshi Ando. Means of positive linear operators. *Mathematische Annalen*, 246(3):205–224, 1980. doi: 10.1007/BF01371042.
- [19] Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire ϵ -variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025. doi: 10.1214/24-AOS2487.
- [20] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer, New York, 2 edition, 2000. doi: 10.1007/978-1-4612-1166-2.
- [21] Jasper C. H. Lee and Paul Valiant. Optimal sub-gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE, 2022. doi: 10.1109/FOCS52979.2021.00071.
- [22] Jasper C. H. Lee and Paul Valiant. Optimal sub-gaussian mean estimation in very high dimensions. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics*, pages 98:1–98:21. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022. doi: 10.4230/LIPIcs.ITCS.2022.98.
- [23] Jasper C. H. Lee, Walter McKelvie, Maoyuan Song, and Paul Valiant. All-purpose mean estimation over $\mathbb{R}$: Optimal sub-gaussianity with outlier robustness and low moments performance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 33435–33475. PMLR, 2025. URL <https://proceedings.mlr.press/v267/lee25w.html>.
- [24] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2598–2612. PMLR, 2020. URL <https://proceedings.mlr.press/v125/lei20a.html>.
- [25] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. doi: 10.1214/17-AOS1639.
- [26] Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 175(3–4):957–973, 2019. doi: 10.1007/s00440-019-00906-4.
- [27] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x.
- [28] Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021. doi: 10.1214/20-AOS1961.
- [29] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. doi: 10.3150/14-BEJ645.
- [30] Stanislav Minsker. U-statistics of growing order and sub-gaussian mean estimators with sharp constants. *Mathematical Statistics and Learning*, 7(1/2):1–39, 2024. doi: 10.4171/MSL/43.
- [31] Stanislav Minsker and Mohamed Ndaoud. Robust and efficient mean estimation: An approach based on the properties of self-normalized sums. *Electronic Journal of Statistics*, 15(2):6036–6070, 2021. doi: 10.1214/21-EJS1925.
- [32] Stanislav Minsker and Nate Strawn. The geometric median and applications to robust mean estimation. *SIAM Journal on Mathematics of Data Science*, 6(2):504–533, 2024. doi: 10.1137/23M1592420.
- [33] Hemant K. Mishra, Michael Nussbaum, and Mark M. Wilde. On the optimal error exponents for classical and quantum antidistinguishability. *Letters in Mathematical Physics*, 114:76, 2024. doi: 10.1007/s11005-024-01821-z.
- [34] Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, New York, 1983.
- [35] Gleb Novikov, David Steurer, and Stefan Tiegel. Robust mean estimation without moments for symmetric distributions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6c59ace4fc4872a14df13d91762ad4f0-Abstract-Conference.html.
- [36] Art B. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, 2001. doi: 10.1201/9781420036152.
- [37] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A robust univariate mean estimator is all you need. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages

- 4034–4044. PMLR, 2020. URL <https://proceedings.mlr.press/v108/prasad20a.html>.
- [38] W. Pusz and S. L. Woronowicz. Functional calculus for sesquilinear forms and the purification map. *Reports on Mathematical Physics*, 8(2):159–170, 1975. doi: 10.1016/0034-4877(75)90061-0.
- [39] Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994. doi: 10.1214/aos/1176325370.
- [40] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part i: Introduction. preliminaries. the geometry of semi-algebraic sets. the decision problem for the existential theory of the reals. *Journal of Symbolic Computation*, 13(3):255–299, 1992. doi: 10.1016/S0747-7171(10)80003-3.
- [41] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part ii: The general decision problem. preliminaries for quantifier elimination. *Journal of Symbolic Computation*, 13(3):301–327, 1992. doi: 10.1016/S0747-7171(10)80004-5.
- [42] James Renegar. On the computational complexity and geometry of the first-order theory of the reals. part iii: Quantifier elimination. *Journal of Symbolic Computation*, 13(3):329–352, 1992. doi: 10.1016/S0747-7171(10)80005-7.
- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
- (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
- (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
- (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
- (a) Citations of the creator If your work uses existing assets. [Not Applicable]
- (b) The license information of the assets, if applicable. [Not Applicable]
- (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
- (d) Information about consent from data providers/curators. [Not Applicable]
- (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

CHECKLIST

- For all models and algorithms presented, check if you include:
 - A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
- For any theoretical claim, check if you include:
 - Statements of the full set of assumptions of all theoretical results. [Yes]
 - Complete proofs of all theoretical results. [Yes]
 - Clear explanations of any assumptions. [Yes]
- For all figures and tables that present empirical results, check if you include:
 - The full text of instructions given to participants and screenshots. [Not Applicable]
 - Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Gaussian Sharp Mean Estimation under Finite Covariance in Fixed Dimension: Supplementary Materials

A	Additional related work	14
B	Quadratic likelihood: proofs and additional details	14
B.1	Moment classes and normalized coordinates	14
B.2	Proof of Lemma 2.1	15
B.3	Multivariate Chernoff divergence and data processing	17
B.4	Proof of Lemma 2.2	17
B.5	Proof of Corollary 2.3	19
C	Proof of the quadratic likelihood maximal inequality	19
D	Proof of the multivariate Chernoff bound	23
D.1	Reduction to at most $d + 1$ distributions	24
D.2	Normalization under the boundary and balance conditions	24
D.3	Bounding affinity by matrix geometric means	25
D.4	Second-order expansion of the matrix geometric mean	27
D.5	An auxiliary functional for the coefficient bound	28
D.6	The sharp coefficient bound	29
D.7	Uniform control of the remainder	29
D.8	Proof of the multivariate Chernoff divergence bound	30
D.9	Proof of the Carathéodory reduction	31
E	Construction with the theoretical threshold	31
E.1	Acceptance of (μ, Σ)	32
E.2	Common intersection and measurability	32
E.3	Comparison with the Gaussian radius	33
E.4	Sharpness and scope of the result	33
F	Technical details for the estimator	34
F.1	Parameter range	34
F.2	Intersection of at most $d + 1$ confidence balls	34
F.3	Details of the comparison with the Gaussian radius	35

G Sharpness: proof and additional discussion	36
G.1 Proof of Proposition G.1	37
H Relative perturbation of the matrix geometric mean	37
I Second-order expansion for two matrices	39
J Properties of the auxiliary functional	40
J.1 A scalar representation of the auxiliary functional	40
J.2 Bounds and monotonicity	40
K Proof of the coefficient propagation inequality	41
L Computational analysis of the estimator	43
L.1 Reduction to scalar covariance bounds	43
L.2 Polynomial characterization of rejection	44
L.3 Intersection and norm minimization	45
L.4 Threshold selection	46
L.5 Measurability and complexity	46
M Threshold computation and complexity bounds	47
M.1 Threshold computation	47
M.2 Polynomial construction and optimization	47

A ADDITIONAL RELATED WORK

Section 1 presents the main comparison with related work. This appendix provides additional details on the regimes and confidence radii most closely related to the leading constant studied in this paper.

Table 1: Results most directly related to the leading confidence constant.

Work	Setting and regime	$(1 - \delta)$ -confidence radius
Lugosi–Mendelson [25]	d arbitrary; Σ unknown; $\mathbb{E} \ X - \mu\ ^2 < \infty$; $n \geq 1$, $\delta \in (0, 1)$	$C \left(\sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sqrt{\frac{\ \Sigma\ _{\text{op}} L}{n}} \right)$; $C < \infty$ universal
Lee–Valiant [21]	$d = 1$; $\text{Var}(X) = \sigma^2 < \infty$ unknown; $L \rightarrow \infty$, $L/n \rightarrow 0$	$(1 + o(1)) \sigma \sqrt{\frac{2L}{n}}$
Lee–Valiant [22]	Σ unknown; $d_{\text{eff}}/L^2 \rightarrow \infty$; $n/L^3 \rightarrow \infty$	$(1 + o(1)) \sqrt{\frac{\text{tr}(\Sigma)}{n}}$
Gupta–Hopkins–Price [13]	$d \geq 2$; $\Sigma \preceq \sigma^2 I_d$ with σ^2 known; $n \geq CL \geq C^2 d$	$(1 - \tau) J_d \sigma \sqrt{\frac{2L}{n}}$; $\tau, C > 0$ universal
Theorem 1	d fixed; Σ unknown; $\mathbb{E} \ X - \mu\ ^2 < \infty$; $L \rightarrow \infty$, $L/n \rightarrow 0$	$\left[1 + C_d \left(\frac{\log L}{L} + \frac{L}{n} \right) \right] \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}}$ $= (1 + o(1)) \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}}$

Several nearby models lead to different benchmarks. Under symmetry on the line, Gupta, Lee and Price [12] obtain guarantees governed by Fisher information. Novikov, Steurer and Tiegel [35] study robust location estimation for symmetric product and elliptical distributions without moment assumptions. Contamination models are considered by Lugosi and Mendelson [28], while Prasad, Balakrishnan and Ravikumar [37] study reductions from multivariate to univariate estimation, and Cherapanamjeri, Tripuraneni, Bartlett and Jordan [8] consider directional moments below order two. These problems use different structural assumptions or pursue different forms of robustness, so their optimal constants need not coincide with the finite covariance benchmark considered here.

Our construction is also connected to empirical likelihood and information projection, although the feasible distributions are different. Classical empirical likelihood enforces moment constraints by reweighting the observations, and the resulting distribution therefore remains supported on the sample [36, 39]. In our construction, the relevant statistic minimizes reverse relative entropy over all distributions with a prescribed mean and a covariance upper bound. The minimizing distribution may place mass outside the observed sample. This gives the statistic a natural interpretation as a reverse information projection of the empirical distribution onto a moment class. Related ideas appear in the work of Larsson, Ramdas and Ruf [19] on reverse information projection and log-optimal e-variables, and in minimum divergence formulations for best-arm identification under heavy tails [1]. The empirical distribution serves as a single reference measure: every accepted candidate admits a representing distribution P for which $D_{\text{KL}}(\hat{P}_n \| P)$ is small, permitting comparison through the multivariate Chernoff divergence.

B QUADRATIC LIKELIHOOD: PROOFS AND ADDITIONAL DETAILS

B.1 Moment classes and normalized coordinates

For each $k \geq 1$, let $\mathcal{P}(\mathbb{R}^k)$ denote the set of Borel probability measures on $\mathbb{R}^k$. We regard $\mathbb{R}^d$ itself as the sample space and let $X : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $X(x) = x$, denote the canonical random vector. Thus, under $P \in \mathcal{P}(\mathbb{R}^d)$, the canonical map X has law P . For a candidate (m, V) , the associated moment class is

$$\mathcal{M}(m, V) := \{P \in \mathcal{P}(\mathbb{R}^d) : P(m + \text{Range}(V)) = 1, \mathbb{E}_P[X] = m, \text{Cov}_P(X) \preceq V\}.$$

The affine support restriction is consistent with the covariance bound. Indeed, if a lies in the null space of V , then every $P \in \mathcal{M}(m, V)$ satisfies $\text{Var}_P(a^\top X) = a^\top \text{Cov}_P(X) a \leq a^\top V a = 0$. Thus $a^\top (X - m) = 0$ almost surely

under P . Applying this observation to a basis of the null space shows that $X - m \in \text{Range}(V)$ almost surely.

We first normalize the proposed moments on this affine support. Let $r = \text{rank}(V)$, choose a matrix $U \in \mathbb{R}^{d \times r}$ whose columns form an orthonormal basis of $\text{Range}(V)$, and write $V = U\Lambda U^\top$, where Λ is positive definite. For $x \in m + \text{Range}(V)$, define

$$z_{m,V}(x) := \Lambda^{-1/2} U^\top (x - m) \in \mathbb{R}^r, \quad y_{m,V}(x) := \begin{pmatrix} 1 \\ z_{m,V}(x) \end{pmatrix} \in \mathbb{R}^{r+1}.$$

The map $x \mapsto z_{m,V}(x)$ is a bijection from $m + \text{Range}(V)$ to $\mathbb{R}^r$, with inverse $z \mapsto m + U\Lambda^{1/2}z$. Under this change of coordinates, a distribution in $\mathcal{M}(m, V)$ becomes a distribution with mean zero and covariance bounded by I_r . Conversely, every such normalized distribution maps back to a member of $\mathcal{M}(m, V)$.

The first coordinate of $y_{m,V}(x)$ is fixed at one. This augmentation allows a single quadratic form in $y_{m,V}(x)$ to include constant, linear, and quadratic terms in the normalized observation. When $r = 0$, the affine support is the singleton $\{m\}$, and we set $y_{m,V}(m) = 1$.

The main result of this section gives an exact interpretation of this statistic. After division by the sample size, it equals the smallest relative entropy from the empirical measure to any probability distribution with the proposed mean and covariance bound. Consequently, every accepted candidate admits a representing distribution P for which $D_{\text{KL}}(\hat{P}_n \| P)$ is small. Any finite collection of accepted candidates can therefore be compared through the multivariate Chernoff divergence, with the empirical measure appearing as the first argument of each relative entropy.

The definition is invariant under the choice of normalized coordinates. Any other choice gives normalized vectors related by an orthogonal transformation. Conjugating A by the corresponding augmented orthogonal matrix preserves positive semidefiniteness, trace, and the values of the quadratic forms, and hence preserves the supremum in (1).

Let $e_1 = (1, 0, \dots, 0)^\top$. The feasible matrix $A = e_1 e_1^\top$ gives $y_{m,V}(x)^\top A y_{m,V}(x) = 1$ throughout the affine support. Therefore $\mathcal{L}_n(m, V) \geq 0$. Moreover, every feasible matrix satisfies $A \preceq I_{r+1}$, so each quadratic factor is at most $\|y_{m,V}(X_j)\|^2$. Thus the statistic is finite whenever all observations lie in the associated affine support.

B.2 Proof of Lemma 2.1

Proof. If $r = 0$, the support assumption implies that every observation equals m . In this case, $\mathcal{M}(m, 0) = \{\delta_m\}$, the statistic is zero, and both sides of (2) equal zero. We therefore assume $r \geq 1$.

We work in normalized coordinates. Put $z_j = z_{m,V}(X_j)$, let $Q_n = n^{-1} \sum_{j=1}^n \delta_{z_j}$, and write $y(z) = (1, z^\top)^\top$. The normalized moment class is

$$\mathcal{P}_r := \{P \in \mathcal{P}(\mathbb{R}^r) : \mathbb{E}_P[Z] = 0, \text{Cov}_P(Z) \preceq I_r\},$$

where $Z(z) = z$ is the canonical random vector on $\mathbb{R}^r$. Because the coordinate map is a measurable bijection with measurable inverse, it preserves relative entropy. It is therefore enough to establish the representation with Q_n and $\mathcal{P}_r$ in place of $\hat{P}_n$ and $\mathcal{M}(m, V)$.

For a probability measure P with finite second moment, define its augmented moment matrix by

$$M_P := \mathbb{E}_P[y(Z)y(Z)^\top] = \begin{pmatrix} 1 & b_P^\top \\ b_P & S_P \end{pmatrix}, \quad b_P := \mathbb{E}_P[Z], \quad S_P := \mathbb{E}_P[ZZ^\top].$$

We claim that $P \in \mathcal{P}_r$ if and only if $M_P \preceq I_{r+1}$. One direction is immediate: if $b_P = 0$ and $S_P \preceq I_r$, then $M_P = \text{diag}(1, S_P) \preceq I_{r+1}$. Conversely, if $M_P \preceq I_{r+1}$, then

$$I_{r+1} - M_P = \begin{pmatrix} 0 & -b_P^\top \\ -b_P & I_r - S_P \end{pmatrix} \succeq 0.$$

For every $v \in \mathbb{R}^r$ and $t \in \mathbb{R}$, its quadratic form at (t, v) is $-2tb_P^\top v + v^\top(I_r - S_P)v$. Nonnegativity for all t forces $b_P^\top v = 0$. Since v is arbitrary, $b_P = 0$, and the lower-right principal block gives $S_P \preceq I_r$. This proves the claim.

We first show that every feasible quadratic form provides a lower bound on relative entropy. Fix $A \succeq 0$ with $\text{tr } A = 1$, and put $f_A(z) = y(z)^\top A y(z)$. For every $P \in \mathcal{P}_r$, $\int_{\mathbb{R}^r} f_A(z) dP(z) = \text{tr}(A M_P) \leq \text{tr } A = 1$. The inequality follows because $A \succeq 0$ and $I_{r+1} - M_P \succeq 0$, which imply $\text{tr}(A(I_{r+1} - M_P)) \geq 0$.

If $Q_n \not\ll P$, the desired entropy lower bound is immediate. Otherwise, let S be the finite support of Q_n , and write $q_z = Q_n(\{z\})$ and $p_z = P(\{z\})$ for $z \in S$. If f_A vanishes at an empirical atom, its empirical logarithmic average is $-\infty$, and the bound again follows. In the remaining case, Jensen's inequality gives

$$\int_{\mathbb{R}^r} \log f_A dQ_n - D_{\text{KL}}(Q_n \| P) = \sum_{z \in S} q_z \log \left(\frac{f_A(z) p_z}{q_z} \right) \leq \log \left(\sum_{z \in S} f_A(z) p_z \right) \leq \log \int_{\mathbb{R}^r} f_A dP \leq 0.$$

Thus $D_{\text{KL}}(Q_n \| P) \geq \int_{\mathbb{R}^r} \log f_A dQ_n$. Taking the supremum over A shows that the optimized quadratic statistic divided by n is no larger than $D_{\text{KL}}(Q_n \| P)$ for any $P \in \mathcal{P}_r$.

To prove the converse, define

$$F(A) := \int_{\mathbb{R}^r} \log(y(z)^\top A y(z)) dQ_n(z) = \frac{1}{n} \sum_{j=1}^n \log(y(z_j)^\top A y(z_j)).$$

The feasible set $\{A \succeq 0 : \text{tr } A = 1\}$ is compact. The function F is upper semicontinuous on this set, with the convention $\log 0 = -\infty$, and is bounded above because $y(z_j)^\top A y(z_j) \leq \|y(z_j)\|^2$ for every j . Hence F attains its maximum at some matrix A^* . Since $F(e_1 e_1^\top) = 0$, we have $F(A^*) \geq 0$. In particular, $y(z_j)^\top A^* y(z_j) > 0$ for every empirical observation: a zero denominator would make $F(A^*) = -\infty$.

Define $G := \int_{\mathbb{R}^r} \frac{y(z) y(z)^\top}{y(z)^\top A^* y(z)} dQ_n(z)$. All denominators are positive, so this is a finite positive semidefinite matrix. For any feasible A , consider the feasible segment $A_t = (1-t)A^* + tA$, where $0 \leq t \leq 1$. Since A^* maximizes F , differentiation at $t = 0$ gives

$$0 \geq \left. \frac{dF(A_t)}{dt} \right|_{t=0^+} = \int_{\mathbb{R}^r} \frac{y(z)^\top (A - A^*) y(z)}{y(z)^\top A^* y(z)} dQ_n(z) = \text{tr}(GA) - \text{tr}(GA^*).$$

The second trace has the explicit value

$$\text{tr}(GA^*) = \int_{\mathbb{R}^r} \frac{y(z)^\top A^* y(z)}{y(z)^\top A^* y(z)} dQ_n(z) = 1.$$

Consequently $\text{tr}(GA) \leq 1$ for every feasible A . Taking $A = vv^\top$ for an arbitrary unit vector $v \in \mathbb{R}^{r+1}$ yields $v^\top G v \leq 1$, and therefore $G \preceq I_{r+1}$.

We now use A^* to construct an admissible probability measure. Define a measure P_0 , supported on the empirical atoms, by $\frac{dP_0}{dQ_n}(z) := \frac{1}{y(z)^\top A^* y(z)}$. Its augmented moment matrix is G . Write

$$G = \begin{pmatrix} g_{11} & g_{x1}^\top \\ g_{x1} & G_{xx} \end{pmatrix}, \quad \tilde{g} := 1 - g_{11}.$$

The entries have a direct interpretation:

$$P_0(\mathbb{R}^r) = g_{11}, \quad \int_{\mathbb{R}^r} z dP_0(z) = g_{x1}, \quad \int_{\mathbb{R}^r} z z^\top dP_0(z) = G_{xx}.$$

Since $G \preceq I_{r+1}$, we have $\tilde{g} \geq 0$, so P_0 is a subprobability measure. It may have less than unit mass and a nonzero first moment. The remaining step is to correct these two defects without exceeding the second moment bound.

If $\tilde{g} = 0$, the matrix $I_{r+1} - G$ is positive semidefinite and has a zero upper-left entry. The same quadratic-form argument used above forces $g_{x1} = 0$. Its lower-right block gives $G_{xx} \preceq I_r$. Hence P_0 already belongs to $\mathcal{P}_r$, and we take $P = P_0$.

Suppose instead that $\tilde{g} > 0$. Since

$$I_{r+1} - G = \begin{pmatrix} \tilde{g} & -g_{x1}^\top \\ -g_{x1} & I_r - G_{xx} \end{pmatrix} \succeq 0,$$

its Schur complement gives $I_r - G_{xx} - g_{x1}g_{x1}^\top/\tilde{g} \succeq 0$. Equivalently, for every $v \in \mathbb{R}^r$, minimizing the quadratic form $\tilde{g}t^2 - 2t g_{x1}^\top v + v^\top (I_r - G_{xx})v$ over $t \in \mathbb{R}$ yields $v^\top (I_r - G_{xx})v - \frac{(g_{x1}^\top v)^2}{\tilde{g}} \geq 0$. This inequality shows that the remaining allowance in the second moment bound is sufficient to restore the missing mass and cancel the first moment.

Specifically, set $z_0 = -g_{x1}/\tilde{g}$ and define $P = P_0 + \tilde{g}\delta_{z_0}$. Then $P(\mathbb{R}^r) = g_{11} + \tilde{g} = 1$ and $\mathbb{E}_P[Z] = g_{x1} + \tilde{g}z_0 = 0$, and, since the mean is zero, $\text{Cov}_P(Z) = G_{xx} + \tilde{g}z_0z_0^\top = G_{xx} + \frac{g_{x1}g_{x1}^\top}{\tilde{g}} \preceq I_r$. Thus $P \in \mathcal{P}_r$ in both cases.

It remains to compare its relative entropy with the optimized likelihood. For every empirical atom $z \in S$, the construction gives $P(\{z\}) \geq P_0(\{z\}) = \frac{q_z}{y(z)^\top A^\star y(z)}$. Therefore

$$D_{\text{KL}}(Q_n \| P) = \sum_{z \in S} q_z \log \frac{q_z}{P(\{z\})} \leq \sum_{z \in S} q_z \log (y(z)^\top A^\star y(z)) = F(A^\star) = \frac{\mathcal{L}_n(m, V)}{n}.$$

This argument remains valid if the residual atom z_0 coincides with an empirical atom, since adding mass only increases its probability. Relative entropy is evaluated after P_0 has been completed to the probability measure P .

The previously established lower bound applies to the constructed $P \in \mathcal{P}_r$, so the last inequality must be an equality. Thus the minimum is attained and equals the optimized statistic. Finally, pushing P forward under $z \mapsto m + U\Lambda^{1/2}z$ produces a distribution in $\mathcal{M}(m, V)$. Invariance of relative entropy under this bijection proves (2). $\square$

The construction explains why allowing mass outside the empirical support matters. The optimizing quadratic form first determines a reweighting of the observed atoms. If these weights sum to less than one, the remaining mass can be placed at a single additional point to restore the proposed mean while respecting the covariance bound. Thus a moment pair can be realized even when the additional location required by the moments is absent from the empirical support. This differs from empirical likelihood formulations that restrict the representing distribution to the observed support [36].

B.3 Multivariate Chernoff divergence and data processing

The affinity, and hence the multivariate Chernoff divergence, is independent of the dominating measure. To see this, let $\tilde{\nu}$ be another dominating probability measure and put $\bar{\nu} = (\nu + \tilde{\nu})/2$. If $a = d\nu/d\bar{\nu}$ and $f_i = dP_i/d\bar{\nu}$, then $f_i = p_i a$ almost everywhere with respect to $\bar{\nu}$. Since the positive weights sum to one,

$$\prod_{i \in I_\beta} f_i^{\beta_i} = a \prod_{i \in I_\beta} p_i^{\beta_i} \quad \bar{\nu}\text{-almost everywhere.}$$

Integrating this identity shows that the affinity computed using $\bar{\nu}$ equals the affinity computed using ν . The same argument with $\tilde{\nu}$ proves the claim. The definition and this argument apply without change to probability measures on any common measurable space.

B.4 Proof of Lemma 2.2

Proof. Fix a weight vector $\beta \in \Delta_N$. Choose a probability measure ν dominating $Q, P_1, \dots, P_N$, and write $q = dQ/d\nu$ and $p_i = dP_i/d\nu$. Set

$$g_\beta(x) := \prod_{i \in I_\beta} p_i(x)^{\beta_i}, \quad \mathcal{W}_\beta := \int_{\mathbb{R}^d} g_\beta(x) d\nu(x).$$

The assumption that each relative entropy is finite implies $Q \ll P_i$ for every i . Consequently, $p_i > 0$ almost surely under Q for every $i \in I_\beta$, and hence $g_\beta > 0$ almost surely under Q . Since $Q \ll \nu$, this implies $\mathcal{W}_\beta > 0$. The affinity bound above also gives $\mathcal{W}_\beta \leq 1$.

We may therefore define a probability measure $\bar{P}_\beta$ by its density

$$\frac{d\bar{P}_\beta}{d\nu}(x) = \bar{p}_\beta(x) := \frac{g_\beta(x)}{\mathcal{W}_\beta}.$$

This distribution is the normalized geometric mixture associated with the weights β . Moreover, $Q \ll \bar{P}_\beta$, because $\bar{p}_\beta > 0$ almost surely under Q .

On the support of Q , the logarithmic density ratio can be written as

$$\log \frac{q}{\bar{p}_\beta} = \log q - \sum_{i \in I_\beta} \beta_i \log p_i + \log \mathcal{W}_\beta = \sum_{i \in I_\beta} \beta_i \log \frac{q}{p_i} + \log \mathcal{W}_\beta.$$

The second equality uses $\sum_{i \in I_\beta} \beta_i = 1$. Integrating with respect to Q gives the identity

$$D_{\text{KL}}(Q \| \bar{P}_\beta) = \sum_{i \in I_\beta} \beta_i D_{\text{KL}}(Q \| P_i) + \log \mathcal{W}_\beta.$$

All terms are well defined: the divergences $D_{\text{KL}}(Q \| P_i)$ are finite by assumption, and $0 < \mathcal{W}_\beta \leq 1$. In particular, the logarithmic density ratios on the right are integrable under Q .

By nonnegativity of relative entropy,

$$-\log \mathcal{W}_\beta = \sum_{i \in I_\beta} \beta_i D_{\text{KL}}(Q \| P_i) - D_{\text{KL}}(Q \| \bar{P}_\beta) \leq \sum_{i \in I_\beta} \beta_i D_{\text{KL}}(Q \| P_i) \leq s.$$

This holds for every weight vector β , so taking the supremum proves the result. $\square$

The common first argument in the relative entropy bounds is essential. Since each $D_{\text{KL}}(Q \| P_i)$ is integrated with respect to Q , the weighted divergences can be combined through the normalized geometric mixture $\bar{P}_\beta$, yielding the entropy identity in the proof. In Corollary 2.3, the common measure is the empirical measure $\hat{P}_n$, and the P_i are the representing distributions associated with accepted candidates.

The classical multivariate Chernoff divergence satisfies the following data-processing inequality; compare [33].

Lemma B.1 (Data-processing inequality). *Let K be a Markov kernel from $\mathbb{R}^d$ to a standard Borel space $\mathcal{Y}$. For a probability measure P on $\mathbb{R}^d$, let KP denote the probability measure on $\mathcal{Y}$ defined by $(KP)(B) = \int_{\mathbb{R}^d} K(x, B) dP(x)$. Then $C^*(KP_1, \dots, KP_N) \leq C^*(P_1, \dots, P_N)$.*

Proof. Fix $\beta \in \Delta_N$, and let ν dominate the input distributions, with densities $p_i = dP_i/d\nu$. Introduce the joint reference probability measure $\Pi(dx, dy) := \nu(dx)K(x, dy)$ on $\mathbb{R}^d \times \mathcal{Y}$, and let $\rho = K\nu$ be its output marginal. Denote the coordinate variables under Π by X and Y .

Every output law KP_i is absolutely continuous with respect to ρ . Indeed, if $\rho(B) = 0$, then $K(x, B) = 0$ for ν -almost every x , and hence also for P_i -almost every x . Let $q_i = d(KP_i)/d\rho$.

We first identify these output densities as conditional expectations. For every bounded measurable function $\varphi : \mathcal{Y} \rightarrow \mathbb{R}$,

$$\int_{\mathcal{Y}} \varphi(y) q_i(y) d\rho(y) = \int_{\mathbb{R}^d} \int_{\mathcal{Y}} \varphi(y) K(x, dy) dP_i(x) = \mathbb{E}_\Pi[p_i(X) \varphi(Y)].$$

Since Y has law ρ under Π , the defining property of conditional expectation therefore gives $q_i(Y) = \mathbb{E}_\Pi[p_i(X) | Y]$ Π -almost surely.

Conditional generalized Hölder's inequality, applied to the indices in I_β , now yields

$$\mathbb{E}_\Pi \left[\prod_{i \in I_\beta} p_i(X)^{\beta_i} \middle| Y \right] \leq \prod_{i \in I_\beta} (\mathbb{E}_\Pi[p_i(X) | Y])^{\beta_i} = \prod_{i \in I_\beta} q_i(Y)^{\beta_i}.$$

For positive weights strictly below one, the Hölder exponents are $1/\beta_i$, whose reciprocals sum to one. If $|I_\beta| = 1$, the displayed inequality is an equality.

Taking expectations and using the input and output marginals of Π gives

$$\int_{\mathbb{R}^d} \prod_{i \in I_\beta} p_i(x)^{\beta_i} d\nu(x) \leq \int_{\mathcal{Y}} \prod_{i \in I_\beta} q_i(y)^{\beta_i} d\rho(y).$$

Thus applying the same kernel to every distribution can only increase the affinity for each fixed weight vector. Since $x \mapsto -\log x$ is decreasing, the corresponding Chernoff exponent can only decrease. Taking the supremum over β proves the lemma, including the case of infinite input multivariate Chernoff divergence. $\square$

A particular kernel used later adds independent Gaussian noise. For $\tau > 0$, take $K(x, \cdot) = \mathcal{N}(x, \tau I_d)$, so that $KP_i = P_i * \mathcal{N}(0, \tau I_d)$, where $*$ denotes convolution of probability measures. The multivariate Chernoff divergence is nonincreasing under this common Gaussian smoothing, while every mean is unchanged and each covariance matrix increases by τI_d .

B.5 Proof of Corollary 2.3

Proof. Because $\mathcal{L}_n(m_i, V_i) \leq T < \infty$ for every i , all observations lie in the corresponding affine support $m_i + \text{Range}(V_i)$. Lemma 2.1 therefore applies to every candidate and provides an attaining distribution $P_i \in \mathcal{M}(m_i, V_i)$ with $D_{\text{KL}}(\hat{P}_n \| P_i) = \frac{\mathcal{L}_n(m_i, V_i)}{n} \leq \frac{T}{n}$. Although the representing distributions may differ, the measure in the first argument is always the same empirical measure $\hat{P}_n$. Lemma 2.2, with $Q = \hat{P}_n$ and $s = T/n$, gives the asserted multivariate Chernoff divergence bound. $\square$

C PROOF OF THE QUADRATIC LIKELIHOOD MAXIMAL INEQUALITY

Let Θ be a nonempty compact convex subset of a finite-dimensional affine space. For each coordinate i , let ν_i be a probability measure on a measurable space $(\mathcal{X}_i, \mathcal{A}_i)$, and for each $\theta \in \Theta$ let $f_{i,\theta}$ be a nonnegative density with respect to ν_i . Assume that, for every i and x_i , the map $\theta \mapsto f_{i,\theta}(x_i)$ is affine, and that $\int_{\mathcal{X}_i} f_{i,\theta} d\nu_i = 1$ for every $\theta \in \Theta$. Set

$$\mathcal{X} := \prod_{i=1}^n \mathcal{X}_i, \quad \nu := \bigotimes_{i=1}^n \nu_i, \quad p_\theta(x_1, \dots, x_n) := \prod_{i=1}^n f_{i,\theta}(x_i).$$

Let P_θ denote the probability measure with density p_θ relative to ν . Thus, for each $\theta \in \Theta$, P_θ is the product law determined by the coordinate densities $f_{1,\theta}, \dots, f_{n,\theta}$. Its coordinates are independent, but need not be identically distributed.

For $\theta, \phi \in \Theta$, define

$$h_i^2(\theta, \phi) := \int_{\mathcal{X}_i} \left(\sqrt{f_{i,\theta}} - \sqrt{f_{i,\phi}} \right)^2 d\nu_i, \quad h^2(\theta, \phi) := \sum_{i=1}^n h_i^2(\theta, \phi).$$

This is the squared Hellinger distance, summed over coordinates; see, for example, [20]. We use the unnormalized convention, without the factor $1/2$. The quantity h is a pseudometric on Θ : distinct points may have distance zero if they determine the same coordinate distributions. It is continuous on $\Theta \times \Theta$. Indeed, choose an affine basis of Θ . Each coordinate density is a linear combination of the densities at the basis points, with coefficients continuous in θ and bounded on Θ . This gives an integrable envelope, and dominated convergence proves continuity of the Hellinger distances.

Two standard properties of this pseudometric will be used repeatedly [20]. First, joint convexity of squared Hellinger distance and the affine dependence of the densities on θ imply

$$h^2((1-t)\theta_0 + t\theta_1, (1-t)\theta_0 + t\theta_2) \leq th^2(\theta_1, \theta_2), \quad 0 \leq t \leq 1. \quad (12)$$

Both mixtures share the component corresponding to θ_0 , while the remaining components have weight t . Second, normalization of the coordinate densities gives

$$\int_{\mathcal{X}_i} \sqrt{f_{i,\theta} f_{i,\phi}} d\nu_i = 1 - \frac{1}{2} h_i^2(\theta, \phi).$$

Independence and the inequality $1 - x \leq e^{-x}$ therefore yield

$$\int_{\mathcal{X}} \sqrt{p_\theta p_\phi} d\nu = \prod_{i=1}^n \left(1 - \frac{1}{2} h_i^2(\theta, \phi) \right) \leq e^{-h^2(\theta, \phi)/2}. \quad (13)$$

Thus separation in h produces exponential decay of the affinity between the corresponding product laws.

Lemma C.1. *For every integer $p_0 \geq 1$, there is a constant $C_0 = C_0(p_0) \geq 1$ with the following property. Let $C \subset \Theta$ be nonempty, compact and convex, with $\dim \text{aff } C \leq p_0$ and $\text{diam}_h(C) \leq 2$. There is $\phi \in \text{relint } C$ such that the envelope $\bar{p} := \sup_{\theta \in C} p_\theta$ satisfies, for every $\gamma \in \Theta$ and $1/2 \leq \alpha < 1$,*

$$\int_{\mathcal{X}} \bar{p} \, d\nu \leq C_0^{1/2}, \quad (14)$$

$$\int_{\mathcal{X}} \sqrt{p_\gamma \bar{p}} \, d\nu \leq C_0^{1/4} e^{-h^2(\gamma, \phi)/4}, \quad (15)$$

$$\int_{\mathcal{X}} p_\gamma^{1-\alpha} \bar{p}^\alpha \, d\nu \leq C_0^{1/2} e^{-(1-\alpha)h^2(\gamma, \phi)/2}. \quad (16)$$

Proof. Write $m = \dim \text{aff } C$. If $m = 0$, then $C = \{\phi\}$ for its unique point ϕ , and $\bar{p} = p_\phi$. The first assertion follows from normalization, the second from (13), and the third by interpolation between the affinity bound and $\int_{\mathcal{X}} p_\phi \, d\nu = 1$. We may therefore assume $m \geq 1$.

Choose an ellipsoid of maximum volume in C , considered within its affine hull, and let ϕ be its center. John's theorem [17] supplies a linear isomorphism $T : \mathbb{R}^m \rightarrow \text{aff } C - \phi$ such that $B_2^m \subset C' := \{z \in \mathbb{R}^m : \phi + Tz \in C\} \subset mB_2^m$, where B_2^m is the closed Euclidean unit ball. In these coordinates, C' contains the unit ball and is contained in the ball of radius m .

Fix a coordinate i and a point x_i . The function $z \mapsto f_{i, \phi + Tz}(x_i)$ is affine and nonnegative on C' . If its value at zero is zero, nonnegativity on the unit ball forces its linear part to vanish. It is then zero throughout C' . Consequently, all coordinate densities indexed by C vanish wherever $f_{i, \phi}$ vanishes.

On the set $\{f_{i, \phi} > 0\}$, we may write $\frac{f_{i, \phi + Tz}}{f_{i, \phi}} = 1 + \langle c_i, z \rangle$, where c_i is a measurable vector-valued function of the i th observation. Nonnegativity for every $\|z\| \leq 1$ implies $\|c_i\| \leq 1$: otherwise, choosing $z = -c_i / \|c_i\|$ would make the right-hand side negative. Normalization of the densities also gives

$$1 = \int_{\mathcal{X}_i} f_{i, \phi + Tz} \, d\nu_i = 1 + \langle \mathbb{E}_\phi c_i, z \rangle, \quad z \in B_2^m,$$

and hence $\mathbb{E}_\phi c_i = 0$, where the expectation is taken under the i th coordinate law corresponding to ϕ .

The bounded Hellinger diameter controls the total variance of these linear coefficients. For $x \in [-1, 1]$, $x^2 = (\sqrt{1+x} - 1)(\sqrt{1+x} + 1)^2 \leq (1 + \sqrt{2})^2 (\sqrt{1+x} - 1)^2$. For a unit vector v , both ϕ and $\phi + Tv$ belong to C , and the common support property gives $h_i^2(\phi, \phi + Tv) = \mathbb{E}_\phi \left(\sqrt{1 + \langle c_i, v \rangle} - 1 \right)^2$. Summing the preceding scalar inequality therefore yields

$$\sum_{i=1}^n \mathbb{E}_\phi \langle c_i, v \rangle^2 \leq (1 + \sqrt{2})^2 h^2(\phi, \phi + Tv) \leq 4(1 + \sqrt{2})^2. \quad (17)$$

Put $S = \sum_{i=1}^n c_i$ and define $W = \bar{p}/p_\phi$ on $\{p_\phi > 0\}$. For every $z \in C'$, the likelihood ratio satisfies

$$\frac{p_{\phi + Tz}}{p_\phi} = \prod_{i=1}^n (1 + \langle c_i, z \rangle) \leq \exp \left(\sum_{i=1}^n \langle c_i, z \rangle \right) = e^{\langle S, z \rangle}.$$

The inequality uses $1 + x \leq e^x$ and remains valid if a factor vanishes. Since $C' \subset mB_2^m$, taking the supremum gives $W \leq e^{m\|S\|}$.

We next bound the second moment of this envelope ratio. If X is centered and $|X| \leq 1$, then

$$\mathbb{E} e^{tX} \leq \exp((e^t - 1 - t)\mathbb{E} X^2), \quad t \geq 0. \quad (18)$$

Indeed, the scalar inequality $e^{tx} \leq 1 + tx + (e^t - 1 - t)x^2$ holds for $-1 \leq x \leq 1$. Taking expectations eliminates the linear term, and $1 + a \leq e^a$ gives (18).

Choose a $1/2$ -net $\mathcal{U}$ of the Euclidean unit sphere with at most 5^m points. For every $s \in \mathbb{R}^m$, this net satisfies $\|s\| \leq 2 \max_{v \in \mathcal{U}} \langle v, s \rangle$. Consequently,

$$W^2 \leq e^{2m\|S\|} \leq \sum_{v \in \mathcal{U}} e^{4m\langle v, S \rangle}.$$

Under P_ϕ , the vectors c_i are independent. Applying (18) to each centered variable $\langle v, c_i \rangle$, and then using (17), gives

$$\begin{aligned} \mathbb{E}_\phi e^{4m\langle v, S \rangle} &= \prod_{i=1}^n \mathbb{E}_\phi e^{4m\langle v, c_i \rangle} \leq \exp \left((e^{4m} - 1 - 4m) \sum_{i=1}^n \mathbb{E}_\phi \langle v, c_i \rangle^2 \right), \\ &\leq \exp(4(e^{4m} - 1 - 4m)(1 + \sqrt{2})^2). \end{aligned}$$

It follows that

$$\mathbb{E}_\phi W^2 \leq 5^m \exp(4(e^{4m} - 1 - 4m)(1 + \sqrt{2})^2) \leq C_0(p_0),$$

where C_0 may be chosen uniformly over $1 \leq m \leq p_0$ and enlarged to cover $m = 0$.

Every density indexed by C vanishes outside $\{p_\phi > 0\}$. Hence

$$\int_{\mathcal{X}} \bar{p} \, d\nu = \mathbb{E}_\phi W \leq (\mathbb{E}_\phi W^2)^{1/2} \leq C_0^{1/2},$$

which proves (14).

To obtain a bound that retains the distance from an arbitrary $\gamma \in \Theta$, define $X = \sqrt{p_\gamma/p_\phi}$ on $\{p_\phi > 0\}$ and set it to zero elsewhere. Then

$$\mathbb{E}_\phi X = \int_{\mathcal{X}} \sqrt{p_\gamma p_\phi} \, d\nu, \quad \mathbb{E}_\phi X^2 = \int_{\{p_\phi > 0\}} p_\gamma \, d\nu \leq 1.$$

Two successive applications of Cauchy–Schwarz give

$$\mathbb{E}_\phi [X\sqrt{W}] = \mathbb{E}_\phi [\sqrt{X}\sqrt{XW}] \leq (\mathbb{E}_\phi X)^{1/2} (\mathbb{E}_\phi XW)^{1/2} \leq (\mathbb{E}_\phi X)^{1/2} (\mathbb{E}_\phi X^2)^{1/4} (\mathbb{E}_\phi W^2)^{1/4}.$$

Since the left-hand side is $\int_{\mathcal{X}} \sqrt{p_\gamma \bar{p}} \, d\nu$, equation (13) implies

$$\int_{\mathcal{X}} \sqrt{p_\gamma \bar{p}} \, d\nu \leq C_0^{1/4} e^{-h^2(\gamma, \phi)/4},$$

proving (15).

Finally, for $1/2 < \alpha < 1$, write

$$p_\gamma^{1-\alpha} \bar{p}^\alpha = (\sqrt{p_\gamma \bar{p}})^{2(1-\alpha)} \bar{p}^{2\alpha-1}.$$

The exponents $2(1-\alpha)$ and $2\alpha-1$ are nonnegative and sum to one. Hölder's inequality, followed by (14) and (15), yields

$$\int_{\mathcal{X}} p_\gamma^{1-\alpha} \bar{p}^\alpha \, d\nu \leq \left(\int_{\mathcal{X}} \sqrt{p_\gamma \bar{p}} \, d\nu \right)^{2(1-\alpha)} \left(\int_{\mathcal{X}} \bar{p} \, d\nu \right)^{2\alpha-1} \leq C_0^{\alpha/2} e^{-(1-\alpha)h^2(\gamma, \phi)/2}.$$

Since $C_0 \geq 1$ and $\alpha < 1$, this implies (16). The endpoint $\alpha = 1/2$ follows directly from (15). $\square$

The local lemma bounds the likelihood envelope on a set of controlled Hellinger diameter. To apply it to all of Θ , we need a covering estimate that depends only on the affine dimension. The next lemma obtains such an estimate directly from convexity.

Lemma C.2. *Let $p = \dim \text{aff } \Theta$, fix $\gamma \in \Theta$, and let $R \geq 1$. Every subset of*

$$\Theta_R := \{\theta \in \Theta : h(\gamma, \theta) \leq R\}$$

whose distinct points have h -distance strictly greater than one contains at most $(4R)^{2p}$ points.

Proof. The set Θ_R is compact by continuity of h and convex by convexity of squared Hellinger distance in its second argument. Consider a finite separated set $\theta_1, \dots, \theta_N \in \Theta_R$. Set $t = (4R)^{-2}$ and define $H_j := (1-t)\theta_j + t\Theta_R$ for $1 \leq j \leq N$. Each H_j is a contracted copy of Θ_R centered toward θ_j , and convexity gives $H_j \subset \Theta_R$.

For every $\beta \in \Theta_R$, equation (12) and the triangle inequality imply

$$h((1-t)\theta_j + t\beta, \theta_j) \leq \sqrt{t}h(\beta, \theta_j) \leq \sqrt{t}(h(\beta, \gamma) + h(\gamma, \theta_j)) \leq 2R\sqrt{t} = \frac{1}{2}.$$

Thus every point of H_j lies within Hellinger distance $1/2$ of θ_j . If H_j and H_k shared a point, the triangle inequality would imply $h(\theta_j, \theta_k) \leq 1$, contrary to the separation assumption. The sets H_j are therefore pairwise disjoint.

Let $r = \dim \text{aff } \Theta_R \leq p$. If $r = 0$, the set Θ_R is a singleton, so $N \leq 1$. Otherwise, Θ_R has finite positive r -dimensional volume within its affine hull. Under dilation by a factor t , the volume is multiplied by t^r . Since the disjoint sets H_j are all contained in Θ_R ,

$$Nt^r \text{vol}_r(\Theta_R) = \sum_{j=1}^N \text{vol}_r(H_j) \leq \text{vol}_r(\Theta_R).$$

Since $\text{vol}_r(\Theta_R) > 0$, it follows that $N \leq t^{-r} = (4R)^{2r} \leq (4R)^{2p}$. Any separated set with more than $(4R)^{2p}$ points would contain a finite separated subset with the same property, contradicting this bound. $\square$

Proposition C.3 (Affine likelihood maximal inequality). *For every integer $p_0 \geq 1$, there is a constant $C_{p_0}^{\text{aff}} < \infty$ such that every affine family above with $\dim \text{aff } \Theta \leq p_0$ satisfies, for every $\gamma \in \Theta$ and $v \geq 0$,*

$$\mathbb{P}_\gamma \left\{ \sup_{\theta \in \Theta} \frac{p_\theta}{p_\gamma} > e^v \right\} \leq C_{p_0}^{\text{aff}} (1+v)^{p_0+1} e^{-v}.$$

The likelihood ratios are evaluated on $\{p_\gamma > 0\}$, which has full P_γ -probability.

Proof. Fix $\gamma \in \Theta$. Partition Θ into Hellinger shells

$$\mathcal{S}_\ell := \{\theta \in \Theta : \ell \leq h(\gamma, \theta) < \ell + 1\}, \quad \ell = 0, 1, 2, \dots$$

In every nonempty shell, choose a maximal collection $\{\theta_{\ell,j} : 1 \leq j \leq N_\ell\}$ whose pairwise distances are strictly greater than one. Lemma C.2, applied with radius $\ell + 1$, gives $N_\ell \leq [4(\ell + 1)]^{2p_0}$. By maximality of the separated family, for every $\theta \in \mathcal{S}_\ell$ there exists an index j such that $h(\theta, \theta_{\ell,j}) \leq 1$; otherwise, θ could be added while preserving pairwise distances strictly greater than one. Hence the sets $C_{\ell,j} := \{\theta \in \Theta : h(\theta, \theta_{\ell,j}) \leq 1\}$ cover $\mathcal{S}_\ell$. Continuity of h and compactness of Θ imply that each $C_{\ell,j}$ is compact, while joint convexity of squared Hellinger distance together with the affine dependence of the densities on θ implies that each $C_{\ell,j}$ is convex. Finally, the triangle inequality gives $\text{diam}_h(C_{\ell,j}) \leq 2$.

Apply Lemma C.1 to each cell. Let $\phi_{\ell,j} \in C_{\ell,j}$ be the resulting reference point and let $\bar{p}_{\ell,j} = \sup_{\theta \in C_{\ell,j}} p_\theta$. Since $h(\phi_{\ell,j}, \theta_{\ell,j}) \leq 1$ and $h(\gamma, \theta_{\ell,j}) \geq \ell$, the triangle inequality gives $h(\gamma, \phi_{\ell,j}) \geq (\ell - 1)_+$, where $a_+ = \max\{a, 0\}$.

Write $\bar{p}_{\max} = \sup_{\theta \in \Theta} p_\theta$. The cells cover Θ , so $\bar{p}_{\max} \leq \sum_{\ell,j} \bar{p}_{\ell,j}$. For $1/2 \leq \alpha < 1$, subadditivity of $x \mapsto x^\alpha$ on nonnegative numbers gives $\bar{p}_{\max}^\alpha \leq \sum_{\ell,j} \bar{p}_{\ell,j}^\alpha$. Integrating and applying (16), we obtain

$$\int_{\mathcal{X}} p_\gamma^{1-\alpha} \bar{p}_{\max}^\alpha d\nu \leq C_0(p_0)^{1/2} \sum_{\ell \geq 0} [4(\ell + 1)]^{2p_0} e^{-(1-\alpha)(\ell-1)_+^2/2}.$$

Put $\varepsilon = 1 - \alpha \in (0, 1/2]$. Throughout the remainder of the proof, C_{p_0} denotes a finite positive constant depending only on p_0 , whose value may change from line to line. After separating the first three terms and comparing the rest with a Gaussian integral,

$$\sum_{\ell \geq 0} (\ell + 1)^{2p_0} e^{-\varepsilon(\ell-1)_+^2/2} \leq C_{p_0} \left(1 + \int_0^\infty (1 + x^{2p_0}) e^{-\varepsilon x^2/8} dx \right) \leq C_{p_0} \varepsilon^{-(p_0+1/2)}.$$

The last inequality follows by substituting $y = \sqrt{\varepsilon}x$. Since $\varepsilon \leq 1$, enlarging C_{p_0} gives

$$\int_{\mathcal{X}} p_\gamma^{1-\alpha} \bar{p}_{\max}^\alpha d\nu \leq C_{p_0} (1 - \alpha)^{-(p_0+1)}.$$

Let $L_{\max} = \bar{p}_{\max}/p_{\gamma}$ on $\{p_{\gamma} > 0\}$. The last integral is $\mathbb{E}_{\gamma} L_{\max}^{\alpha}$. If $v \geq 1$, choose $\alpha = 1 - (1 + v)^{-1} \in [1/2, 1)$. Markov's inequality gives

$$\mathbb{P}_{\gamma}\{L_{\max} > e^v\} \leq e^{-\alpha v} \mathbb{E}_{\gamma} L_{\max}^{\alpha} \leq C_{p_0}(1 + v)^{p_0+1} e^{-v/(1+v)} \leq C_{p_0}(1 + v)^{p_0+1} e^{-v}.$$

For $0 \leq v < 1$, the same form follows from the trivial bound $\mathbb{P}_{\gamma}\{L_{\max} > e^v\} \leq 1$ after enlarging the constant. Denoting the resulting constant by $C_{p_0}^{\text{aff}}$ proves the proposition. $\square$

All likelihood envelopes used above are measurable because the dependence on θ is continuous and their suprema can be taken over countable dense subsets of the relevant compact subsets of Θ .

Proof of Lemma 2.4. For each i , let ν_i be the actual distribution of Y_i . To apply Proposition C.3, we need an affine family of normalized densities that contains the constant density one. We therefore enlarge the set of trace one matrices to the trace ball of positive semidefinite matrices

$$\mathbb{S}_{+, \leq 1}^k := \{M = M^{\top} \succeq 0 : \text{tr } M \leq 1\}, \quad f_{i,M}(y) := 1 - \text{tr } M + y^{\top} M y.$$

These functions are nonnegative because $1 - \text{tr } M \geq 0$ and $y^{\top} M y \geq 0$. They depend affinely on M , and the second moment assumption gives

$$\int_{\mathbb{R}^k} f_{i,M}(y) d\nu_i(y) = 1 - \text{tr } M + \text{tr}(M \mathbb{E}[Y_i Y_i^{\top}]) = 1 - \text{tr } M + \text{tr } M = 1.$$

Hence they are densities relative to the coordinate laws ν_i . At $M = 0$, every density equals one, so the corresponding product law is exactly the true joint distribution of the observations.

The space of symmetric $k \times k$ matrices has dimension $k(k+1)/2$. Put $p_d = (d+1)(d+2)/2$. Since $k \leq d+1$, Proposition C.3 gives

$$\mathbb{P}\left\{\sup_{M \in \mathbb{S}_{+, \leq 1}^k} \prod_{i=1}^n f_{i,M}(Y_i) > e^v\right\} \leq C_{p_d}^{\text{aff}}(1 + v)^{p_d+1} e^{-v}, \quad v \geq 0.$$

Every feasible matrix A in the theorem belongs to $\mathbb{S}_{+, \leq 1}^k$ and satisfies $f_{i,A}(Y_i) = Y_i^{\top} A Y_i$. Therefore the event in the theorem, with threshold v , is contained in the event on the left-hand side of the preceding inequality.

It remains to absorb the polynomial factor into the logarithmic correction. Choose $a_d \geq p_d + 1$ sufficiently large that $C_{p_d}^{\text{aff}}(a_d + 2)^{p_d+1} e^{p_d+1-a_d} \leq 1$. Such a choice is possible because exponential decay in a_d dominates polynomial growth. For $u \geq e$, set $v = u + a_d \log u$. Since $1 \leq u$ and $\log u \leq u$, one has $1 + v \leq (a_d + 2)u$. Consequently,

$$C_{p_d}^{\text{aff}}(1 + v)^{p_d+1} e^{-v} \leq e^{-u} C_{p_d}^{\text{aff}}(a_d + 2)^{p_d+1} u^{p_d+1-a_d} \leq e^{-u}.$$

The last inequality uses $u \geq e$ and the choice of a_d . This proves the theorem. $\square$

D PROOF OF THE MULTIVARIATE CHERNOFF BOUND

Let $P_1, \dots, P_N$ be Borel probability measures on $\mathbb{R}^d$ with finite second moments, and let X denote the canonical random vector on $\mathbb{R}^d$. Write

$$\mu_i := \mathbb{E}_{P_i}[X], \quad \Sigma_i := \text{Cov}_{P_i}(X), \quad \lambda_i := \|\Sigma_i\|_{\text{op}} > 0.$$

For a proposed common center z , the quantity $\|\mu_i - z\|^2 / \lambda_i$ is its squared distance from the i th mean in units of that distribution's largest variance. Define

$$R := \min_{z \in \mathbb{R}^d} \max_{1 \leq i \leq N} \frac{\|\mu_i - z\|^2}{\lambda_i}.$$

Thus $\sqrt{R}$ is the smallest $\rho \geq 0$ such that the closed balls $\bar{B}(\mu_i, \rho\sqrt{\lambda_i})$, $1 \leq i \leq N$, have a common point. The minimum is attained because the objective is continuous and tends to infinity as $\|z\| \rightarrow \infty$.

We first consider the setting of Lemma 3.1, in which every normalized distance equals the common maximum at a minimizing center and strictly positive weights make the weighted displacements from that center sum to zero. The boundary equalities place z^* on every optimally inflated ball, while the balance relation is equivalent to the first-order optimality condition for the maximum of the normalized squared distances: the corresponding positive gradient multipliers may be taken proportional to $\gamma_i \lambda_i$.

The argument applies to arbitrary configurations of the means, including coincident points and configurations contained in a proper affine subspace. Convex optimality together with Carathéodory's theorem [2, Section I.2] will later identify a subset of at most $d+1$ distributions that satisfies the same boundary and balance conditions while preserving the optimal squared normalized radius.

The proof of Lemma 3.1 separates the leading coefficient from the remainder. We first bound a weighted distributional affinity by a scalar entry of a matrix constructed from the first two moments. We then expand this matrix entry around zero separation. A recursive coefficient estimate will give the loss $R/2$, while a relative perturbation argument will bound all higher order terms uniformly over the covariance scales.

D.1 Reduction to at most $d+1$ distributions

Lemma 3.1 applies when every constraint in the family is tight at a minimizing center and strictly positive weights satisfy the required balance relation. A general finite family need not have these properties for every distribution. Convex optimality and Carathéodory's theorem allow us to retain at most $d+1$ distributions while preserving the optimal squared normalized radius and the required balance relation.

Corollary D.1 (Carathéodory reduction and Chernoff bound). *Let $P_1, \dots, P_N$ be Borel probability measures on $\mathbb{R}^d$ with finite second moments, means μ_i , covariance matrices Σ_i , and positive covariance scales $\lambda_i = \|\Sigma_i\|_{\text{op}}$. Define*

$$R := \min_{z \in \mathbb{R}^d} \max_{1 \leq i \leq N} \frac{\|\mu_i - z\|^2}{\lambda_i}.$$

If $R > 0$, there exist a minimizing center z^ , a subset $I \subset \{1, \dots, N\}$ with $2 \leq |I| \leq \min\{N, d+1\}$, and weights $\gamma_i > 0$ for $i \in I$, with $\sum_{i \in I} \gamma_i = 1$, such that*

$$\|\mu_i - z^*\|^2 = R\lambda_i \quad (i \in I), \quad \sum_{i \in I} \gamma_i (\mu_i - z^*) = 0,$$

and the subfamily $(P_i)_{i \in I}$ has the same optimal squared normalized radius R .

Consequently, for every finite N , if $0 < R \leq [1024(d+1)^2]^{-1}$, then

$$C^*(P_1, \dots, P_N) \geq \frac{R}{2} - 7000(d+1)^3 R^2.$$

The proof is given in Appendix D.9.

This lower bound complements the upper bound in Corollary 2.3. Section 4.1 uses their incompatibility under an empty-intersection assumption to construct a common center.

D.2 Normalization under the boundary and balance conditions

Under the assumptions of Lemma 3.1, translate every distribution by $-z^*$, so that the minimizing center becomes the origin. This common translation preserves both R and the multivariate Chernoff divergence. Define the normalized mean vectors by $m_i := \mu_i / \sqrt{R}$. The boundary equalities now give $\|m_i\| = \sqrt{\lambda_i}$. Thus the normalized mean vectors retain the individual covariance scales, while the observed separation enters explicitly through $\sqrt{R}$.

Dividing the barycentric balance relation by $\sqrt{R}$ gives

$$\gamma_i > 0, \quad \sum_{i=1}^N \gamma_i = 1, \quad \sum_{i=1}^N \gamma_i m_i = 0. \quad (19)$$

Relabel the distributions so that $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$. We will combine their moment matrices in this order.

We next introduce the cumulative weights and barycenters that determine the matrix interpolation parameters. Define $\Gamma_j = \sum_{i=1}^j \gamma_i$ for $1 \leq j \leq N$, and let $t_j = \gamma_j/\Gamma_j$ for $j \geq 2$. Since every γ_i is positive, $0 < t_j < 1$. Define

$$a_1 = \lambda_1, \quad b_1 = m_1, \quad a_j = (1 - t_j)a_{j-1} + t_j\lambda_j, \quad b_j = (1 - t_j)b_{j-1} + t_jm_j. \quad (20)$$

The quantity a_j is a scalar covariance scale, and $b_j \in \mathbb{R}^d$ is the normalized barycenter of the first j mean vectors with weights proportional to $\gamma_1, \dots, \gamma_j$. Induction gives

$$a_j = \frac{1}{\Gamma_j} \sum_{i=1}^j \gamma_i \lambda_i, \quad b_j = \frac{1}{\Gamma_j} \sum_{i=1}^j \gamma_i m_i.$$

In particular, the balance relation (19) implies $b_N = 0$. Jensen's inequality and $\|m_i\|^2 = \lambda_i$ give

$$\|b_j\|^2 \leq \frac{1}{\Gamma_j} \sum_{i=1}^j \gamma_i \|m_i\|^2 = a_j. \quad (21)$$

The scale ordering also implies $a_j \leq \lambda_j$.

D.3 Bounding affinity by matrix geometric means

For positive definite matrices A, B and $0 < \theta < 1$, the weighted matrix geometric mean is $A \#_\theta B := A^{1/2}(A^{-1/2}BA^{-1/2})^\theta A^{1/2}$; see [4, 18, 38]. For positive scalars, this reduces to $a \#_\theta b = a^{1-\theta}b^\theta$. We use the standard integral representation [4]

$$A \#_\theta B = c_\theta \int_0^\infty t^{\theta-1} (A^{-1} + tB^{-1})^{-1} dt, \quad c_\theta := \frac{\sin(\pi\theta)}{\pi}. \quad (22)$$

The map $(A, B) \mapsto A \#_\theta B$ is monotone in each argument and jointly concave [4, 18]. These properties can also be seen directly from the variational formula

$$v^\top (A^{-1} + tB^{-1})^{-1} v = \inf_{y \in \mathbb{R}^{d+1}} \{y^\top A y + t^{-1}(v - y)^\top B(v - y)\}, \quad t > 0.$$

For fixed y , the expression inside braces is linear in (A, B) ; its infimum is therefore concave. Increasing either matrix increases every expression being minimized, which gives monotonicity. Integrating preserves both properties. The scalar inequality $x^\theta \leq (1 - \theta) + \theta x$, applied through spectral calculus, also yields $A \#_\theta B \preceq (1 - \theta)A + \theta B$.

For a real parameter w , define the augmented upper moment matrix

$$M_i(w) := \begin{pmatrix} 1 & w m_i^\top \\ w m_i & \lambda_i I_d + w^2 m_i m_i^\top \end{pmatrix}.$$

The parameter w varies in the expansion; the actual distributional moments correspond to $w = \sqrt{R}$. The factorization

$$M_i(w) = \begin{pmatrix} 1 & 0 \\ w m_i & I_d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \lambda_i I_d \end{pmatrix} \begin{pmatrix} 1 & w m_i^\top \\ 0 & I_d \end{pmatrix}$$

shows that $M_i(w)$ is positive definite for every real w .

Let $v(x) = (1, x^\top)^\top$. For block notation on $(d+1) \times (d+1)$ matrices, we index the leading scalar coordinate by 1 and write x for the remaining d -dimensional block; thus H_{11} , H_{x1} , and H_{xx} denote the corresponding blocks of a matrix H . The actual augmented moment matrix of P_i is

$$S_i := \mathbb{E}_{P_i}[v(X)v(X)^\top] = \begin{pmatrix} 1 & \mu_i^\top \\ \mu_i & \Sigma_i + \mu_i \mu_i^\top \end{pmatrix}.$$

Since $\mu_i = \sqrt{R} m_i$ and $\Sigma_i \preceq \lambda_i I_d$,

$$M_i(\sqrt{R}) - S_i = \begin{pmatrix} 0 & 0 \\ 0 & \lambda_i I_d - \Sigma_i \end{pmatrix} \succeq 0.$$

Thus the matrices $M_i(\sqrt{R})$ retain the means exactly and replace each covariance by its isotropic upper bound.

The interpolation weights used for the matrix geometric means must be chosen carefully. Given $0 < r \leq s$ and an averaging weight $0 < t < 1$, let $\ell = (1-t)r + ts$. We choose a geometric mean exponent θ so that its scalar output equals this arithmetic average: $r^{1-\theta}s^\theta = \ell$. For $r < s$, taking logarithms determines θ uniquely. Define

$$\vartheta(r, s; t) := \begin{cases} \frac{\log(\ell/r)}{\log(s/r)}, & r < s, \\ t, & r = s, \end{cases} \quad \Psi(r, s; t) := \begin{cases} \frac{\vartheta(r, s; t) - t}{\frac{s-r}{2r}}, & r < s, \\ \frac{t(1-t)}{2r}, & r = s. \end{cases}$$

The value assigned to Ψ at equal scales is its continuous limit. Indeed, with $\varepsilon = (s-r)/r$,

$$\vartheta(r, s; t) = \frac{\log(1+t\varepsilon)}{\log(1+\varepsilon)} = t + \frac{t(1-t)}{2}\varepsilon + O(\varepsilon^2),$$

so $(\vartheta(r, s; t) - t)/(s-r) \rightarrow t(1-t)/(2r)$. Concavity of the logarithm gives $\vartheta(r, s; t) \geq t$, and hence $\Psi(r, s; t) \geq 0$.

For $j \geq 2$, set $\theta_j = \vartheta(a_{j-1}, \lambda_j; t_j)$. This is well defined because $a_{j-1} \leq \lambda_{j-1} \leq \lambda_j$. Now define

$$F_1(w) := M_1(w), \quad F_j(w) := F_{j-1}(w) \#_{\theta_j} M_j(w), \quad 2 \leq j \leq N. \quad (23)$$

At $w = 0$, the matrices are block diagonal, and the choice of θ_j ensures $F_j(0) = \text{diag}(1, a_j I_d)$. Thus the zeroth order covariance block follows the arithmetic scale average in (20). The next subsection will show that the first-order mean block follows the corresponding vector average b_j .

Let $\mathcal{G}_j(B_1, \dots, B_j)$ denote the same nested mean applied to arbitrary positive definite inputs, with the exponents $\theta_2, \dots, \theta_j$ fixed. Then $F_j(w) = \mathcal{G}_j(M_1(w), \dots, M_j(w))$. Joint concavity and monotonicity of the binary mean imply, by induction, that $\mathcal{G}_j$ is jointly concave and monotone in all its arguments.

The scalar weights associated with the final nested mean form the vector $\beta = (\beta_1, \dots, \beta_N)$, where

$$\beta_1 := \prod_{k=2}^N (1 - \theta_k), \quad \beta_i := \theta_i \prod_{k=i+1}^N (1 - \theta_k), \quad 2 \leq i \leq N.$$

They are positive and sum to one. At each stage, the previous weights are multiplied by $1 - \theta_j$, and the new input receives weight θ_j . In particular, for positive scalars x_i , $\mathcal{G}_N(x_1, \dots, x_N) = \prod_{i=1}^N x_i^{\beta_i}$. Repeated application of the matrix arithmetic-geometric mean inequality also gives $\mathcal{G}_N(B_1, \dots, B_N) \preceq \sum_{i=1}^N \beta_i B_i$.

Lemma D.2. *For the weight vector β defined above,*

$$\int_{\mathbb{R}^d} \prod_{i=1}^N p_i(x)^{\beta_i} d\nu(x) \leq [F_N(\sqrt{R})]_{11},$$

where $\nu = N^{-1} \sum_{i=1}^N P_i$ and $p_i = dP_i/d\nu$.

Proof. We regularize the rank-one matrices appearing in the integrand so that every matrix mean is evaluated on positive definite inputs. For $\eta > 0$, define $B_{i,\eta}(x) := p_i(x)v(x)v(x)^\top + \eta I_{d+1}$. Joint concavity gives

$$\int_{\mathbb{R}^d} \mathcal{G}_N(B_{1,\eta}(x), \dots, B_{N,\eta}(x)) d\nu(x) \preceq \mathcal{G}_N(S_1 + \eta I_{d+1}, \dots, S_N + \eta I_{d+1}).$$

Since $S_i \preceq M_i(\sqrt{R})$, put $\widetilde{M}_{i,\eta} := M_i(\sqrt{R}) + \eta I_{d+1}$. Monotonicity bounds the right-hand side by $\mathcal{G}_N(\widetilde{M}_{1,\eta}, \dots, \widetilde{M}_{N,\eta})$.

For each fixed x , all the matrices $B_{i,\eta}(x)$ commute. On the line spanned by $v(x)$, their eigenvalues are $p_i(x)\|v(x)\|^2 + \eta$, while on its orthogonal complement every eigenvalue equals η . The scalar formula for the nested mean therefore shows that

$$\mathcal{G}_N(B_{1,\eta}(x), \dots, B_{N,\eta}(x)) \longrightarrow \left(\prod_{i=1}^N p_i(x)^{\beta_i} \right) v(x)v(x)^\top \quad (\eta \downarrow 0).$$

The factor $\|v(x)\|^2$ has the correct power because $\sum_i \beta_i = 1$.

For $0 < \eta \leq 1$, the arithmetic upper bound gives

$$\|\mathcal{G}_N(B_{1,\eta}(x), \dots, B_{N,\eta}(x))\|_{\text{op}} \leq \sum_{i=1}^N \beta_i p_i(x) \|v(x)\|^2 + 1.$$

This dominating function is integrable, since $\int_{\mathbb{R}^d} p_i(x) \|v(x)\|^2 d\nu(x) = 1 + \mathbb{E}_{P_i}[\|X\|^2] < \infty$. Dominated convergence on the left and continuity of the matrix mean on positive definite inputs on the right yield

$$\int_{\mathbb{R}^d} \left(\prod_{i=1}^N p_i(x)^{\beta_i} \right) v(x) v(x)^\top d\nu(x) \preceq F_N(\sqrt{R}).$$

The first coordinate of $v(x)$ is one, so taking the $(1, 1)$ entry gives the desired affinity bound. $\square$

The remainder of the section is organized around the scalar certificate $f_N(w) := [F_N(w)]_{11}$:

$$\mathcal{W}_\beta \leq f_N(\sqrt{R}), \quad f_N(w) = 1 - c_N w^2 + O(N^3 w^4), \quad c_N \geq \frac{1}{2}.$$

The next three subsections derive the coefficient recursion, introduce the scalar functional used to propagate it, and prove the sharp bound $c_N \geq 1/2$; the subsequent perturbation estimate makes the fourth order term uniform over all ratios of covariance scales.

D.4 Second-order expansion of the matrix geometric mean

The affinity certificate reduces the problem to the behavior of $[F_N(w)]_{11}$ near $w = 0$. To compute its quadratic term, we first analyze one binary geometric mean. The calculation distinguishes the exponent θ , which weights the incoming quadratic coefficients, from the averaging weight t , which determines the first-order mean coordinate.

Lemma D.3. *Suppose $\mathcal{A}(w)$ and $\mathcal{B}(w)$ are analytic matrix-valued functions near zero, with even diagonal blocks and odd off-diagonal blocks, and, for some real coefficients c_A, c_B ,*

$$\begin{aligned} \mathcal{A}(w) &= \begin{pmatrix} 1 - c_A w^2 + O(w^4) & w(u)^\top + O(w^3) \\ wu + O(w^3) & rI_d + O(w^2) \end{pmatrix}, \\ \mathcal{B}(w) &= \begin{pmatrix} 1 - c_B w^2 + O(w^4) & w(v)^\top + O(w^3) \\ wv + O(w^3) & sI_d + O(w^2) \end{pmatrix}, \end{aligned}$$

where $0 < r \leq s$ and $u, v \in \mathbb{R}^d$. Let $0 < t < 1$, $\ell = (1-t)r + ts$, and $\theta = \vartheta(r, s; t)$. Then

$$\begin{aligned} (\mathcal{A} \#_\theta \mathcal{B})_{11} &= 1 - \left[(1-\theta)c_A + \theta c_B + \Psi(r, s; t) \|u - v\|^2 \right] w^2 + O(w^4), \\ (\mathcal{A} \#_\theta \mathcal{B})_{x1} &= w((1-t)u + tv) + O(w^3), \\ (\mathcal{A} \#_\theta \mathcal{B})_{xx} &= \ell I_d + O(w^2). \end{aligned}$$

The proof is given in Appendix I.

The hypotheses of the lemma hold for the nested matrices near zero. Each input $M_i(w)$ has even diagonal blocks and odd off-diagonal blocks, and this parity is preserved by the geometric mean. Starting from $F_1 = M_1$ and applying the lemma successively gives the full block expansion

$$F_j(w) = \begin{pmatrix} 1 - c_j w^2 + O(w^4) & w b_j^\top + O(w^3) \\ w b_j + O(w^3) & a_j I_d + O(w^2) \end{pmatrix}.$$

In particular,

$$[F_j(w)]_{11} = 1 - c_j w^2 + O(w^4), \tag{24}$$

where

$$c_1 = 0, \quad c_j = (1 - \theta_j) c_{j-1} + \Psi(a_{j-1}, \lambda_j; t_j) \|b_{j-1} - m_j\|^2. \tag{25}$$

The recursion has a direct interpretation. Each step retains a fraction of the coefficient already accumulated and adds a nonnegative contribution from the discrepancy between the current average mean and the next mean.

The remaining difficulty is quantitative: nonnegativity of the increments in (25) does not by itself imply $c_N \geq 1/2$. The obstruction is that the geometric mean exponent θ_j and the barycentric weight t_j differ when the covariance scales are unequal. The needed lower bound can instead be encoded by a scalar functional determined by two moment constraints.

D.5 An auxiliary functional for the coefficient bound

Fix $0 < r \leq s$ and $u \in \mathbb{R}^d$ with $\|u\|^2 \leq r$. There is a unique pair $(\omega, \chi) \in [0, \sqrt{r}] \times [0, 1]$ satisfying

$$(1 - \chi)\omega^2 + \chi s = r, \quad (1 - \chi)\omega - \chi\sqrt{s} = \|u\|. \quad (26)$$

Equivalently, the two-point distribution supported at ω and $-\sqrt{s}$, with masses $1 - \chi$ and χ , has first moment $\|u\|$ and second moment r . When $u \neq 0$, orienting the supporting line along u gives vector mean u . For $r < s$, the solution is

$$\omega = \frac{r + \|u\| \sqrt{s}}{\sqrt{s} + \|u\|}, \quad \chi = \frac{r - \omega^2}{s - \omega^2},$$

while for $r = s$ one has $\omega = \sqrt{r}$ and $\chi = (1 - \|u\|/\sqrt{r})/2$.

These constraints define the associated functional

$$\Phi_s(r, u) := \Psi(\omega^2, s; \chi)(\omega + \sqrt{s})^2,$$

where (ω, χ) is the pair in (26) and Ψ is extended continuously to $\chi = 0, 1$. For fixed upper scale s , the functional records the quadratic coefficient contributed by the corresponding two-point distribution.

For $s > 1$ and $0 \leq \rho < 1$, define the normalized coefficient

$$\Omega(s, \rho) := \frac{\Phi_s(1, \rho \mathbf{e})}{1 - \rho^2}, \quad (27)$$

where $\mathbf{e}$ is any unit vector. Rotational invariance of the defining two-point constraints makes this definition independent of the choice of $\mathbf{e}$. The scalar properties needed for the induction are collected in the following lemma and proved in Appendix J.

Lemma D.4. *For $0 < r \leq s$ and $\|u\|^2 \leq r$,*

$$0 \leq \Phi_s(r, u) \leq \frac{1}{2} \left(1 - \frac{\|u\|^2}{r} \right).$$

For fixed (r, u) , the map $s \mapsto \Phi_s(r, u)$ is nonincreasing on $[r, \infty)$. In the normalization (27), for each fixed $s > 1$ the map $\rho \mapsto \Omega(s, \rho)$ is nonincreasing on $[0, 1]$. Moreover,

$$\Phi_s(r, 0) = \frac{1}{2}, \quad \Phi_s(r, u) = 0 \quad \text{when } \|u\|^2 = r. \quad (28)$$

At equal scales,

$$\Phi_r(r, u) = \frac{1}{2} \left(1 - \frac{\|u\|^2}{r} \right). \quad (29)$$

The proof is given in Appendix J. The normalized coefficient makes explicit the monotonicity in the residual mean magnitude that is used in the propagation argument.

The induction also uses the following propagation property.

Lemma D.5. *Let $0 < r \leq s$, $0 < t < 1$, and put $\ell = (1 - t)r + ts$ and $w = (1 - t)u + tv$. If $\|u\|^2 \leq r$ and $\|v\|^2 = s$, then $\|w\|^2 \leq \ell$ and*

$$(1 - \vartheta(r, s; t))\Phi_s(r, u) + \Psi(r, s; t)\|u - v\|^2 \geq \Phi_s(\ell, w). \quad (30)$$

The proof is given in Appendix K. The key reduction is one-dimensional: among vectors v with $\|v\|^2 = s$, the antiparallel orientation gives the smallest updated norm. In that configuration, the representation defined by (26) is preserved by the update, and the coefficient identity reduces to the composition rule for Ψ .

D.6 The sharp coefficient bound

The functional yields an invariant directly in the variables (a_j, b_j, λ_j) .

Proposition D.6. *The coefficients defined in (24) satisfy*

$$c_j \geq \Phi_{\lambda_j}(a_j, b_j), \quad 1 \leq j \leq N. \quad (31)$$

In particular, $c_N \geq 1/2$.

Proof. At $j = 1$, one has $a_1 = \lambda_1$ and $\|b_1\|^2 = \|m_1\|^2 = \lambda_1$. Hence (28) gives $\Phi_{\lambda_1}(a_1, b_1) = 0 = c_1$.

Assume (31) holds at stage $j - 1$. Since $\lambda_j \geq \lambda_{j-1}$, Lemma D.4 gives

$$c_{j-1} \geq \Phi_{\lambda_{j-1}}(a_{j-1}, b_{j-1}) \geq \Phi_{\lambda_j}(a_{j-1}, b_{j-1}).$$

Apply Lemma D.5 with

$$r = a_{j-1}, \quad s = \lambda_j, \quad t = t_j, \quad u = b_{j-1}, \quad v = m_j.$$

The assumptions hold by (21) and $\|m_j\|^2 = \lambda_j$. Moreover, (20) gives $\ell = a_j$ and $w = (1 - t_j)b_{j-1} + t_j m_j = b_j$. Combining the coefficient recursion (25) with the preceding lower bound and (30) yields

$$\begin{aligned} c_j &\geq (1 - \theta_j) \Phi_{\lambda_j}(a_{j-1}, b_{j-1}) + \Psi(a_{j-1}, \lambda_j; t_j) \|b_{j-1} - m_j\|^2 \\ &\geq \Phi_{\lambda_j}(a_j, b_j), \end{aligned}$$

which proves (31) by induction.

Finally, the balance relation gives $b_N = 0$. Therefore (28) implies $c_N \geq \Phi_{\lambda_N}(a_N, 0) = \frac{1}{2}$. $\square$

At the terminal step the balance condition forces the residual mean to vanish, and the functional then equals $1/2$ for every remaining scale ratio. The resulting coefficient is therefore independent of both the ambient dimension and the heterogeneity of the covariance scales.

D.7 Uniform control of the remainder

The coefficient bound must be paired with a remainder estimate that is uniform over all ratios of covariance scales. The following proposition provides the required analytic control.

Proposition D.7. *For the ordered matrices in (23), the function $f_N(w) = [F_N(w)]_{11}$ extends holomorphically to a neighborhood of $|w| \leq 1/(16N)$ and is even. With c_N as in (24),*

$$|f_N(w) - (1 - c_N w^2)| \leq 7000N^3 |w|^4, \quad |w| \leq \frac{1}{32N}.$$

The bound is independent of the ambient dimension and the ratios of covariance scales.

Proof of Proposition D.7. We use the relative error $\text{err}_{A_0}(A) = \|A_0^{-1/2}(A - A_0)A_0^{-1/2}\|_{\text{op}}$ and apply Lemma H.1 to the matrices $M_i(w)$. Put $u_i = m_i/\sqrt{\lambda_i}$, so that $\|u_i\| = 1$. Direct normalization gives

$$M_i(0)^{-1/2} M_i(w) M_i(0)^{-1/2} = \begin{pmatrix} 1 & w(u_i)^\top \\ w u_i & I_d + w^2 u_i(u_i)^\top \end{pmatrix}.$$

The off-diagonal perturbation has operator norm $|w|$, and the lower-right perturbation has operator norm $|w|^2$. Hence, also for complex w ,

$$\text{err}_{M_i(0)}(M_i(w)) \leq |w| + |w|^2. \quad (32)$$

The right-hand side is the same for every i , regardless of λ_i .

The error map $T(\varepsilon) = \varepsilon/(1 - 2\varepsilon)$ in Lemma H.1 has the explicit iteration formula $T^{\circ k}(\varepsilon) = \frac{\varepsilon}{1 - 2k\varepsilon}$, as long as the denominator is positive. Indeed, the formula is immediate for $k = 0$, and applying T once more replaces $1 - 2k\varepsilon$ by $1 - 2(k + 1)\varepsilon$.

Choose $\rho = 1/(16N)$ and $\varepsilon_0 = 17/(256N)$. For $|w| \leq \rho$, inequality (32) gives $|w| + |w|^2 \leq \frac{1}{16N} + \frac{1}{256N^2} \leq \varepsilon_0$. At the j th stage, the first input to the binary mean has undergone $j-2$ previous perturbation steps, while the new input still has error at most ε_0 . Since $T(\varepsilon) \geq \varepsilon$, both inputs can be bounded by $T^{(j-2)}(\varepsilon_0)$. Induction using Lemma H.1 therefore gives

$$\text{err}_{F_j(0)}(F_j(w)) \leq T^{(j-1)}(\varepsilon_0) = \frac{\varepsilon_0}{1 - 2(j-1)\varepsilon_0}, \quad |w| \leq \rho.$$

All intermediate errors are less than $1/2$, so each application of the lemma is valid. At the final stage,

$$\text{err}_{F_N(0)}(F_N(w)) \leq \frac{17}{222N + 34} \leq \frac{17}{222N}, \quad |w| \leq \rho.$$

The baseline matrix is $F_N(0) = \text{diag}(1, a_N I_d)$. Its first diagonal entry is one, so the relative-error bound directly controls the scalar function $f_N(w) = [F_N(w)]_{11}$. We obtain $|f_N(w) - 1| \leq 17/(222N)$ for $|w| \leq \rho$. The strict margins in the perturbation estimates also show that f_N is holomorphic on a neighborhood of this closed disk.

To identify the powers that can occur in its Taylor expansion, let $J = \text{diag}(1, -I_d)$. The identity $M_i(-w) = JM_i(w)J$ and invariance of the geometric mean under conjugation by J imply $F_N(-w) = JF_N(w)J$ and $f_N(-w) = f_N(w)$. Thus f_N is even. Its second-order coefficient is the one computed in (24), and we may write

$$f_N(w) = 1 - c_N w^2 + \sum_{k=2}^{\infty} a_{2k} w^{2k}.$$

Cauchy's coefficient estimate on the circle $|w| = \rho$ gives $|a_{2k}| \leq \frac{17}{222N} \rho^{-2k}$ for $k \geq 2$. For $|w| \leq \rho/2$, put $x = |w|/\rho = 16N|w| \leq 1/2$. Summing the resulting geometric series yields

$$|f_N(w) - (1 - c_N w^2)| \leq \frac{17}{222N} \sum_{k=2}^{\infty} x^{2k} = \frac{17}{222N} \cdot \frac{x^4}{1 - x^2}.$$

Since $1 - x^2 \geq 3/4$, we conclude that

$$|f_N(w) - (1 - c_N w^2)| \leq 7000N^3 |w|^4, \quad |w| \leq \frac{1}{32N}.$$

Indeed, the coefficient obtained directly from the preceding estimate is $\frac{17}{222N} \cdot \frac{(16N)^4}{1-1/4} = \frac{4456448}{666} N^3 < 7000N^3$. This establishes the required fourth order remainder uniformly over all ratios of covariance scales. $\square$

D.8 Proof of the multivariate Chernoff divergence bound

Proof of Lemma 3.1. Let β be the weight vector generated by the ordered nested mean. If $0 < R \leq \frac{1}{1024N^2}$, then Proposition D.7 applies at the actual separation parameter $w = \sqrt{R}$.

Write $\mathcal{W}_\beta = \int_{\mathbb{R}^d} \prod_{i=1}^N p_i^{\beta_i} d\nu$. Lemma D.2, Propositions D.6 and D.7 give

$$\mathcal{W}_\beta \leq f_N(\sqrt{R}) \leq 1 - c_N R + 7000N^3 R^2 \leq 1 - \frac{R}{2} + 7000N^3 R^2.$$

If $\mathcal{W}_\beta = 0$, then $C^*(P_1, \dots, P_N) = +\infty$, and the conclusion is immediate. Otherwise, the definition of the multivariate Chernoff divergence and the inequality $-\log x \geq 1 - x$ for $x > 0$ imply

$$C^*(P_1, \dots, P_N) \geq -\log \mathcal{W}_\beta \geq 1 - \mathcal{W}_\beta \geq \frac{R}{2} - 7000N^3 R^2.$$

This is the asserted bound. $\square$

D.9 Proof of the Carathéodory reduction

Proof. For each i , write $f_i(z) = \frac{\|\mu_i - z\|^2}{\lambda_i}$ and $f(z) = \max_{1 \leq i \leq N} f_i(z)$. The function f is continuous and coercive, so it attains its minimum at some z^* . Let $I_{\text{act}} = \{i : f_i(z^*) = R\}$ be the active index set at z^* .

Each f_i is differentiable and convex, with gradient $\nabla f_i(z) = \frac{2(z - \mu_i)}{\lambda_i}$. The subdifferential of a finite maximum of differentiable convex functions is the convex hull of the gradients corresponding to the active indices. Since z^* minimizes f , we therefore have

$$0 \in \text{conv} \left\{ \frac{\mu_i - z^*}{\lambda_i} : i \in I_{\text{act}} \right\}.$$

Carathéodory's theorem [2, Section I.2] gives a subset $I \subset I_{\text{act}}$ containing at most $d + 1$ indices and coefficients $\alpha_i \geq 0$ such that

$$\sum_{i \in I} \alpha_i = 1, \quad \sum_{i \in I} \alpha_i \frac{\mu_i - z^*}{\lambda_i} = 0.$$

Discarding any index with zero coefficient leaves strictly positive weights. Define $\gamma_i := (\alpha_i / \lambda_i) / \sum_{j \in I} (\alpha_j / \lambda_j)$ for $i \in I$. Then $\gamma_i > 0$, $\sum_{i \in I} \gamma_i = 1$, and $\sum_{i \in I} \gamma_i (\mu_i - z^*) = 0$.

Because every retained index is active, $\frac{\|\mu_i - z^*\|^2}{\lambda_i} = R$ for $i \in I$. We next verify that the retained subfamily has the same minimum. For any $z \in \mathbb{R}^d$, convexity gives $f_i(z) \geq f_i(z^*) + \nabla f_i(z^*)^\top (z - z^*)$. Multiply by α_i and sum over I . The gradient terms cancel, while $f_i(z^*) = R$ for every $i \in I$, so $\sum_{i \in I} \alpha_i f_i(z) \geq R$. Since a maximum is at least any convex combination, $\max_{i \in I} f_i(z) \geq \sum_{i \in I} \alpha_i f_i(z) \geq R$. At $z = z^*$, every retained constraint equals R . Hence

$$\min_{z \in \mathbb{R}^d} \max_{i \in I} \frac{\|\mu_i - z\|^2}{\lambda_i} = R.$$

Thus the subfamily has the same minimizing center and optimal squared normalized radius and satisfies the boundary and balance conditions of Lemma 3.1.

The subset contains at least two indices: a singleton would force $\mu_i = z^*$ by the balance relation, which would give $f_i(z^*) = 0$ and contradict $R > 0$. Therefore $2 \leq |I| \leq \min\{N, d + 1\}$.

Under the stated bound on R , $\sqrt{R} \leq \frac{1}{32(d+1)} \leq \frac{1}{32|I|}$. Lemma 3.1 therefore applies to the retained subfamily and gives

$$C^*((P_i)_{i \in I}) \geq \frac{R}{2} - 7000|I|^3 R^2.$$

Every admissible weight vector for the multivariate Chernoff divergence on this subfamily extends to the full family by assigning zero weight to indices outside I . Because factors with zero weight are omitted,

$$C^*(P_1, \dots, P_N) \geq C^*((P_i)_{i \in I}).$$

Since $|I| \leq d + 1$, the claimed bound follows. $\square$

E CONSTRUCTION WITH THE THEORETICAL THRESHOLD

We record the construction with the theoretical threshold used to expose the statistical mechanism in detail. Each accepted candidate specifies a Euclidean ball of possible estimates. The main point is to prove that these balls have a common intersection, even though their centers and covariance scales may differ arbitrarily. The computational construction in Section 5 uses the same deterministic intersection argument with computable thresholds.

Throughout this section, let a_d be the likelihood calibration constant from Lemma 2.4, and put

$$L = \log(1/\delta), \quad T_d = L + a_d \log L, \quad s_d = \frac{T_d}{n}, \quad b_d = 7000(d + 1)^3. \quad (33)$$

Define the squared radius factor

$$\kappa_d(x) = 2x(1 + 16b_d x), \quad x \geq 0. \quad (34)$$

The leading term $2x$ corresponds to the coefficient $1/2$ in the Chernoff divergence lower bound. The additional term compensates for the quadratic remainder in that bound.

At threshold T_d , a candidate (m, V) is accepted if $\mathcal{L}_n(m, V) \leq T_d$, and we associate with it the ball

$$B(m, V) = \overline{B}\left(m, \sqrt{\kappa_d(s_d) \|V\|_{\text{op}}}\right).$$

This is the specialization of the main construction to $(T, s) = (T_d, s_d)$. When $V = 0$, the ball is the singleton $\{m\}$.

The set of accepted candidates is nonempty for every realized sample. Indeed, define the empirical mean and covariance

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j, \quad \hat{\Sigma}_n = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)(X_j - \bar{X}_n)^\top.$$

The centered sample vectors span $\text{Range}(\hat{\Sigma}_n)$, so every observation belongs to $\bar{X}_n + \text{Range}(\hat{\Sigma}_n)$. Moreover, the empirical measure $\hat{P}_n$ has mean $\bar{X}_n$ and covariance $\hat{\Sigma}_n$, hence $\hat{P}_n \in \mathcal{M}(\bar{X}_n, \hat{\Sigma}_n)$. Lemma 2.1 therefore gives

$$\mathcal{L}_n(\bar{X}_n, \hat{\Sigma}_n) = 0. \quad (35)$$

Thus $(\bar{X}_n, \hat{\Sigma}_n)$ is always accepted.

We may take a_d to be a positive integer by increasing it if necessary. The required parameter restrictions are verified in Appendix F.1.

E.1 Acceptance of (μ, Σ)

Let the observations have mean μ and covariance $\Sigma \succeq 0$, and write $r = \text{rank}(\Sigma)$. We first verify that the pair (μ, Σ) satisfies the hypotheses of the likelihood maximal inequality, including when Σ is singular.

If $r \geq 1$, write $\Sigma = U\Lambda U^\top$ with $U^\top U = I_r$ and $\Lambda \succ 0$. For any vector a in the null space of Σ , $\mathbb{E}[(a^\top (X_i - \mu))^2] = a^\top \Sigma a = 0$. Apply this identity to a finite basis of the null space. It follows that $X_i - \mu$ is orthogonal to that null space, and hence belongs to $\text{Range}(\Sigma)$ almost surely.

The normalized observations $Z_i = \Lambda^{-1/2} U^\top (X_i - \mu)$ therefore satisfy $\mathbb{E}Z_i = 0$ and $\mathbb{E}[Z_i Z_i^\top] = I_r$. After adjoining the constant coordinate, the vectors $Y_i = \begin{pmatrix} 1 \\ Z_i \end{pmatrix} \in \mathbb{R}^{r+1}$ satisfy

$$\mathbb{E}[Y_i Y_i^\top] = \begin{pmatrix} 1 & 0 \\ 0 & I_r \end{pmatrix} = I_{r+1}.$$

Since $r + 1 \leq d + 1$ and $L \geq e$, Lemma 2.4 gives

$$\mathbb{P}\{\mathcal{L}_n(\mu, \Sigma) > T_d\} \leq e^{-L} = \delta. \quad (36)$$

If $\Sigma = 0$, then $X_i = \mu$ almost surely. The rank zero statistic is identically zero, so (36) also holds in this case. Thus the pair (μ, Σ) is accepted with probability at least $1 - \delta$, including when Σ is singular.

E.2 Common intersection and measurability

The intersection of every family of at most $d + 1$ balls associated with accepted candidates is established in Appendix F.2.

The argument compares an upper and a lower bound on the multivariate Chernoff divergence. Put $s = s_d$ and suppose that at most $d + 1$ balls associated with accepted candidates have empty intersection. Corollary 2.3 supplies representing distributions with $C^* \leq s$. Lemma F.1 allows a common Gaussian convolution with squared normalized radius $\kappa_d(s)$. Data processing (Lemma B.1) preserves the upper bound, whereas Corollary D.1 gives, for the smoothed family,

$$C^* \geq \frac{\kappa_d(s)}{2} - b_d \kappa_d(s)^2 > s.$$

The strict inequality follows from the chosen inflation in (34) and the parameter restrictions. This contradiction establishes the finite intersections needed below.

Helly's theorem in $\mathbb{R}^d$ [2, 14] now implies that every finite subfamily of balls associated with accepted candidates intersects. By (35), the empirical pair is always accepted; fix its ball

$$B_0 := B(\bar{X}_n, \hat{\Sigma}_n).$$

Every finite collection of the closed sets $B_0 \cap B(m, V)$, with (m, V) accepted, has nonempty intersection. Since B_0 is compact, the finite intersection property gives a nonempty full intersection.

Thus the intersection obtained with the theoretical threshold is

$$\mathcal{C}(T_d, s_d) = \bigcap_{(m, V): \mathcal{L}_n(m, V) \leq T_d} B(m, V).$$

It is nonempty, closed, convex, and contained in the compact ball B_0 , hence compact. Its minimum-norm selector is

$$\hat{\mu}(T_d, s_d) := \arg \min_{z \in \mathcal{C}(T_d, s_d)} \|z\|^2.$$

The minimizer exists by compactness and is unique because the squared Euclidean norm is strictly convex.

On the event that (μ, Σ) is accepted, the corresponding ball is one of the sets defining $\mathcal{C}(T_d, s_d)$. Hence

$$\|\hat{\mu}(T_d, s_d) - \mu\| \leq \sqrt{\kappa_d(s_d) \|\Sigma\|_{\text{op}}}. \quad (37)$$

By (36), this event has probability at least $1 - \delta$. In particular, when $\Sigma = 0$, the ball $B(\mu, \Sigma)$ is $\{\mu\}$, so the estimator equals μ almost surely.

E.3 Comparison with the Gaussian radius

It remains to express the radius in (37) in terms of $\mathcal{Q}_\delta(\Sigma)$. Assume first that $\lambda = \|\Sigma\|_{\text{op}} > 0$. From (33) and (34),

$$\sqrt{\kappa_d(s_d)\lambda} = \sqrt{\frac{2\lambda L}{n}} \sqrt{\frac{T_d}{L}} \sqrt{1 + 16b_d \frac{T_d}{n}}.$$

See Appendix F.3 for the comparison bounds.

Combining (37), (40), and (41), we conclude that, on an event of probability at least $1 - \delta$,

$$\|\hat{\mu}(T_d, s_d) - \mu\| \leq \left[1 + C_d \left(\frac{\log L}{L} + \frac{L}{n} \right) \right] \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}}.$$

Products of the remainders have again been absorbed into C_d , after enlarging L_d and decreasing c_d if necessary. The case $\Sigma = 0$ was already established. This proves the stated statistical bound for the selector defined with the theoretical threshold. The selector is defined from the sample and the confidence level, while (μ, Σ) enters the analysis through the event that it is accepted.

E.4 Sharpness and scope of the result

We begin with the statistical lower bound used in the introduction. It shows that the leading radius in Theorem 1 cannot be improved even after restricting the model to a Gaussian location family.

The lower benchmark is proved in Proposition G.1.

The estimator may depend on the confidence level. For a fixed candidate (m, V) , computing the quadratic statistic amounts to maximizing a concave function over a finite-dimensional convex matrix set. Section 5 addresses the additional task of computing the intersection over all accepted candidates. An isotropic reduction and a polynomial characterization of acceptance give a polynomial-time algorithm for fixed d for the estimator used in Theorem 1.

Finally, the comparison with $\mathcal{Q}_\delta(\Sigma)$ uses the top eigenvalue scale and a scalar Gaussian quantile bound. It establishes the first-order Gaussian radius in the high-confidence regime, uniformly over covariance shapes, including singular matrices. Determining the next order contribution of the full covariance spectrum and Gaussian sharp accuracy at fixed confidence remains open.

Remark E.1 (Fixed dimension versus growing dimension). The constants C_d , L_d , and c_d may depend on d . The theorem therefore applies to each fixed dimension as $L \rightarrow \infty$ and $L/n \rightarrow 0$; regimes in which d grows with n or L are outside its scope.

F TECHNICAL DETAILS FOR THE ESTIMATOR

F.1 Parameter range

We first specify a parameter range in which the geometric argument applies. Choose $L_d \geq e$ sufficiently large that $a_d \log L \leq L$ for every $L \geq L_d$, and choose

$$c_d \leq \frac{1}{2} \min \left\{ \frac{1}{64b_d}, \frac{1}{4096(d+1)^2} \right\}.$$

Then $L \geq L_d$ and $L/n \leq c_d$ imply

$$0 < s_d = \frac{T_d}{n} \leq \frac{2L}{n} \leq \min \left\{ \frac{1}{64b_d}, \frac{1}{4096(d+1)^2} \right\}.$$

In particular, $b_d s_d \leq \frac{1}{64}$ and $\kappa_d(s_d) \leq \frac{5}{2} s_d < \frac{1}{1024(d+1)^2}$. We may enlarge L_d or decrease c_d later when comparing the resulting radius with the Gaussian benchmark.

F.2 Intersection of at most $d+1$ confidence balls

We next fix the observed sample and argue deterministically. Suppose that a collection of balls associated with accepted candidates has cardinality at most $d+1$. A single ball is nonempty, so it suffices to consider $2 \leq N \leq d+1$.

Assume, toward a contradiction, that

$$\bigcap_{i=1}^N B(m_i, V_i) = \emptyset.$$

Write $v_i = \|V_i\|_{\text{op}}$ and define

$$R_V := \inf_{z \in \mathbb{R}^d} \max_{1 \leq i \leq N} \frac{\|m_i - z\|^2}{v_i}.$$

For $v_i = 0$, interpret the corresponding ratio as zero when $z = m_i$ and as $+\infty$ otherwise. This convention expresses exactly the constraint imposed by a ball of radius zero.

The objective defining R_V is lower semicontinuous and has compact finite sublevel sets. Consequently, whenever its infimum is finite, that infimum is attained. Moreover,

$$\left\{ z : \max_{1 \leq i \leq N} \frac{\|m_i - z\|^2}{v_i} \leq \kappa_d(s_d) \right\} = \bigcap_{i=1}^N B(m_i, V_i).$$

The assumed empty intersection therefore implies

$$R_V > \kappa_d(s_d). \tag{38}$$

By Lemma 2.1, each accepted candidate admits a probability measure $P_i \in \mathcal{M}(m_i, V_i)$ such that $D_{\text{KL}}(\hat{P}_n \| P_i) \leq s_d$. The empirical measure $\hat{P}_n$ is the first argument in each of these relative entropy bounds. Corollary 2.3 therefore gives

$$C^*(P_1, \dots, P_N) \leq s_d. \tag{39}$$

Let Σ_i be the covariance matrix of P_i , and put $\lambda_i = \|\Sigma_i\|_{\text{op}} \leq \|V_i\|_{\text{op}} = v_i$.

Two obstacles prevent a direct application of Lemma 3.2. Some λ_i may be zero, and the separation may lie outside the range where the local quadratic bound is useful. Convolving every distribution with the same Gaussian resolves both issues while preserving the upper bound (39).

For $\tau > 0$, define $P_i^{(\tau)} = P_i * \mathcal{N}(0, \tau I_d)$. The smoothed law has mean m_i , covariance $\Sigma_i + \tau I_d$, and covariance scale $\lambda_i + \tau > 0$. For $\tau \geq 0$, define the squared normalized radius

$$R(\tau) := \inf_{z \in \mathbb{R}^d} \max_{1 \leq i \leq N} \frac{\|m_i - z\|^2}{\lambda_i + \tau},$$

using at $\tau = 0$ the same convention as above when $\lambda_i = 0$.

Lemma F.1. *The function $R : [0, \infty) \rightarrow [0, \infty]$ is nonincreasing, continuous on $(0, \infty)$, and satisfies $\lim_{\tau \downarrow 0} R(\tau) = R(0)$ and $\lim_{\tau \rightarrow \infty} R(\tau) = 0$.*

Proof. For each $\tau > 0$, the objective is continuous and coercive in z , so the infimum is attained. On a compact interval of positive τ values, evaluation at a fixed center gives a uniform upper bound on the minimum. Any minimizing center then lies in a fixed compact set because $\|m_i - z\|^2 \leq R(\tau)(\lambda_i + \tau)$ bounds its distance from m_i . Joint continuity of the objective on this compact set shows that $R(\tau)$ is continuous for $\tau > 0$.

Increasing τ increases every denominator, so $R(\tau)$ is nonincreasing, and evaluation at any fixed center gives $\lim_{\tau \rightarrow \infty} R(\tau) = 0$. The inequality $R(\tau) \leq R(0)$ is immediate. For the reverse endpoint inequality, suppose that the increasing limit as $\tau \downarrow 0$ is finite and choose minimizing centers z_j for a sequence $\tau_j \downarrow 0$. Their objective values are bounded, so the centers are bounded. Along a subsequence, $z_j \rightarrow z^*$, and $\|m_i - z_j\|^2 \leq R(\tau_j)(\lambda_i + \tau_j)$ passes to the limit for every i . If $\lambda_i > 0$, this bounds the limiting ratio by $\lim_{\tau \downarrow 0} R(\tau)$; if $\lambda_i = 0$, it forces $z^* = m_i$. Hence $R(0)$ is no larger than that limit. If the limit is infinite, the reverse inequality is automatic. $\square$

Since $\lambda_i \leq v_i$, we have $R(0) \geq R_V$. Together with (38), the continuity and endpoint behavior of $R(\tau)$ imply that some $\tau > 0$ satisfies $R(\tau) = \kappa_d(s_d)$.

The smoothed laws now have positive covariance scales and separation within the range of Corollary D.1. That corollary gives

$$C^*(P_1^{(\tau)}, \dots, P_N^{(\tau)}) \geq \frac{\kappa_d(s_d)}{2} - b_d \kappa_d(s_d)^2.$$

The chosen inflation makes this lower bound strictly greater than s_d . Indeed,

$$\frac{\kappa_d(s_d)}{2} - b_d \kappa_d(s_d)^2 - s_d = b_d s_d^2 [16 - 4(1 + 16b_d s_d)^2].$$

Because $b_d s_d \leq 1/64$, the bracket is at least $16 - 4(5/4)^2 > 0$.

On the other hand, the same Gaussian convolution is applied to every P_i . The data-processing inequality from Lemma B.1, followed by (39), gives

$$C^*(P_1^{(\tau)}, \dots, P_N^{(\tau)}) \leq C^*(P_1, \dots, P_N) \leq s_d.$$

This contradiction proves that every family of at most $d + 1$ balls associated with accepted candidates has nonempty intersection.

F.3 Details of the comparison with the Gaussian radius

We first bound the two inflation factors. Since $T_d/L = 1 + a_d \log L/L$ and $T_d \leq 2L$, the elementary inequality $\sqrt{1+x} \leq 1 + x/2$ for $x \geq 0$ gives

$$\sqrt{\frac{T_d}{L}} \leq 1 + \frac{a_d \log L}{2L}, \quad \sqrt{1 + 16b_d \frac{T_d}{n}} \leq 1 + 16b_d \frac{L}{n}.$$

Multiplying these inequalities and absorbing the product term into a constant depending only on d , we obtain

$$\sqrt{\frac{T_d}{L}} \sqrt{1 + 16b_d \frac{T_d}{n}} \leq 1 + C_d \left(\frac{\log L}{L} + \frac{L}{n} \right). \quad (40)$$

For example, the product term is bounded by a constant multiple of L/n because $\log L/L \leq 1$ for $L \geq e$.

Next, let u be a unit eigenvector of Σ corresponding to λ . For $Y \sim \mathcal{N}(0, \Sigma)$, one has $\langle u, Y \rangle \sim \mathcal{N}(0, \lambda)$ and $|\langle u, Y \rangle| \leq \|Y\|$. If z_δ is the upper δ -quantile of $|Z|$ for $Z \sim \mathcal{N}(0, 1)$, stochastic domination therefore gives $\mathcal{Q}_\delta(\Sigma) \geq \sqrt{\lambda} z_\delta$.

We use the following scalar estimate. For all sufficiently large L , one has $z_\delta \geq 1$, and the Gaussian tail bounds imply

$$c z_\delta^{-1} e^{-z_\delta^2/2} \leq \mathbb{P}\{|Z| > z_\delta\} = e^{-L} \leq 2e^{-z_\delta^2/2}$$

for a universal constant $c > 0$. The upper tail bound first gives $z_\delta^2 \leq 2(L + \log 2)$. Taking logarithms in the lower tail bound then yields

$$z_\delta^2 \geq 2L - 2 \log z_\delta + 2 \log c \geq 2L - C \log L.$$

After increasing the lower threshold on L , the quantity $C \log L/(2L)$ is at most $1/2$. Taking square roots and using $(1 - x)^{-1/2} \leq 1 + C'x$ on that interval gives

$$\sqrt{2\lambda L} \leq \left(1 + C \frac{\log L}{L}\right) \mathcal{Q}_\delta(\Sigma). \quad (41)$$

For completeness, the matching upper comparison used in the introduction is also uniform over covariance shapes. Write $Y = \Sigma^{1/2}G$ with $G \sim \mathcal{N}(0, I_d)$. Since $\|Y\| \leq \sqrt{\lambda} \|G\|$ pointwise,

$$\mathcal{Q}_\delta(\Sigma) \leq \sqrt{\lambda} q_{d,\delta},$$

where $q_{d,\delta}$ is the upper δ -quantile of $\|G\|$. Put $R_G = \|G\|^2 \sim \chi_d^2$. For $L > 1$, Chernoff's bound with $t = (1 - 1/L)/2$ gives

$$\mathbb{P}\{R_G > x\} \leq (1 - 2t)^{-d/2} e^{-tx} = L^{d/2} \exp\left\{-\frac{1 - 1/L}{2}x\right\}.$$

Taking $x = 2L + (d + 4) \log L + 4$, the right-hand side is at most e^{-L} once L is sufficiently large depending only on d . Hence

$$q_{d,\delta}^2 \leq 2L + O_d(\log L), \quad \mathcal{Q}_\delta(\Sigma) \leq \sqrt{2\lambda L} \left[1 + O_d\left(\frac{\log L}{L}\right)\right].$$

Combining this with (41) yields, uniformly over all nonzero $\Sigma \succeq 0$,

$$\mathcal{Q}_\delta(\Sigma) = \sqrt{2\|\Sigma\|_{\text{op}} L} \left[1 + O_d\left(\frac{\log L}{L}\right)\right].$$

The lower comparison has a universal constant; only the upper comparison depends on the fixed dimension.

G SHARPNESS: PROOF AND ADDITIONAL DISCUSSION

Proposition G.1 (Gaussian lower benchmark). *Fix $d \geq 1$, $n \geq 1$, $0 < \delta < 1/2$, and a nonzero covariance matrix $\Sigma \succeq 0$. Let $\lambda = \|\Sigma\|_{\text{op}}$ and let Φ denote the standard normal distribution function. Suppose an estimator $\tilde{\mu}$ and a deterministic radius $r \geq 0$ satisfy $\sup_{\mu \in \mathbb{R}^d} \mathbb{P}_{\mathcal{N}(\mu, \Sigma)^{\otimes n}}\{\|\tilde{\mu} - \mu\| > r\} \leq \delta$. Then $r \geq \sqrt{\lambda/n} \Phi^{-1}(1 - \delta)$. Consequently, for fixed d and $L = \log(1/\delta) \rightarrow \infty$, one has $r \geq [1 - O_d(\log L/L)] \mathcal{Q}_\delta(\Sigma)/\sqrt{n}$ uniformly over nonzero $\Sigma \succeq 0$.*

G.1 Proof of Proposition G.1

Proof. Let u be a unit eigenvector of Σ with eigenvalue λ . If the first claim were false, choose h with $r < h < \sqrt{\lambda/n} \Phi^{-1}(1 - \delta)$ and consider the Gaussian laws with means $\mu_+ = hu$ and $\mu_- = -hu$ and common covariance Σ . The Euclidean balls of radius r around these two means are disjoint, so the assumed confidence guarantee induces a test between the two hypotheses with error at most δ under each hypothesis.

For n observations, the optimal equal prior test depends only on the projection onto u and has error $\Phi(-h\sqrt{n/\lambda})$, which is strictly larger than δ by the choice of h . This contradiction proves the first claim. For the consequence, both $\Phi^{-1}(1 - \delta)$ and the upper δ quantile of $\|G\|$ for $G \sim \mathcal{N}(0, I_d)$ equal $\sqrt{2L}[1 + O_d(\log L/L)]$ in fixed dimension. Since $\|\Sigma^{1/2}G\| \leq \sqrt{\lambda}\|G\|$, the stated uniform comparison follows. $\square$

There are two distinct sharpness statements in the paper. Proposition G.1 shows that the final confidence radius in Theorem 1 is first order unavoidable, already within the Gaussian location model. We now record the separate fact that the leading coefficient $1/2$ in the multivariate Chernoff divergence lower bound used by the construction is itself exact, again within a Gaussian family. To see this, let $u_1, \dots, u_{d+1}$ be the vertices of a centered regular simplex in $\mathbb{R}^d$, normalized so that $\|u_i\| = 1$ and $\sum_{i=1}^{d+1} u_i = 0$. For $R > 0$, take $P_i = \mathcal{N}(\sqrt{R}u_i, I_d)$ for $1 \leq i \leq d+1$. Each covariance scale equals one. For any center z ,

$$\max_i \left\| \sqrt{R}u_i - z \right\|^2 \geq \frac{1}{d+1} \sum_{i=1}^{d+1} \left\| \sqrt{R}u_i - z \right\|^2 = R + \|z\|^2.$$

At $z = 0$, all squared distances equal R . Thus the optimal squared normalized radius is exactly R .

Let φ_i denote the Lebesgue density of P_i , and fix $\beta \in \Delta_{d+1}$. Write $\bar{u}_\beta = \sum_i \beta_i u_i$. Completing the square gives

$$\sum_{i=1}^{d+1} \beta_i \left\| x - \sqrt{R}u_i \right\|^2 = \left\| x - \sqrt{R}\bar{u}_\beta \right\|^2 + R(1 - \|\bar{u}_\beta\|^2).$$

Since $\sum_i \beta_i = 1$, integration of the Gaussian density therefore yields

$$\int_{\mathbb{R}^d} \prod_{i=1}^{d+1} \varphi_i(x)^{\beta_i} dx = \exp \left\{ -\frac{R}{2} (1 - \|\bar{u}_\beta\|^2) \right\}.$$

The corresponding Chernoff exponent is at most $R/2$, and the uniform weights attain this value because their weighted mean is zero. Hence

$$C^*(P_1, \dots, P_{d+1}) = \frac{R}{2}.$$

Thus the coefficient $1/2$ in the leading term is optimal for a multivariate Chernoff divergence inequality with a remainder of order R^2 . The regular simplex extends the equilateral triangle example from dimension two to every dimension.

The sources of the dependence on the dimension can be isolated as follows. The relative entropy representation in Lemma 2.1 is exact in every dimension. The quantitative remainder in the binary matrix mean estimates and the propagation inequality depends on the number of distributions being combined and is independent of the ambient dimension. Carathéodory's theorem reduces that number to at most $d + 1$. Helly's theorem then uses intersections of at most $d + 1$ balls associated with accepted candidates to obtain the global intersection. Separately, the likelihood calibration depends on the affine dimension of the matrix parameter space, which is bounded by $p_d = \frac{(d+1)(d+2)}{2}$. These are the sources of the dimension dependence in the final constants.

H RELATIVE PERTURBATION OF THE MATRIX GEOMETRIC MEAN

This appendix proves the relative perturbation estimate used in the proof of Proposition D.7. Each perturbation is measured relative to its own positive definite baseline, which makes the constants uniform over ratios of covariance scales.

For a positive definite matrix A_0 and a possibly complex matrix A of the same size, define $\text{err}_{A_0}(A) = \left\| A_0^{-1/2}(A - A_0)A_0^{-1/2} \right\|_{\text{op}}$. This normalization removes the scale and condition number of A_0 from the perturbation estimate.

Lemma H.1. *Let $A_0, B_0 \succ 0$, let $0 < \theta < 1$, and suppose that*

$$\text{err}_{A_0}(A) \leq \varepsilon, \quad \text{err}_{B_0}(B) \leq \varepsilon, \quad 0 \leq \varepsilon < \frac{1}{2}.$$

The integral representation (22) defines an analytic continuation of $A \#_{\theta} B$ to this complex neighborhood, and

$$\text{err}_{A_0 \#_{\theta} B_0}(A \#_{\theta} B) \leq T(\varepsilon), \quad T(\varepsilon) = \frac{\varepsilon}{1 - 2\varepsilon}.$$

The bound is independent of the matrix size, the baseline matrices, and θ .

Proof. We first record how relative error behaves under inversion. Write $A = A_0^{1/2}(I + E_A)A_0^{1/2}$ with $\|E_A\|_{\text{op}} \leq \varepsilon < 1$. The Neumann series gives

$$(I + E_A)^{-1} - I = -E_A + E_A^2 - E_A^3 + \cdots, \quad \|(I + E_A)^{-1} - I\|_{\text{op}} \leq \frac{\varepsilon}{1 - \varepsilon}.$$

Since $A^{-1} = A_0^{-1/2}(I + E_A)^{-1}A_0^{-1/2}$, we obtain

$$\text{err}_{A_0^{-1}}(A^{-1}) \leq \frac{\varepsilon}{1 - \varepsilon}.$$

The same estimate holds for B^{-1} .

Next, relative error bounds are preserved by positive sums. To verify this, suppose that $X_{j,0} \succ 0$ and $X_j - X_{j,0} = X_{j,0}^{1/2} E_j X_{j,0}^{1/2}$ with $\|E_j\|_{\text{op}} \leq \eta$. For arbitrary conformable vectors u and v , the operator norm bound and Cauchy-Schwarz give

$$\left| u^* \sum_j (X_j - X_{j,0}) v \right| \leq \eta \sum_j \left\| X_{j,0}^{1/2} u \right\| \left\| X_{j,0}^{1/2} v \right\| \leq \eta \left(\sum_j u^* X_{j,0} u \right)^{1/2} \left(\sum_j v^* X_{j,0} v \right)^{1/2}.$$

The symbol $*$ denotes conjugate transpose. After normalization by $\sum_j X_{j,0}$, this yields the relative-error bound η . The same argument, with sums replaced by integrals, applies to positive weighted integrals whenever the baseline integral is finite.

For each $s > 0$, put $Q(s) = A^{-1} + sB^{-1}$ and $Q_0(s) = A_0^{-1} + sB_0^{-1}$. The first inversion and the property for positive sums give

$$\text{err}_{Q_0(s)}(Q(s)) \leq \frac{\varepsilon}{1 - \varepsilon} < 1.$$

A second inversion therefore gives

$$\text{err}_{Q_0(s)^{-1}}(Q(s)^{-1}) \leq \frac{\varepsilon}{1 - 2\varepsilon} = T(\varepsilon).$$

This estimate is uniform in s .

Finally, integrate against the positive weight $c_{\theta} s^{\theta-1}$ in (22). The baseline integral is $A_0 \#_{\theta} B_0$, so the property for positive integrals yields the stated bound.

All inverses above are well defined because their normalized perturbations have norm strictly less than one. On any smaller closed complex neighborhood, the same estimates dominate the integrand by a constant multiple of the integrable baseline. Thus the integral depends holomorphically on analytic matrix inputs and agrees with the usual geometric mean for positive definite inputs. $\square$

I SECOND-ORDER EXPANSION FOR TWO MATRICES

This appendix verifies the block expansion used to obtain the recursion (25). The notation is that of Lemma D.3; the analytic justification uses Lemma H.1.

Proof of Lemma D.3. Writing τ for the integration variable in (22), block inversion gives

$$\begin{aligned} (\mathcal{A}^{-1})_{11} &= 1 + \left(c_A + \frac{\|u\|^2}{r} \right) w^2 + O(w^4), \\ (\mathcal{A}^{-1})_{x1} &= -\frac{w}{r} u + O(w^3), \quad (\mathcal{A}^{-1})_{xx} = r^{-1} I_d + O(w^2). \end{aligned}$$

The formulas for $\mathcal{B}^{-1}$ are obtained by replacing (c_A, u, r) with (c_B, v, s) .

Let $Q = \mathcal{A}^{-1} + \tau \mathcal{B}^{-1}$. Its upper-left inverse entry is the reciprocal of the Schur complement $Q_{11} - Q_{1x} Q_{xx}^{-1} Q_{x1}$. At order w^2 , the terms involving the mean vectors simplify as follows:

$$\frac{\|u\|^2}{r} + \frac{\tau \|v\|^2}{s} - \frac{\|u/r + \tau v/s\|^2}{1/r + \tau/s} = \frac{\tau}{s + r\tau} \left(\|u\|^2 - 2\langle u, v \rangle + \|v\|^2 \right) = \frac{\tau \|u - v\|^2}{s + r\tau}.$$

Consequently,

$$(Q^{-1})_{11} = \frac{1}{1 + \tau} - \frac{w^2}{(1 + \tau)^2} \left[c_A + \tau c_B + \frac{\tau \|u - v\|^2}{s + r\tau} \right] + O(w^4).$$

The off-diagonal inverse formula $(Q^{-1})_{x1} = -Q_{xx}^{-1} Q_{x1} (Q^{-1})_{11}$ gives

$$(Q^{-1})_{x1} = \frac{w}{1 + \tau} \left(u + \frac{r\tau}{s + r\tau} (v - u) \right) + O(w^3).$$

We integrate these coefficients against $c_\theta \tau^{\theta-1} d\tau$. This coefficientwise integration is justified by Lemma H.1 in Appendix H: on a sufficiently small complex circle, the integrand is dominated by an integrable multiple of its positive definite baseline. Cauchy's coefficient formula then permits interchange of coefficient extraction and integration.

The coefficients of c_A and c_B follow from

$$c_\theta \int_0^\infty \frac{\tau^{\theta-1}}{(1 + \tau)^2} d\tau = 1 - \theta, \quad c_\theta \int_0^\infty \frac{\tau^\theta}{(1 + \tau)^2} d\tau = \theta.$$

For $r < s$, the coefficient determining the mean interpolation is

$$c_\theta \int_0^\infty \frac{r\tau^\theta}{(s + r\tau)(1 + \tau)} d\tau = \frac{r^{1-\theta} s^\theta - r}{s - r} = \frac{\ell - r}{s - r} = t.$$

The stated expression follows from the identities $\frac{r\tau}{(s+r\tau)(1+\tau)} = \frac{r}{s-r} \left(\frac{s}{s+r\tau} - \frac{1}{1+\tau} \right)$ and $c_\theta \int_0^\infty \tau^{\theta-1} / (c + \tau) d\tau = c^{\theta-1}$, where the second identity holds for any $c > 0$.

Finally, the coefficient of $\|u - v\|^2$ is

$$c_\theta \int_0^\infty \frac{\tau^\theta}{(s + r\tau)(1 + \tau)^2} d\tau = \frac{\theta - t}{s - r} = \Psi(r, s; t).$$

Indeed, multiplying the identity $\frac{1}{(1+\tau)^2} - \frac{r}{(s+r\tau)(1+\tau)} = \frac{s-r}{(s+r\tau)(1+\tau)^2}$ by $c_\theta \tau^\theta$ and integrating gives the claimed value from the preceding two integrals.

These calculations establish the upper-left and off-diagonal expansions. At $w = 0$, the lower-right block is $r^{1-\theta} s^\theta I_d = \ell I_d$, which proves the remaining expansion. When $r = s$, all formulas follow by continuity, with $\theta = t$ and $\Psi(r, r; t) = t(1 - t)/(2r)$. $\square$

J PROPERTIES OF THE AUXILIARY FUNCTIONAL

We retain the scale variables r, s from Section 3. The defining pair (ω, χ) is determined by the moment constraints (26); a logarithmic reparametrization is used only in the monotonicity argument.

J.1 A scalar representation of the auxiliary functional

For $r < s$ and $\|u\| < \sqrt{r}$, the constraints (26) give $\omega = \frac{r+\|u\|\sqrt{s}}{\sqrt{s}+\|u\|}$ and $\chi = \frac{r-\omega^2}{s-\omega^2}$. The identities $(1-\chi)\omega^2 + \chi s = r$ and $(1-\chi)\omega - \chi\sqrt{s} = \|u\|$ are immediate, and the variance identity for a two-point distribution gives

$$r - \|u\|^2 = \chi(1-\chi)(\omega + \sqrt{s})^2. \quad (42)$$

The boundary cases follow by continuity. If $r = s$, then $\omega = \sqrt{r}$ and $\chi = (1 - \|u\|/\sqrt{r})/2$, which yields (29) directly from $\Psi(r, r; \chi) = \chi(1-\chi)/(2r)$.

The functional is scale invariant:

$$\Phi_{cs}(cr, \sqrt{c}u) = \Phi_s(r, u), \quad c > 0. \quad (43)$$

Indeed, the corresponding ω is multiplied by $\sqrt{c}$, χ is unchanged, and $\Psi(cr, cs; t) = c^{-1}\Psi(r, s; t)$.

For the remaining assertions, scale invariance allows us to normalize $r = 1$ while retaining $s > 1$ as the upper scale. Put $\rho := \|u\| \in [0, 1)$ and recall the normalized coefficient Ω from (27); thus $\Omega(s, \rho) = \Phi_s(1, u)/(1 - \rho^2)$. Introduce the auxiliary scalar coordinates

$$\sigma := \frac{1}{4} \log s, \quad \sigma_\rho := \operatorname{artanh}\left(\tanh(\sigma) \frac{1-\rho}{1+\rho}\right), \quad g(x) := x \coth(2x), \quad g(0) := \frac{1}{2}.$$

We now derive the representation used below. The definition of σ_ρ is equivalent to $\omega = e^{-2\sigma_\rho}$, $s = e^{4\sigma}$, where ω is the positive support point in (26). Since $(1-\chi)\omega^2 + \chi s = 1$, the interpolation exponent $\alpha := \vartheta(\omega^2, s; \chi)$ satisfies

$$\alpha = \frac{\log(1/\omega^2)}{\log(s/\omega^2)} = \frac{\sigma_\rho}{\sigma + \sigma_\rho}, \quad \chi = \frac{1 - \omega^2}{s - \omega^2}.$$

By (42), $1 - \rho^2 = \chi(1-\chi)(\omega + \sqrt{s})^2$. Therefore the definition of Φ_s gives

$$\Omega(s, \rho) = \frac{\Psi(\omega^2, s; \chi)}{\chi(1-\chi)} = \frac{(\alpha - \chi)(s - \omega^2)}{(1 - \omega^2)(s - 1)}.$$

Using $s = e^{4\sigma}$ and $\omega^2 = e^{-4\sigma_\rho}$, the last expression becomes

$$\Omega(s, \rho) = \frac{\sigma_\rho}{2(\sigma + \sigma_\rho)} (\coth(2\sigma) + \coth(2\sigma_\rho)) - \frac{1}{2} (\coth(2\sigma) - 1).$$

Collecting the two coth terms yields the convenient form

$$\Omega(s, \rho) = \frac{1}{2} - \frac{g(\sigma) - g(\sigma_\rho)}{2(\sigma + \sigma_\rho)}. \quad (44)$$

J.2 Bounds and monotonicity

Proof of Lemma D.4. It is enough to consider $s > 1$ and $0 < \rho < 1$; the boundary cases follow by continuity. Differentiation gives

$$\frac{dg(x)}{dx} = \coth(2x) - \frac{2x}{\sinh^2(2x)} = \frac{\frac{1}{2} \sinh(4x) - 2x}{\sinh^2(2x)} \geq 0.$$

Since $0 < \sigma_\rho < \sigma$, formula (44) implies $\Omega(s, \rho) \leq 1/2$.

Set $J(\sigma, \sigma_\rho) := \frac{g(\sigma) - g(\sigma_\rho)}{\sigma + \sigma_\rho}$, so $\Omega = \frac{1}{2} - \frac{1}{2}J$. For fixed σ ,

$$\frac{\partial J}{\partial \sigma_\rho} = -\frac{(\sigma + \sigma_\rho) \frac{dg(\sigma_\rho)}{d\sigma_\rho} + g(\sigma) - g(\sigma_\rho)}{(\sigma + \sigma_\rho)^2} \leq 0.$$

As ρ increases, σ_ρ decreases, so J is nondecreasing and Ω is nonincreasing in ρ .

For fixed ρ , the identity $\tanh(\sigma_\rho) = \frac{1-\rho}{1+\rho} \tanh(\sigma)$ gives $\frac{d\sigma_\rho}{d\sigma} = \frac{\sinh(2\sigma_\rho)}{\sinh(2\sigma)}$. A direct simplification yields

$$\frac{dJ(\sigma, \sigma_\rho(\sigma))}{d\sigma} = \frac{(\sigma_\rho \sinh(2\sigma) - \sigma \sinh(2\sigma_\rho))(\sinh(2(\sigma + \sigma_\rho)) + 2(\sigma + \sigma_\rho))}{(\sigma + \sigma_\rho)^2 \sinh^2(2\sigma) \sinh(2\sigma_\rho)} \geq 0.$$

The last inequality follows because $x \mapsto \sinh(2x)/x$ is increasing on $(0, \infty)$. Hence Ω is nonincreasing in σ , equivalently in s .

By (27) and (43), returning to general r replaces the normalized upper scale by s/r . Hence $s \mapsto \Phi_s(r, u)$ is nonincreasing for fixed (r, u) . Nonnegativity follows directly from $\Psi \geq 0$, while $\Omega(s, \rho) \leq 1/2$ and scale invariance give $\Phi_s(r, u) \leq \frac{1}{2} \left(1 - \frac{\|u\|^2}{r}\right)$. If $\rho = 0$, then $\sigma_\rho = \sigma$, so (44) gives $\Omega(s, 0) = 1/2$ and therefore $\Phi_s(r, 0) = 1/2$. If $\rho \uparrow 1$, the factor $1 - \rho^2$ in (27) tends to zero while Ω stays bounded by $1/2$, giving $\Phi_s(r, u) = 0$ when $\|u\|^2 = r$. Together with the equal-scale calculation above, all claims follow. $\square$

We record two scalar facts used in the propagation proof. First, the endpoint value of the normalized coefficient is

$$\Omega(s, 1) := \lim_{\rho \uparrow 1} \Omega(s, \rho) = \frac{1}{\log s} - \frac{1}{s-1}, \quad s > 1. \quad (45)$$

This follows from (44) by setting $\sigma_\rho = 0$.

Second, the coefficient generated by a binary interpolation can be compared with $1/2$ through the position of its arithmetic mean relative to the geometric mean.

Lemma J.1. *Let $0 < a \leq b$, $0 < t < 1$, and $\ell = (1-t)a + tb$. Define $\Upsilon(a, b; t) := \frac{\ell \Psi(a, b; t)}{t(1-t)}$. Then*

$$\Upsilon(a, b; t) \geq \frac{1}{2} \iff \ell \geq \sqrt{ab},$$

with equality at $\ell = \sqrt{ab}$; for $a = b$, one has $\Upsilon(a, a; t) = 1/2$.

Proof. The equal-scale case is immediate. Suppose $a < b$ and write $a = \sqrt{ab} e^{-\xi}$, $b = \sqrt{ab} e^{\xi}$, and $\ell = \sqrt{ab} e^{\kappa}$, where $\xi > 0$ and $-\xi < \kappa < \xi$. Solving for t and substituting into the definition of Υ gives

$$\Upsilon(a, b; t) = \frac{1}{2} + \frac{(\kappa/\xi) \sinh \xi - \sinh \kappa}{2(\cosh \xi - \cosh \kappa)}.$$

The denominator is positive, while

$$(\kappa/\xi) \sinh \xi - \sinh \kappa = \kappa \left(\frac{\sinh \xi}{\xi} - \frac{\sinh \kappa}{\kappa} \right)$$

has the sign of κ because $x \mapsto \sinh x/x$ is even and strictly increasing for $x > 0$. Since $\kappa \geq 0$ is equivalent to $\ell \geq \sqrt{ab}$, the claim follows. $\square$

K PROOF OF THE COEFFICIENT PROPAGATION INEQUALITY

The argument has two steps. First, the angular dependence is reduced to the antiparallel configuration. In that one-dimensional case, the two-point distribution determined by (26) is preserved by the update, and the two binary coefficients compose exactly.

Proof of Lemma D.5. Set $w = (1-t)u + tv$. The elementary variance identity

$$\|w\|^2 = (1-t)\|u\|^2 + t\|v\|^2 - t(1-t)\|u-v\|^2$$

gives

$$\ell - \|w\|^2 = (1-t)(r - \|u\|^2) + t(1-t)\|u-v\|^2 \geq 0. \quad (46)$$

Hence $\Phi_s(\ell, w)$ is well defined.

If $r = s$, then $\vartheta(r, r; t) = t$ and $\Psi(r, r; t) = t(1 - t)/(2r)$. By (29), the desired inequality is exactly (46) divided by $2r$, so equality holds.

Assume $r < s$. By scale invariance (43), normalize $r = 1$ and retain $s > 1$ for the rescaled upper scale. Write $\rho := \|u\| \in [0, 1]$ and $q := \|w\|/\sqrt{\ell}$, where

$$\ell = 1 - t + ts, \quad \theta = \frac{\log \ell}{\log s}.$$

We first take $\rho < 1$; the endpoint $\rho = 1$ follows by continuity. By (27), $\Phi_s(1, u) = (1 - \rho^2)\Omega(s, \rho)$. Moreover, (46) and $\|w\|^2 = \ell q^2$ give

$$t(1 - t)\|u - v\|^2 = \ell(1 - q^2) - (1 - t)(1 - \rho^2).$$

Multiplying this identity by $\Psi(1, s; t)/(t(1 - t))$ and using $\Upsilon(1, s; t) = \ell\Psi(1, s; t)/(t(1 - t))$ gives

$$\Psi(1, s; t)\|u - v\|^2 = \Upsilon(1, s; t)(1 - q^2) - \frac{\Psi(1, s; t)}{t}(1 - \rho^2).$$

Adding $(1 - \theta)\Phi_s(1, u)$ and dividing by $1 - q^2 > 0$ therefore yields

$$\frac{(1 - \theta)\Phi_s(1, u) + \Psi(1, s; t)\|u - v\|^2}{1 - q^2} = \Upsilon(1, s; t) + \frac{1 - \rho^2}{1 - q^2} \left[(1 - \theta)\Omega(s, \rho) - \frac{\Psi(1, s; t)}{t} \right]. \quad (47)$$

The bracket is nonnegative. Indeed, $\Omega(s, \rho)$ is nonincreasing in ρ , so (45) yields $\Omega(s, \rho) \geq \frac{1}{\log s} - \frac{1}{s-1}$. It therefore suffices to show $\frac{\Psi(1, s; t)}{t(1 - \theta)} \leq \frac{1}{\log s} - \frac{1}{s-1}$. With $y = t(s - 1)$, so that $\ell = 1 + y$, this becomes $\frac{1}{\log(1 + y)} - \frac{1}{y} \geq \frac{1}{\log s} - \frac{1}{s-1}$. The function on the left is decreasing on $(0, \infty)$, since

$$\frac{d}{dy} \left(\frac{1}{\log(1 + y)} - \frac{1}{y} \right) = \frac{1}{y^2} - \frac{1}{(1 + y)\log^2(1 + y)} < 0.$$

The last inequality is equivalent to $\log(1 + y) < y/\sqrt{1 + y}$, which follows after writing $1 + y = e^{2\eta}$ from $\eta < \sinh \eta$.

For fixed s, ρ, t , the right-hand side of (47) is nondecreasing in q , whereas $\Omega(s/\ell, q)$ is nonincreasing in q . Thus the most restrictive orientation minimizes $\|w\|$. By the reverse triangle inequality, it is enough to consider the antiparallel orientation. When $\rho > 0$, write $\hat{u} = u/\rho$; when $\rho = 0$, choose any unit vector $\hat{u}$. Thus we may take

$$u = \rho\hat{u}, \quad v = -\sqrt{s}\hat{u}, \quad \|\hat{u}\| = 1.$$

Let $\omega \in (0, 1]$ and $\chi \in [0, 1)$ be the parameters determined by (26) for $(r, s, u) = (1, s, \rho\hat{u})$. Then

$$(1 - \chi)\omega^2 + \chi s = 1, \quad (1 - \chi)\omega - \chi\sqrt{s} = \rho, \quad (48)$$

and

$$\Phi_s(1, u) = \Psi(\omega^2, s; \chi)(\omega + \sqrt{s})^2. \quad (49)$$

After the update, the weight at the endpoint $-\sqrt{s}$ is $\bar{\chi} := (1 - t)\chi + t$. Consequently,

$$(1 - \bar{\chi})\omega^2 + \bar{\chi}s = \ell, \quad (1 - \bar{\chi})\omega - \bar{\chi}\sqrt{s} = (1 - t)\rho - t\sqrt{s}. \quad (50)$$

To identify the accumulated coefficient, set $\alpha := \vartheta(\omega^2, s; \chi)$ and $\bar{\alpha} := \vartheta(\omega^2, s; \bar{\chi})$. Because $1 = \omega^2 \#_{\alpha} s$ and $\ell = 1 \#_{\theta} s$, expanding the scalar powers gives $\ell = (\omega^2)^{(1 - \alpha)(1 - \theta)} s^{\alpha(1 - \theta) + \theta} = \omega^2 \#_{(1 - \theta)\alpha + \theta} s$. Uniqueness of the exponent in the scalar geometric mean therefore gives

$$\bar{\alpha} = (1 - \theta)\alpha + \theta.$$

The two relations in (48) also imply

$$s - 1 = (1 - \chi)(s - \omega^2), \quad \rho + \sqrt{s} = (1 - \chi)(\omega + \sqrt{s}).$$

Hence, from the definition of Ψ ,

$$\Psi(1, s; t)(\rho + \sqrt{s})^2 = \frac{(\theta - t)(1 - \chi)}{s - \omega^2}(\omega + \sqrt{s})^2.$$

The first coefficient contributes

$$(1 - \theta)\Psi(\omega^2, s; \chi)(\omega + \sqrt{s})^2 = \frac{(1 - \theta)(\alpha - \chi)}{s - \omega^2}(\omega + \sqrt{s})^2.$$

Adding the last two displays, the numerator is

$$(1 - \theta)(\alpha - \chi) + (\theta - t)(1 - \chi) = \bar{\alpha} - \bar{\chi},$$

because $\bar{\chi} = (1 - t)\chi + t$ and $\bar{\alpha} = (1 - \theta)\alpha + \theta$. We have therefore proved the exact composition identity

$$(1 - \theta)\Psi(\omega^2, s; \chi)(\omega + \sqrt{s})^2 + \Psi(1, s; t)(\rho + \sqrt{s})^2 = \Psi(\omega^2, s; \bar{\chi})(\omega + \sqrt{s})^2. \quad (51)$$

By (49), the left-hand side is exactly the left-hand side of (30) in the antiparallel configuration.

It remains to compare the right-hand side of (51) with $\Phi_s(\ell, w)$. From (50),

$$\ell - ((1 - t)\rho - t\sqrt{s})^2 = \bar{\chi}(1 - \bar{\chi})(\omega + \sqrt{s})^2. \quad (52)$$

If $\ell < \omega\sqrt{s}$, then the signed mean in (50) is positive, and that system is exactly the representation defining $\Phi_s(\ell, w)$. Hence (51) gives equality in (30).

If $\ell \geq \omega\sqrt{s}$, Lemma J.1 and (52) imply

$$\Psi(\omega^2, s; \bar{\chi})(\omega + \sqrt{s})^2 \geq \frac{\ell - ((1 - t)\rho - t\sqrt{s})^2}{2\ell}.$$

By Lemma D.4,

$$\Phi_s(\ell, w) \leq \frac{1}{2} \left(1 - \frac{\|w\|^2}{\ell} \right) = \frac{\ell - ((1 - t)\rho - t\sqrt{s})^2}{2\ell}.$$

Thus (30) also holds in this case, and the proof is complete. $\square$

L COMPUTATIONAL ANALYSIS OF THE ESTIMATOR

The computational analysis has four components: threshold computation, reduction to scalar covariance bounds, a polynomial characterization of rejection, and optimization over the resulting intersection. The analysis below also establishes measurability and the statistical guarantee. Appendix M derives the computational cost of these components.

L.1 Reduction to scalar covariance bounds

Given $c > 1$ and $\zeta > 0$, accept (m, V) when $\mathcal{L}_n(m, V) \leq \log c$, and attach the ball

$$\bar{B}\left(m, \sqrt{\zeta \|V\|_{\text{op}}}\right).$$

The exact statistical threshold corresponds to $c = \exp(T_d)$ and $\zeta = \kappa_d(T_d/n)$. The thresholds constructed below satisfy the same conditions used in the statistical proof.

Lemma L.1 (Isotropic reduction). *If (m, V) is accepted and $v = \|V\|_{\text{op}}$, then (m, vI_d) is accepted and gives the same ball. If the observations are not all identical, the full intersection equals*

$$\mathcal{K}(c, \zeta) := \bigcap_{\substack{m \in \mathbb{R}^d, v > 0 \\ \mathcal{L}_n(m, vI_d) \leq \log c}} \bar{B}(m, \sqrt{\zeta v}). \quad (53)$$

If all observations equal x , the original intersection, including candidates with $V = 0$, is $\{x\}$ whenever it is nonempty.

Proof. Since $V \preceq vI_d$, every distribution in $\mathcal{M}(m, V)$ also belongs to $\mathcal{M}(m, vI_d)$. For $v > 0$, the latter moment class has no affine support restriction; for $v = 0$, the two classes are identical. The relative entropy representation in Lemma 2.1 consequently gives

$$\mathcal{L}_n(m, vI_d) \leq \mathcal{L}_n(m, V).$$

Replacing V by vI_d preserves both the center and radius of the associated ball. Since the original family already includes every isotropic candidate, restricting to these candidates leaves the intersection unchanged.

When $V = 0$, the only distribution in $\mathcal{M}(m, V)$ is the point mass at m , and $D_{\text{KL}}(\hat{P}_n \| \delta_m)$ is finite only when $X_1 = \dots = X_n = m$. Hence a zero-covariance candidate is accepted precisely when all observations equal its center. If their common value is x , then $(x, 0)$ is accepted with statistic zero and its ball is $\{x\}$. The full intersection is therefore contained in $\{x\}$ and equals this singleton whenever it is nonempty. $\square$

For the remaining construction, assume that the observations are not all equal. For the parameters used in Algorithm 1, $\zeta = \kappa_d(s)$, and Lemma L.1 gives $\mathcal{K}(c, \zeta) = \mathcal{C}(\log c, s)$. The candidates are $(m, v) \in \mathbb{R}^d \times (0, \infty)$. The isotropic reduction concerns the candidate covariance bound; the underlying distribution remains arbitrary.

L.2 Polynomial characterization of rejection

For the fixed sample, put

$$w_i = \begin{pmatrix} 1 \\ X_i \end{pmatrix}, \quad S(m, v) = \begin{pmatrix} 1 & m^\top \\ m & mm^\top + vI_d \end{pmatrix}.$$

For $v > 0$, this matrix is positive definite: for $t \in \mathbb{R}$ and $u \in \mathbb{R}^d$,

$$(t, u^\top) S(m, v) \begin{pmatrix} t \\ u \end{pmatrix} = (t + m^\top u)^2 + v \|u\|^2,$$

which is positive unless $(t, u) = 0$.

Lemma L.2 (Likelihood representation). *For $v > 0$,*

$$\exp\{\mathcal{L}_n(m, vI_d)\} = \max_{\substack{H \succeq 0 \\ \text{tr } H = 1}} \frac{\prod_{i=1}^n w_i^\top H w_i}{[\text{tr}(HS(m, v))]^n}. \quad (54)$$

Consequently, acceptance at threshold $\log c$ is equivalent to

$$\prod_{i=1}^n w_i^\top H w_i \leq c [\text{tr}(HS(m, v))]^n \quad (55)$$

for every positive semidefinite H with trace one.

Proof. The standardized vectors in (1) are $y_i = Jw_i$, where

$$J = \begin{pmatrix} 1 & 0 \\ -m/\sqrt{v} & I_d/\sqrt{v} \end{pmatrix}.$$

Since J is invertible, the change of variables $H = J^\top A J$, with inverse $A = J^{-\top} H J^{-1}$, gives

$$A \succeq 0 \iff H \succeq 0, \quad y_i^\top A y_i = w_i^\top H w_i, \quad \text{tr } A = \text{tr}(HS(m, v)).$$

The trace identity follows from $J^{-1} J^{-\top} = S(m, v)$. Thus $\{A \succeq 0 : \text{tr } A = 1\}$ corresponds bijectively to $\{H \succeq 0 : \text{tr}(HS(m, v)) = 1\}$. Exponentiating the sum of logarithms in the statistic gives

$$\exp\{\mathcal{L}_n(m, vI_d)\} = \max_{\substack{H \succeq 0 \\ \text{tr}(HS(m, v)) = 1}} \prod_{i=1}^n w_i^\top H w_i.$$

The identity remains valid when some factors vanish, under the convention $\log 0 = -\infty$. The product is continuous on the compact feasible set, and a positive product is available from a positive definite A .

For any nonzero $H \succeq 0$, positivity of $S(m, v)$ gives $\text{tr}(HS(m, v)) > 0$. Dividing H by this trace imposes the displayed normalization. The resulting ratio is unchanged by multiplying H by any positive scalar, since numerator and denominator both have degree n in H . The ratio can therefore be maximized under the normalization $\text{tr } H = 1$, which gives (54). Its positive denominator can be cleared without changing the inequality, which proves (55). $\square$

Put $p_d = (d+1)(d+2)/2$. For $\theta \in \mathbb{R}^{p_d}$, define L_θ by placing its coordinates in the lower triangular entries of a $(d+1) \times (d+1)$ matrix, ordered row by row. Set $H_\theta = L_\theta L_\theta^\top$. Every positive semidefinite matrix has such a factorization, including singular matrices. We allow arbitrary diagonal entries in L_θ , and $\text{tr } H_\theta = \|\theta\|^2$.

Put $D(m, v, \theta) = \text{tr}(H_\theta S(m, v))$ and define

$$P_c(m, v, \theta) = \prod_{i=1}^n w_i^\top H_\theta w_i - cD(m, v, \theta)^n. \quad (56)$$

For fixed c and a fixed sample, $P_c(m, v, \theta)$ is a polynomial in (m, v, θ) . By Lemma L.2, for $v > 0$,

$$\mathcal{L}_n(m, vI_d) > \log c \iff \exists \theta \in \mathbb{R}^{p_d} : \|\theta\|^2 = 1 \wedge P_c(m, v, \theta) > 0. \quad (57)$$

Indeed, a matrix with trace one that violates (55) admits the representation H_θ with $\|\theta\|^2 = 1$. Conversely, any such θ gives a positive semidefinite matrix with trace one that violates the acceptance inequality. The strict inequality preserves acceptance at the threshold.

All conditions in (57) are polynomial. To check the degree, partition

$$H_\theta = \begin{pmatrix} h_{11} & h^\top \\ h & H_{xx} \end{pmatrix}.$$

Then $D = h_{11} + 2h^\top m + m^\top H_{xx} m + v \text{tr } H_{xx}$. Each entry of H_θ is quadratic in θ , so D has total degree at most four in (m, v, θ) . The product $\prod_{i=1}^n w_i^\top H_\theta w_i$ has degree $2n$, and P_c has degree at most $4n$.

L.3 Intersection and norm minimization

A proposed estimate z belongs to $\mathcal{K}(c, \zeta)$ if and only if every candidate with $v > 0$ whose ball excludes z is rejected. By (57), this is the quantified polynomial condition

$$z \in \mathcal{K}(c, \zeta) \iff \forall (m, v) \in \mathbb{R}^{d+1} : \quad (58)$$

$$[v > 0 \wedge \|z - m\|^2 > \zeta v] \implies \exists \theta \in \mathbb{R}^{p_d} : \|\theta\|^2 = 1 \wedge P_c(m, v, \theta) > 0.$$

The strict inequality in the rejection condition ensures that candidates at the acceptance threshold remain accepted.

Formula (58) has d free variables, a universally quantified block of $d+1$ candidate variables, and an existential block of p_d witness variables after conversion to prenex form. It contains four polynomial comparisons of degree at most $4n$. Quantifier elimination produces a quantifier-free formula $\varphi(z)$ involving only polynomial comparisons in z and satisfying $\{z \in \mathbb{R}^d : \varphi(z)\} = \mathcal{K}(c, \zeta)$.

For fixed d , standard quantifier elimination produces a quantifier-free description of $\mathcal{K}(c, \zeta)$ in polynomial time; Appendix M gives the detailed complexity bound. When the intersection is nonempty, compact, and convex, the squared norm attains its minimum by compactness, and strict convexity makes the minimizer unique. The same appendix gives a semialgebraic procedure that computes this point in polynomial time.

Lemma L.3 (Error bound for selected points). *Suppose that the full intersection of balls associated with accepted candidates at threshold $\log c$ and radius coefficient ζ is nonempty. If (μ, Σ) is accepted, every point z of that intersection satisfies*

$$\|z - \mu\| \leq \sqrt{\zeta \|\Sigma\|_{\text{op}}}.$$

Proof. The ball associated with (μ, Σ) is one of the balls in the intersection. Every selected point must belong to it. Lemma L.1 identifies this intersection with (53) for a nonconstant sample, and identifies it with the common observation when all observations coincide. $\square$

L.4 Threshold selection

Let $\delta \in (0, 1)$, and take the positive integer a_d from Lemma 2.4. Let $\ell_\delta := \min\{m \geq 1 : 2^{-m} \leq \delta\}$ and $r_\delta := \min\{m \geq 0 : 2^m \geq \ell_\delta\}$, with both minima over integers, and put $h_\delta = \ell_\delta + a_d r_\delta$ and $c = 2^{h_\delta}$. Successive halving and doubling find these integers using arithmetic operations and comparisons. For analysis only, put $L = \log(1/\delta)$ and $T_\delta = \log c = h_\delta \log 2$. Minimality gives

$$L \leq \ell_\delta \log 2 < L + \log 2, \quad \log \ell_\delta \leq r_\delta \log 2 < \log \ell_\delta + \log 2.$$

For $L \geq e$, $\ell_\delta \geq L/\log 2 > L$. Hence there is a constant $\epsilon_d < \infty$, depending only on d , such that $T_d \leq T_\delta \leq T_d + \epsilon_d$. The maximal inequality therefore gives $\mathbb{P}\{\mathcal{L}_n(\mu, \Sigma) > T_\delta\} \leq \delta$. The acceptance test itself uses $c = 2^{h_\delta}$; $T_\delta = h_\delta \log 2$ is needed only to verify the statistical threshold.

To compute the radius coefficient, put $S_M = 2 \sum_{r=0}^{M-1} [(2r+1)3^{2r+1}]^{-1}$ and $E_M = 3^{1-2M}/4$. Take the least $M \geq 1$ with $h_\delta E_M \leq 1$, and set $\bar{T}_\delta = h_\delta(S_M + E_M)$, so that $T_\delta \leq \bar{T}_\delta \leq T_\delta + 1$; Appendix M proves these inequalities from the series for $\log 2$ and its remainder bound. Define $s_\delta = \bar{T}_\delta/n$ and $\zeta = \kappa_d(s_\delta) = 2s_\delta(1+16b_d s_\delta)$. This is the pair (T_δ, s_δ) fixed in Section 4.

By Lemma 2.1, every accepted candidate admits a distribution P in its moment class satisfying $D_{\text{KL}}(\hat{P}_n \| P) \leq T_\delta/n \leq s_\delta$. The deterministic intersection proof in Section 4.1 therefore applies when $s_\delta \leq s_0 := \min\{(64b_d)^{-1}, [4096(d+1)^2]^{-1}\}$. Since $T_d \leq \bar{T}_\delta \leq T_d + \epsilon_d + 1$, choosing L_d sufficiently large and c_d sufficiently small ensures this condition throughout the statistical regime. On the event that (μ, Σ) is accepted, Lemma L.3 gives the radius $\sqrt{\zeta \|\Sigma\|_{\text{op}}}$. The comparison in Section 4.1 then yields

$$\sqrt{\zeta \|\Sigma\|_{\text{op}}} \leq \left[1 + C_d \left(\frac{\log L}{L} + \frac{L}{n}\right)\right] \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}}.$$

L.5 Measurability and complexity

Algorithm 1 is defined for every sample and every $\delta \in (0, 1)$. After handling constant samples and computing $s = s_\delta = \bar{T}_\delta/n$, it returns the empirical mean when $s > s_0$. When $s \leq s_0$, the deterministic proof guarantees a nonempty compact intersection, so the point of minimum norm is well defined. In the range of Theorem 1, the constants are chosen so that $s \leq s_0$, and the algorithm therefore uses the intersection selector analyzed above.

Proposition L.4 (Statistical and computational guarantees). *Fix d . Algorithm 1 defines a Borel estimator $\hat{\mu}_{n,\delta}^{(d)}$ for every sample $X_1, \dots, X_n \in \mathbb{R}^d$ and every $\delta \in (0, 1)$. There are constants C_d, L_d, c_d depending only on d such that, if $L \geq L_d$ and $L/n \leq c_d$, then for every Borel distribution with mean μ and finite covariance Σ ,*

$$\mathbb{P}\left\{\left\|\hat{\mu}_{n,\delta}^{(d)} - \mu\right\| > \left[1 + C_d \left(\frac{\log L}{L} + \frac{L}{n}\right)\right] \frac{\mathcal{Q}_\delta(\Sigma)}{\sqrt{n}}\right\} \leq \delta.$$

Algorithm 1 computes this estimator in time $(n + \ell_\delta)^{O(d^5)}$. In the regime above, $\ell_\delta = O(n)$, so the running time is $n^{O(d^5)}$.

Proof. The threshold calculation gives $T_\delta \geq T_d$, $\bar{T}_\delta \geq T_\delta$, and $\bar{T}_\delta \leq T_d + \epsilon_d + 1$. The constants may therefore be chosen so that $s \leq s_0$ in the stated regime. Section 4.1 supplies nonemptiness, compactness, uniqueness of the selector, and the Gaussian comparison. The acceptance and radius bounds above prove the statistical bound. If $\Sigma = 0$, all observations equal μ almost surely, so the algorithm returns μ .

For fixed n, d, δ , the coefficients c, ζ are constant, and polynomial formulas describe the graph of the selected point in the sample and estimator coordinates. This graph is semialgebraic, so the selected point is a Borel function of the sample. The cases in which the algorithm returns the common observation or the empirical mean also define Borel functions on Borel subsets of the input space. As δ varies, each value of ℓ_δ determines a dyadic interval on which $r_\delta, h_\delta, M, c, \bar{T}_\delta, \zeta$ are constant. These countably many intervals are Borel, which proves joint Borel measurability in the sample and confidence input.

The complexity analysis begins with threshold computation. Since $r_\delta = O(\log(\ell_\delta + 1))$ and $M = O(\log(h_\delta + 1))$, this step has cost polynomial in ℓ_δ . If $s > s_0$, only computation of the empirical mean remains. Otherwise, the

algorithm constructs a polynomial in $p_d + d + 1 = (d + 1)(d + 4)/2$ variables with degree at most $4n$. Quantifier elimination then produces a quantifier-free formula defining the intersection in $n^{O(d(d+1)p_d)}$ arithmetic operations and comparisons. Minimizing the squared norm over this set requires semialgebraic optimization in d variables, giving total cost $(n + \ell_\delta)^{O(d^2(d+1)p_d)} = (n + \ell_\delta)^{O(d^5)}$. In the theorem's regime, $\ell_\delta = O(L + 1) = O(n)$, so the bound reduces to $n^{O(d^5)}$. Appendix M derives these complexity bounds. $\square$

M THRESHOLD COMPUTATION AND COMPLEXITY BOUNDS

M.1 Threshold computation

The procedure initializes $\ell_\delta = 1$ and $t = 1/2$, then halves t and increments ℓ_δ until $t \leq \delta$. Starting from $r_\delta = 0$ and $p = 1$, it doubles p and increments r_δ until $p \geq \ell_\delta$. The resulting integers satisfy the minimality conditions in Section L.4. It then computes $h_\delta = \ell_\delta + a_d r_\delta$ and obtains $c = 2^{h_\delta}$ by repeated doubling or repeated squaring. The comparisons remain exact when δ is a dyadic endpoint.

The series identity

$$\log 2 = 2 \sum_{r=0}^{\infty} \frac{1}{(2r+1)3^{2r+1}}$$

follows by integrating the geometric series for $2/(1-t^2)$ from zero to $1/3$. Its tail after M terms obeys

$$0 \leq \log 2 - S_M \leq 2 \sum_{r=M}^{\infty} 3^{-(2r+1)} = \frac{3^{1-2M}}{4} = E_M.$$

Thus $h_\delta E_M \leq 1$ gives $h_\delta \log 2 \leq \bar{T}_\delta = h_\delta(S_M + E_M) \leq h_\delta \log 2 + 1$. The procedure updates the powers in the series and remainder by multiplying by $1/9$. It reaches $h_\delta E_M \leq 1$ after $M = O(\log(h_\delta + 1))$ terms.

Threshold computation requires $O(\ell_\delta)$ halving steps, $O(\log(\ell_\delta + 1))$ doubling steps, and $M = O(\log(h_\delta + 1))$ series terms. In the theorem's regime, $\ell_\delta = O(n)$ and $h_\delta = O(n)$, so this cost is polynomial in n . The threshold is computed using arithmetic operations and comparisons; logarithms enter the bounds used to analyze it.

M.2 Polynomial construction and optimization

For fixed d , the polynomial $P_c(m, v, \theta)$ in (56) has $\nu_d = p_d + d + 1 = (d + 1)(d + 4)/2$ variables and total degree at most $4n$. Its dense representation has at most $\binom{4n+\nu_d}{\nu_d} = n^{O(d^2)}$ coefficients for $n \geq 2$, and multiplying the n quadratic factors $w_i^\top H_\theta w_i$ together with the n copies of D constructs these coefficients in polynomially many arithmetic operations.

We count exact arithmetic operations and comparisons in the ordered field generated by the sample coordinates, the confidence input, and the threshold coefficients, with unit cost per operation. The output is represented by a standard real univariate representation, consisting of a univariate polynomial, a Thom encoding of one real root, and rational functions of that root for the coordinates [3, Chapters 12–13].

Formula (58) contains four polynomial comparisons of degree at most $4n$, with d free variables and quantified blocks of sizes $d + 1$ and p_d . For fixed d , the quantifier elimination bound of Basu et al. [3, Theorem 14.16] therefore gives a quantifier-free formula $\varphi(z)$ defining $\mathcal{K}(c, \zeta)$ using $n^{O(d^4)}$ arithmetic operations and comparisons. The length of this formula is $n^{O(d^4)}$, and the degrees of its polynomials are $n^{O(d^3)}$.

For the optimization step, introduce a scalar ρ and eliminate z from the condition $\Theta(\rho) \iff \exists z \in \mathbb{R}^d : \varphi(z) \wedge \|z\|^2 \leq \rho$. Since d is fixed, a second application of quantifier elimination has cost $n^{O(d^5)}$. Because the intersection is nonempty and compact, the feasible set of ρ is $[\rho^*, \infty)$, where $\rho^* = \min_{\varphi(z)} \|z\|^2$.

Univariate sign determination and root ordering compute ρ^* and a Thom encoding of this algebraic number [3, Chapter 10]. Adding the constraint $\|z\|^2 = \rho^*$ to $\varphi(z)$ leaves the singleton consisting of the unique point of minimum norm. Standard semialgebraic sampling then computes a real univariate representation of this point [3, Chapter 13], again within $n^{O(d^5)}$ operations for fixed d . Including threshold computation and construction of P_c , the total cost is $(n + \ell_\delta)^{O(d^5)}$, and hence $n^{O(d^5)}$ in the regime of Theorem 1.