FAST LEARNING BEYOND FINITE VARIANCE: FRACTIONAL BERNSTEIN GEOMETRY UNDER HEAVY-TAILED LOSSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Fast rates under heavy-tailed losses are known when sufficiently high moments and Bernstein-type geometry remain available. We ask what survives when even the second moment of the *relative* loss may not exist. For a finite class, suppose the centered relative loss satisfies

$$\mathbb{E}|Z_f|^q \le B\Delta_f^\gamma, \qquad 1 < q \le 2.$$

We show that an observable pairwise median-of-means selector obeys

$$\Delta(\widehat{f}) \lesssim B^{1/(q-\gamma)} \left(\frac{\log(M/\delta)}{n}\right)^{\frac{q-1}{q-\gamma}},$$

and prove a matching minimax lower bound, yielding a sharp interpolation between weak-moment estimation and Bernstein fast-rate geometry. The same selector requires no knowledge of q, B, or γ. We further show that L^q increment regularity of order a together with risk growth of order m implies $\gamma = aq/m$; Lipschitz increments and quadratic growth therefore give the exponent $2(q-1)/q$.

The framework is formulated directly through relative losses, so it can remain meaningful even when individual population risks are infinite. For squared vector quantization, relative losses require only a first moment although every absolute population distortion may be infinite. We establish local quadratic relative-risk growth from centroid stationarity and a weighted Voronoi margin condition, and for a symmetric two-component Student-t family prove $A_{\mathrm{V}}(a) = O_{\nu,d}((s/a)^\nu)$. Hence, for $1 < q < \nu \le 2$, sufficiently separated mixtures admit local finite-dictionary relative-quantization rates of order $(\log(M/\delta)/n)^{2(q-1)/q}$ despite $\mathbb{E}\|X\|^2 = \infty$.

1 INTRODUCTION

Fast-rate learning theory exploits a feedback loop between stochastic fluctuations and excess risk: near an optimum, favorable geometry makes excess-loss fluctuations shrink, yielding rates faster than the geometry-free scale. Existing heavy-tail fast-rate theory already shows that unbounded losses can enjoy acceleration when sufficiently high moments and Bernstein-type structure remain available (Dinh et al., 2016). We study a different boundary: *what rate geometry remains when even the second moment of the relative loss need not exist?*

We answer this question for finite hypothesis classes using *relative losses*. Fix a relative-risk minimizer $f^\star$, let $\Delta_f = P(\ell_f - \ell_{f^\star})$, and center the relative loss as $Z_f = (\ell_f - \ell_{f^\star}) - \Delta_f$. For $q \in (1, 2]$ we study

$$\mathbb{E}|Z_f|^q \le B\Delta_f^\gamma, \qquad \gamma \in [0, 1]. \tag{1.1}$$

This fractional Bernstein geometry yields the rate exponent

$$\beta(q, \gamma) = \frac{q-1}{q-\gamma}. \tag{1.2}$$

It recovers the weak-moment exponent $(q-1)/q$ at $\gamma = 0$ and the classical Bernstein exponent $1/(2-\gamma)$ at $q = 2$. A robust pairwise median-of-means selector attains this rate without using q, B, γ

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

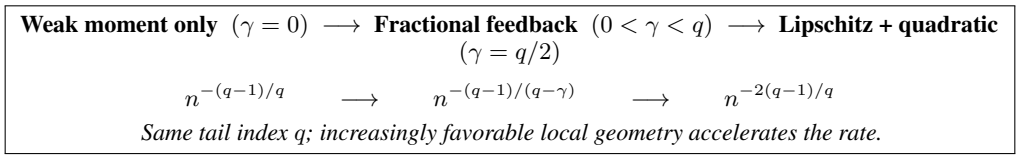


Figure 1: Phase map for the sharp exponent $\beta(q, \gamma) = (q-1)/(q-\gamma)$. At $q = 2$ the law recovers the usual second-moment Bernstein exponent.

or $f^\star$ in its optimization rule, and a matching minimax lower bound shows that the interpolation is sharp for finite classes. Equivalently, the same selector satisfies the corresponding guarantee for every parameter triple (q, B, γ) for which the distribution obeys the condition.

The abstract condition follows from a simple geometric mechanism. If centered relative-loss increments satisfy $\|Z_f\|_q \lesssim d(f, f^\star)^a$ while relative excess risk grows as $\Delta_f \gtrsim d(f, f^\star)^m$, then $\gamma = aq/m$. In the important Lipschitz–quadratic case $(a, m) = (1, 2)$, $\gamma = q/2$ and

$$\beta(q, q/2) = \frac{2(q-1)}{q}. \tag{1.3}$$

Thus, fast learning can survive below finite variance whenever relative fluctuations contract linearly and the objective has local quadratic growth.

Why relative losses? Absolute risks may cease to be useful under heavy tails even when comparisons remain well defined. Squared Euclidean vector quantization makes this distinction explicit: for bounded codebooks, $\ell_C(x) = \min_j \|x - c_j\|^2$ is quadratic in $\|x\|$, but $\ell_C - \ell_D$ is only linear. Hence $P\ell_C$ can equal $+\infty$ for every C while $P(\ell_C - \ell_D)$ is finite under $\mathbb{E}\|X\| < \infty$. We prove a relative Pollard theorem in which local quadratic growth follows from centroid stationarity and a weighted Voronoi margin condition, with no finite second moment.

We then make the geometry explicit for Student-t mixtures. Generic finite mixtures with tail index $\nu > 1$ have finite weighted mass along Voronoi hyperplanes. For the symmetric two-component family $\frac{1}{2}t_\nu(-ae_1, s^2 I_d) + \frac{1}{2}t_\nu(ae_1, s^2 I_d)$, we further prove

$$A_{\mathrm{V}}(a) \lesssim_{\nu,d} (s/a)^\nu. \tag{1.4}$$

Thus sufficient separation itself verifies positive local curvature. Combining this with $1 < q < \nu \le 2$ gives a concrete infinite-variance model with fast exponent $2(q-1)/q$.

Statistical picture. Figure 1 summarizes the rate law. The tail parameter q controls the robust mean-estimation scale $n^{-(q-1)/q}$, while the feedback exponent γ converts that scale into a fixed point. The gain is continuous rather than binary: one does not need to cross a finite-variance threshold before local geometry becomes useful. This is particularly relevant for $q = 1 + \epsilon$: recent finite-$(1+\epsilon)$-moment generalization theory exhibits the geometry-free scale $n^{-\epsilon/(1+\epsilon)}$ (Aminian, 2026), while Lipschitz–quadratic geometry in our model doubles the exponent to $n^{-2\epsilon/(1+\epsilon)}$.

What is and is not assumed in the quantization example. For a generic Student-t mixture, we prove that the weighted boundary constant is finite; this alone is insufficient for positive curvature. The stronger symmetric example is designed to close exactly that gap: the mixture parameters imply a quantitative decay of the boundary term, which in turn implies local curvature. Conversely, our learning theorem remains finite-class: it selects among a local finite dictionary containing the centroidal reference codebook, rather than solving continuous robust k-means globally. We keep this distinction explicit because it separates the new geometric verification from the separate problem of continuous-class localization.

Proof strategy in one line. The paper deliberately separates probability from geometry: robust pairwise mean estimation supplies the weak-moment stochastic scale, a fixed-point argument converts fractional Bernstein feedback into the upper rate, a rare-event testing family proves sharpness, and the quantization sections verify the required feedback through deterministic relative-loss cancellation plus weighted Voronoi stability. This modularity is what allows the same minimax law to cover both finite-variance and genuinely infinite-variance regimes.

Contributions. Our contributions are threefold. The claims below are deliberately narrower than prior heavy-tail fast-rate results: the novelty is the below-second-moment interpolation law and its relative-risk realization, not heavy-tailed fast learning or heavy-tailed k-means per se.

1. First, we establish a sharp minimax theory for finite-class learning when the available relative-loss moment may have order strictly below two. Under a finite q-th moment, $1 < q \le 2$, and a fractional Bernstein condition with exponent γ, a robust pairwise selector achieves
$$\Delta(\widehat{f}) \lesssim B^{1/(q-\gamma)} \left(\frac{\log(M/\delta)}{n} \right)^{\frac{q-1}{q-\gamma}}.$$
A matching lower bound shows that this dependence on n, M, δ, and B is minimax-sharp up to constants.
2. Second, we give primitive conditions that make the abstract fractional Bernstein geometry verifiable. If relative losses are locally L^q-Hölder of order a while the excess risk grows as distance to the power m, then the fractional Bernstein exponent is
$$\gamma = \frac{aq}{m}.$$
Thus, Lipschitz relative losses combined with quadratic local growth yield $\gamma = q/2$ and the fast-rate exponent $2(q-1)/q$.
3. Third, we realize this below-second-moment regime in vector quantization through relative rather than absolute population risk. We formulate quantization through relative losses, develop a weighted relative Pollard geometry requiring only low-order moments, and prove for a symmetric two-component Student-t family that separation itself yields
$$A_{\mathrm{V}}(a) \lesssim_{\nu,d} (s/a)^{\nu}.$$
Consequently, for $1 < q < \nu \le 2$, sufficiently separated mixtures admit local finite-dictionary relative-quantization rates of order
$$\left(\frac{\log(M/\delta)}{n} \right)^{2(q-1)/q},$$
even though the ordinary population k-means distortion is infinite.

2 Relative Risk and Fractional Bernstein Geometry

Let P be the law of Z and $\mathcal{F}$ a loss class. We only require pairwise integrability $\ell_f - \ell_g \in L^1(P)$; individual $P\ell_f$ need not be finite. Fix any reference f_0 and define $\mathcal{R}_{f_0}(f) = P(\ell_f - \ell_{f_0})$.

Definition 2.1 (Relative risk)**.** A *relative-risk minimizer* is $f^\star \in \operatorname{argmin}_f \mathcal{R}_{f_0}(f)$, and $\Delta_f := P(\ell_f - \ell_{f^\star}) \ge 0$ is its relative excess risk.

Proposition 2.2 (Reference invariance)**.** *If all pairwise loss differences are integrable, the set of minimizers and all excess differences are independent of the reference f_0.*

This follows from the cocycle identity $P(\ell_f - \ell_g) + P(\ell_g - \ell_h) = P(\ell_f - \ell_h)$; see Appendix E. Define the centered relative loss $Z_f = (\ell_f - \ell_{f^\star}) - \Delta_f$.

Definition 2.3 (Fractional Bernstein condition)**.** For $q \in (1, 2]$, $\gamma \in [0, 1]$, and $B > 0$, the class satisfies $\mathrm{FB}(q, B, \gamma)$ if
$$\mathbb{E}|Z_f|^q \le B\Delta_f^\gamma, \qquad f \in \mathcal{F}, \tag{2.1}$$
with $\gamma = 0$ interpreted as a uniform qth-moment bound.

The condition may also hold only in a neighborhood of $f^\star$; Section 4 gives primitive sufficient conditions and Appendix D gives a compactness-based globalization principle. The key point is that (2.1) controls only centered *relative* losses, not absolute losses or their variances.

Interpretation. The parameters q and γ capture distinct effects: tail integrability and contraction of fluctuations near the optimum. At $\gamma = 0$ the rate reduces to the weak-moment scale $n^{-(q-1)/q}$, while $q = 2$ recovers the classical second-moment Bernstein interpolation. Thus, for $1 < q < 2$ and $\gamma > 0$, learning can accelerate despite infinite variance.

Crucially, $\Delta_f = P(\ell_f - \ell_{f^\star})$ is defined directly from the relative loss, not as $P\ell_f - P\ell_{f^\star}$. This remains meaningful when both absolute risks are infinite, as in the quantization setting of Section 5.

The fixed-point exponent. The rate in Section 3 can already be anticipated from Definition 2.3. A robust mean estimate based on a qth moment has error scale $\|Z_f\|_q(\log M/n)^{(q-1)/q}$. Under fractional Bernstein geometry this becomes $B^{1/q}\Delta_f^{\gamma/q}(\log M/n)^{(q-1)/q}$. Balancing this stochastic error with Δ_f gives

$$\Delta_f^{1-\gamma/q} \asymp B^{1/q}(\log M/n)^{(q-1)/q},$$

which yields exactly $B^{1/(q-\gamma)}(\log M/n)^{(q-1)/(q-\gamma)}$. Theorem 3.1 makes this heuristic high-probability and Theorem 3.2 proves it cannot generally be improved.

3 Sharp Finite-Class Minimax Theory

Let $\mathcal{F} = \{f_1, \dots, f_M\}$, $M \geq 2$. For $f, g \in \mathcal{F}$, estimate $D(f, g) = P(\ell_f - \ell_g)$ by a median of K block averages,

$$\widehat{D}(f, g) = \mathrm{Med}_{k \leq K} \frac{1}{|I_k|} \sum_{i \in I_k} (\ell_f(Z_i) - \ell_g(Z_i)), \tag{3.1}$$

with odd $K \asymp \log(2M/\delta)$. Define $S_n(f) = \max_g \widehat{D}(f, g)$ and choose

$$\widehat{f} \in \mathrm{argmin}_{f \in \mathcal{F}} S_n(f). \tag{3.2}$$

The construction is observable and uses neither $f^\star$ nor (q, B, γ). The score has a useful tournament interpretation: $S_n(f)$ records the largest robust evidence that f is worse than a competitor, and the selector chooses the hypothesis with the smallest worst defeat. Because K is odd, $\widehat{D}(g, f) = -\widehat{D}(f, g)$ exactly.

Theorem 3.1 (Finite-class upper bound)**.** *Assume* $\mathrm{FB}(q, B, \gamma)$ *for* $q \in (1, 2]$ *and* $\gamma \in [0, 1]$. *There are constants* $c_q, C_{q,\gamma}$ *such that, for* $\delta \in (0, 1/2)$ *and* $n \geq c_q \log(2M/\delta)$*, with probability at least* $1 - \delta$*,*

$$\Delta_{\widehat{f}} \leq C_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{\log(2M/\delta)}{n} \right)^{\frac{q-1}{q-\gamma}}. \tag{3.3}$$

Proof idea. A von Bahr–Esseen bound gives block-mean deviations of order $B^{1/q}\Delta_f^{\gamma/q}(K/n)^{(q-1)/q}$. MOM converts this into exponential confidence. The analysis needs only the M comparisons against the unknown anchor $f^\star$; antisymmetry and the minimax score transfer those bounds to $\widehat{f}$. The resulting inequality $x \lesssim ax^{\gamma/q} + a^{q/(q-\gamma)}$ closes by an elementary fixed-point lemma. Appendix B gives the full proof. The important point is that the optimization rule does not use (q, B, γ). The same selector therefore satisfies the corresponding guarantee for every admissible parameter triple for which the distribution obeys the condition; the parameters enter only the analysis. We use this simultaneous-guarantee formulation rather than claiming a separate adaptive-minimax theorem over nested model classes.

Why only $\log M$ **comparisons appear.** Although the score uses all pairwise statistics, the proof does not union-bound over M^2 pairs. On the good event it controls $\widehat{D}(f, f^\star)$ simultaneously for $f \in \mathcal{F}$. Since $S_n(\widehat{f}) \leq S_n(f^\star)$ and MOM is antisymmetric, these anchored comparisons are enough to upper-bound the excess risk of the selected hypothesis. This keeps the complexity term at $\log M$ and also clarifies why the unknown optimum is an analysis device rather than an input to the algorithm.

For sharpness, let $\mathcal{P}_q(B, \gamma; \mathcal{F})$ be the distributions admitting a relative minimizer for which (2.1) holds.

Theorem 3.2 (Matching finite-class minimax lower bound)**.** *For every $q \in (1, 2]$ and $\gamma \in [0, 1]$, there are constants $c, c', \delta_0 > 0$ such that for every $M \geq 2$ there is a fixed M-element loss class $\mathcal{F}_M$ for which, whenever $\delta \leq \delta_0$ and $\log M + \log(1/\delta) \leq c'n$, every learning rule $\widetilde{f} : \mathcal{Z}^n \to \mathcal{F}_M$ satisfies*

$$\sup_{P \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M)} P^n \left[\Delta_P(\widetilde{f}) \geq cB^{1/(q-\gamma)} \left(\frac{\log(M/\delta)}{n} \right)^{\frac{q-1}{q-\gamma}} \right] \geq \delta. \tag{3.4}$$

The lower bound is stated on a common observation space with a fixed loss class; only the data-generating distribution varies. This is important for a genuine learning lower bound: the learner is not being challenged by a class that changes with the unknown model or with the sample size. For large M we use a binary packing of size M; for the finitely many small values of M, a two-point subexperiment embedded in the same M-hypothesis class closes the gap left by asymptotic Fano constants.

The construction uses rare informative observations of probability p and amplitude A. Calibrating $A^{q-\gamma}p^{1-\gamma} \asymp B$ makes the excess gap $Ap \asymp B^{1/(q-\gamma)}p^{(q-1)/(q-\gamma)}$, while the information divergence remains $O(p)$. Fano handles the complexity regime and a two-point Bretagnolle–Huber argument handles confidence and small M; Appendix C gives the complete arbitrary-M proof. Together, Theorems 3.1–3.2 identify the sharp exponent

$$\beta(q, \gamma) = \frac{q-1}{q-\gamma}. \tag{3.5}$$

The law has four useful checkpoints. With no Bernstein gain ($\gamma = 0$), it gives the weak-moment exponent $(q-1)/q$. With finite variance ($q = 2$), it gives $1/(2-\gamma)$. Under quadratic geometry ($\gamma = q/2$), it gives $2(q-1)/q$. Finally, as $q \downarrow 1$, the rate necessarily degenerates, reflecting the difficulty of high-confidence mean estimation near the first-moment boundary. The lower bound shows that these transitions are statistical, not artifacts of the MOM proof.

4 Primitive Verification

The fractional Bernstein condition follows from two model-level properties. Let d be a pseudo-metric and $\mathcal{U}$ a neighborhood of $f^\star$.

Assumption 4.1 (L^q increment regularity)**.** For some $a, L > 0$, $\|Z_f\|_{L^q(P)} \leq Ld(f, f^\star)^a$ for all $f \in \mathcal{U}$.

Assumption 4.2 (Relative-risk growth)**.** For some $m, c > 0$, $\Delta_f \geq cd(f, f^\star)^m$ for all $f \in \mathcal{U}$.

Proposition 4.3 (Primitive verification principle)**.** *Under Assumptions 4.1–4.2,*

$$\mathbb{E}|Z_f|^q \leq L^q c^{-aq/m} \Delta_f^{aq/m}. \tag{4.1}$$

Hence, if $aq \leq m$, the local fractional Bernstein parameters are $\gamma = aq/m$ and $B = L^q c^{-aq/m}$.

In particular, Lipschitz increments and quadratic growth give the regime used below.

Corollary 4.4 (Lipschitz–quadratic geometry)**.** *If $\|Z_f\|_q \leq Ld(f, f^\star)$ and $\Delta_f \geq cd(f, f^\star)^2$, then*

$$\mathbb{E}|Z_f|^q \leq L^q c^{-q/2} \Delta_f^{q/2}, \qquad \gamma = q/2. \tag{4.2}$$

For later use we also record that pairwise regularity suffices.

Proposition 4.5 (Pairwise verification)**.** *If $\|\ell_f - \ell_g\|_q \leq Ld(f, g)^a$ locally, then $\|Z_f\|_q \leq 2Ld(f, f^\star)^a$; combining with risk growth yields Proposition 4.3 up to the factor 2^q.*

A random envelope $|\ell_f - \ell_g| \leq A(Z)d(f, g)^a$ with $A \in L^q$ is an immediate special case. A local condition can be globalized on compact classes given a positive far-region risk gap and uniform qth moments.

Proposition 4.6 (Globalization from a local condition)**.** *Under those compactness, gap, and uniform-moment conditions, a local fractional Bernstein bound extends to the full class with a finite enlarged constant.*

Full statements and proofs are in Appendix D.

Why this matters. Proposition 4.3 isolates the source of the fast exponent. The moment order alone does not determine the rate: the same q can lead to different γ depending on the contraction of relative-loss increments and the growth of the objective. In the Lipschitz–quadratic case, combining $\gamma = q/2$ with Theorem 3.1 gives

$$\Delta_{\widehat{f}} \lesssim B^{2/q} \left(\frac{\log(M/\delta)}{n} \right)^{2(q-1)/q}.$$

For $q < 2$, this is strictly faster than the geometry-free weak-moment exponent $(q-1)/q$, without requiring finite variance. This is the mechanism we verify for relative vector quantization.

Local versus global geometry. Quadratic growth is often local. For a finite candidate family, one may either restrict the dictionary to this verified basin, as in our Student-t result, or globalize the condition. On a compact parameter space, an isolated relative optimum and uniform qth-moment control away from it yield a positive far-region risk gap and hence a finite Bernstein constant. Proposition 4.6 formalizes this two-region argument.

Pairwise versus anchored assumptions. Although $|Z_f|_q \lesssim d(f, f^\star)^a$ involves the unknown $f^\star$, Proposition 4.5 reduces its verification to pairwise increments, with centering costing only a factor two in L^q. For squared quantization, deterministic pairwise codebook comparisons directly provide this regularity.

5 Relative Vector Quantization Beyond Finite Variance

Let $C = (c_1, \dots, c_k) \in \overline{B}(0,\rho)^k / S_k =: \mathcal{C}_\rho$ and $\ell_C(x) = \min_j \|x - c_j\|^2$. Write d_{perm} for optimal permutation alignment. When $\mathbb{E}\|X\|^2 = \infty$, $P\ell_C = \infty$ for every bounded codebook, so we optimize relative distortion $P(\ell_C - \ell_{C_0})$ instead.

Lemma 5.1 (Relative-loss Lipschitz bound)**.** *For $C, D \in \mathcal{C}_\rho$,*

$$|\ell_C(x) - \ell_D(x)| \le 2(\|x\| + \rho) d_{\mathrm{perm}}(C, D). \tag{5.1}$$

Consequently, if $m_q = (\mathbb{E}\|X\|^q)^{1/q} < \infty$, then $\|\ell_C - \ell_D\|_q \le 2(m_q + \rho) d_{\mathrm{perm}}(C, D)$.

The common quadratic term cancels, leaving only linear growth in $\|x\|$. More explicitly, for fixed centers c, d, the difference $\|x - c\|^2 - \|x - d\|^2$ is affine in x. Taking minima preserves a Lipschitz comparison after optimal center alignment. This elementary cancellation is what lowers the required data moment from order two to order one for well-posedness, and to order q for L^q control.

Proposition 5.2 (Well-defined relative quantization)**.** *If $\mathbb{E}\|X\| < \infty$, all pairwise loss differences are integrable, the relative minimizers are reference invariant, the relative-risk map is Lipschitz on $\mathcal{C}_\rho$, and a relative-optimal codebook exists.*

It remains to verify quadratic growth. Classical Pollard-type arguments are often written for finite population distortion and therefore hide second-moment integrability inside the objective. We instead expand only relative losses and separate the deterministic within-cell curvature from the stochastic instability caused by points that switch Voronoi cells. Let $C^\star = (c_1^\star, \dots, c_k^\star)$ have distinct centers, Voronoi cells $V_j^\star$, zero boundary mass, cell masses p_j with $p_{\min} > 0$, and centroid identities $\mathbb{E}[(X - c_j^\star)\mathbf{1}_{V_j^\star}] = 0$. For $x \in V_j^\star$, let $m_\star(x) = \min_{\ell \ne j}(\|x - c_\ell^\star\|^2 - \|x - c_j^\star\|^2)$.

Assumption 5.3 (Weighted Voronoi margin)**.** For some $A_{\mathrm{V}}, t_0 > 0$ and $0 < t \le t_0$,

$$\mathbb{E}[(\|X\| + \rho)\mathbf{1}\{m_\star(X) \le t(\|X\| + \rho)\}] \le A_{\mathrm{V}} t. \tag{5.2}$$

The condition can be read as a stability requirement. If a codebook is perturbed by radius r, only observations whose optimal Voronoi margin is of order $r(\|X\| + \rho)$ can switch assignments. The expectation in (5.2) weights exactly this switching region by the size of the possible loss perturbation. This is why a first moment can suffice even though the underlying loss is quadratic.

Theorem 5.4 (Relative Pollard growth)**.** *Assume $\mathbb{E}\|X\| < \infty$ and the conditions above. If $16A_{\mathrm{V}} < p_{\min}$, then for some $r_0 > 0$,*

$$\Delta(C) := P(\ell_C - \ell_{C^\star}) \ge \lambda d_{\mathrm{perm}}(C, C^\star)^2, \quad \lambda = p_{\min} - 16A_{\mathrm{V}} > 0 \tag{5.3}$$

whenever $d_{\mathrm{perm}}(C, C^\star) \le r_0$.

Mechanism. After aligning centers, the fixed-cell expansion contributes at least $p_{\min} r^2$. Assignment changes can occur only in a variable-width tube $m_\star(X) \lesssim r(\|X\| + \rho)$, and the weighted margin bounds their total gain by $16 A_{\mathrm{V}} r^2$. No second moment appears; the full proof is in Appendix G. The weighting by $\|X\| + \rho$ is not cosmetic: a perturbation of a center changes a pairwise squared-distance discriminant by an amount proportional to the perturbation radius times $\|X\| + \rho$. Thus the relevant boundary neighborhood has variable width in the tails, and the margin condition measures exactly the mass that can destabilize assignments.

The decomposition behind the theorem is exact at the relative-loss level. Holding the original Voronoi partition fixed, the centroid identities eliminate the linear terms and produce $\sum_j p_j \|h_j\|^2$. Reassigning points can only decrease this fixed-partition comparison; the weighted margin controls that decrease. Hence the curvature constant has the transparent form $p_{\min} - 16 A_{\mathrm{V}}$, separating within-cell mass from boundary instability.

Combining Lemma 5.1 with Theorem 5.4 gives the fractional geometry.

Corollary 5.5 (Fractional Bernstein geometry for quantization)**.** *If $q \in (1, 2]$, $m_q < \infty$, and Theorem 5.4 holds, then locally*

$$\mathbb{E}|Z_C|^q \le [4(m_q + \rho)]^q \lambda^{-q/2} \Delta(C)^{q/2}. \tag{5.4}$$

Thus $\gamma = q/2$ and the finite-class rate exponent is $2(q-1)/q$.

For any finite local dictionary containing $C^\star$, Theorem 3.1 therefore gives

$$\Delta(\widehat{C}) \lesssim_q \frac{(m_q + \rho)^2}{\lambda} \left(\frac{\log(2M/\delta)}{n} \right)^{2(q-1)/q}. \tag{5.5}$$

The finite-dictionary restriction is discussed in Section 8; Appendix F gives the relative-risk and compactness details. Importantly, the result is meaningful even when $P\ell_C = +\infty$ for every bounded C: the learned quantity is the finite relative excess $P(\ell_C - \ell_{C^\star})$, not a subtraction of two infinite expectations.

6 Separated Student-t Mixtures: An Infinite-Variance Example

We now verify the preceding geometry in a concrete heavy-tailed model. First, codimension-one Student-t tails are sufficiently light for weighted Voronoi tubes.

Proposition 6.1 (Weighted Voronoi regularity for Student-t mixtures)**.** *For every finite d-dimensional Student-t_ν mixture with $\nu > 1$ and every fixed bounded codebook with distinct centers, the weighted Voronoi margin condition holds with some finite A_{V} and $t_0 > 0$.*

The reason is dimensional: on an affine hyperplane the relevant weighted radial tail is $r^{-\nu} dr$, integrable for $\nu > 1$, whereas the full-space second moment requires $\nu > 2$; see Appendix H. Finiteness alone does not imply $16 A_{\mathrm{V}} < p_{\min}$, so generic Student-$t$ tails by themselves do not guarantee positive curvature. The next step is therefore quantitative rather than merely integrability-based: in a symmetric family we show that the boundary constant decays explicitly with signal-to-scale separation.

Consider

$$P_a = \tfrac{1}{2} t_\nu(-a e_1, s^2 I_d) + \tfrac{1}{2} t_\nu(a e_1, s^2 I_d), \qquad \nu > 1, \tag{6.1}$$

and the symmetric centroidal codebook $C_a^\circ = \{-c_a e_1, c_a e_1\}$, where $c_a = \mathbb{E}[X_1 \mid X_1 > 0] = \mathbb{E}|a + s T_\nu|$. Then $|c_a - a| \le s\mathbb{E}|T_\nu|$, so $c_a \asymp a$ at large separation; symmetry gives $p_{\min} = 1/2$, exact centroid identities, zero boundary mass, and the separating hyperplane $\{x_1 = 0\}$. These facts reduce the entire curvature question to the weighted mass of a tube around one hyperplane.

This model is deliberately simple enough that every term in the geometry can be evaluated. The codebook separation is $2c_a$, the two Voronoi cells have equal mass, and the boundary is fixed by symmetry. The only nontrivial quantity is therefore the weighted tube mass. This makes the family a useful stress test for the theory: it has genuinely infinite second moment when $\nu \le 2$, yet its codimension-one boundary geometry becomes increasingly benign as a/s grows.

Theorem 6.2 (Separation-induced Voronoi sparsity)**.** *There is $C_{\nu,d} < \infty$ such that, for all sufficiently large a/s, a valid weighted margin constant satisfies*

$$A_{\mathrm{V}}(a) \le C_{\nu,d}(s/a)^{\nu}. \tag{6.2}$$

Hence there is $D_0(\nu, d) < \infty$ such that $a/s \ge D_0(\nu, d)$ implies $16A_{\mathrm{V}}(a) < 1/2 = p_{\min}$.

The proof first bounds the weighted boundary functional by $\mathcal{B}_a \lesssim_{\nu,d} s^{\nu}a^{1-\nu}$. On $x_1 = 0$, both mixture components have the same density, and scaling the tangential integral gives precisely this power of a. Since the discriminant is $4c_a x_1$ and $c_a \asymp a$, a uniform variable-width tube argument yields $A_{\mathrm{V}}(a) \lesssim \mathcal{B}_a/c_a \lesssim (s/a)^{\nu}$.

Corollary 6.3 (Local relative curvature for separated Student-t)**.** *If $a/s \ge D_0(\nu, d)$, then for some $r_a > 0$ and all $C \in \mathcal{C}_{\rho_a}$ with $d_{\mathrm{perm}}(C, C_a^{\circ}) \le r_a$,*

$$P_a(\ell_C - \ell_{C_a^{\circ}}) \ge \lambda_a d_{\mathrm{perm}}(C, C_a^{\circ})^2, \qquad \lambda_a = \tfrac{1}{2} - 16A_{\mathrm{V}}(a) > 0. \tag{6.3}$$

Thus C_a° is a strict local relative-risk minimizer; no global characterization of the population two-means optimum is required. This distinction is useful because the statistical theorem below only needs a finite candidate family contained in the verified local basin.

Combining the previous corollary with the relative-loss Lipschitz bound gives $\|Z_C\|_q \lesssim (m_q + \rho_a)d_{\mathrm{perm}}(C, C_a^{\circ})$ and hence a local fractional Bernstein exponent $q/2$. The following theorem packages the full implication into a statistical statement.

Theorem 6.4 (Explicit infinite-variance fast relative quantization)**.** *Let $1 < q < \nu \le 2$ and $a/s \ge D_0(\nu, d)$. Then $\mathbb{E}\|X\|^q < \infty$ but $\mathbb{E}\|X\|^2 = \infty$. For any finite family $\mathcal{C}_M \subseteq \mathcal{C}_{\rho_a} \cap \{C : d_{\mathrm{perm}}(C, C_a^{\circ}) \le r_a\}$ of cardinality M containing C_a°, the robust pairwise selector satisfies, with probability at least $1 - \delta$,*

$$P_a(\ell_{\widehat{C}} - \ell_{C_a^{\circ}}) \le C_q \frac{(m_q + \rho_a)^2}{\lambda_a} \left(\frac{\log(2M/\delta)}{n}\right)^{2(q-1)/q} \tag{6.4}$$

for $n \gtrsim_q \log(2M/\delta)$. Meanwhile $P_a\ell_C = \infty$ for every bounded codebook C.

The attainable sample exponent approaches $2(\nu - 1)/\nu$ as $q \uparrow \nu$; for $\nu = 2$ it can be arbitrarily close to one, although the prefactor is not uniform in this limit. The example therefore separates two effects cleanly: the tail index limits which fractional moments are available, while geometric separation controls whether those moments can be converted into positive local curvature.

7 Related Work and Positioning

Heavy-tailed fast rates and the below-second-moment boundary. Fast learning with heavy-tailed losses predates our work. In particular, Dinh et al. (2016) combine an L^r-integrable envelope ($r \ge 2$) with a multi-scale Bernstein condition and derive fast rates, including for vector quantization. Generalized-Bayes and central/witness-condition approaches also obtain fast rates for unbounded losses (Ho et al., 2020; Grunwald & Mehta, 2020), while the broader Bernstein/central-condition literature explains several routes from local stochastic geometry to acceleration (van Erven et al., 2015). We therefore do not claim novelty for heavy-tailed fast rates, Bernstein-type feedback, or heavy-tailed k-means themselves. Our boundary is different: the available relative-loss moment can have order $q \in (1, 2)$, so the second moment used by the standard Bernstein relation need not exist. In this regime we identify and match minimax upper and lower bounds with exponent

$$\frac{q-1}{q-\gamma}.$$

At $q = 2$ this recovers the usual second-moment Bernstein exponent, while at $\gamma = 0$ it reduces to the weak-moment estimation exponent.

Robust learning under weak moments. Robust risk minimization under heavy tails has been developed using Catoni-type estimators (Brownlees et al., 2015), small-ball methods (Mendelson, 2018), median-of-means tournaments and minmax MOM procedures (Lugosi & Mendelson, 2020;

Lecué & Lerasle, 2020), and robust empirical-risk proxies (Mathieu & Minsker, 2021); see Lugosi & Mendelson (2019) for a survey. Local Bernstein assumptions have also been combined with minmax MOM under weak design moments (Chinot et al., 2020). PAC-Bayes methods provide complementary guarantees under hostile tails (Alquier & Guedj, 2018; Holland, 2019; Haddouche & Guedj, 2023). Recent work under a finite $(1+\epsilon)$-moment gives the geometry-free scale $n^{-\epsilon/(1+\epsilon)}$ (Aminian, 2026), which coincides with our law at $q = 1+\epsilon$ and $\gamma = 0$. Thus our contribution is not a new robust mean primitive or the combination "MOM + Bernstein": pairwise MOM supplies high-confidence comparisons, while the below-second-moment feedback condition determines the accelerated fixed point and its sharp minimax law.

Relative risk when absolute risks are infinite. Rather than requiring finite individual risks, we take the integrable comparison $P(\ell_f - \ell_{f^\star})$ as primitive. This remains meaningful when $P\ell_f = P\ell_{f^\star} = +\infty$. In squared quantization, cancellation of the common quadratic term leaves relative losses with only linear growth in $|X|$, allowing statistical analysis under a first moment even when all absolute squared distortions are infinite.

Vector quantization and clustering. Classical vector quantization traces to Pollard (1981; 1982), with fast rates under local or margin conditions (Levrard, 2013; 2015) and unbounded-loss analyses under moment assumptions (Bachem et al., 2017). While Dinh et al. (2016) already establish fast rates for heavy-tailed vector quantization, we study the stricter regime $\mathbb{E}|X| < \infty$ but $\mathbb{E}|X|^2 = \infty$, where absolute squared distortions may all be infinite. We instead use relative losses and a weighted Voronoi boundary condition, verified for symmetric Student-t models via $A_{\mathrm{V}}(a) = O_{\nu,d}((s/a)^\nu)$. Our result concerns a local finite dictionary containing the reference codebook, not global continuous-class robust k-means.

Broader heavy-tail mechanisms. Related work studies dependent heavy-tailed ERM (Roy et al., 2021), finite-p-moment stochastic optimization (Cutkosky & Mehta, 2021), unbounded-loss stability (Li et al., 2024), and confidence amplification (Qian et al., 2025). Our result complements these directions by isolating a sharp finite-class minimax law below the second-moment boundary and realizing it in a relative-risk problem where absolute population objectives need not be finite.

8 Discussion, Limitations, and Future Work

The rate law

$$\beta(q,\gamma) = \frac{q-1}{q-\gamma}$$

separates two distinct resources for heavy-tailed learning: integrability, encoded by q, and local risk geometry, encoded by γ. Weak moments alone give the baseline $n^{-(q-1)/q}$ scale, while favorable geometry can accelerate this rate without restoring finite variance. Relative losses are crucial when absolute population objectives are infinite.

The main limitation is the finite-class scope: our quantization result selects from a local finite dictionary containing the reference optimum, rather than the full continuous codebook space. Natural extensions include infinite-class localization, oracle guarantees with approximation error, and data-driven discovery of the favorable local basin. More broadly, the relation $\gamma = aq/m$ suggests fractional fast-rate laws beyond quadratic geometry and vector quantization.

Interpretation of the Student-t example. Separation improves the geometric part of the guarantee by driving $A_{\mathrm{V}}(a)$ down and making the local curvature λ_a positive; it need not improve every constant in the statistical bound. Likewise, for $\nu = 2$, the exponent $2(q-1)/q$ can approach one as $q \uparrow 2$, but the prefactor is not uniform in this limit.

Broader message. Our results suggest that finite variance is not the fundamental dividing line for fast learning. The relevant question is whether the stochastic fluctuations of *relative* losses contract sufficiently rapidly with excess risk near a favorable optimum. Fractional Bernstein geometry makes this tradeoff explicit and provides a target for verification in heavy-tailed problems where absolute objectives are poorly behaved but pairwise comparisons remain integrable.

AI USE STATEMENT

Generative AI tools were used to assist in language editing, literature search and organization, and manuscript organization. The authors take full responsibility for the manuscript and independently reviewed and verified all mathematical arguments, proofs, and citations.

REPRODUCIBILITY STATEMENT

All assumptions and complete proofs of the theoretical results are provided in Appendices A–H. The paper contains no empirical experiments, datasets, or implementation-dependent results.

REFERENCES

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018. doi: 10.1007/s10994-017-5690-0.

Gholamali Aminian. Generalization bounds under heavy-tailed losses. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Olivier Bachem, Mario Lucic, S. Hamed Hassani, and Andreas Krause. Uniform deviation bounds for k-Means clustering. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 283–291. PMLR, 2017.

Christian Brownlees, Emilien Joly, and Gabor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015. doi: 10.1214/15-AOS1350.

Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Statistical learning with Lipschitz and convex loss functions. *Electronic Journal of Statistics*, 14(1):1902–1953, 2020.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021.

Vu C. Dinh, Lam Si Tung Ho, Binh T. Nguyen, and Duy M. H. Nguyen. Fast learning rates with heavy-tailed losses. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Peter D. Grunwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: From ERM to generalized bayes. *Journal of Machine Learning Research*, 21, 2020.

Maxime Haddouche and Benjamin Guedj. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023.

Lam Si Tung Ho, Binh T. Nguyen, Vu C. Dinh, and Duy M. H. Nguyen. Posterior concentration and fast convergence rates for generalized bayesian learning. *Information Sciences*, 538:372–383, 2020. doi: 10.1016/j.ins.2020.05.138.

Matthew J. Holland. PAC-Bayes under potentially heavy tails. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2711–2720, 2019.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020. doi: 10.1214/19-AOS1828.

Clément Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7: 1716–1746, 2013. doi: 10.1214/13-EJS822.

Clément Levrard. Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, 43(2):592–619, 2015. doi: 10.1214/14-AOS1293.

Shaojie Li, Bowei Zhu, and Yong Liu. Algorithmic stability unleashed: Generalization bounds with unbounded losses. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29489–29510. PMLR, 2024.

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.

Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2020.

Timothée Mathieu and Stanislav Minsker. Excess risk bounds in robust empirical risk minimization. *Information and Inference: A Journal of the IMA*, 10(4):1423–1490, 2021. doi: 10.1093/imaiai/iaab004.

Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1–2):459–502, 2018. doi: 10.1007/s00440-017-0784-y.

David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981. doi: 10.1214/aos/1176345339.

David Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28(2):199–205, 1982.

Huajie Qian, Donghao Ying, Henry Lam, and Wotao Yin. Subsampled ensemble can improve generalization tail exponentially. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A. Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861, 2015.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

A TECHNICAL PRELIMINARIES

This appendix contains the elementary probabilistic and analytic results used in the proof of the finite-class upper bound. The only heavy-tail concentration ingredient is the von Bahr–Esseen inequality, combined with median-of-means amplification; the remaining results are elementary analytic inequalities used to resolve the resulting fractional fixed-point relation.

Throughout, $q \in (1, 2]$.

A.1 THE VON BAHR–ESSEEN INEQUALITY

We use the following classical moment inequality for sums of independent centered random variables.

Lemma A.1 (von Bahr–Esseen inequality)**.** *Let $1 \le q \le 2$, and let $X_1, \dots, X_m$ be independent real-valued random variables satisfying*

$$\mathbb{E}X_i = 0, \qquad \mathbb{E}|X_i|^q < \infty.$$

Then,

$$\mathbb{E}\left|\sum_{i=1}^{m} X_i\right|^q \le C_q \sum_{i=1}^{m} \mathbb{E}|X_i|^q, \tag{A.1}$$

where $C_q < \infty$ depends only on q. In particular, one may use a universal numerical upper bound for C_q over $q \in [1, 2]$.

We only use the dependence on the moment order and do not optimize the numerical constant; see von Bahr & Esseen (1965). An immediate consequence is the weak-moment scaling of a sample average.

Corollary A.2 (Moment of a block average)**.** *Let $Y_1, \dots, Y_m$ be i.i.d. copies of Y, with*

$$\mu := \mathbb{E}Y$$

and

$$\mathbb{E}|Y - \mu|^q \le v^q. \tag{A.2}$$

Then,

$$\mathbb{E}\left|\frac{1}{m}\sum_{i=1}^{m} Y_i - \mu\right|^q \le C_q m^{-(q-1)} v^q. \tag{A.3}$$

Consequently,

$$\left\|\frac{1}{m}\sum_{i=1}^{m} Y_i - \mu\right\|_{L^q} \le C_q^{1/q} v\, m^{-(q-1)/q}. \tag{A.4}$$

Proof. Apply Lemma A.1 to

$$X_i := Y_i - \mu.$$

Then,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{m}\sum_{i=1}^{m}(Y_i - \mu)\right|^q &= m^{-q}\mathbb{E}\left|\sum_{i=1}^{m}(Y_i - \mu)\right|^q \\ &\le C_q m^{-q} \sum_{i=1}^{m} \mathbb{E}|Y_i - \mu|^q \\ &\le C_q m^{-(q-1)} v^q.\end{aligned}$$

Taking the q-th root gives (A.4). □

A.2 SCALAR MEDIAN-OF-MEANS CONCENTRATION

We now derive the robust scalar mean-estimation result used in Theorem 3.1. Let $Y_1, \dots, Y_n$ be i.i.d. observations. Partition $\{1, \dots, n\}$ into K disjoint blocks

$$I_1, \dots, I_K$$

whose sizes differ by at most one. Define the block means

$$\overline{Y}_k := \frac{1}{|I_k|} \sum_{i \in I_k} Y_i,$$

and, for odd K,

$$\mathrm{MOM}_K(Y_1, \dots, Y_n) := \mathrm{Med}\{\overline{Y}_1, \dots, \overline{Y}_K\}. \tag{A.5}$$

Lemma A.3 (Scalar MOM concentration under a finite q-th moment)**.** *Let*

$$1 < q \le 2,$$

and suppose

$$\mu := \mathbb{E}Y, \qquad \mathbb{E}|Y - \mu|^q \le v^q. \tag{A.6}$$

Let K be odd and satisfy

$$1 \le K \le \frac{n}{2}.$$

Then, there exist constants $a_q < \infty$ and universal $c_{\mathrm{mom}} > 0$ such that

$$\mathbb{P}\left[|\mathrm{MOM}_K(Y_1, \dots, Y_n) - \mu| > a_q v \left(\frac{K}{n}\right)^{(q-1)/q} \right] \le 2e^{-c_{\mathrm{mom}} K}. \tag{A.7}$$

Proof. Let

$$m_k := |I_k|.$$

Since the block sizes differ by at most one and $K \le n/2$,

$$m_k \ge \left\lfloor \frac{n}{K} \right\rfloor \ge \frac{n}{2K}. \tag{A.8}$$

By Corollary A.2,

$$\begin{aligned} \mathbb{E}|\overline{Y}_k - \mu|^q &\le C_q m_k^{-(q-1)} v^q \\ &\le C_q' v^q \left(\frac{K}{n}\right)^{q-1}. \end{aligned} \tag{A.9}$$

Choose a_q sufficiently large that Markov's inequality yields

$$\mathbb{P}\left[|\overline{Y}_k - \mu| > a_q v \left(\frac{K}{n}\right)^{(q-1)/q} \right] \le \frac{1}{4}. \tag{A.10}$$

Call a block *bad* if the event in (A.10) occurs. Since the blocks are disjoint, the corresponding bad-block indicators are independent, and each has success probability at most $1/4$.

The median can violate the same deviation threshold only if at least half of the blocks are bad. If N_{bad} denotes the number of bad blocks, Hoeffding's inequality therefore gives

$$\mathbb{P}\left[N_{\mathrm{bad}} \ge \frac{K}{2} \right] \le \exp(-K/8), \tag{A.11}$$

up to an immaterial adjustment of the numerical constant. Thus (A.7) follows. □

Corollary A.4 (Confidence calibration)**.** *Under the assumptions of Lemma A.3, there exists a universal constant $c > 0$ such that choosing an odd*

$$K \ge c \log \frac{2N}{\delta} \tag{A.12}$$

gives, for each of N random variables to which Lemma A.3 is applied,

$$\mathbb{P}\left[|\widehat{\mu}_j - \mu_j| > a_q v_j \left(\frac{K}{n}\right)^{(q-1)/q} \text{ for some } j \right] \le \delta. \tag{A.13}$$

Proof. Apply Lemma A.3 to each j and take a union bound. No independence across the N estimators is required. □

For Theorem 3.1, we take

$$N = M,$$

because the proof requires only the M comparisons against the optimal anchor $f^\star$, rather than all M^2 ordered pairs.

A.3 Elementary power inequalities

The next calculus lemma is used to control the score of the optimal hypothesis.

Lemma A.5 (Power supremum)**.** *Let*

$$A \geq 0, \qquad \nu \in [0, 1).$$

For $\nu = 0$, interpret x^ν as 1. Then,

$$\sup_{x \geq 0} \{-x + Ax^\nu\} \leq A^{1/(1-\nu)}. \tag{A.14}$$

If $0 < \nu < 1$, the exact value is

$$\sup_{x \geq 0} \{-x + Ax^\nu\} = (1-\nu)\nu^{\nu/(1-\nu)} A^{1/(1-\nu)}. \tag{A.15}$$

Proof. When $\nu = 0$,

$$-x + A \leq A,$$

with equality at $x = 0$.

Now suppose

$$0 < \nu < 1.$$

For

$$\phi(x) := -x + Ax^\nu,$$

the unique positive critical point satisfies

$$-1 + A\nu x^{\nu-1} = 0,$$

hence

$$x_\star = (A\nu)^{1/(1-\nu)}.$$

At this point,

$$\begin{aligned} \phi(x_\star) &= -x_\star + Ax_\star^\nu \\ &= (1-\nu)\nu^{\nu/(1-\nu)} A^{1/(1-\nu)}. \end{aligned}$$

Since

$$(1-\nu)\nu^{\nu/(1-\nu)} \leq 1,$$

the upper bound follows. □

A.4 A fixed-point inequality

We finally record the deterministic fixed-point inequality used to close the excess-risk bound.

Lemma A.6 (Fixed-point inequality)**.** *Let*

$$a, b \geq 0, \qquad \nu \in [0, 1).$$

Suppose $x \geq 0$ satisfies

$$x \leq ax^\nu + b. \tag{A.16}$$

When $\nu = 0$, interpret ax^ν as a. Then

$$x \leq 2b + (2a)^{1/(1-\nu)}. \tag{A.17}$$

Proof. If

$$x \le 2b,$$

the conclusion is immediate.

Suppose instead that

$$x > 2b.$$

Then

$$b < \frac{x}{2}.$$

Using (A.16),

$$x \le ax^\nu + b < ax^\nu + \frac{x}{2}.$$

Therefore,

$$\frac{x}{2} < ax^\nu.$$

If $x = 0$, there is nothing to prove, so we may divide by x^ν and obtain

$$x^{1-\nu} < 2a.$$

Hence,

$$x < (2a)^{1/(1-\nu)}.$$

Combining this case with $x \le 2b$ gives

$$x \le \max\left\{2b, (2a)^{1/(1-\nu)}\right\} \le 2b + (2a)^{1/(1-\nu)}.$$

□

Corollary A.7 (Fractional fixed-point scale)**.** *Suppose*

$$x \le Ax^{\gamma/q} + CA^{q/(q-\gamma)}, \tag{A.18}$$

where

$$q > 1, \qquad 0 \le \gamma \le 1.$$

Then

$$x \le C_{q,\gamma,C} A^{q/(q-\gamma)}. \tag{A.19}$$

Proof. Set

$$\nu := \frac{\gamma}{q}.$$

Since $q > 1$ and $\gamma \le 1$,

$$0 \le \nu < 1.$$

Apply Lemma A.6 with

$$a = A, \qquad b = CA^{q/(q-\gamma)}.$$

Since

$$\frac{1}{1-\gamma/q} = \frac{q}{q-\gamma},$$

both terms in (A.17) have the same power of A. □

How these preliminary results enter the main theorem. Under the fractional Bernstein condition

$$\mathbb{E}|Z_f|^q \le B\Delta_f^\gamma,$$

Lemma A.3 yields the anchored deviation scale

$$A_n \asymp_q B^{1/q} \left(\frac{\log(M/\delta)}{n}\right)^{(q-1)/q}. \tag{A.20}$$

Lemma A.5 and Corollary A.7 then resolve the resulting fractional fixed-point inequality, giving

$$A_n^{q/(q-\gamma)} \asymp_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{\log(M/\delta)}{n}\right)^{\frac{q-1}{q-\gamma}}, \tag{A.21}$$

which is precisely the fast-rate scale appearing in Theorem 3.1.

B Proof of the Finite-Class Upper Bound

We prove Theorem 3.1. The key observation is that the analysis does not require uniform concentration over all M^2 pairwise comparisons. It is sufficient to control the M comparisons against a fixed optimal anchor $f^\star$. The anchor is used only in the proof and is not required by the learning rule. Throughout this appendix, pairwise loss differences are understood as measurable real-valued random variables. Their expectations are finite by the integrability assumption in Section 3.

Let

$$\mathcal{F} = \{f_1, \dots, f_M\},$$

and fix a relative-risk minimizer $f^\star \in \mathcal{F}$. Recall

$$\Delta_f := P(\ell_f - \ell_{f^\star}),$$

and define

$$W_f := \ell_f - \ell_{f^\star}, \qquad Z_f := W_f - \Delta_f.$$

Thus,

$$PZ_f = 0.$$

The fractional Bernstein assumption of Theorem 3.1 is

$$\mathbb{E}|Z_f|^q \le B\Delta_f^\gamma, \qquad f \in \mathcal{F}, \tag{B.1}$$

with the case $\gamma = 0$ interpreted as

$$\mathbb{E}|Z_f|^q \le B.$$

B.1 Scalar median-of-means input

We use the scalar median-of-means concentration result established in Lemma A.3 of Appendix A. In particular, if $Y_1, \dots, Y_n$ are i.i.d. with mean μ and

$$\mathbb{E}|Y - \mu|^q \le v^q,$$

then for an odd number of blocks $K \le n/2$,

$$\mathbb{P}\left[|\operatorname{MOM}_K(Y_1, \dots, Y_n) - \mu| > a_q v \left(\frac{K}{n}\right)^{(q-1)/q} \right] \le 2e^{-c_{\mathrm{mom}} K}.$$

B.2 Uniform control against the optimal anchor

For each $f \in \mathcal{F}$, the statistic

$$\widehat{D}(f, f^\star)$$

defined in (3.1) is exactly the median-of-means estimator of the mean

$$D(f, f^\star) = \Delta_f$$

based on the random variable

$$W_f = \ell_f - \ell_{f^\star}.$$

Set

$$\nu := \frac{\gamma}{q}. \tag{B.2}$$

Since

$$0 \le \gamma \le 1 \qquad \text{and} \qquad q > 1,$$

we have

$$0 \le \nu < 1.$$

Define

$$A_n := a_q B^{1/q} \left(\frac{K}{n}\right)^{(q-1)/q}. \tag{B.3}$$

When $\gamma > 0$, condition (B.1) gives

$$\|Z_f\|_{L^q(P)} \le B^{1/q} \Delta_f^{\gamma/q}.$$

Applying Lemma A.3 to W_f therefore yields

$$\mathbb{P}\left[\left|\widehat{D}(f, f^\star) - \Delta_f\right| > A_n \Delta_f^\nu\right] \le c_0 e^{-c_1 K}. \tag{B.4}$$

When $\gamma = 0$, the same argument yields

$$\mathbb{P}\left[\left|\widehat{D}(f, f^\star) - \Delta_f\right| > A_n\right] \le c_0 e^{-c_1 K}. \tag{B.5}$$

For notational convenience, in the remainder of the proof we use the convention

$$\Delta_f^0 := 1$$

when $\gamma = 0$. With this convention, (B.4) and (B.5) can be written uniformly as

$$\mathbb{P}\left[\left|\widehat{D}(f, f^\star) - \Delta_f\right| > A_n \Delta_f^\nu\right] \le c_0 e^{-c_1 K}. \tag{B.6}$$

Define the event

$$\mathcal{E} := \bigcap_{f \in \mathcal{F}} \left\{\left|\widehat{D}(f, f^\star) - \Delta_f\right| \le A_n \Delta_f^\nu\right\}. \tag{B.7}$$

A union bound over the M hypotheses gives

$$\mathbb{P}(\mathcal{E}^c) \le M c_0 e^{-c_1 K}. \tag{B.8}$$

Hence, there exists a universal constant $c > 0$ such that choosing an odd integer

$$K \ge c \log \frac{2M}{\delta} \tag{B.9}$$

ensures

$$\mathbb{P}(\mathcal{E}) \ge 1 - \delta. \tag{B.10}$$

Observe that this argument uses only M anchored comparisons, rather than all M^2 ordered pairs. Importantly, the optimal anchor $f^\star$ appears only in the proof event $\mathcal{E}$; the estimator itself does not know $f^\star$.

B.3 Score of the optimal hypothesis

We now work on the event $\mathcal{E}$.

For every $g \in \mathcal{F}$,

$$\widehat{D}(g, f^\star) \ge \Delta_g - A_n \Delta_g^\nu. \tag{B.11}$$

Since the number of blocks is odd,

$$\widehat{D}(f^\star, g) = -\widehat{D}(g, f^\star).$$

Therefore,

$$\widehat{D}(f^\star, g) \le -\Delta_g + A_n \Delta_g^\nu. \tag{B.12}$$

Recalling

$$S_n(f) = \max_{g \in \mathcal{F}} \widehat{D}(f, g),$$

we obtain

$$S_n(f^\star) \le \sup_{x \ge 0} \left\{-x + A_n x^\nu\right\}. \tag{B.13}$$

If $\nu = 0$, then

$$\sup_{x \ge 0} \{-x + A_n\} = A_n.$$

If $0 < \nu < 1$, elementary optimization gives

$$\sup_{x \geq 0} \{-x + A_n x^\nu\} = (1-\nu)\nu^{\nu/(1-\nu)} A_n^{1/(1-\nu)} \leq A_n^{1/(1-\nu)}. \tag{B.14}$$

Since

$$\frac{1}{1-\nu} = \frac{q}{q-\gamma},$$

both cases yield

$$S_n(f^\star) \leq C_{q,\gamma} A_n^{q/(q-\gamma)}. \tag{B.15}$$

B.4 Closing the fixed point

By definition,

$$\widehat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} S_n(f),$$

and hence,

$$S_n(\widehat{f}) \leq S_n(f^\star). \tag{B.16}$$

On the other hand, because the maximum defining $S_n(\widehat{f})$ includes the candidate $g = f^\star$,

$$S_n(\widehat{f}) \geq \widehat{D}(\widehat{f}, f^\star). \tag{B.17}$$

The event $\mathcal{E}$ is uniform over all $f \in \mathcal{F}$, so it applies in particular to the data-dependent hypothesis $\widehat{f}$. Therefore,

$$\widehat{D}(\widehat{f}, f^\star) \geq \Delta_{\widehat{f}} - A_n \Delta_{\widehat{f}}^\nu. \tag{B.18}$$

Combining (B.15)–(B.18),

$$\Delta_{\widehat{f}} - A_n \Delta_{\widehat{f}}^\nu \leq C_{q,\gamma} A_n^{1/(1-\nu)}. \tag{B.19}$$

Equivalently,

$$\Delta_{\widehat{f}} \leq A_n \Delta_{\widehat{f}}^\nu + C_{q,\gamma} A_n^{1/(1-\nu)}. \tag{B.20}$$

We record a simple deterministic consequence. If $x, a, b \geq 0$, $0 \leq \nu < 1$, and

$$x \leq a x^\nu + b,$$

then either $x \leq 2b$, or $x > 2b$, in which case

$$\frac{x}{2} \leq a x^\nu$$

and hence,

$$x \leq (2a)^{1/(1-\nu)}.$$

Thus,

$$x \leq 2b + (2a)^{1/(1-\nu)}. \tag{B.21}$$

Applying (B.21) to (B.20) with

$$x = \Delta_{\widehat{f}}, \qquad a = A_n, \qquad b = C_{q,\gamma} A_n^{1/(1-\nu)},$$

gives

$$\Delta_{\widehat{f}} \leq C'_{q,\gamma} A_n^{1/(1-\nu)}. \tag{B.22}$$

Since

$$\frac{1}{1-\nu} = \frac{q}{q-\gamma},$$

we conclude that

$$\Delta_{\widehat{f}} \leq C'_{q,\gamma} A_n^{q/(q-\gamma)}. \tag{B.23}$$

B.5 Proof of Theorem 3.1

Proof of Theorem 3.1. Choose K as the smallest odd integer satisfying

$$K \geq c \log \frac{2M}{\delta},$$

where c is sufficiently large for (B.10). The sample-size assumption in Theorem 3.1 guarantees that $K \leq n/2$, after adjusting the constant c_q if necessary. By (B.10), the event $\mathcal{E}$ has probability at least $1 - \delta$. On this event, (B.23) and the definition (B.3) give

$$\begin{aligned} \Delta_{\widehat{f}} &\leq C_{q,\gamma} \left[B^{1/q} \left(\frac{K}{n} \right)^{(q-1)/q} \right]^{q/(q-\gamma)} \\ &= C_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{K}{n} \right)^{(q-1)/(q-\gamma)}. \end{aligned} \tag{B.24}$$

Finally,

$$K \lesssim \log \frac{2M}{\delta},$$

so

$$\Delta_{\widehat{f}} \leq C_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{\log(2M/\delta)}{n} \right)^{\frac{q-1}{q-\gamma}}$$

with probability at least $1 - \delta$.

The construction of the selector uses only the pairwise observed loss differences and the block count K. In particular, it does not require knowledge of $f^\star$, q, B, and γ. This completes the proof. □

C Minimax Lower Bounds

We prove Theorem 3.2. The construction is based on a fixed finite loss class defined on a common observation space. Only the data-generating distribution varies. In particular, neither the loss functions nor the observation space depend on the sample size n, the confidence level δ, or the rare-event probability used in the lower-bound construction.

Throughout this appendix, let

$$\alpha := \frac{q-1}{q-\gamma}, \qquad q \in (1, 2], \qquad \gamma \in [0, 1]. \tag{C.1}$$

C.1 A fixed finite loss class

We first construct the hypothesis class used for both the complexity and confidence lower bounds.

Lemma C.1 (Binary packing)**.** *There exist universal constants*

$$\rho \in (0, 1/2), \qquad C_{\mathrm{pack}} < \infty,$$

such that, for every $M \geq 2$, there exist an integer

$$d \leq C_{\mathrm{pack}} \log(2M)$$

and a set

$$\mathcal{V} \subseteq \{-1, +1\}^d$$

satisfying

$$|\mathcal{V}| = M$$

and

$$d_H(u, v) \geq \rho d, \qquad \forall u, v \in \mathcal{V}, \quad u \neq v. \tag{C.2}$$

Here, d_H denotes the Hamming distance.

Lemma C.1 is a standard consequence of the Varshamov–Gilbert packing argument; see, e.g., Tsybakov (2009). We use only the existence of a packing with $d \asymp \log M$ and a constant relative Hamming separation.

Fix such a code $\mathcal{V}$. Define the common observation space

$$\mathcal{Z} := \{1, \dots, d\} \times \{-1, +1\} \times [0, \infty). \tag{C.3}$$

For every $u \in \mathcal{V}$, define a hypothesis f_u through the loss

$$\ell_{f_u}(j, y, t) := t\, \mathbf{1}_{\{u_j \neq y\}}. \tag{C.4}$$

Let

$$\mathcal{F}_M := \{f_u : u \in \mathcal{V}\}. \tag{C.5}$$

The class $\mathcal{F}_M$ and all its loss functions are now fixed. The amplitude appearing below is encoded in the distribution of the third observation coordinate T, rather than in the loss definition.

C.2 A RARE-EVENT FAMILY

Fix once and for all a numerical bias parameter

$$\eta \in (0, 1);$$

for concreteness one may take $\eta = 1/2$.

For $v \in \mathcal{V}$, $p \in (0, 1/2]$, and $A > 0$, define a distribution

$$P_{v,p,A}$$

on $\mathcal{Z}$ as follows.

First draw

$$J \sim \mathrm{Unif}\{1, \dots, d\}.$$

Independently draw

$$R \sim \mathrm{Bernoulli}(p),$$

and set

$$T := AR. \tag{C.6}$$

If $R = 0$, let Y be an independent Rademacher random variable, so that its distribution is common to all models. If $R = 1$, then, conditionally on $J = j$,

$$P_{v,p,A}(Y = v_j \mid J = j, R = 1) = \frac{1 + \eta}{2}, \tag{C.7}$$

and

$$P_{v,p,A}(Y = -v_j \mid J = j, R = 1) = \frac{1 - \eta}{2}. \tag{C.8}$$

Thus, only a fraction p of observations carries information about the unknown index v.

For $u, v \in \mathcal{V}$, write

$$h(u, v) := \frac{d_H(u, v)}{d}. \tag{C.9}$$

By Lemma C.1,

$$h(u, v) \geq \rho \qquad \text{whenever } u \neq v. \tag{C.10}$$

C.3 EXCESS-RISK SEPARATION

We first compute the exact excess risk for this hard family.

Lemma C.2 (Risk gap)**.** *Under $P_{v,p,A}$, the unique risk minimizer in $\mathcal{F}_M$ is f_v. Moreover, for every $u \in \mathcal{V}$,*

$$\Delta_v(u) := P_{v,p,A}(\ell_{f_u} - \ell_{f_v}) = \eta A p\, h(u, v). \tag{C.11}$$

In particular, if $u \neq v$,

$$\Delta_v(u) \geq \rho \eta A p. \tag{C.12}$$

Proof. Let

$$X_{u,v} := \ell_{f_u}(J, Y, T) - \ell_{f_v}(J, Y, T).$$

If $u_J = v_J$, then

$$X_{u,v} = 0.$$

If $u_J \neq v_J$, then necessarily $u_J = -v_J$. On the event $R = 1$,

$$X_{u,v} = \begin{cases} +A, & Y = v_J, \\ -A, & Y = -v_J. \end{cases}$$

When $R = 0$, $T = 0$ and hence $X_{u,v} = 0$.

Therefore,

$$\begin{aligned} P_{v,p,A}(X_{u,v} = +A) &= p\, h(u,v) \frac{1+\eta}{2}, \\ P_{v,p,A}(X_{u,v} = -A) &= p\, h(u,v) \frac{1-\eta}{2}. \end{aligned} \tag{C.13}$$

It follows that

$$P_{v,p,A} X_{u,v} = \eta A p\, h(u,v),$$

which proves (C.11).

For $u \neq v$, the packing property gives $h(u,v) \geq \rho$, so the excess risk is strictly positive. Thus f_v is the unique minimizer. □

C.4 Calibration of the fractional Bernstein condition

We next choose the amplitude A so that every distribution in the hard family satisfies exactly the same fractional Bernstein condition as in the upper bound.

Lemma C.3 (Fractional Bernstein calibration)**.** *Let $p \in (0, 1/2]$. Define*

$$A := \left(\frac{B\eta^\gamma}{2^q p^{1-\gamma}} \right)^{1/(q-\gamma)}. \tag{C.14}$$

Then, for every $v \in \mathcal{V}$, the distribution $P_{v,p,A}$ belongs to

$$\mathcal{P}_q(B, \gamma; \mathcal{F}_M).$$

More precisely, for every $u \in \mathcal{V}$,

$$\mathbb{E}_{v,p,A} \left| (\ell_{f_u} - \ell_{f_v}) - \Delta_v(u) \right|^q \leq B \Delta_v(u)^\gamma, \tag{C.15}$$

with the usual interpretation when $\gamma = 0$.

Proof. Fix $u, v \in \mathcal{V}$, and abbreviate

$$h := h(u,v), \qquad X := X_{u,v}, \qquad \Delta := \Delta_v(u).$$

If $u = v$, then $X = \Delta = 0$, and the claim is immediate.

Suppose $u \neq v$. From (C.13),

$$\mathbb{E}|X|^q = A^q p h. \tag{C.16}$$

Using

$$|a - b|^q \leq 2^{q-1} \big(|a|^q + |b|^q \big),$$

we obtain

$$\mathbb{E}|X - \Delta|^q \leq 2^{q-1} \left(\mathbb{E}|X|^q + \Delta^q \right). \tag{C.17}$$

By Lemma C.2,

$$\Delta = \eta A p h.$$

Since

$$0 < \eta \leq 1, \qquad 0 < ph \leq 1,$$

we have

$$\Delta^q = \eta^q A^q (ph)^q \leq A^q p h.$$

Consequently,

$$\mathbb{E}|X - \Delta|^q \le 2^q A^q p h. \tag{C.18}$$

On the other hand,

$$B\Delta^\gamma = B\eta^\gamma A^\gamma p^\gamma h^\gamma.$$

Thus, it suffices to show

$$2^q A^{q-\gamma} p^{1-\gamma} h^{1-\gamma} \le B\eta^\gamma. \tag{C.19}$$

Because

$$\gamma \in [0, 1]$$

and $h \le 1$,

$$h^{1-\gamma} \le 1.$$

The choice (C.14) gives

$$2^q A^{q-\gamma} p^{1-\gamma} = B\eta^\gamma.$$

Therefore, (C.19) holds, proving the fractional Bernstein inequality. □

The preceding calibration immediately yields the desired statistical scale.

Corollary C.4 (Calibrated separation)**.** *Under the choice* (C.14)*, every $u \neq v$ satisfies*

$$\Delta_v(u) \ge c_{q,\gamma} B^{1/(q-\gamma)} p^\alpha, \tag{C.20}$$

where

$$\alpha = \frac{q-1}{q-\gamma}$$

and $c_{q,\gamma} > 0$ depends only on q, γ (and on the fixed numerical choices ρ, η).

Proof. By (C.12),

$$\Delta_v(u) \ge \rho\eta A p.$$

Substituting (C.14),

$$Ap = \left(\frac{B\eta^\gamma}{2^q}\right)^{1/(q-\gamma)} p^{1-\frac{1-\gamma}{q-\gamma}}.$$

Since

$$1 - \frac{1-\gamma}{q-\gamma} = \frac{q-1}{q-\gamma} = \alpha,$$

the claim follows. □

C.5 Information divergence

The information carried by one observation is only of order p.

Lemma C.5 (Pairwise Kullback–Leibler divergence)**.** *For every $u, v \in \mathcal{V}$,*

$$D_{\mathrm{KL}}\left(P_{v,p,A} \,\|\, P_{u,p,A}\right) = p\, h(u,v)\, \kappa_\eta, \tag{C.21}$$

where

$$\kappa_\eta := \eta \log \frac{1+\eta}{1-\eta}. \tag{C.22}$$

In particular,

$$D_{\mathrm{KL}}\left(P_{v,p,A}^n \,\|\, P_{u,p,A}^n\right) \le np\kappa_\eta. \tag{C.23}$$

Proof. The models coincide whenever $R = 0$. Conditional on $R = 1$, they also coincide on coordinates j for which $u_j = v_j$. If $u_j \neq v_j$, the conditional distribution of Y under P_v is

$$\left(\frac{1+\eta}{2}, \frac{1-\eta}{2}\right)$$

with respect to the labels $(v_j, -v_j)$, whereas under P_u these two probabilities are exchanged. Their KL divergence equals

$$\frac{1+\eta}{2}\log\frac{1+\eta}{1-\eta} + \frac{1-\eta}{2}\log\frac{1-\eta}{1+\eta}$$
$$= \eta\log\frac{1+\eta}{1-\eta} = \kappa_\eta.$$

A fraction $h(u,v)$ of coordinates differ and the informative event occurs with probability p, proving (C.21). The product formula for KL divergence gives (C.23). □

Notice that the divergence does not depend on the amplitude A. The amplitude determines the loss separation, whereas p determines the information available for distinguishing the models.

C.6 The complexity obstruction

We first obtain the $\log M$ component of the minimax rate. For sufficiently large M, the packing construction is combined with Fano's inequality. For the finitely many smaller values of M, we instead use a two-point Le Cam argument inside the same fixed M-hypothesis class. This separation avoids applying Fano's inequality in the small-cardinality regime where its usual additive constant does not yield a uniform positive testing-error lower bound.

Proposition C.6 (Arbitrary-M complexity lower bound)**.** *There exist constants*

$$c^{\mathrm{comp}}_{q,\gamma} > 0, \qquad c_{\mathrm{F}} > 0, \qquad c_{\mathrm{p}} > 0,$$

where c_{F} and c_{p} may be chosen universally after fixing η, such that the following holds for every $M \geq 2$.

Whenever

$$\log M \leq c_{\mathrm{p}} n, \tag{C.24}$$

every measurable learning rule

$$\widetilde{f} : \mathcal{Z}^n \to \mathcal{F}_M$$

satisfies

$$\sup_{P \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M)} P^n\left[\Delta_P(\widetilde{f}) \geq c^{\mathrm{comp}}_{q,\gamma} B^{1/(q-\gamma)}\left(\frac{\log M}{n}\right)^{\alpha}\right] \geq c_{\mathrm{F}}. \tag{C.25}$$

The constants are uniform in M, B, and n.

Proof. We divide the argument into two regimes.

Case 1: sufficiently large M. Let $M_0 \geq 2$ be a sufficiently large universal integer, fixed below, and suppose first that

$$M \geq M_0.$$

Choose

$$p_M := c_* \frac{\log M}{n}, \tag{C.26}$$

where $c_* > 0$ is a sufficiently small numerical constant. Under (C.24), after reducing c_{p} if necessary,

$$0 < p_M \leq \frac{1}{2}.$$

Let A_M be the amplitude from (C.14) with $p = p_M$.

For every $v \in \mathcal{V}$, define

$$P_v := P_{v,p_M,A_M}.$$

By Lemma C.3,

$$P_v \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M).$$

By Lemma C.5,

$$D_{\mathrm{KL}}(P_v^n \| P_u^n) \leq n p_M \kappa_\eta = c_* \kappa_\eta \log M, \qquad u, v \in \mathcal{V}. \tag{C.27}$$

Since

$$|\mathcal{V}| = M,$$

Fano's inequality gives

$$\inf_{\widehat{v}} \sup_{v \in \mathcal{V}} P_v^n(\widehat{v} \neq v) \geq 1 - \frac{c_* \kappa_\eta \log M + \log 2}{\log M}. \tag{C.28}$$

Here, the infimum ranges over all estimators taking values in $\mathcal{V}$. Choose first M_0 sufficiently large that

$$\frac{\log 2}{\log M_0} \leq \frac{1}{4},$$

and then choose $c_* > 0$ sufficiently small that

$$c_* \kappa_\eta \leq \frac{1}{4}.$$

It follows that, uniformly for every $M \geq M_0$,

$$\inf_{\widehat{v}} \sup_{v \in \mathcal{V}} P_v^n(\widehat{v} \neq v) \geq \frac{1}{2}. \tag{C.29}$$

Given a learning rule $\widetilde{f}$, let $\widehat{v} \in \mathcal{V}$ be the unique codeword such that

$$\widetilde{f} = f_{\widehat{v}}.$$

Whenever

$$\widehat{v} \neq v,$$

Corollary C.4 gives

$$\begin{aligned} \Delta_{P_v}(\widetilde{f}) &\geq c_{q,\gamma} B^{1/(q-\gamma)} p_M^\alpha \\ &= c_{q,\gamma} c_*^\alpha B^{1/(q-\gamma)} \left(\frac{\log M}{n} \right)^\alpha. \end{aligned}$$

Therefore, for every $M \geq M_0$,

$$\sup_{P \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M)} P^n \left[\Delta_P(\widetilde{f}) \geq c_{q,\gamma}^{\mathrm{large}} B^{1/(q-\gamma)} \left(\frac{\log M}{n} \right)^\alpha \right] \geq \frac{1}{2}, \tag{C.30}$$

where

$$c_{q,\gamma}^{\mathrm{large}} := c_{q,\gamma} c_*^\alpha.$$

Case 2: $2 \leq M < M_0$**.** We now treat the finitely many small cardinalities by a two-point argument inside the same class $\mathcal{F}_M$.

Choose any two distinct codewords

$$v_0, v_1 \in \mathcal{V}.$$

Set

$$p_0 := \frac{c_{**}}{n}, \tag{C.31}$$

where $c_{**} > 0$ is a sufficiently small universal constant. The regime (C.24), after reducing c_{p} if necessary, implies n is bounded below by a positive universal constant for $M \geq 2$. Thus, c_{**} can be chosen so that

$$0 < p_0 \leq \frac{1}{2}.$$

Let A_0 be given by (C.14) with $p = p_0$, and define

$$P_0 := P_{v_0,p_0,A_0}, \qquad P_1 := P_{v_1,p_0,A_0}.$$

By Lemma C.3,

$$P_0, P_1 \in \mathcal{P}_q(B, \gamma; \mathcal{F}_M).$$

Lemma C.5 gives

$$D_{\mathrm{KL}}(P_0^n \| P_1^n) \leq n p_0 \kappa_\eta = c_{**} \kappa_\eta. \tag{C.32}$$

By the Bretagnolle–Huber inequality, every binary test $\psi \in \{0,1\}$ satisfies

$$\begin{aligned}\max\{P_0^n(\psi=1), P_1^n(\psi=0)\} &\geq \frac{1}{4}\exp\left\{-D_{\mathrm{KL}}(P_0^n \| P_1^n)\right\} \\ &\geq \frac{1}{4}\exp\{-c_{**}\kappa_\eta\}.\end{aligned} \tag{C.33}$$

Choose c_{**} sufficiently small that the last quantity is at least a universal constant $c_{\mathrm{L}} > 0$.

For an arbitrary learning rule $\widetilde{f} : \mathcal{Z}^n \to \mathcal{F}_M$, define

$$\psi := \begin{cases} 0, & \widetilde{f} = f_{v_0}, \\ 1, & \text{otherwise.} \end{cases}$$

Then

$$\{\psi = 1\} = \{\widetilde{f} \neq f_{v_0}\}$$

under P_0, while under P_1,

$$\{\psi = 0\} = \{\widetilde{f} = f_{v_0}\} \subseteq \{\widetilde{f} \neq f_{v_1}\}.$$

Hence,

$$\max\left\{P_0^n(\widetilde{f} \neq f_{v_0}), P_1^n(\widetilde{f} \neq f_{v_1})\right\} \geq c_{\mathrm{L}}. \tag{C.34}$$

The important point is that the learner still takes values in the full M-element class $\mathcal{F}_M$. By the packing property, under either P_i, every incorrect output $f_u \neq f_{v_i}$ satisfies

$$h(u, v_i) \geq \rho.$$

Therefore, Corollary C.4 implies that every learning error incurs

$$\begin{aligned}\Delta_{P_i}(\widetilde{f}) &\geq c_{q,\gamma} B^{1/(q-\gamma)} p_0^\alpha \\ &= c_{q,\gamma} c_{**}^\alpha B^{1/(q-\gamma)} n^{-\alpha}.\end{aligned} \tag{C.35}$$

Since

$$2 \leq M < M_0,$$

we have

$$\log 2 \leq \log M \leq \log M_0.$$

Thus,

$$\log M \asymp 1$$

uniformly over this regime, and in particular

$$n^{-\alpha} \geq (\log M_0)^{-\alpha} \left(\frac{\log M}{n}\right)^\alpha.$$

Consequently,

$$\sup_{P \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M)} P^n\left[\Delta_P(\widetilde{f}) \geq c_{q,\gamma}^{\mathrm{small}} B^{1/(q-\gamma)} \left(\frac{\log M}{n}\right)^\alpha\right] \geq c_{\mathrm{L}}, \tag{C.36}$$

where

$$c_{q,\gamma}^{\mathrm{small}} := c_{q,\gamma} c_{**}^\alpha (\log M_0)^{-\alpha}.$$

Finally, set

$$c_{q,\gamma}^{\mathrm{comp}} := \min\left\{c_{q,\gamma}^{\mathrm{large}}, c_{q,\gamma}^{\mathrm{small}}\right\},$$

and

$$c_{\mathrm{F}} := \min\left\{\frac{1}{2}, c_{\mathrm{L}}\right\}.$$

Both constants are independent of M, B, and n. Combining the two cases proves (C.25) for every $M \geq 2$. □

C.7 THE CONFIDENCE OBSTRUCTION

We next derive the $\log(1/\delta)$ component.

Proposition C.7 (Two-point confidence lower bound)**.** *There exist constants*

$$c_{q,\gamma}^{(\delta)} > 0, \qquad c_\delta > 0,$$

and a universal $\delta_0 \in (0, 1/4)$ *such that the following holds. For every*

$$\delta \in (0, \delta_0]$$

satisfying

$$\log(1/\delta) \le c_\delta n, \tag{C.37}$$

every measurable learning rule

$$\widetilde{f} : \mathcal{Z}^n \to \mathcal{F}_M$$

satisfies

$$\sup_{P \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M)} P^n \left[\Delta_P(\widetilde{f}) \ge c_{q,\gamma}^{(\delta)} B^{1/(q-\gamma)} \left(\frac{\log(1/\delta)}{n} \right)^{\alpha} \right] \ge \delta. \tag{C.38}$$

Proof. Choose any two distinct codewords

$$v_0, v_1 \in \mathcal{V}.$$

Set

$$p_\delta := c_* \frac{\log(1/\delta)}{n}, \tag{C.39}$$

where $c_* > 0$ is chosen below. Under (C.37), c_δ may be chosen so that

$$p_\delta \le \frac{1}{2}.$$

Let A_δ be given by (C.14), and write

$$P_0 := P_{v_0, p_\delta, A_\delta}, \qquad P_1 := P_{v_1, p_\delta, A_\delta}.$$

Both distributions belong to $\mathcal{P}_q(B, \gamma; \mathcal{F}_M)$.

By (C.23),

$$D_{\mathrm{KL}}(P_0^n \| P_1^n) \le n p_\delta \kappa_\eta = c_* \kappa_\eta \log(1/\delta). \tag{C.40}$$

For any binary test $\psi \in \{0, 1\}$, the Bretagnolle–Huber inequality gives

$$P_0^n(\psi = 1) + P_1^n(\psi = 0) \ge \frac{1}{2} \exp\left\{ -D_{\mathrm{KL}}(P_0^n \| P_1^n) \right\}. \tag{C.41}$$

Hence,

$$\begin{aligned} \max\{P_0^n(\psi = 1), P_1^n(\psi = 0)\} &\ge \frac{1}{4} \exp\left\{ -D_{\mathrm{KL}}(P_0^n \| P_1^n) \right\} \\ &\ge \frac{1}{4} \delta^{c_* \kappa_\eta}. \end{aligned} \tag{C.42}$$

Choose $c_* > 0$ sufficiently small that

$$c_* \kappa_\eta \le \frac{1}{2}.$$

Then, for every sufficiently small universal δ_0,

$$\frac{1}{4} \delta^{c_* \kappa_\eta} \ge \delta, \qquad 0 < \delta \le \delta_0. \tag{C.43}$$

Now, let $\widetilde{f}$ be any learning rule. Construct a binary test by

$$\psi := \begin{cases} 0, & \widetilde{f} = f_{v_0}, \\ 1, & \text{otherwise.} \end{cases} \tag{C.44}$$

Under P_0,

$$\{\psi = 1\} = \{\widetilde{f} \neq f_{v_0}\}.$$

Under P_1,

$$\{\psi = 0\} = \{\widetilde{f} = f_{v_0}\} \subseteq \{\widetilde{f} \neq f_{v_1}\}.$$

Therefore,

$$\begin{aligned} \max\left\{P_0^n(\widetilde{f} \neq f_{v_0}), P_1^n(\widetilde{f} \neq f_{v_1})\right\} \\ \geq \max\left\{P_0^n(\psi = 1), P_1^n(\psi = 0)\right\} \geq \delta. \end{aligned} \tag{C.45}$$

For either model, any incorrect hypothesis has Hamming distance at least ρd from the optimal codeword. By Corollary C.4,

$$\Delta_P(\widetilde{f}) \geq c_{q,\gamma} B^{1/(q-\gamma)} p_\delta^\alpha$$

whenever the learner makes an error. Substituting (C.39) proves (C.38). □

C.8 Combining Complexity and Confidence

We now combine the complexity and confidence obstructions. The distributions witnessing these two lower bounds need not coincide: the minimax risk takes a supremum over $P \in \mathcal{P}_q(B, \gamma; \mathcal{F}_M)$, so each obstruction may be witnessed by a different admissible distribution. This is fully consistent with the minimax quantifier

$$\sup_{P \in \mathcal{P}_q(B,\gamma;\mathcal{F}_M)}.$$

Proof of Theorem 3.2. Let

$$a := \log M, \qquad b := \log(1/\delta).$$

Choose the universal constant δ_0 in the theorem no larger than the confidence threshold in Proposition C.7, and such that

$$\delta_0 \leq c_{\mathrm{F}}, \tag{C.46}$$

where c_{F} is the uniform testing-error constant from Proposition C.6.

Assume

$$a + b \leq c'_{q,\gamma} n, \tag{C.47}$$

where $c'_{q,\gamma} > 0$ is sufficiently small. Then both

$$a \lesssim n \qquad \text{and} \qquad b \lesssim n,$$

so the complexity and confidence constructions above are valid.

There are two cases.

Case 1: $a \geq b$. Proposition C.6 gives a distribution $P \in \mathcal{P}_q(B, \gamma; \mathcal{F}_M)$ for which

$$P^n\left[\Delta_P(\widetilde{f}) \geq c_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{a}{n}\right)^\alpha\right] \geq c_{\mathrm{F}}.$$

Since $\delta \leq \delta_0 \leq c_{\mathrm{F}}$, the probability is at least δ. Moreover,

$$a \geq \frac{a+b}{2},$$

so

$$a^\alpha \geq 2^{-\alpha}(a+b)^\alpha.$$

Hence,

$$P^n\left[\Delta_P(\widetilde{f}) \geq c'_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{a+b}{n}\right)^\alpha\right] \geq \delta. \tag{C.48}$$

Case 2: $b > a$**.** Proposition C.7 gives a distribution $P \in \mathcal{P}_q(B, \gamma; \mathcal{F}_M)$ such that

$$P^n \left[\Delta_P(\widetilde{f}) \geq c_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{b}{n} \right)^{\alpha} \right] \geq \delta.$$

Since

$$b \geq \frac{a+b}{2},$$

the same comparison gives

$$P^n \left[\Delta_P(\widetilde{f}) \geq c'_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{a+b}{n} \right)^{\alpha} \right] \geq \delta. \tag{C.49}$$

Finally,

$$a + b = \log M + \log(1/\delta) = \log(M/\delta).$$

Thus, for every measurable learning rule

$$\widetilde{f} : \mathcal{Z}^n \to \mathcal{F}_M,$$

there exists

$$P \in \mathcal{P}_q(B, \gamma; \mathcal{F}_M)$$

such that

$$P^n \left[\Delta_P(\widetilde{f}) \geq c_{q,\gamma} B^{1/(q-\gamma)} \left(\frac{\log(M/\delta)}{n} \right)^{\frac{q-1}{q-\gamma}} \right] \geq \delta. \tag{C.50}$$

This proves Theorem 3.2. □

C.9 Remarks on the construction

The loss class is fixed. For each M, the observation space (C.3) and the loss class (C.5) are fixed before n, δ, and the hard distribution are chosen. The rare-event probability p and amplitude A enter only through the data-generating distribution. Thus the lower bound does not rely on an n-dependent loss function.

The hard distributions have finite absolute risks. For every fixed choice of p and A,

$$T \in \{0, A\},$$

so all losses in the lower-bound construction are bounded by A. Consequently, the minimax obstruction does not arise from any $+\infty - (+\infty)$ pathology. The lower bound already holds in a finite-risk subclass of the relative-risk framework.

Why the fractional exponent appears. The moment constraint forces the calibration

$$A^{q-\gamma} p^{1-\gamma} \asymp B.$$

The signal carried by a wrong hypothesis is of order Ap, hence

$$\begin{aligned} Ap &\asymp B^{1/(q-\gamma)} p^{1-\frac{1-\gamma}{q-\gamma}} \\ &= B^{1/(q-\gamma)} p^{\frac{q-1}{q-\gamma}}. \end{aligned}$$

Meanwhile the statistical information distinguishing the models is only $O(p)$ per sample. Taking

$$np \asymp \log M$$

or

$$np \asymp \log(1/\delta)$$

therefore yields exactly the minimax exponent

$$\frac{q-1}{q-\gamma}.$$

Endpoint checks. When $\gamma = 0$, the calibration gives

$$A \asymp B^{1/q} p^{-1/q}$$

and hence

$$Ap \asymp B^{1/q} p^{(q-1)/q},$$

the ordinary weak-q-moment scale.

When $\gamma = 1$,

$$A \asymp B^{1/(q-1)}$$

is independent of p, and therefore

$$Ap \asymp B^{1/(q-1)} p,$$

yielding the n^{-1} exponent predicted by the fractional Bernstein rate law.

Finally, when $q = 2$,

$$\alpha = \frac{1}{2-\gamma},$$

recovering the classical Bernstein interpolation.

D Proofs for Primitive Verification

This appendix provides complete proofs for the verification results of Section 4. We distinguish carefully between *local* fractional Bernstein geometry, which follows from local regularity and risk growth, and a *global* fractional Bernstein condition over the full candidate class. The latter requires either a localized candidate family or an additional globalization argument.

Throughout this appendix, recall

$$L_f := \ell_f - \ell_{f^\star}, \qquad \Delta_f := PL_f, \qquad Z_f := L_f - \Delta_f.$$

Thus,

$$PZ_f = 0.$$

D.1 Primitive verification from regularity and risk growth

We first prove Proposition 4.3.

Proof of Proposition 4.3. Suppose Assumptions 4.1 and 4.2 hold on $\mathcal{U} \subseteq \mathcal{F}$. Thus, for every $f \in \mathcal{U}$,

$$\|Z_f\|_{L^q(P)} \le L\, d(f, f^\star)^a$$

and

$$\Delta_f \ge c\, d(f, f^\star)^m.$$

Raising the first inequality to the q-th power gives

$$\mathbb{E}|Z_f|^q \le L^q d(f, f^\star)^{aq}. \tag{D.1}$$

The relative-risk growth condition implies

$$d(f, f^\star)^m \le c^{-1} \Delta_f,$$

and hence,

$$d(f, f^\star) \le c^{-1/m} \Delta_f^{1/m}. \tag{D.2}$$

Substituting (D.2) into (D.1),

$$\begin{aligned} \mathbb{E}|Z_f|^q &\le L^q \left(c^{-1/m} \Delta_f^{1/m} \right)^{aq} \\ &= L^q c^{-aq/m} \Delta_f^{aq/m}. \end{aligned} \tag{D.3}$$

This proves

$$\mathbb{E}|Z_f|^q \le L^q c^{-aq/m} \Delta_f^{aq/m}, \qquad f \in \mathcal{U}.$$

The inequality itself holds for every $a, m > 0$. If in addition

$$aq \le m,$$

then

$$\gamma := \frac{aq}{m} \in [0, 1].$$

Therefore, the local fractional Bernstein condition of Definition 2.3 holds on $\mathcal{U}$ with

$$\gamma = \frac{aq}{m}, \qquad B = L^q c^{-aq/m}.$$

□

Remark D.1 (Why $aq \le m$ appears). The restriction $aq \le m$ is not needed to derive (D.3); it is imposed only to ensure that the resulting exponent

$$\gamma = \frac{aq}{m}$$

lies in the range $[0, 1]$ used in Definition 2.3. If $aq/m > 1$, the primitive assumptions still yield the stronger power $\Delta_f^{aq/m}$, but this does not by itself imply a fractional Bernstein condition with exponent in $[0, 1]$. Such a reduction may be possible under additional localization or an a priori upper bound on the excess risk, but it does not follow from Assumptions 4.1 and 4.2 alone.

D.2 The Lipschitz–quadratic specialization

We next prove Corollary 4.4.

Proof of Corollary 4.4. Assume

$$\|Z_f\|_{L^q(P)} \le L\, d(f, f^\star) \tag{D.4}$$

and

$$\Delta_f \ge c\, d(f, f^\star)^2 \tag{D.5}$$

for every $f \in \mathcal{U}$.

This is Proposition 4.3 with

$$a = 1, \qquad m = 2.$$

Hence,

$$\mathbb{E}|Z_f|^q \le L^q c^{-q/2} \Delta_f^{q/2}, \tag{D.6}$$

for all $f \in \mathcal{U}$.

Since $q \in (1, 2]$,

$$\frac{q}{2} \in (1/2, 1],$$

so the corresponding fractional Bernstein exponent is

$$\gamma = \frac{q}{2}.$$

□

Remark D.2 (Resulting statistical exponent). Corollary 4.4 establishes a *local* fractional Bernstein condition with

$$\gamma = \frac{q}{2}.$$

If this condition holds throughout the finite candidate family, or is extended globally via Proposition 4.6, then Theorem 3.1 yields the statistical exponent

$$\frac{q-1}{q - q/2} = \frac{2(q-1)}{q}. \tag{D.7}$$

Equivalently, writing $q = 1 + \epsilon$, the exponent becomes

$$\frac{2\epsilon}{1+\epsilon}.$$

Thus, quadratic relative-risk growth doubles the geometry-free weak-moment exponent $(q-1)/q$, while requiring no second-moment assumption.

D.3 Verification from pairwise loss regularity

We now prove Proposition 4.5.

Lemma D.3 (Centering inequality)**.** *Let $Y \in L^q(P)$ for some $q \geq 1$. Then*

$$\|Y - PY\|_{L^q(P)} \leq 2\|Y\|_{L^q(P)}. \tag{D.8}$$

Proof. By Minkowski's inequality,

$$\|Y - PY\|_{L^q(P)} \leq \|Y\|_{L^q(P)} + |PY|.$$

By Jensen's inequality,

$$|PY| \leq P|Y| \leq \|Y\|_{L^q(P)}.$$

Therefore,

$$\|Y - PY\|_{L^q(P)} \leq 2\|Y\|_{L^q(P)}.$$

□

Proof of Proposition 4.5. For $f \in \mathcal{U}$, define

$$Y_f := \ell_f - \ell_{f^\star}.$$

Then,

$$Z_f = Y_f - PY_f.$$

By Lemma D.3,

$$\|Z_f\|_{L^q(P)} \leq 2\|Y_f\|_{L^q(P)}. \tag{D.9}$$

Applying the assumed pairwise regularity with $g = f^\star$,

$$\|Y_f\|_{L^q(P)} = \|\ell_f - \ell_{f^\star}\|_{L^q(P)} \leq L_0 d(f, f^\star)^a. \tag{D.10}$$

Hence,

$$\|Z_f\|_{L^q(P)} \leq 2L_0 d(f, f^\star)^a. \tag{D.11}$$

If, in addition,

$$\Delta_f \geq c\, d(f, f^\star)^m,$$

then Proposition 4.3, with $L = 2L_0$, gives

$$\mathbb{E}|Z_f|^q \leq (2L_0)^q c^{-aq/m} \Delta_f^{aq/m}. \tag{D.12}$$

This proves the proposition. □

D.4 Random Lipschitz envelopes

The random-envelope criterion used in Section 4 is an immediate consequence of the preceding proposition.

Corollary D.4 (Random-envelope verification)**.** *Suppose that for some measurable random variable $A : \mathcal{Z} \to [0, \infty)$,*

$$|\ell_f(Z) - \ell_g(Z)| \leq A(Z) d(f, g)^a, \qquad \forall f, g \in \mathcal{U}, \tag{D.13}$$

and assume

$$A \in L^q(P).$$

Then,

$$\|\ell_f - \ell_g\|_{L^q(P)} \leq \|A\|_{L^q(P)} d(f, g)^a. \tag{D.14}$$

Consequently,

$$\|Z_f\|_{L^q(P)} \leq 2\|A\|_{L^q(P)} d(f, f^\star)^a. \tag{D.15}$$

If, moreover,

$$\Delta_f \geq c\, d(f, f^\star)^m,$$

then

$$\mathbb{E}|Z_f|^q \leq 2^q \|A\|_{L^q(P)}^q c^{-aq/m} \Delta_f^{aq/m}. \tag{D.16}$$

Proof. Taking $L^q(P)$ norms in (D.13) immediately gives (D.14). The remaining conclusions follow from Proposition 4.5. □

D.5 Globalization of a local condition

We now give the full proof of Proposition 4.6. This step is logically distinct from the primitive verification results above: the latter produce a local fractional Bernstein inequality, whereas Theorem 3.1 requires the inequality to hold over the candidate class to which it is applied.

Proof of Proposition 4.6. Let

$$\mathcal{U}_0 := \{f \in \mathcal{F} : d(f, f^\star) \le r_0\}.$$

For $f \in \mathcal{U}_0$, the assumed local fractional Bernstein condition gives

$$\mathbb{E}|Z_f|^q \le B_{\mathrm{loc}} \Delta_f^\gamma. \tag{D.17}$$

Now consider $f \notin \mathcal{U}_0$. By assumption,

$$\Delta_f \ge \Delta_0 > 0$$

and

$$\mathbb{E}|Z_f|^q \le V_0.$$

Since $\gamma \ge 0$,

$$\Delta_f^\gamma \ge \Delta_0^\gamma.$$

Therefore,

$$\begin{aligned} \mathbb{E}|Z_f|^q &\le V_0 \\ &= V_0 \Delta_0^{-\gamma} \Delta_0^\gamma \\ &\le V_0 \Delta_0^{-\gamma} \Delta_f^\gamma. \end{aligned} \tag{D.18}$$

Combining (D.17) and (D.18),

$$\mathbb{E}|Z_f|^q \le B \Delta_f^\gamma, \qquad \forall f \in \mathcal{F},$$

where

$$B := \max\left\{B_{\mathrm{loc}}, V_0 \Delta_0^{-\gamma}\right\}. \tag{D.19}$$

This proves the proposition. □

D.6 Finite-class globalization

For completeness, we isolate the finite-class consequence used throughout the main text.

Corollary D.5 (Finite-class globalization)**.** *Let $\mathcal{F}$ be finite, and suppose $f^\star$ is the unique relative-risk minimizer in the parameterization under consideration. Assume*

$$\mathbb{E}|Z_f|^q < \infty, \qquad \forall f \in \mathcal{F}. \tag{D.20}$$

Suppose further that, for some $r_0 > 0$,

$$\mathbb{E}|Z_f|^q \le B_{\mathrm{loc}} \Delta_f^\gamma$$

whenever

$$d(f, f^\star) \le r_0.$$

Then, a global fractional Bernstein condition holds over $\mathcal{F}$ with some finite constant B.

Proof. Define the far set

$$\mathcal{F}_{\mathrm{far}} := \{f \in \mathcal{F} : d(f, f^\star) > r_0\}.$$

If

$$\mathcal{F}_{\mathrm{far}} = \varnothing,$$

the local condition already holds over the full class.

Otherwise, since $\mathcal{F}_{\mathrm{far}}$ is finite and $f^\star$ is the unique minimizer,

$$\Delta_f > 0, \qquad \forall f \in \mathcal{F}_{\mathrm{far}}.$$

Hence,

$$\Delta_0 := \min_{f \in \mathcal{F}_{\mathrm{far}}} \Delta_f > 0. \tag{D.21}$$

Likewise, by (D.20),

$$V_0 := \max_{f \in \mathcal{F}_{\mathrm{far}}} \mathbb{E}|Z_f|^q < \infty. \tag{D.22}$$

Proposition 4.6 applies with these choices of Δ_0 and V_0. □

Remark D.6 (Symmetries and quotient parameterizations). The uniqueness assumption in Corollary D.5 should be understood on the parameter space modulo symmetries that leave the statistical object unchanged. In vector quantization, permuting the centers does not change the loss, so codebooks that differ only by a permutation represent the same point in the corresponding quotient space. Accordingly, uniqueness modulo permutations is exactly uniqueness in this quotient space, with distance induced by d_{perm}.

D.7 Compact-class globalization

Although the main minimax theorem concerns finite classes, the vector-quantization application uses compactness to pass from local geometry to the uniform control needed away from the reference minimizer, before restricting to a finite candidate family. We record this elementary globalization argument explicitly.

Lemma D.7 (Positive gap away from a unique optimum)**.** *Let (Θ, d) be compact, and let*

$$\Delta : \Theta \to [0, \infty)$$

be continuous. Suppose

$$\Delta(\theta^\star) = 0$$

and $\theta^\star$ is the unique minimizer in Θ. Then, for every $r_0 > 0$,

$$\inf_{\{\theta : d(\theta, \theta^\star) \geq r_0\}} \Delta(\theta) > 0, \tag{D.23}$$

provided the set over which the infimum is taken is nonempty.

Proof. The set

$$K_{r_0} := \{\theta \in \Theta : d(\theta, \theta^\star) \geq r_0\}$$

is closed in a compact space and hence compact. By continuity, Δ attains its minimum on K_{r_0}. If that minimum were zero, there would exist $\theta \in K_{r_0}$ with

$$\Delta(\theta) = 0,$$

contradicting uniqueness of $\theta^\star$. □

Lemma D.8 (Uniform relative-loss moment on bounded codebooks)**.** *Let*

$$\mathcal{C}_\rho = \left\{ C : \max_j \|c_j\|_2 \leq \rho \right\},$$

and suppose

$$\mathbb{E}\|X\|_2^q < \infty$$

for some $q \geq 1$. Fix $C^\star \in \mathcal{C}_\rho$. Then,

$$\sup_{C \in \mathcal{C}_\rho} \mathbb{E}\,|(\ell_C - \ell_{C^\star}) - P(\ell_C - \ell_{C^\star})|^q < \infty. \tag{D.24}$$

Proof. By Lemma 5.1,

$$|\ell_C(X) - \ell_{C^\star}(X)| \leq 2(\|X\|_2 + \rho) d_{\mathrm{perm}}(C, C^\star).$$

For $C, C^\star \in \mathcal{C}_\rho$,

$$d_{\mathrm{perm}}(C, C^\star) \leq 2\rho\sqrt{k}.$$

Therefore,

$$|\ell_C(X) - \ell_{C^\star}(X)| \leq 4\rho\sqrt{k}(\|X\|_2 + \rho). \tag{D.25}$$

The right-hand side belongs to $L^q(P)$. Hence,

$$\sup_{C \in \mathcal{C}_\rho} \|\ell_C - \ell_{C^\star}\|_{L^q(P)} < \infty.$$

Applying Lemma D.3,

$$\sup_{C \in \mathcal{C}_\rho} \|(\ell_C - \ell_{C^\star}) - P(\ell_C - \ell_{C^\star})\|_{L^q(P)} < \infty,$$

which proves the result. □

Corollary D.9 (Globalization for bounded vector quantization)**.** *Suppose*

$$\mathbb{E}\|X\|_2^q < \infty,$$

and let $C^\star \in \mathcal{C}_\rho$ be unique modulo permutations. Assume that, for some $r_0 > 0$,

$$\mathbb{E}|Z_C|^q \le B_{\mathrm{loc}} \Delta(C)^\gamma$$

whenever

$$d_{\mathrm{perm}}(C, C^\star) \le r_0.$$

Then, there exists a finite constant B such that

$$\mathbb{E}|Z_C|^q \le B \Delta(C)^\gamma, \qquad \forall C \in \mathcal{C}_\rho. \tag{D.26}$$

Proof. Work in the quotient space

$$\mathcal{C}_\rho / S_k$$

equipped with d_{perm}. This quotient is compact. Continuity of the relative distortion follows from the relative-loss Lipschitz bound and $\mathbb{E}\|X\|_2 < \infty$.

By uniqueness modulo permutations and Lemma D.7, there exists

$$\Delta_0 > 0$$

such that

$$\Delta(C) \ge \Delta_0$$

whenever

$$d_{\mathrm{perm}}(C, C^\star) > r_0.$$

By Lemma D.8, there exists $V_0 < \infty$ such that

$$\mathbb{E}|Z_C|^q \le V_0$$

uniformly over $\mathcal{C}_\rho$. Proposition 4.6 now yields (D.26). □

Summary. The results of this appendix provide a direct route from primitive local regularity to the fractional Bernstein geometry used in the main theory. Suppose that, near a reference minimizer $f^\star$, the centered relative-loss increments satisfy an L^q bound of order a,

$$\|Z_f\|_q \le L\, d(f, f^\star)^a,$$

while the relative excess risk grows at order m,

$$\Delta_f \ge c\, d(f, f^\star)^m.$$

Combining these two inequalities eliminates the parameter distance and yields

$$\mathbb{E}|Z_f|^q \le L^q c^{-aq/m}\, \Delta_f^{aq/m}.$$

Thus, whenever $aq \le m$, the fractional Bernstein condition holds locally with

$$\gamma = \frac{aq}{m}, \qquad B = L^q c^{-aq/m}.$$

This identifies the Bernstein exponent as the quantitative balance between stochastic increment regularity and deterministic risk growth, rather than as an independent assumption. When the primitive conditions are only local, so is the resulting fractional Bernstein condition. A finite candidate family contained entirely in the favorable neighborhood can therefore invoke the local condition directly. If a global condition is instead required, the globalization result shows that local geometry can be extended over a compact parameter space provided the reference minimizer is isolated, the excess risk is separated away from zero outside the local neighborhood, and the relevant q-moments are uniformly controlled. Hence, primitive verification and globalization are logically separate steps: the former determines the exponent $\gamma = aq/m$, while the latter determines whether that local geometry can be used globally.

E Relative-Risk Formalism

This appendix formalizes the relative-risk framework used throughout the paper. The central observation is elementary but important: pairwise population comparisons may remain well defined even when the individual population risks are infinite.

We distinguish three objects that should not be conflated:

1. the pointwise loss $\ell_f(Z)$;
2. the population risk $P\ell_f$, which may be infinite;
3. the pairwise loss difference $\ell_f - \ell_g$, whose expectation may nevertheless be finite.

The theory developed in this paper is formulated in terms of the third object. In particular, relative population comparisons are defined directly through $P(\ell_f - \ell_g)$, rather than as differences of the possibly infinite quantities $P\ell_f$ and $P\ell_g$.

E.1 Basic assumptions

Let

$$(\mathcal{Z}, \mathcal{A}, P)$$

be a probability space, and let $\mathcal{F}$ be a hypothesis class.

Throughout the statistical results of the paper, we impose the following operational assumption.

Assumption E.1 (Well-defined observed losses)**.** For every $f \in \mathcal{F}$,

$$\ell_f(Z) \in \mathbb{R} \qquad P\text{-almost surely.} \tag{E.1}$$

The population expectation $P\ell_f$ is allowed to equal $+\infty$.

Assumption E.1 ensures that observed pairwise differences

$$\ell_f(Z) - \ell_g(Z)$$

are pointwise meaningful. It does *not* impose any integrability condition on the individual losses.

For the relative-risk formalism, we additionally assume

$$\ell_f - \ell_g \in L^1(P), \qquad \forall f, g \in \mathcal{F}. \tag{E.2}$$

Condition (E.2) is strictly different from requiring

$$\ell_f \in L^1(P).$$

In particular, two losses may share the same non-integrable component while their difference is integrable.

E.2 Reference-based relative risk

Fix any reference hypothesis

$$f_0 \in \mathcal{F}.$$

Under (E.2), define

$$\mathcal{R}_{f_0}(f) := P(\ell_f - \ell_{f_0}), \qquad f \in \mathcal{F}. \tag{E.3}$$

Every quantity in (E.3) is a finite real number even when

$$P\ell_f = P\ell_{f_0} = +\infty.$$

The following identity is the basic algebraic property of relative risk.

Lemma E.2 (Cocycle identity)**.** *Suppose* (E.2) *holds. Then, for every $f, g, h \in \mathcal{F}$,*

$$P(\ell_f - \ell_g) + P(\ell_g - \ell_h) = P(\ell_f - \ell_h). \tag{E.4}$$

In particular,

$$P(\ell_f - \ell_g) = -P(\ell_g - \ell_f). \tag{E.5}$$

Proof. Pointwise,

$$(\ell_f - \ell_g) + (\ell_g - \ell_h) = \ell_f - \ell_h.$$

All three differences belong to $L^1(P)$, so expectations may be taken termwise, proving (E.4). Equation (E.5) follows by taking $h = f$. □

E.3 Reference invariance

We now prove Proposition 2.2.

Proposition E.3 (Affine change of reference)**.** *For any two references $f_0, g_0 \in \mathcal{F}$,*

$$\mathcal{R}_{g_0}(f) = \mathcal{R}_{f_0}(f) - \mathcal{R}_{f_0}(g_0), \qquad \forall f \in \mathcal{F}. \tag{E.6}$$

Thus, changing the reference adds the same constant to the relative risk of every hypothesis.

Proof. By Lemma E.2,

$$\begin{aligned}\mathcal{R}_{g_0}(f) &= P(\ell_f - \ell_{g_0}) \\ &= P(\ell_f - \ell_{f_0}) + P(\ell_{f_0} - \ell_{g_0}) \\ &= \mathcal{R}_{f_0}(f) - \mathcal{R}_{f_0}(g_0).\end{aligned}$$

□

Proof of Proposition 2.2. Let $f_0, g_0 \in \mathcal{F}$ be arbitrary references. Proposition E.3 shows that

$$\mathcal{R}_{g_0}(f) = \mathcal{R}_{f_0}(f) + c$$

for all f, where

$$c := -\mathcal{R}_{f_0}(g_0)$$

does not depend on f. Hence,

$$\operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_{g_0}(f) = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}_{f_0}(f).$$

Thus, the set of relative-risk minimizers is independent of the reference.

Now let $f^\star$ be any relative-risk minimizer. For every $f \in \mathcal{F}$,

$$\begin{aligned}P(\ell_f - \ell_{f^\star}) &= \mathcal{R}_{f_0}(f) - \mathcal{R}_{f_0}(f^\star) \\ &\geq 0.\end{aligned}$$

Therefore,

$$\Delta_f := P(\ell_f - \ell_{f^\star}) \geq 0. \tag{E.7}$$

Finally, suppose all ordinary population risks are finite. Then linearity of expectation gives

$$P(\ell_f - \ell_{f^\star}) = P\ell_f - P\ell_{f^\star}.$$

Thus, relative excess risk coincides with ordinary excess risk whenever the latter is well defined. □

E.4 Relative optimality without absolute risks

The preceding result allows optimality to be defined without ever forming an expression of the type

$$+\infty - (+\infty).$$

Indeed, a hypothesis $f^\star$ is relatively optimal if and only if

$$P(\ell_f - \ell_{f^\star}) \geq 0, \qquad \forall f \in \mathcal{F}. \tag{E.8}$$

Proposition E.4 (Pairwise characterization of a minimizer)**.** *Under* (E.2)*, the following are equivalent:*

1. *$f^\star$ minimizes $\mathcal{R}_{f_0}$ for one, and hence every, reference f_0;*
2. $$P(\ell_f - \ell_{f^\star}) \geq 0 \qquad \forall f \in \mathcal{F}.$$

Proof. If $f^\star$ minimizes $\mathcal{R}_{f_0}$, then

$$P(\ell_f - \ell_{f^\star}) = \mathcal{R}_{f_0}(f) - \mathcal{R}_{f_0}(f^\star) \geq 0.$$

Conversely, if the pairwise inequalities hold, then

$$\mathcal{R}_{f_0}(f) - \mathcal{R}_{f_0}(f^\star) = P(\ell_f - \ell_{f^\star}) \geq 0,$$

so $f^\star$ minimizes $\mathcal{R}_{f_0}$. □

This characterization is particularly useful when the ordinary population objective is infinite everywhere: optimality is then expressed entirely through finite pairwise expectations.

E.5 Existence of relative-risk minimizers

Reference invariance alone does not guarantee the existence of a relative-risk minimizer. In the settings considered in this paper, existence follows from standard finite-class or compactness arguments, which we record explicitly below.

Lemma E.5 (Existence for finite classes)**.** *Suppose $\mathcal{F}$ is finite and* (E.2) *holds. Then a relative-risk minimizer exists.*

Proof. For a fixed f_0, the map

$$f \mapsto \mathcal{R}_{f_0}(f)$$

is finite-valued on a finite set and therefore attains its minimum. □

The corresponding result for compact parameter spaces is equally standard.

Lemma E.6 (Existence on compact classes)**.** *Let (Θ, d) be compact and suppose that hypotheses are indexed by $\theta \in \Theta$. Fix $\theta_0 \in \Theta$. If*

$$\theta \longmapsto P(\ell_\theta - \ell_{\theta_0})$$

is continuous, then a relative-risk minimizer exists.

Proof. The relative-risk map is a continuous real-valued function on a compact space, so it attains its minimum. □

E.6 Continuity from an integrable increment envelope

A convenient sufficient condition for the continuity required above is an integrable Lipschitz envelope.

Lemma E.7 (Continuity of relative risk)**.** *Suppose that for some pseudo-metric d and some $A \in L^1(P)$,*

$$|\ell_f(Z) - \ell_g(Z)| \leq A(Z)d(f,g), \qquad \forall f, g \in \mathcal{F}. \tag{E.9}$$

Then,

$$|P(\ell_f - \ell_h) - P(\ell_g - \ell_h)| \leq \|A\|_{L^1(P)} d(f,g) \tag{E.10}$$

for every fixed reference h. In particular, relative risk is Lipschitz continuous.

Proof. Using the cocycle identity,

$$\begin{aligned} &|P(\ell_f - \ell_h) - P(\ell_g - \ell_h)| \\ &= |P(\ell_f - \ell_g)| \\ &\leq P|\ell_f - \ell_g| \\ &\leq P[A]\, d(f,g). \end{aligned}$$

□

E.7 Centered relative losses

Once a relative-risk minimizer $f^\star$ has been fixed, define

$$L_f := \ell_f - \ell_{f^\star}, \qquad \Delta_f := PL_f,$$

and

$$Z_f := L_f - \Delta_f. \tag{E.11}$$

Under (E.2), $L_f \in L^1(P)$, so

$$\Delta_f \in [0, \infty)$$

is a finite real number. If additionally

$$L_f \in L^q(P),$$

then

$$Z_f \in L^q(P).$$

No moment assumption on the individual losses ℓ_f or $\ell_{f^\star}$ is needed.

The pairwise difference between two centered relative losses is

$$Z_f - Z_g = (\ell_f - \ell_g) - P(\ell_f - \ell_g). \tag{E.12}$$

Thus, the stochastic process underlying the theory depends only on centered pairwise loss increments.

E.8 Observable pairwise comparisons

The robust selector of Section 3 uses only observations of

$$\ell_f(Z_i) - \ell_g(Z_i).$$

Under Assumption E.1, these quantities are real-valued almost surely even if

$$P\ell_f = +\infty.$$

The corresponding population target is

$$D(f, g) := P(\ell_f - \ell_g),$$

which is finite under (E.2). By the cocycle identity,

$$D(f, g) = \Delta_f - \Delta_g. \tag{E.13}$$

Thus, neither the estimator nor its population target requires the absolute risk $P\ell_f$ to be finite.

Remark E.8 (Why pointwise finiteness is separated from integrability). The condition

$$\ell_f(Z) \in \mathbb{R} \quad P\text{-a.s.}$$

ensures that the observed loss, and hence pairwise loss differences, are well defined on the sample. It does not imply

$$P|\ell_f| < \infty.$$

Indeed, a random variable may be finite almost surely while having an infinite absolute expectation. This distinction is essential for the heavy-tailed vector-quantization setting considered in this paper: squared Euclidean losses are finite pointwise, while their population expectations may be infinite.

E.9 Common non-integrable components

The simplest mechanism producing finite relative losses and infinite absolute risks is cancellation of a common component.

Proposition E.9 (Cancellation principle)**.** *Suppose that*

$$\ell_f(Z) = H(Z) + r_f(Z), \qquad f \in \mathcal{F}, \tag{E.14}$$

where H is measurable but need not be integrable. If

$$r_f - r_g \in L^1(P), \qquad \forall f, g \in \mathcal{F},$$

then

$$\ell_f - \ell_g = r_f - r_g \in L^1(P).$$

Consequently, the entire relative-risk ordering is determined by the remainder family $\{r_f\}$, irrespective of whether PH is finite.

Proof. The common term cancels pointwise:

$$\ell_f - \ell_g = (H + r_f) - (H + r_g) = r_f - r_g.$$

□

This proposition captures the mechanism exploited by squared Euclidean vector quantization: the leading $\|X\|_2^2$ term is common to all codebooks and disappears from relative comparisons.

E.10 Specialization to vector quantization

We now verify the relative-risk formalism used in Section 5.

For a bounded codebook

$$C = \{c_1, \dots, c_k\} \in \mathcal{C}_\rho,$$

recall

$$\ell_C(x) = \min_j \|x - c_j\|_2^2.$$

By Lemma 5.1,

$$|\ell_C(x) - \ell_D(x)| \le 2(\|x\|_2 + \rho) d_{\mathrm{perm}}(C, D). \tag{E.15}$$

Proof of Proposition 5.2. Assume

$$\mathbb{E}\|X\|_2 < \infty.$$

For every $C, D \in \mathcal{C}_\rho$, (E.15) gives

$$P|\ell_C - \ell_D| \le 2\big(\mathbb{E}\|X\|_2 + \rho\big) d_{\mathrm{perm}}(C, D) < \infty. \tag{E.16}$$

Thus, all pairwise loss differences are integrable.

Reference invariance now follows from Proposition 2.2.

Moreover, for any fixed reference C_0,

$$\begin{aligned} |\mathcal{R}_{C_0}(C) - \mathcal{R}_{C_0}(D)| &= |P(\ell_C - \ell_D)| \\ &\le 2\big(\mathbb{E}\|X\|_2 + \rho\big) d_{\mathrm{perm}}(C, D). \end{aligned} \tag{E.17}$$

Hence, the relative-risk map is Lipschitz continuous.

Finally, the quotient

$$\mathcal{C}_\rho / S_k$$

is compact. By Lemma E.6, the continuous relative-risk map attains its minimum. Therefore at least one relative-optimal codebook exists. □

E.11 Infinite absolute distortion

We also justify explicitly the claim that the ordinary squared quantization risk is infinite whenever the data have infinite second moment.

Lemma E.10 (Infinite squared distortion for bounded codebooks)**.** *Let $C \in \mathcal{C}_\rho$. Then*

$$\ell_C(x) \ge (\|x\|_2 - \rho)_+^2. \tag{E.18}$$

Consequently, if

$$\mathbb{E}\|X\|_2^2 = +\infty,$$

then

$$P\ell_C = +\infty, \qquad \forall C \in \mathcal{C}_\rho. \tag{E.19}$$

Proof. For every center c_j with $\|c_j\|_2 \le \rho$, the reverse triangle inequality gives

$$\|x - c_j\|_2 \ge (\|x\|_2 - \rho)_+.$$

Taking the minimum over j,

$$\ell_C(x) = \min_j \|x - c_j\|_2^2 \ge (\|x\|_2 - \rho)_+^2.$$

It remains to note that

$$\mathbb{E}\|X\|_2^2 = +\infty$$

implies

$$\mathbb{E}(\|X\|_2 - \rho)_+^2 = +\infty.$$

Indeed, on the event

$$\{\|X\|_2 \ge 2\rho\},$$

we have

$$(\|X\|_2 - \rho)^2 \ge \frac{1}{4}\|X\|_2^2.$$

Therefore, the truncated tail still has infinite second moment. Equation (E.19) follows. □

Combining Lemma E.10 with (E.16) yields the regime emphasized in the main text:

$$P\ell_C = +\infty \quad \text{for every bounded codebook } C,$$

while simultaneously

$$P|\ell_C - \ell_D| < \infty \quad \text{for every bounded pair } C, D,$$

provided only

$$\mathbb{E}\|X\|_2 < \infty.$$

Thus, relative vector quantization is a genuinely finite-valued statistical problem even when the conventional population distortion is infinite everywhere.

E.12 Summary

The relative-risk framework replaces potentially undefined differences of absolute population risks by expectations of pairwise loss differences. If the observed losses are pointwise finite and $\ell_f - \ell_g \in L^1(P)$ for every $f, g \in \mathcal{F}$, then all pairwise population comparisons

$$D(f, g) := P(\ell_f - \ell_g)$$

are finite and satisfy the cocycle identity $D(f, g) + D(g, h) = D(f, h)$. Consequently, choosing any reference $f_0 \in \mathcal{F}$ defines a finite relative objective

$$R_{f_0}(f) := P(\ell_f - \ell_{f_0}),$$

whose minimizers are independent of the reference. For any relative minimizer $f^\star$, the relative excess risk

$$\Delta_f := P(\ell_f - \ell_{f^\star}) = D(f, f^\star)$$

is therefore well-defined and nonnegative, without requiring the individual expectations $P\ell_f$ to be finite. Whenever the ordinary population risks are finite, this construction reduces exactly to standard excess-risk minimization; when they are infinite, it still provides finite population comparisons and observable pairwise empirical differences. Thus, relative risk supplies the common foundation for both the robust pairwise selector and the fractional Bernstein analysis, allowing the learning theory to remain meaningful even when the absolute population objective itself is undefined or infinite.

F VECTOR QUANTIZATION: RELATIVE-LOSS BOUNDS

This appendix collects the geometric and moment bounds for squared Euclidean vector quantization used in Section 5. We first justify the permutation-invariant parameterization, then prove the relative-loss Lipschitz bound and its L^q consequences, and finally record the compactness and continuity properties used in the globalization argument.

Throughout, let

$$\widetilde{\mathcal{C}}_\rho := \overline{B}(0,\rho)^k, \qquad \mathcal{C}_\rho := \widetilde{\mathcal{C}}_\rho / S_k.$$

Thus, a codebook is represented by an ordered tuple $C = (c_1, \dots, c_k)$, with tuples differing only by a permutation identified. Coincident centers are allowed in the ambient parameter space; this is essential for compactness. The loss is

$$\ell_C(x) := \min_{1 \le j \le k} \|x - c_j\|_2^2.$$

Since the loss is invariant under permutations of the centers, we write

$$d_{\mathrm{perm}}(C, D) := \min_{\pi \in S_k} \left(\sum_{j=1}^{k} \|c_j - d_{\pi(j)}\|_2^2 \right)^{1/2}. \tag{F.1}$$

F.1 PERMUTATION ALIGNMENT AND QUOTIENT GEOMETRY

We begin with several elementary facts about d_{perm}.

Lemma F.1 (Existence of an optimal matching)**.** *For every $C, D \in \mathcal{C}_\rho$, there exists $\pi^\star \in S_k$ such that*

$$d_{\mathrm{perm}}(C, D)^2 = \sum_{j=1}^{k} \|c_j - d_{\pi^\star(j)}\|_2^2. \tag{F.2}$$

Proof. The permutation group S_k is finite. Therefore the minimum in (F.1) is attained. □

After choosing an optimal permutation, we will repeatedly relabel D and write

$$D = (d_1, \dots, d_k)$$

so that

$$d_{\mathrm{perm}}(C, D)^2 = \sum_{j=1}^{k} \|c_j - d_j\|_2^2. \tag{F.3}$$

In particular,

$$\max_j \|c_j - d_j\|_2 \le d_{\mathrm{perm}}(C, D). \tag{F.4}$$

Lemma F.2 (Diameter of the bounded codebook space)**.** *For every $C, D \in \mathcal{C}_\rho$,*

$$d_{\mathrm{perm}}(C, D) \le 2\rho\sqrt{k}. \tag{F.5}$$

Proof. For any permutation π,

$$\|c_j - d_{\pi(j)}\|_2 \le \|c_j\|_2 + \|d_{\pi(j)}\|_2 \le 2\rho.$$

Hence,

$$d_{\mathrm{perm}}(C, D)^2 \le \sum_{j=1}^{k} (2\rho)^2 = 4k\rho^2.$$

□

The quantity d_{perm} is a metric on the quotient of ordered center tuples by permutations. Equivalently,

$$d_{\mathrm{perm}}(C, D) = 0$$

if and only if C and D coincide modulo a permutation of their centers.

F.2 Relative-loss Lipschitz continuity

We now give the complete proof of Lemma 5.1.

Proof of Lemma 5.1. Let $C, D \in \mathcal{C}_\rho$, and choose an optimal permutation as in Lemma F.1. Relabel D so that (F.3) holds.

For any two real vectors $(a_1, \dots, a_k)$ and $(b_1, \dots, b_k)$,

$$\left| \min_j a_j - \min_j b_j \right| \le \max_j |a_j - b_j|. \tag{F.6}$$

Applying this with

$$a_j = \|x - c_j\|_2^2, \qquad b_j = \|x - d_j\|_2^2,$$

gives

$$|\ell_C(x) - \ell_D(x)| \le \max_j \left| \|x - c_j\|_2^2 - \|x - d_j\|_2^2 \right|. \tag{F.7}$$

For each j,

$$\begin{aligned} &\left| \|x - c_j\|_2^2 - \|x - d_j\|_2^2 \right| \\ &= \left| 2\langle x, d_j - c_j \rangle + \|c_j\|_2^2 - \|d_j\|_2^2 \right| \\ &\le 2\|x\|_2 \|c_j - d_j\|_2 + (\|c_j\|_2 + \|d_j\|_2)\|c_j - d_j\|_2 \\ &\le 2(\|x\|_2 + \rho)\|c_j - d_j\|_2. \end{aligned} \tag{F.8}$$

Combining (F.7), (F.8), and (F.4),

$$\begin{aligned} |\ell_C(x) - \ell_D(x)| &\le 2(\|x\|_2 + \rho) \max_j \|c_j - d_j\|_2 \\ &\le 2(\|x\|_2 + \rho) d_{\mathrm{perm}}(C, D). \end{aligned}$$

Thus,

$$|\ell_C(x) - \ell_D(x)| \le 2(\|x\|_2 + \rho) d_{\mathrm{perm}}(C, D). \tag{F.9}$$

Now suppose

$$m_q := (\mathbb{E}\|X\|_2^q)^{1/q} < \infty.$$

Taking $L^q(P)$ norms in (F.9) and using Minkowski's inequality,

$$\begin{aligned} \|\ell_C - \ell_D\|_{L^q(P)} &\le 2\big\|\|X\|_2 + \rho\big\|_{L^q(P)} d_{\mathrm{perm}}(C, D) \\ &\le 2(m_q + \rho) d_{\mathrm{perm}}(C, D). \end{aligned}$$

This proves both assertions of the lemma. □

F.3 Centered relative-loss increments

Fix a relative-optimal codebook $C^\star$, and define

$$\Delta(C) := P(\ell_C - \ell_{C^\star}),$$

and

$$Z_C := (\ell_C - \ell_{C^\star}) - \Delta(C). \tag{F.10}$$

Proposition F.3 (Centered L^q regularity)**.** *Suppose*

$$\mathbb{E}\|X\|_2^q < \infty$$

for some $q \ge 1$. Then, for every $C \in \mathcal{C}_\rho$,

$$\|Z_C\|_{L^q(P)} \le 4(m_q + \rho) d_{\mathrm{perm}}(C, C^\star). \tag{F.11}$$

Consequently,

$$\mathbb{E}|Z_C|^q \le [4(m_q + \rho)]^q d_{\mathrm{perm}}(C, C^\star)^q. \tag{F.12}$$

Proof. Let

$$Y_C := \ell_C - \ell_{C^\star}.$$

By the centering inequality Lemma D.3,

$$\|Z_C\|_{L^q(P)} = \|Y_C - PY_C\|_{L^q(P)} \le 2\|Y_C\|_{L^q(P)}. \tag{F.13}$$

Lemma 5.1 gives

$$\|Y_C\|_{L^q(P)} \le 2(m_q + \rho) d_{\mathrm{perm}}(C, C^\star). \tag{F.14}$$

Combining (F.13) and (F.14) yields (F.11). Raising both sides to the q-th power gives (F.12). □

Thus, the stochastic regularity exponent for vector quantization is

$$a = 1.$$

Combined with the local quadratic-growth result of Theorem 5.4, this is exactly the input used in Corollary 5.5.

F.4 Integrability and continuity of relative distortion

The same Lipschitz inequality also implies that relative distortion is well-defined under only a first moment.

Lemma F.4 (Pairwise integrability)**.** *If*

$$\mathbb{E}\|X\|_2 < \infty,$$

then, for every $C, D \in \mathcal{C}_\rho$,

$$P|\ell_C - \ell_D| < \infty. \tag{F.15}$$

Moreover,

$$|P(\ell_C - \ell_D)| \le 2\big(\mathbb{E}\|X\|_2 + \rho\big) d_{\mathrm{perm}}(C, D). \tag{F.16}$$

Proof. Taking expectations in (F.9),

$$\begin{aligned} P|\ell_C - \ell_D| &\le 2P(\|X\|_2 + \rho) d_{\mathrm{perm}}(C, D) \\ &= 2\big(\mathbb{E}\|X\|_2 + \rho\big) d_{\mathrm{perm}}(C, D). \end{aligned}$$

The right-hand side is finite. □

Fix a reference codebook C_0. Then

$$\mathcal{R}_{C_0}(C) := P(\ell_C - \ell_{C_0})$$

satisfies

$$\begin{aligned} |\mathcal{R}_{C_0}(C) - \mathcal{R}_{C_0}(D)| &= |P(\ell_C - \ell_D)| \\ &\le 2\big(\mathbb{E}\|X\|_2 + \rho\big) d_{\mathrm{perm}}(C, D). \end{aligned} \tag{F.17}$$

Thus, the relative-risk map is globally Lipschitz on the bounded codebook space.

F.5 Compactness of the codebook quotient

We now justify the compactness assertion used in Proposition 5.2 and in the globalization argument.

Let

$$\widetilde{\mathcal{C}}_\rho := \big(\overline{B}(0, \rho)\big)^k \subset (\mathbb{R}^d)^k$$

be the space of ordered k-tuples of centers.

Lemma F.5 (Compact quotient)**.** *The quotient*

$$\widetilde{\mathcal{C}}_\rho / S_k$$

is compact under the metric induced by d_{perm}. Equivalently, the unordered bounded-codebook space $\mathcal{C}_\rho$ is compact with respect to d_{perm}.

Proof. The closed Euclidean ball

$$\overline{B}(0,\rho)$$

is compact in $\mathbb{R}^d$. Therefore, the finite product

$$\widetilde{\mathcal{C}}_\rho = \overline{B}(0,\rho)^k$$

is compact.

The finite group S_k acts continuously on $\widetilde{\mathcal{C}}_\rho$ by permutation of coordinates. Let

$$\pi : \widetilde{\mathcal{C}}_\rho \longrightarrow \widetilde{\mathcal{C}}_\rho / S_k$$

denote the quotient map. The continuous image of a compact space is compact, so the quotient is compact in its quotient topology.

Moreover, the quotient metric is precisely

$$d_{\mathrm{perm}}([C],[D]) = \min_{\pi \in S_k} \|C - \pi D\|_{(\mathbb{R}^d)^k},$$

which induces the same quotient topology because S_k is finite. Hence, the quotient is compact under d_{perm}. □

Combining Lemma F.5 with the continuity bound (F.17) proves existence of a relative-optimal codebook whenever

$$\mathbb{E}\|X\|_2 < \infty.$$

F.6 Local permutation alignment near a separated codebook

The proof of the Relative Pollard theorem writes a nearby codebook as

$$c_j = c_j^\star + h_j.$$

We justify the corresponding local labeling.

Suppose

$$C^\star = \{c_1^\star, \dots, c_k^\star\}$$

has distinct centers and define

$$s_\star := \min_{j \neq \ell} \|c_j^\star - c_\ell^\star\|_2 > 0. \tag{F.18}$$

Lemma F.6 (Stable local matching)**.** *If*

$$d_{\mathrm{perm}}(C, C^\star) < \frac{s_\star}{4}, \tag{F.19}$$

then, after an optimal permutation matching,

$$\|c_j - c_j^\star\|_2 < \frac{s_\star}{4}, \qquad j = 1, \dots, k. \tag{F.20}$$

Moreover, each perturbed center c_j is strictly closer to its matched center $c_j^\star$ than to any other optimal center $c_\ell^\star$, $\ell \neq j$.

Proof. Choose an optimal matching. By (F.4),

$$\|c_j - c_j^\star\|_2 \le d_{\mathrm{perm}}(C, C^\star) < \frac{s_\star}{4}.$$

For $\ell \neq j$,

$$\begin{aligned}\|c_j - c_\ell^\star\|_2 &\ge \|c_j^\star - c_\ell^\star\|_2 - \|c_j - c_j^\star\|_2 \\ &> s_\star - \frac{s_\star}{4} = \frac{3s_\star}{4},\end{aligned}$$

whereas

$$\|c_j - c_j^\star\|_2 < \frac{s_\star}{4}.$$

Thus, the matched optimal center is uniquely closest. □

The local radius restriction

$$r_0 \le \frac{s_\star}{4}$$

used in Appendix G therefore gives a canonical local correspondence between perturbed and optimal centers.

F.7 Uniform L^q control on the bounded codebook class

For globalization, we also need a uniform bound on centered relative losses away from the optimum.

Proposition F.7 (Uniform centered relative-loss moment)**.** *Suppose*

$$\mathbb{E}\|X\|_2^q < \infty$$

for some $q \geq 1$. Fix $C^\star \in \mathcal{C}_\rho$. Then,

$$\sup_{C \in \mathcal{C}_\rho} \|Z_C\|_{L^q(P)} \leq 8\rho\sqrt{k}\,(m_q + \rho). \tag{F.21}$$

Consequently,

$$\sup_{C \in \mathcal{C}_\rho} \mathbb{E}|Z_C|^q < \infty. \tag{F.22}$$

Proof. By Proposition F.3,

$$\|Z_C\|_{L^q(P)} \leq 4(m_q + \rho) d_{\mathrm{perm}}(C, C^\star).$$

Lemma F.2 gives

$$d_{\mathrm{perm}}(C, C^\star) \leq 2\rho\sqrt{k}.$$

Hence,

$$\|Z_C\|_{L^q(P)} \leq 8\rho\sqrt{k}(m_q + \rho),$$

uniformly over C. □

F.8 Positive excess-risk gap away from a unique optimum

Assume now that $C^\star$ is unique modulo permutations.

Because relative risk is continuous and the quotient $\mathcal{C}_\rho / S_k$ is compact, uniqueness implies strict separation from the optimum outside every fixed neighborhood.

Proposition F.8 (Positive relative-risk gap)**.** *Suppose*

$$\mathbb{E}\|X\|_2 < \infty$$

and $C^\star$ is the unique relative-optimal codebook modulo permutations. Then, for every $r > 0$ such that

$$\{C \in \mathcal{C}_\rho : d_{\mathrm{perm}}(C, C^\star) \geq r\} \neq \varnothing,$$

there exists

$$\Delta_r > 0$$

such that

$$\Delta(C) \geq \Delta_r \qquad \textit{whenever } d_{\mathrm{perm}}(C, C^\star) \geq r. \tag{F.23}$$

Proof. The set

$$K_r := \{C : d_{\mathrm{perm}}(C, C^\star) \geq r\}$$

is closed in the compact quotient space and is therefore compact.

The map

$$C \mapsto \Delta(C)$$

is continuous by (F.17). Hence it attains its minimum on K_r.

If this minimum were zero, there would exist $C \in K_r$ with

$$\Delta(C) = 0,$$

so C would also be relative-optimal. Uniqueness modulo permutations would imply

$$d_{\mathrm{perm}}(C, C^\star) = 0,$$

contradicting $C \in K_r$. Thus the minimum is strictly positive. □

F.9 Connection to globalization

Suppose the local Relative Pollard theorem yields, for

$$d_{\mathrm{perm}}(C, C^\star) \le r_0,$$

the local fractional Bernstein condition

$$\mathbb{E}|Z_C|^q \le B_{\mathrm{loc}} \Delta(C)^\gamma. \tag{F.24}$$

If $C^\star$ is unique modulo permutations, then Proposition F.8 gives

$$\Delta(C) \ge \Delta_0 > 0$$

outside this neighborhood, while Proposition F.7 gives

$$\mathbb{E}|Z_C|^q \le V_0 < \infty$$

uniformly over $\mathcal{C}_\rho$.

Therefore, Proposition 4.6 applies and yields

$$\mathbb{E}|Z_C|^q \le B\Delta(C)^\gamma, \qquad \forall C \in \mathcal{C}_\rho, \tag{F.25}$$

with

$$B = \max\left\{B_{\mathrm{loc}}, V_0 \Delta_0^{-\gamma}\right\}.$$

This argument clarifies the role of uniqueness: it is not needed for the local Relative Pollard theorem or for the local fractional Bernstein condition. It is invoked only to obtain the positive far-region excess-risk gap needed for globalization.

Summary. The vector-quantization reduction rests on a cancellation property of relative squared losses. For any $C, D \in \mathcal{C}_\rho$, the common quadratic term in the two distortions cancels, yielding the deterministic increment bound

$$|\ell_C(x) - \ell_D(x)| \le 2(\|x\| + \rho)\, d_{\mathrm{perm}}(C, D).$$

Consequently, if $m_q = (\mathbb{E}\|X\|^q)^{1/q} < \infty$, then the centered relative loss satisfies

$$\|Z_C\|_q \le 4(m_q + \rho)\, d_{\mathrm{perm}}(C, C^\star).$$

Thus, the stochastic regularity needed by the fractional Bernstein theory requires only a q-th moment of $\|X\|$, rather than the $2q$-th moment naturally associated with controlling the absolute squared loss itself. In particular, relative squared losses remain integrable under only a first moment, even in regimes where the absolute population distortions are infinite. When this linear L^q increment control is combined with local quadratic relative-risk growth,

$$\Delta(C) \ge \lambda\, d_{\mathrm{perm}}(C, C^\star)^2,$$

the primitive verification theorem gives the fractional Bernstein exponent

$$\gamma = \frac{q}{2}.$$

Hence, the role of Appendix F is to isolate the probabilistic half of the vector-quantization argument: cancellation reduces the required tail order, while the geometric half, the quadratic lower bound on relative risk, is supplied by the relative Pollard analysis in Appendix G.

G Relative Pollard Geometry

This appendix gives a complete proof of Theorem 5.4. The argument separates two competing effects. If the Voronoi assignments induced by the reference codebook remain unchanged, perturbing the centers produces an exact positive quadratic contribution to the relative distortion. The only possible loss of this curvature comes from points whose nearest-center assignments change under the perturbation. The weighted Voronoi margin condition controls precisely this switching contribution, allowing the positive quadratic term to dominate locally.

Throughout, let

$$C^\star = \{c_1^\star, \dots, c_k^\star\} \in \mathcal{C}_\rho$$

have distinct centers, and let

$$V_1^\star, \dots, V_k^\star$$

be its Voronoi cells under the measurable tie-breaking convention used in Section 5. Recall

$$p_j := P(V_j^\star), \qquad p_{\min} := \min_j p_j > 0. \tag{G.1}$$

We assume

$$P(\partial V^\star) = 0$$

and the centroid conditions

$$\mathbb{E}\left[(X - c_j^\star)\mathbf{1}_{\{X \in V_j^\star\}}\right] = 0, \qquad j = 1, \dots, k. \tag{G.2}$$

For $x \in V_j^\star$, define

$$m_\star(x) := \min_{\ell \neq j} \left[\|x - c_\ell^\star\|_2^2 - \|x - c_j^\star\|_2^2\right]. \tag{G.3}$$

The Weighted Voronoi Margin Condition states that there exist $A_{\mathrm{V}} \geq 0$ and $t_0 > 0$ such that

$$\mathbb{E}\left[(\|X\|_2 + \rho)\mathbf{1}\left\{m_\star(X) \leq t(\|X\|_2 + \rho)\right\}\right] \leq A_{\mathrm{V}} t, \qquad 0 < t \leq t_0. \tag{G.4}$$

G.1 Local alignment of perturbed codebooks

Let

$$C = \{c_1, \dots, c_k\} \in \mathcal{C}_\rho$$

be a codebook close to $C^\star$. Choose a permutation attaining $d_{\mathrm{perm}}(C, C^\star)$, relabel the centers of C accordingly, and write

$$c_j = c_j^\star + h_j, \qquad j = 1, \dots, k. \tag{G.5}$$

Set

$$r := d_{\mathrm{perm}}(C, C^\star) = \left(\sum_{j=1}^k \|h_j\|_2^2\right)^{1/2}. \tag{G.6}$$

Then

$$\|h_j\|_2 \leq r, \qquad j = 1, \dots, k. \tag{G.7}$$

Let

$$s_\star := \min_{j \neq \ell} \|c_j^\star - c_\ell^\star\|_2 > 0. \tag{G.8}$$

By Lemma F.6, if

$$r < \frac{s_\star}{4}, \tag{G.9}$$

the optimal matching gives a stable local labeling of the perturbed centers. We use this labeling throughout the proof.

G.2 Perturbation of Voronoi discriminants

For $j \neq \ell$, define the optimal and perturbed pairwise discriminants

$$g_{j\ell}^\star(x) := \|x - c_\ell^\star\|_2^2 - \|x - c_j^\star\|_2^2 \tag{G.10}$$

and

$$g_{j\ell}^C(x) := \|x - c_\ell\|_2^2 - \|x - c_j\|_2^2. \tag{G.11}$$

Lemma G.1 (Discriminant perturbation)**.** *For every $x \in \mathbb{R}^d$ and every $j \neq \ell$,*

$$\left|g_{j\ell}^C(x) - g_{j\ell}^\star(x)\right| \leq 4r(\|x\|_2 + \rho). \tag{G.12}$$

Proof. For every j,

$$\begin{aligned}&\left|\|x-c_j\|_2^2-\|x-c_j^\star\|_2^2\right|\\&=\left|2\langle x,c_j^\star-c_j\rangle+\|c_j\|_2^2-\|c_j^\star\|_2^2\right|\\&\le 2\|x\|_2\|h_j\|_2+(\|c_j\|_2+\|c_j^\star\|_2)\|h_j\|_2.\end{aligned}$$

Since both codebooks lie in $\mathcal{C}_\rho$,

$$\|c_j\|_2,\|c_j^\star\|_2\le\rho,$$

and hence,

$$\left|\|x-c_j\|_2^2-\|x-c_j^\star\|_2^2\right|\le 2(\|x\|_2+\rho)\|h_j\|_2. \tag{G.13}$$

Therefore,

$$\begin{aligned}&\left|g_{j\ell}^C(x)-g_{j\ell}^\star(x)\right|\\&\le 2(\|x\|_2+\rho)\left(\|h_j\|_2+\|h_\ell\|_2\right)\\&\le 4r(\|x\|_2+\rho),\end{aligned}$$

using (G.7). □

G.3 Switching localization

For $x\in V_j^\star$, define the gain obtained by allowing the perturbed codebook to reassign x:

$$G_C(x):=\|x-c_j\|_2^2-\min_{1\le\ell\le k}\|x-c_\ell\|_2^2. \tag{G.14}$$

Clearly,

$$G_C(x)\ge 0.$$

The next lemma is the key geometric step.

Lemma G.2 (Switching localization and gain)**.** *For every $x\in V_j^\star$,*

$$G_C(x)\le 4r(\|x\|_2+\rho)\mathbf{1}\left\{m_\star(x)\le 4r(\|x\|_2+\rho)\right\}. \tag{G.15}$$

Proof. If the perturbed nearest center is still j, then

$$G_C(x)=0,$$

and the claim is trivial.

Suppose instead that some $\ell\neq j$ attains the perturbed minimum. Then

$$g_{j\ell}^C(x)=\|x-c_\ell\|_2^2-\|x-c_j\|_2^2\le 0. \tag{G.16}$$

Because $x\in V_j^\star$,

$$g_{j\ell}^\star(x)\ge 0. \tag{G.17}$$

Hence,

$$\begin{aligned}g_{j\ell}^\star(x)&\le g_{j\ell}^\star(x)-g_{j\ell}^C(x)\\&\le\left|g_{j\ell}^\star(x)-g_{j\ell}^C(x)\right|\\&\le 4r(\|x\|_2+\rho)\end{aligned}$$

by Lemma G.1.

Since

$$m_\star(x)\le g_{j\ell}^\star(x),$$

any reassignment implies

$$m_\star(x)\le 4r(\|x\|_2+\rho). \tag{G.18}$$

Moreover, when ℓ wins,

$$\begin{aligned}G_C(x)&=\|x-c_j\|_2^2-\|x-c_\ell\|_2^2\\&=-g_{j\ell}^C(x)\\&=-g_{j\ell}^\star(x)+\left(g_{j\ell}^\star(x)-g_{j\ell}^C(x)\right)\\&\le\left|g_{j\ell}^\star(x)-g_{j\ell}^C(x)\right|\\&\le 4r(\|x\|_2+\rho).\end{aligned} \tag{G.19}$$

Combining (G.18) and (G.19) proves the result. □

A useful feature of Lemma G.2 is that no additional factor depending on k appears. Although each observation has multiple possible competing centers, only its winning competitor contributes to $G_C(x)$, while $m_\star(x)$ already minimizes over all competitors.

G.4 Exact fixed-partition quadratic term

We next evaluate the relative distortion that would result if the optimal Voronoi partition were kept fixed.

Lemma G.3 (Fixed-partition expansion)**.** *Under the centroid conditions* (G.2)*,*

$$\sum_{j=1}^{k} \mathbb{E}\left[\mathbf{1}_{\{X\in V_j^\star\}}\left(\|X-c_j\|_2^2-\|X-c_j^\star\|_2^2\right)\right]=\sum_{j=1}^{k} p_j\|h_j\|_2^2. \tag{G.20}$$

Consequently,

$$\sum_{j=1}^{k} p_j\|h_j\|_2^2 \geq p_{\min} r^2. \tag{G.21}$$

Proof. For $x \in V_j^\star$,

$$\begin{aligned}
&\|x-c_j\|_2^2-\|x-c_j^\star\|_2^2\\
&=\|x-c_j^\star-h_j\|_2^2-\|x-c_j^\star\|_2^2\\
&=-2\langle x-c_j^\star,h_j\rangle+\|h_j\|_2^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\{X\in V_j^\star\}}\left(\|X-c_j\|_2^2-\|X-c_j^\star\|_2^2\right)\right]\\
&=-2\left\langle \mathbb{E}\left[(X-c_j^\star)\mathbf{1}_{\{X\in V_j^\star\}}\right],h_j\right\rangle+p_j\|h_j\|_2^2.
\end{aligned}$$

The centroid condition makes the linear term vanish, so

$$\mathbb{E}\left[\mathbf{1}_{\{X\in V_j^\star\}}\left(\|X-c_j\|_2^2-\|X-c_j^\star\|_2^2\right)\right]=p_j\|h_j\|_2^2.$$

Summing over j proves (G.20). Finally,

$$\sum_j p_j\|h_j\|_2^2 \geq p_{\min}\sum_j \|h_j\|_2^2 = p_{\min} r^2.$$

□

Importantly, this is an exact identity rather than a second-order Taylor approximation. It requires only the first moments appearing in the centroid conditions.

G.5 Exact decomposition of relative distortion

For $x \in V_j^\star$,

$$\ell_{C^\star}(x)=\|x-c_j^\star\|_2^2$$

outside a P-null boundary set, while by definition

$$\ell_C(x)=\|x-c_j\|_2^2-G_C(x).$$

Thus,

$$\ell_C(x)-\ell_{C^\star}(x)=\|x-c_j\|_2^2-\|x-c_j^\star\|_2^2-G_C(x). \tag{G.22}$$

Taking expectations and summing over the optimal Voronoi cells, Lemma G.3 yields the exact decomposition

$$\Delta(C)=\sum_{j=1}^{k} p_j\|h_j\|_2^2-\mathbb{E}G_C(X). \tag{G.23}$$

This identity makes the mechanism behind local curvature explicit: the increase in relative distortion consists of a positive quadratic within-cell contribution, corresponding to perturbations that preserve the Voronoi assignments of the reference codebook, offset by the gain that may arise from points whose nearest-center assignments change under the perturbation. Thus, the local quadratic lower bound follows once the reassignment gain is sufficiently small relative to the within-cell quadratic contribution, precisely the role played by the weighted Voronoi margin condition.

G.6 Control of the switching contribution

Suppose

$$r \le \frac{t_0}{4}. \tag{G.24}$$

Then

$$4r \le t_0,$$

so the Weighted Voronoi Margin Condition can be applied at

$$t = 4r.$$

By Lemma G.2,

$$\begin{aligned} \mathbb{E}G_C(X) &\le 4r\mathbb{E}\left[(\|X\|_2+\rho)\mathbf{1}\left\{m_\star(X) \le 4r(\|X\|_2+\rho)\right\}\right] \\ &\le 4rA_{\mathrm{V}}(4r) \\ &= 16A_{\mathrm{V}}r^2. \end{aligned} \tag{G.25}$$

This is the only place where the Weighted Voronoi Margin Condition is used.

G.7 Proof of the Relative Pollard theorem

Proof of Theorem 5.4. Let

$$C \in \mathcal{C}_\rho$$

and choose an optimal permutation alignment with $C^\star$. Write

$$c_j = c_j^\star + h_j$$

and

$$r = d_{\mathrm{perm}}(C, C^\star).$$

Choose

$$r_0 \le \min\left\{\frac{t_0}{4}, \frac{s_\star}{4}\right\}. \tag{G.26}$$

Suppose

$$r \le r_0.$$

By the exact decomposition (G.23),

$$\Delta(C) = \sum_{j=1}^{k} p_j \|h_j\|_2^2 - \mathbb{E}G_C(X).$$

Lemma G.3 gives

$$\sum_{j=1}^{k} p_j \|h_j\|_2^2 \ge p_{\min} r^2,$$

while (G.25) gives

$$\mathbb{E}G_C(X) \le 16A_{\mathrm{V}}r^2.$$

Therefore,

$$\begin{aligned} \Delta(C) &\ge (p_{\min} - 16A_{\mathrm{V}})\, r^2 \\ &= (p_{\min} - 16A_{\mathrm{V}})\, d_{\mathrm{perm}}(C, C^\star)^2. \end{aligned} \tag{G.27}$$

Under the assumed boundary-sparsity condition

$$16A_{\mathrm{V}} < p_{\min},$$

define

$$\lambda := p_{\min} - 16A_{\mathrm{V}} > 0. \tag{G.28}$$

Then,

$$\Delta(C) \geq \lambda d_{\mathrm{perm}}(C, C^\star)^2,$$

which proves the theorem. □

G.8 First-order origin of the centroid condition

For completeness, we record a sufficient condition under which the centroid identities (G.2) follow from relative optimality. This lemma is not needed when the centroid condition is assumed directly in Theorem 5.4.

Lemma G.4 (Centroid condition from interior stationarity)**.** *Suppose*

$$\mathbb{E}\|X\|_2 < \infty, \qquad P(\partial V^\star) = 0,$$

and suppose $C^\star$ is an interior local minimizer of the relative distortion over ordered center tuples. Then,

$$\mathbb{E}\left[(X - c_j^\star)\mathbf{1}_{\{X \in V_j^\star\}}\right] = 0, \qquad j = 1, \ldots, k. \tag{G.29}$$

Proof. Fix j and a direction $u \in \mathbb{R}^d$. Perturb only the j-th center:

$$c_j(t) = c_j^\star + tu,$$

leaving the remaining centers unchanged.

For $x \notin \partial V^\star$, the identity of the nearest optimal center is locally constant in t for sufficiently small t. Thus, the pointwise derivative of the relative-loss difference at $t = 0$ is

$$-2\langle x - c_j^\star, u\rangle \mathbf{1}_{\{x \in V_j^\star\}}. \tag{G.30}$$

The corresponding difference quotients are bounded, for sufficiently small $|t|$, by

$$C\|u\|_2 \left(\|x\|_2 + \rho + \|u\|_2\right), \tag{G.31}$$

which is integrable because

$$\mathbb{E}\|X\|_2 < \infty.$$

Dominated convergence therefore allows differentiation under the expectation:

$$\frac{d}{dt}\mathcal{R}(C(t))\bigg|_{t=0} = -2\left\langle \mathbb{E}\left[(X - c_j^\star)\mathbf{1}_{\{X \in V_j^\star\}}\right], u\right\rangle. \tag{G.32}$$

Since $C^\star$ is an interior local minimizer, this derivative vanishes for every direction u. Hence

$$\mathbb{E}\left[(X - c_j^\star)\mathbf{1}_{\{X \in V_j^\star\}}\right] = 0.$$

□

Remark G.5 (Boundary-constrained centers). If an optimal center lies on the boundary of the constraint set,

$$\|c_j^\star\|_2 = \rho,$$

interior stationarity need not hold in all perturbation directions, and the usual centroid identity does not follow automatically from constrained optimality. Accordingly, when boundary centers are allowed, the centroid equality must be imposed as a separate assumption, as in Theorem 5.4.

G.9 Sanity check: one-center quantization

When $k = 1$, there are no Voronoi boundaries, so

$$A_{\mathrm{V}} = 0, \qquad p_{\min} = 1.$$

The theorem gives

$$\lambda = 1.$$

Indeed, if

$$C^\star = \{c^\star\}$$

satisfies the centroid condition

$$c^\star = \mathbb{E}X,$$

then

$$P\left[\|X - c\|_2^2 - \|X - c^\star\|_2^2\right] = \|c - c^\star\|_2^2,$$

provided only

$$\mathbb{E}\|X\|_2 < \infty.$$

Thus, the constant predicted by Theorem 5.4 is exact in this degenerate no-switching case.

Interpretation. The proof establishes the explicit curvature decomposition

$$\lambda = p_{\min} - 16A_{\mathrm{V}}. \tag{G.33}$$

The first term measures the deterministic within-cell quadratic cost of perturbing the centers, while the second quantifies the maximal loss of curvature caused by Voronoi assignment instability. This decomposition requires only first-order integrability and therefore remains meaningful even when the ordinary squared population distortion is infinite.

H Student-t Boundary Geometry

This appendix proves the weighted boundary regularity result used in Section 6. The main technical point is that the relevant neighborhoods of Voronoi boundaries do not have fixed Euclidean width: their width grows proportionally to $\|x\|_2 + \rho$. We therefore establish a variable-width hyperplane tube bound rather than invoke a standard fixed-width tube formula.

Throughout, let

$$w_\rho(x) := \|x\|_2 + \rho. \tag{H.1}$$

H.1 Polynomial envelopes for Student-t densities

We begin with a simple global envelope.

Lemma H.1 (Polynomial envelope)**.** *Let*

$$f_{\nu,\mu,s}(x) = c_{\nu,d} s^{-d} \left(1 + \frac{\|x - \mu\|_2^2}{\nu s^2}\right)^{-(\nu+d)/2},$$

where

$$\nu > 0, \qquad s > 0.$$

Then, there exists a finite constant

$$C_{\nu,d,\mu,s} < \infty$$

such that

$$f_{\nu,\mu,s}(x) \le C_{\nu,d,\mu,s}(1 + \|x\|_2)^{-(\nu+d)}, \qquad x \in \mathbb{R}^d. \tag{H.2}$$

Consequently, every finite mixture

$$f(x) = \sum_{r=1}^{R} \pi_r f_{\nu,\mu_r,s_r}(x)$$

with common degrees of freedom ν satisfies

$$f(x) \le C_{\mathrm{mix}}(1 + \|x\|_2)^{-(\nu+d)} \tag{H.3}$$

for some finite C_{mix}.

Proof. For

$$\|x\|_2 \geq 2\|\mu\|_2,$$

the reverse triangle inequality gives

$$\|x-\mu\|_2 \geq \|x\|_2 - \|\mu\|_2 \geq \frac{1}{2}\|x\|_2.$$

Hence, outside a fixed bounded region,

$$\begin{aligned} f_{\nu,\mu,s}(x) &\leq c_{\nu,d} s^{-d} \left(1 + \frac{\|x\|_2^2}{4\nu s^2}\right)^{-(\nu+d)/2} \\ &\leq C(1+\|x\|_2)^{-(\nu+d)}. \end{aligned}$$

On the bounded set

$$\{\|x\|_2 < 2\|\mu\|_2\},$$

the Student-t density is bounded and $(1+\|x\|_2)^{-(\nu+d)}$ is bounded away from zero. Increasing the constant gives the global inequality.

The finite-mixture statement follows by summing the componentwise bounds. □

H.2 Weighted mass on affine hyperplanes

Lemma H.2 (Weighted Student-t mass on affine hyperplanes)**.** *Let $f_{\nu,\mu,s}$ be a d-dimensional Student-t density with*

$$\nu > 1, \qquad s > 0,$$

and let $H \subset \mathbb{R}^d$ be any affine hyperplane. Then, for every $\rho \geq 0$,

$$\int_H (\|x\|_2 + \rho)^2 f_{\nu,\mu,s}(x)\, d\mathcal{H}^{d-1}(x) < \infty. \tag{H.4}$$

The same conclusion holds for every finite mixture of such Student-t_ν components.

We now prove Lemma H.2.

Proof of Lemma H.2. Let $H \subset \mathbb{R}^d$ be an affine hyperplane and let

$$f = f_{\nu,\mu,s}.$$

By Lemma H.1,

$$\begin{aligned} &\int_H w_\rho(x)^2 f(x)\, d\mathcal{H}^{d-1}(x) \\ &\leq C \int_H (\|x\|_2 + \rho)^2 (1+\|x\|_2)^{-(\nu+d)}\, d\mathcal{H}^{d-1}(x). \end{aligned} \tag{H.5}$$

For $d \geq 2$, choose any point $x_H \in H$ of minimum norm and identify H with

$$x_H + \mathbb{R}^{d-1}.$$

At large tangential radius r,

$$\|x\|_2 \asymp r,$$

and the surface element has radial order

$$r^{d-2}\, dr.$$

Thus, the tail of the right-hand side of (H.5) is bounded by a constant multiple of

$$\int_1^\infty r^2 r^{-(\nu+d)} r^{d-2}\, dr = \int_1^\infty r^{-\nu}\, dr. \tag{H.6}$$

This is finite when

$$\nu > 1.$$

The contribution from bounded r is finite because the density is bounded.

When $d = 1$, an affine hyperplane is a single point and its $\mathcal{H}^0$-integral is automatically finite.

Therefore, for every $d \geq 1$ and every $\nu > 1$,

$$\int_H (\|x\|_2 + \rho)^2 f_{\nu,\mu,s}(x)\, d\mathcal{H}^{d-1}(x) < \infty.$$

A finite mixture inherits the result by linearity. □

Remark H.3 (Codimension-one versus full-space tails). For $d \geq 2$, the tail of the weighted hyperplane integral reduces, up to multiplicative constants, to a one-dimensional radial integral with integrand of order

$$r^{-\nu}\, dr,$$

whereas the full-space second moment has radial tail of order

$$r^{1-\nu}\, dr.$$

Consequently, the weighted hyperplane integral is finite for $\nu > 1$, while the full-space second moment is finite only for $\nu > 2$. This one-power improvement reflects the codimension-one nature of the Voronoi boundary and explains why weighted boundary regularity can remain finite even when the ambient distribution has infinite variance.

H.3 A VARIABLE-WIDTH HYPERPLANE TUBE LEMMA

We next establish the quantitative result needed for the Weighted Voronoi Margin Condition.

Let $H \subset \mathbb{R}^d$ be an affine hyperplane. Choose a unit normal

$$u \in \mathbb{S}^{d-1}.$$

Every $x \in \mathbb{R}^d$ can be written uniquely as

$$x = y + zu, \qquad y \in H,$$

where

$$|z| = \operatorname{dist}(x, H).$$

Lemma H.4 (Variable-width hyperplane tube)**.** *Let f be a density satisfying*

$$f(x) \leq C_0(1 + \|x\|_2)^{-(\nu+d)} \tag{H.7}$$

for some

$$\nu > 1.$$

Fix

$$\rho \geq 0 \qquad \textit{and} \qquad a > 0.$$

Then, there exist constants

$$t_H > 0, \qquad C_H < \infty$$

such that

$$\begin{aligned} \int_{\mathbb{R}^d} w_\rho(x) \mathbf{1}\left\{\operatorname{dist}(x, H) \leq at\, w_\rho(x)\right\} f(x)\, dx \\ \leq C_H t, \qquad 0 < t \leq t_H. \end{aligned} \tag{H.8}$$

Proof. Write

$$x = y + zu, \qquad y \in H.$$

Lebesgue measure decomposes as

$$dx = d\mathcal{H}^{d-1}(y)\, dz. \tag{H.9}$$

On the tube event,

$$|z| \leq at\big(\|y + zu\|_2 + \rho\big).$$

Using

$$\|y + zu\|_2 \leq \|y\|_2 + |z|,$$

we obtain

$$|z| \le at\big(\|y\|_2 + |z| + \rho\big).$$

Choose t_H initially so that

$$at_H \le \frac{1}{4}. \tag{H.10}$$

Then,

$$(1 - at)|z| \le at(\|y\|_2 + \rho),$$

and hence,

$$|z| \le 2at(\|y\|_2 + \rho). \tag{H.11}$$

Consequently,

$$\begin{aligned} w_\rho(y + zu) &= \|y + zu\|_2 + \rho \\ &\le \|y\|_2 + |z| + \rho \\ &\le (1 + 2at)(\|y\|_2 + \rho) \\ &\le 2(\|y\|_2 + \rho), \end{aligned} \tag{H.12}$$

after decreasing t_H if necessary.

We next compare the density in the allowed normal strip with a tangential polynomial envelope. By (H.11),

$$|z| \le 2at_H(\|y\|_2 + \rho).$$

Choose t_H still smaller so that

$$2at_H \le \frac{1}{4}. \tag{H.13}$$

For sufficiently large $\|y\|_2$, say

$$\|y\|_2 \ge 4\rho,$$

we then have

$$\begin{aligned} \|y + zu\|_2 &\ge \|y\|_2 - |z| \\ &\ge \|y\|_2 - \frac{1}{4}(\|y\|_2 + \rho) \\ &\ge \frac{1}{2}\|y\|_2. \end{aligned} \tag{H.14}$$

Hence, by (H.7),

$$f(y + zu) \le C_1(1 + \|y\|_2)^{-(\nu+d)} \tag{H.15}$$

for all sufficiently large $\|y\|_2$.

For bounded $\|y\|_2$, the region allowed by (H.11) is also bounded, and (H.7) supplies a finite uniform bound. Increasing C_1 therefore makes (H.15) valid for all $y \in H$ and all admissible z.

Using (H.11), (H.12), and (H.15), we obtain

$$\begin{aligned} I_t &:= \int_{\mathbb{R}^d} w_\rho(x)\mathbf{1}\left\{\operatorname{dist}(x, H) \le atw_\rho(x)\right\} f(x)\, dx \\ &\le 2C_1 \int_H (\|y\|_2 + \rho) \int_{|z| \le 2at(\|y\|_2+\rho)} (1 + \|y\|_2)^{-(\nu+d)}\, dz\, d\mathcal{H}^{d-1}(y) \\ &= 8aC_1 t \int_H (\|y\|_2 + \rho)^2 (1 + \|y\|_2)^{-(\nu+d)}\, d\mathcal{H}^{d-1}(y). \end{aligned} \tag{H.16}$$

The final surface integral is finite by the same argument as in Lemma H.2, because $\nu > 1$.

Thus,

$$I_t \le C_H t$$

for a finite C_H, proving the lemma. □

Remark H.5 (Why fixed-width tube formulas are insufficient). At large radius,

$$at(\|x\|_2 + \rho) \asymp at\|x\|_2,$$

so the Euclidean width of the relevant boundary tube grows linearly with $\|x\|_2$. Equivalently, at large radius the tube behaves as an angular sector with aperture of order t, rather than as a strip of uniformly bounded width. Standard fixed-width tube estimates therefore do not directly capture the geometry required here. The variable-width argument above tracks this radial dependence explicitly and yields the uniform tube-to-margin control needed in the relative Pollard analysis.

H.4 Student-t mixtures satisfy the tube bound

Lemma H.1 gives immediately:

Corollary H.6 (Student-t variable-width tubes)**.** *Let*

$$P = \sum_{r=1}^{R} \pi_r t_\nu(\mu_r, s_r^2 I_d), \qquad \nu > 1.$$

For every affine hyperplane $H \subset \mathbb{R}^d$, every $\rho \geq 0$, and every $a > 0$, there exist

$$t_H > 0, \qquad C_H < \infty$$

such that

$$\mathbb{E}\left[w_\rho(X)\mathbf{1}\left\{\operatorname{dist}(X, H) \leq a t w_\rho(X)\right\}\right] \leq C_H t \tag{H.17}$$

for every

$$0 < t \leq t_H.$$

Proof. By Lemma H.1, the finite mixture density satisfies the polynomial envelope (H.7). Apply Lemma H.4. □

H.5 From bisector tubes to the Weighted Voronoi Margin Condition

Let

$$C^\star = \{c_1^\star, \dots, c_k^\star\} \subset B(0, \rho)$$

have distinct centers.

For $j \neq \ell$, recall

$$g_{j\ell}^\star(x) := \|x - c_\ell^\star\|_2^2 - \|x - c_j^\star\|_2^2, \tag{H.18}$$

and let

$$H_{j\ell} := \{x : g_{j\ell}^\star(x) = 0\}$$

be the corresponding bisector hyperplane.

Set

$$s_{j\ell} := \|c_j^\star - c_\ell^\star\|_2 > 0. \tag{H.19}$$

Because $g_{j\ell}^\star$ is affine with

$$\|\nabla g_{j\ell}^\star\|_2 = 2 s_{j\ell},$$

we have the exact identity

$$|g_{j\ell}^\star(x)| = 2 s_{j\ell} \operatorname{dist}(x, H_{j\ell}). \tag{H.20}$$

We can now prove Proposition 6.1.

Proof of Proposition 6.1. For $x \in V_j^\star$,

$$m_\star(x) = \min_{\ell \neq j} g_{j\ell}^\star(x).$$

Therefore, if

$$m_\star(x) \leq t w_\rho(x),$$

there exists some $\ell \neq j$ such that

$$0 \leq g_{j\ell}^\star(x) \leq t w_\rho(x).$$

Consequently,

$$\begin{aligned} &\{m_\star(x) \leq t w_\rho(x)\} \\ &\subseteq \bigcup_{1 \leq j < \ell \leq k} \left\{|g_{j\ell}^\star(x)| \leq t w_\rho(x)\right\}. \end{aligned} \tag{H.21}$$

By (H.20),

$$\begin{aligned}&\left\{|g^\star_{j\ell}(X)| \le t w_\rho(X)\right\}\\&= \left\{\operatorname{dist}(X, H_{j\ell}) \le \frac{t}{2s_{j\ell}} w_\rho(X)\right\}.\end{aligned} \tag{H.22}$$

Apply Corollary H.6 to each pair (j, ℓ), with

$$a_{j\ell} := \frac{1}{2s_{j\ell}}.$$

Because the number of pairs is finite, there exist

$$t_0 := \min_{j<\ell} t_{j\ell} > 0$$

and finite constants $C_{j\ell}$ such that, for $0 < t \le t_0$,

$$\begin{aligned}&\mathbb{E}\left[w_\rho(X)\mathbf{1}\left\{|g^\star_{j\ell}(X)| \le t w_\rho(X)\right\}\right]\\&\le C_{j\ell} t.\end{aligned} \tag{H.23}$$

Using the union bound at the level of indicators,

$$\begin{aligned}&\mathbb{E}\left[w_\rho(X)\mathbf{1}\left\{m_\star(X) \le t w_\rho(X)\right\}\right]\\&\le \sum_{j<\ell} \mathbb{E}\left[w_\rho(X)\mathbf{1}\left\{|g^\star_{j\ell}(X)| \le t w_\rho(X)\right\}\right]\\&\le t \sum_{j<\ell} C_{j\ell}.\end{aligned} \tag{H.24}$$

Setting

$$A_{\mathrm{V}} := \sum_{j<\ell} C_{j\ell} < \infty \tag{H.25}$$

proves

$$\mathbb{E}\left[(\|X\|_2 + \rho)\mathbf{1}\left\{m_\star(X) \le t(\|X\|_2 + \rho)\right\}\right] \le A_{\mathrm{V}} t, \qquad 0 < t \le t_0.$$

Thus every fixed finite Student-t_ν mixture with $\nu > 1$ satisfies the Weighted Voronoi Margin Condition for every fixed finite codebook with distinct centers. □

H.6 Finiteness versus smallness of the boundary constant

Proposition 6.1 establishes that

$$A_{\mathrm{V}} < \infty.$$

This finiteness statement should be distinguished from the strictly stronger smallness condition required by Theorem 5.4,

$$16 A_{\mathrm{V}} < p_{\min}.$$

The former is a tail-regularity property: for Student-t mixtures with $\nu > 1$, the weighted Voronoi boundary functional is finite. The latter is a geometric boundary-sparsity condition ensuring that the switching contribution is sufficiently small for the positive within-cell quadratic term to dominate, and hence for the relative risk to have positive local curvature.

For a generic finite mixture, a large minimum separation between the reference centers,

$$s_\star = \min_{j \ne \ell} \|c^\star_j - c^\star_\ell\|_2,$$

does not by itself imply $16 A_{\mathrm{V}} < p_{\min}$. Indeed, a Voronoi bisector associated with two reference centers may still pass through a region of substantial probability density contributed by another mixture component. Thus, Student-t tails with $\nu > 1$ guarantee $A_{\mathrm{V}} < \infty$, but positive local quadratic growth requires the additional geometric smallness condition $16 A_{\mathrm{V}} < p_{\min}$. The next subsection shows that this condition need not remain an external assumption: for the symmetric two-component Student-t family considered there, A_{V} decays quantitatively with the separation-to-scale ratio, so sufficiently large component separation implies the required boundary sparsity.

Remark H.7 (Generic versus explicit boundary sparsity). For a generic finite mixture, deriving verifiable conditions for

$$16A_{\mathrm{V}} < p_{\min}$$

directly from the component locations, scales, and weights is a separate geometric problem. The symmetric two-component family considered below has sufficiently explicit geometry to obtain the quantitative bound

$$A_{\mathrm{V}}(a) = O_{\nu,d}((s/a)^{\nu}).$$

Consequently, the boundary contribution vanishes as the separation-to-scale ratio a/s tends to infinity, and sufficiently large separation ensures $16A_{\mathrm{V}}(a) < p_{\min} = 1/2$, hence positive local relative-risk curvature.

H.7 Explicit boundary sparsity for a symmetric two-component Student-t mixture

We now strengthen the preceding generic boundary-regularity result by exhibiting a concrete family for which the Weighted Voronoi Margin constant becomes arbitrarily small as the mixture separation grows.

Fix

$$\nu > 1, \qquad s > 0,$$

and, for $a > 0$, consider the symmetric two-component mixture

$$P_a := \frac{1}{2}t_\nu(-ae_1, s^2 I_d) + \frac{1}{2}t_\nu(+ae_1, s^2 I_d). \tag{H.26}$$

Let f_a denote its density.

The distribution P_a is invariant under the reflection

$$x_1 \mapsto -x_1$$

and under rotations of the remaining $d-1$ coordinates.

Define

$$c_a := \mathbb{E}_{P_a}[X_1 \mid X_1 > 0]. \tag{H.27}$$

Since $\nu > 1$, the first moment exists. We consider the symmetric centroidal codebook

$$C_a^\circ := \{-c_a e_1, +c_a e_1\}. \tag{H.28}$$

For convenience, throughout this subsection we take the ambient codebook radius to be

$$\rho_a := 2c_a. \tag{H.29}$$

Thus, the centers of C_a° lie strictly inside the admissible ball $B(0, \rho_a)$. The Voronoi cells of C_a° are

$$V_-^\circ = \{x_1 < 0\}, \qquad V_+^\circ = \{x_1 > 0\}, \tag{H.30}$$

up to the null hyperplane

$$H := \{x_1 = 0\}. \tag{H.31}$$

By symmetry,

$$P_a(V_-^\circ) = P_a(V_+^\circ) = \frac{1}{2}, \qquad p_{\min} = \frac{1}{2}. \tag{H.32}$$

Moreover,

$$\mathbb{E}[X \mid X_1 > 0] = c_a e_1,$$

and, by reflection symmetry,

$$\mathbb{E}[X \mid X_1 < 0] = -c_a e_1.$$

Hence, the centroid identities hold:

$$\mathbb{E}\left[(X - c_a e_1)\mathbf{1}_{\{X_1 > 0\}}\right] = 0, \tag{H.33}$$

and

$$\mathbb{E}\left[(X + c_a e_1)\mathbf{1}_{\{X_1 < 0\}}\right] = 0. \tag{H.34}$$

Thus, the only nontrivial quantitative ingredient required by the relative Pollard theorem is the smallness of the weighted boundary constant.

H.7.1 Asymptotics of the centroid location

We first show that the centroid location tracks the mixture separation.

Lemma H.8 (Centroid asymptotics)**.** *Let T_ν denote a standard univariate Student-t_ν random variable. Then*

$$c_a = \mathbb{E}|a + sT_\nu|. \tag{H.35}$$

Consequently,

$$|c_a - a| \le s\,\mathbb{E}|T_\nu|. \tag{H.36}$$

In particular,

$$\frac{c_a}{a} \longrightarrow 1 \qquad \text{as } a/s \to \infty. \tag{H.37}$$

Moreover, if

$$a \ge 2s\,\mathbb{E}|T_\nu|, \tag{H.38}$$

then

$$\frac{a}{2} \le c_a \le \frac{3a}{2}. \tag{H.39}$$

Proof. The first marginal of P_a has the distribution

$$X_1 \overset{d}{=} Sa + sT_\nu,$$

where S is an independent Rademacher random variable.

Because X_1 is symmetrically distributed around zero,

$$\begin{aligned} c_a &= \mathbb{E}[X_1 \mid X_1 > 0] \\ &= \mathbb{E}|X_1|. \end{aligned}$$

Moreover,

$$|Sa + sT_\nu| \overset{d}{=} |a + sT_\nu|,$$

so

$$c_a = \mathbb{E}|a + sT_\nu|.$$

By the reverse triangle inequality,

$$\big||a + sT_\nu| - a\big| \le s|T_\nu|.$$

Taking expectations gives

$$|c_a - a| \le s\mathbb{E}|T_\nu|.$$

Since

$$\nu > 1,$$

the first absolute moment of T_ν is finite. Dividing by a and sending $a/s \to \infty$ proves (H.37). The two-sided bound (H.39) follows immediately from (H.36). □

H.7.2 Decay of the weighted boundary functional

Define the weighted hyperplane functional

$$\mathcal{B}_a := \int_H (\|x\|_2 + \rho_a)^2 f_a(x)\, d\mathcal{H}^{d-1}(x). \tag{H.40}$$

The next proposition quantifies how this functional decreases with the mixture separation.

Proposition H.9 (Boundary decay under symmetric separation)**.** *There exists a finite constant*

$$C_{\nu,d} < \infty$$

such that, for all sufficiently large a/s,

$$\mathcal{B}_a \le C_{\nu,d} s^\nu a^{1-\nu}. \tag{H.41}$$

Consequently, for every fixed $s > 0$,

$$\mathcal{B}_a \longrightarrow 0 \qquad \text{as } a/s \to \infty. \tag{H.42}$$

Proof. Write a point on the hyperplane H as

$$x = (0, y), \qquad y \in \mathbb{R}^{d-1},$$

and put

$$r := \|y\|_2.$$

On H, the two mixture components have the same density. Therefore

$$f_a(0, y) = c_{\nu,d} s^{-d} \left(1 + \frac{a^2 + r^2}{\nu s^2}\right)^{-(\nu+d)/2}. \tag{H.43}$$

Set

$$p := \frac{\nu + d}{2}$$

and

$$A^2 := a^2 + \nu s^2. \tag{H.44}$$

Then,

$$f_a(0, y) = K_{\nu,d} s^{\nu} (A^2 + r^2)^{-p}, \tag{H.45}$$

where

$$K_{\nu,d} := c_{\nu,d} \nu^{(\nu+d)/2}$$

depends only on (ν, d). By definition,

$$\mathcal{B}_a = K_{\nu,d} s^{\nu} \int_{\mathbb{R}^{d-1}} (r + \rho_a)^2 (A^2 + r^2)^{-p} \, dy. \tag{H.46}$$

Since

$$\begin{aligned} \rho_a &= 2c_a, \\ (r + 2c_a)^2 &\le 2r^2 + 8c_a^2. \end{aligned} \tag{H.47}$$

Hence,

$$\begin{aligned} \mathcal{B}_a \le{}& 2K_{\nu,d} s^{\nu} \int_{\mathbb{R}^{d-1}} r^2 (A^2 + r^2)^{-p} \, dy \\ &+ 8K_{\nu,d} s^{\nu} c_a^2 \int_{\mathbb{R}^{d-1}} (A^2 + r^2)^{-p} \, dy. \end{aligned} \tag{H.48}$$

The standard scaling identity

$$y = Az$$

gives

$$\int_{\mathbb{R}^{d-1}} (A^2 + \|y\|_2^2)^{-p} \, dy = C_{\nu,d}^{(0)} A^{-(\nu+1)}, \tag{H.49}$$

where

$$C_{\nu,d}^{(0)} < \infty.$$

Similarly,

$$\int_{\mathbb{R}^{d-1}} \|y\|_2^2 (A^2 + \|y\|_2^2)^{-p} \, dy = C_{\nu,d}^{(2)} A^{1-\nu}. \tag{H.50}$$

For $d \ge 2$, the latter integral is finite precisely because

$$\nu > 1.$$

When $d = 1$, the integral containing $\|y\|_2^2$ vanishes identically. For sufficiently large a/s, Lemma H.8 gives

$$c_a \le \frac{3a}{2} \le \frac{3A}{2}.$$

Consequently,

$$c_a^2 A^{-(\nu+1)} \lesssim A^{1-\nu}.$$

Substituting (H.49) and (H.50) into (H.48),

$$\mathcal{B}_a \le C_{\nu,d} s^{\nu} A^{1-\nu}.$$

Since

$$\nu > 1$$

and

$$A \geq a,$$

we have

$$A^{1-\nu} \leq a^{1-\nu}.$$

Therefore,

$$\mathcal{B}_a \leq C_{\nu,d} s^\nu a^{1-\nu},$$

which proves the proposition. □

Remark H.10 (Explicit surface integrals). The surface integrals appearing in the preceding proof admit explicit closed forms. For $A > 0$,

$$\begin{aligned} &\int_{\mathbb{R}^{d-1}} (A^2 + \|y\|_2^2)^{-(\nu+d)/2}\, dy \\ &\quad = \pi^{(d-1)/2} \frac{\Gamma((\nu+1)/2)}{\Gamma((\nu+d)/2)} A^{-(\nu+1)}, \end{aligned} \tag{H.51}$$

whereas, for $\nu > 1$,

$$\begin{aligned} &\int_{\mathbb{R}^{d-1}} \|y\|_2^2 (A^2 + \|y\|_2^2)^{-(\nu+d)/2}\, dy \\ &\quad = \frac{d-1}{2} \pi^{(d-1)/2} \frac{\Gamma((\nu-1)/2)}{\Gamma((\nu+d)/2)} A^{1-\nu}. \end{aligned} \tag{H.52}$$

The second identity makes the threshold $\nu > 1$ explicit through the factor $\Gamma((\nu-1)/2)$. This is precisely the integrability threshold for the weighted codimension-one boundary term used in the Student-t margin analysis.

H.7.3 Uniform tube-to-margin control

For the symmetric codebook C_a°, the unique pairwise discriminant is

$$\begin{aligned} g_a(x) &= \|x + c_a e_1\|_2^2 - \|x - c_a e_1\|_2^2 \\ &= 4 c_a x_1. \end{aligned} \tag{H.53}$$

Hence, the optimal Voronoi margin is simply

$$m_a(x) = 4 c_a |x_1|. \tag{H.54}$$

We next turn the boundary functional into a quantitative Weighted Voronoi Margin constant.

Proposition H.11 (Uniform tube-to-margin bound)**.** *There exist a universal constant*

$$\kappa > 0$$

and a finite constant

$$C_{\nu,d}^{\mathrm{tube}} < \infty$$

such that, for all sufficiently large a/s and every

$$\begin{aligned} &0 < t \leq \kappa c_a, \\ &\mathbb{E}_{P_a}\left[(\|X\|_2 + \rho_a)\mathbf{1}\left\{m_a(X) \leq t(\|X\|_2 + \rho_a)\right\}\right] \\ &\quad \leq C_{\nu,d}^{\mathrm{tube}} \frac{\mathcal{B}_a}{c_a} t. \end{aligned} \tag{H.55}$$

Consequently, a valid Weighted Voronoi Margin constant satisfies

$$A_{\mathrm{V}}(a) \leq C_{\nu,d}^{\mathrm{tube}} \frac{\mathcal{B}_a}{c_a}. \tag{H.56}$$

In particular, there exists $C_{\nu,d} < \infty$ such that, for all sufficiently large a/s,

$$A_{\mathrm{V}}(a) \leq C_{\nu,d} \left(\frac{s}{a}\right)^\nu. \tag{H.57}$$

Proof. Write

$$x = (z, y), \qquad z = x_1, \qquad y \in \mathbb{R}^{d-1},$$

and set

$$r := \|y\|_2.$$

Define

$$w_a(z, y) := \sqrt{z^2 + r^2} + 2c_a. \tag{H.58}$$

By (H.54), the margin event is

$$4c_a|z| \le t w_a(z, y). \tag{H.59}$$

Let

$$\tau := \frac{t}{4c_a}.$$

Using

$$\sqrt{z^2 + r^2} \le |z| + r,$$

(H.59) implies

$$|z| \le \tau(|z| + r + 2c_a).$$

If

$$\tau \le \frac{1}{2},$$

then

$$\begin{aligned} |z| &\le \frac{\tau}{1-\tau}(r + 2c_a) \\ &\le 2\tau(r + 2c_a) \\ &= \frac{t}{2c_a}(r + 2c_a). \end{aligned} \tag{H.60}$$

We next compare the density inside this strip to the density on the bisector. Define

$$Q := \nu s^2 + a^2 + r^2. \tag{H.61}$$

For the two mixture components,

$$\begin{aligned} Q_\pm &:= \nu s^2 + r^2 + (z \mp a)^2 \\ &= Q + z^2 \mp 2az. \end{aligned} \tag{H.62}$$

For sufficiently large a/s, Lemma H.8 gives

$$c_a \le 2a.$$

If

$$t \le \kappa c_a,$$

then (H.60) gives

$$\begin{aligned} |z| &\le \frac{\kappa}{2}(r + 2c_a) \\ &\le \frac{\kappa}{2}(r + 4a). \end{aligned} \tag{H.63}$$

Since

$$r + 4a \le \sqrt{17}\sqrt{r^2 + a^2} \le \sqrt{17}\sqrt{Q},$$

we have

$$|z| \le \frac{\kappa\sqrt{17}}{2}\sqrt{Q}. \tag{H.64}$$

Moreover,

$$a \le \sqrt{Q},$$

so

$$2a|z| \le \kappa\sqrt{17}\, Q.$$

Choose

$$\kappa \le \frac{1}{2\sqrt{17}}. \tag{H.65}$$

Then,

$$Q_{\pm} \geq Q - 2a|z| \geq \frac{Q}{2}. \tag{H.66}$$

Let

$$p = \frac{\nu + d}{2}.$$

Since each Student-t component density is proportional to $Q_{\pm}^{-p}$, (H.66) implies

$$Q_{\pm}^{-p} \leq 2^p Q^{-p}.$$

Averaging the two components therefore gives

$$f_a(z, y) \leq 2^{(\nu+d)/2} f_a(0, y) \tag{H.67}$$

throughout the relevant tube. The comparison constant is independent of a, s, r, and t.

We also have, using (H.60) and reducing κ if necessary,

$$\begin{aligned} w_a(z, y) &= \sqrt{z^2 + r^2} + 2c_a \\ &\leq r + |z| + 2c_a \\ &\leq 2(r + 2c_a). \end{aligned} \tag{H.68}$$

For each fixed y, the admissible normal interval has length at most

$$\frac{t}{c_a}(r + 2c_a). \tag{H.69}$$

Combining (H.67), (H.68), and (H.69),

$$\begin{aligned} &\int_{\mathbb{R}} w_a(z, y) \mathbf{1}\left\{m_a(z, y) \leq t w_a(z, y)\right\} f_a(z, y)\, dz \\ &\leq C_{\nu,d} \frac{t}{c_a}(r + 2c_a)^2 f_a(0, y). \end{aligned} \tag{H.70}$$

Integrating over $y \in \mathbb{R}^{d-1}$,

$$\begin{aligned} &\mathbb{E}_{P_a}\left[(\|X\|_2 + \rho_a)\mathbf{1}\left\{m_a(X) \leq t(\|X\|_2 + \rho_a)\right\}\right] \\ &\leq C_{\nu,d} \frac{t}{c_a} \int_H (\|x\|_2 + \rho_a)^2 f_a(x)\, d\mathcal{H}^{d-1}(x) \\ &= C_{\nu,d} \frac{\mathcal{B}_a}{c_a} t. \end{aligned}$$

This proves (H.55) and (H.56).

Finally, Proposition H.9 gives

$$\mathcal{B}_a \leq C_{\nu,d} s^{\nu} a^{1-\nu},$$

whereas Lemma H.8 gives, for sufficiently large a/s,

$$c_a \geq \frac{a}{2}.$$

Therefore,

$$\begin{aligned} A_{\mathrm{V}}(a) &\leq C_{\nu,d} \frac{s^{\nu} a^{1-\nu}}{a} \\ &= C_{\nu,d} \left(\frac{s}{a}\right)^{\nu}, \end{aligned}$$

which proves (H.57). □

H.7.4 Explicit separation implies positive curvature

The preceding result removes the abstract boundary-sparsity assumption for sufficiently separated members of the symmetric family.

Corollary H.12 (Explicit boundary sparsity)**.** *There exists a finite constant*

$$D_0(\nu, d) < \infty$$

such that

$$\frac{a}{s} \geq D_0(\nu, d) \tag{H.71}$$

implies

$$16A_{\mathrm{V}}(a) < \frac{1}{2} = p_{\min}. \tag{H.72}$$

Consequently,

$$\lambda_a := \frac{1}{2} - 16A_{\mathrm{V}}(a) > 0. \tag{H.73}$$

Proof. By Proposition H.11,

$$A_{\mathrm{V}}(a) \leq C_{\nu,d} \left(\frac{s}{a}\right)^{\nu}$$

for all sufficiently large a/s.

Hence it is sufficient that

$$C_{\nu,d} \left(\frac{s}{a}\right)^{\nu} < \frac{1}{32}.$$

Equivalently,

$$\frac{a}{s} > (32C_{\nu,d})^{1/\nu}.$$

Increasing the threshold if necessary so that the large-separation conditions used in Lemma H.8 and Propositions H.9 and H.11 also hold yields a finite constant $D_0(\nu, d)$ with the desired property. □

Remark H.13 (Separation threshold). The proof yields a finite separation threshold $D_0(\nu, d)$ by taking the maximum of the finitely many thresholds required in Lemma H.8 and Propositions H.9 and H.11, together with a threshold ensuring

$$C_{\nu,d}(s/a)^{\nu} < \frac{1}{32}.$$

We do not attempt to optimize the numerical value of $D_0(\nu, d)$, since only its finiteness is needed for the curvature argument. The relevant quantitative conclusion is instead the scale-free decay

$$A_{\mathrm{V}}(a) = O_{\nu,d}((s/a)^{\nu}),$$

which shows that the weighted boundary contribution vanishes as the separation-to-scale ratio a/s tends to infinity.

H.7.5 An explicit infinite-variance local model

The previous estimates verify all ingredients needed for a local Pollard argument around the centroidal reference codebook C_a°. To avoid assuming global optimality of that codebook, we record the local criterion directly.

Proposition H.14 (Local centroidal-reference Pollard criterion)**.** *Let $C^{\circ} = (c_1^{\circ}, \ldots, c_k^{\circ}) \in \mathcal{C}_{\rho}$ have distinct centers. Suppose*

$$\mathbb{E}\|X\|_2 < \infty.$$

Assume

$$P(\partial V^{\circ}) = 0, \qquad p_{\min} := \min_j P(V_j^{\circ}) > 0,$$

the centroid identities

$$\mathbb{E}\left[(X - c_j^{\circ})\mathbf{1}_{\{X \in V_j^{\circ}\}}\right] = 0, \qquad j = 1, \ldots, k,$$

and the Weighted Voronoi Margin Condition

$$\mathbb{E}\left[(\|X\|_2+\rho)\mathbf{1}\left\{m_\circ(X)\le t(\|X\|_2+\rho)\right\}\right]\le A_{\mathrm{V}}t,\qquad 0<t\le t_0.$$

If

$$16A_{\mathrm{V}}<p_{\min},$$

then there exists $r_0>0$ *such that*

$$P(\ell_C-\ell_{C^\circ})\ge \big(p_{\min}-16A_{\mathrm{V}}\big)d_{\mathrm{perm}}(C,C^\circ)^2 \tag{H.74}$$

whenever

$$d_{\mathrm{perm}}(C,C^\circ)\le r_0.$$

Consequently, C° *is a strict local relative-risk minimizer.*

Proof. Choose a permutation realizing $d_{\mathrm{perm}}(C,C^\circ)$ and write

$$c_j=c_j^\circ+h_j,\qquad r^2:=\sum_{j=1}^k\|h_j\|_2^2.$$

For r smaller than one quarter of the minimum separation of the centers of C°, this alignment is locally unambiguous.

Keeping the C°-Voronoi assignments fixed and using the centroid identities gives the exact within-cell contribution

$$\sum_{j=1}^k\mathbb{E}\left[(\|X-c_j\|_2^2-\|X-c_j^\circ\|_2^2)\mathbf{1}_{\{X\in V_j^\circ\}}\right]=\sum_{j=1}^k p_j\|h_j\|_2^2\ge p_{\min}r^2.$$

The only possible reduction comes from observations whose Voronoi assignment changes. As in the switching argument of Appendix G, every pairwise discriminant changes by at most

$$4r(\|x\|_2+\rho).$$

Hence, switching can occur only on

$$\left\{m_\circ(x)\le 4r(\|x\|_2+\rho)\right\},$$

and the corresponding gain is bounded by

$$4r(\|x\|_2+\rho)\mathbf{1}\left\{m_\circ(x)\le 4r(\|x\|_2+\rho)\right\}.$$

Choose

$$r_0\le\min\left\{\frac{t_0}{4},\frac{1}{4}\min_{j\ne\ell}\|c_j^\circ-c_\ell^\circ\|_2\right\}.$$

Then, the Weighted Voronoi Margin Condition with $t=4r$ yields a total switching contribution at most

$$16A_{\mathrm{V}}r^2.$$

Subtracting this contribution from the within-cell quadratic term proves

$$P(\ell_C-\ell_{C^\circ})\ge(p_{\min}-16A_{\mathrm{V}})r^2,$$

which is (H.74). Positivity of the coefficient implies strict local minimality. □

Corollary H.15 (Explicit symmetric Student-t local curvature)**.** *Suppose*

$$\nu>1,\qquad \frac{a}{s}\ge D_0(\nu,d).$$

Then, there exists $r_a>0$ *such that*

$$P_a\left(\ell_C-\ell_{C_a^\circ}\right)\ge\lambda_a d_{\mathrm{perm}}(C,C_a^\circ)^2 \tag{H.75}$$

whenever

$$d_{\mathrm{perm}}(C,C_a^\circ)\le r_a,$$

where

$$\lambda_a:=\frac{1}{2}-16A_{\mathrm{V}}(a)>0.$$

In particular, C_a° *is a strict local relative-risk minimizer.*

Proof. The symmetry of P_a gives

$$P_a(\partial V^\circ) = 0, \qquad p_{\min} = \frac{1}{2},$$

and the centroid identities were verified above. Corollary H.12 gives

$$16A_{\mathrm{V}}(a) < \frac{1}{2}.$$

Apply Proposition H.14. □

Remark H.16 (No global-optimality characterization is needed). The preceding corollary is deliberately local and does not require a characterization of the global population two-means optimum under P_a. The centroid identities, positive cell masses, and explicit boundary-sparsity bound already yield strict local relative-risk curvature around C_a°. Consequently, for any finite candidate family contained in this local neighborhood and containing C_a°, the codebook C_a° is its unique relative-risk minimizer, modulo permutations.

If

$$1 < q < \nu \le 2,$$

then

$$\mathbb{E}_{P_a}\|X\|_2^q < \infty, \qquad \mathbb{E}_{P_a}\|X\|_2^2 = +\infty. \tag{H.76}$$

Thus, the q-moment required for relative-loss concentration is finite even though the ordinary squared-distortion objective is in the infinite-variance regime. Combining Corollary H.15 with the vector-quantization increment bound of Appendix F yields, locally around C_a°,

$$\mathbb{E}\left|(\ell_C - \ell_{C_a^\circ}) - P_a(\ell_C - \ell_{C_a^\circ})\right|^q \le B_{q,a}\left[P_a(\ell_C - \ell_{C_a^\circ})\right]^{q/2}, \tag{H.77}$$

where

$$B_{q,a} = [4(m_q + \rho_a)]^q \lambda_a^{-q/2}. \tag{H.78}$$

Hence, the explicit symmetric Student-t family realizes the fractional Bernstein exponent

$$\gamma = \frac{q}{2}$$

in a regime where $\mathbb{E}_{P_a}\|X\|_2^2 = \infty$.

H.8 Consequences for relative Pollard geometry

Combining the preceding result with Theorem 5.4 gives the precise application chain.

Corollary H.17 (Student-t relative Pollard geometry)**.** *Let*

$$P = \sum_{r=1}^{R} \pi_r t_\nu(\mu_r, s_r^2 I_d), \qquad \nu > 1.$$

Let $C^\star \in \mathcal{C}_\rho$ be a relative-risk minimizer with distinct centers and suppose

$$P(\partial V^\star) = 0,$$

$$p_{\min} > 0,$$

the centroid conditions

$$\mathbb{E}\left[(X - c_j^\star)\mathbf{1}_{\{X \in V_j^\star\}}\right] = 0, \qquad j = 1, \dots, k,$$

hold, and a valid Weighted Voronoi Margin constant satisfies

$$16A_{\mathrm{V}} < p_{\min}.$$

Then, there exist $r_0 > 0$ and

$$\lambda = p_{\min} - 16A_{\mathrm{V}} > 0$$

such that

$$\Delta(C) \ge \lambda d_{\mathrm{perm}}(C, C^\star)^2 \tag{H.79}$$

whenever

$$d_{\mathrm{perm}}(C, C^\star) \le r_0.$$

Proof. Proposition 6.1 supplies a finite Weighted Voronoi Margin constant. The remaining assumptions are exactly those of Theorem 5.4, which gives the conclusion. □

Because Student-t mixtures possess Lebesgue densities,

$$P(\partial V^\star) = 0$$

holds automatically for codebooks with distinct centers: the Voronoi boundary is contained in a finite union of affine hyperplanes.

H.9 Infinite-variance regime

For completeness, we record how the boundary result interacts with the moment assumptions used by the learning theorem.

A d-dimensional Student-t_ν distribution satisfies

$$\mathbb{E}\|X\|_2^r < \infty \qquad \text{for every } 0 < r < \nu, \tag{H.80}$$

whereas

$$\mathbb{E}\|X\|_2^r = +\infty \qquad \text{for } r \geq \nu. \tag{H.81}$$

The same tail threshold holds for a finite mixture with common degrees of freedom ν.

Hence, in the regime

$$1 < q < \nu \leq 2, \tag{H.82}$$

we simultaneously have

$$\mathbb{E}\|X\|_2^q < \infty,$$

and

$$\mathbb{E}\|X\|_2^2 = +\infty.$$

By Appendix F, the first condition gives

$$\|Z_C\|_{L^q(P)} \leq 4(m_q + \rho) d_{\mathrm{perm}}(C, C^\star). \tag{H.83}$$

Under the boundary-sparsity assumptions of Corollary H.17,

$$\Delta(C) \geq \lambda d_{\mathrm{perm}}(C, C^\star)^2. \tag{H.84}$$

Therefore,

$$\begin{aligned} \mathbb{E}|Z_C|^q &\leq [4(m_q + \rho)]^q d_{\mathrm{perm}}(C, C^\star)^q \\ &\leq [4(m_q + \rho)]^q \lambda^{-q/2} \Delta(C)^{q/2}. \end{aligned} \tag{H.85}$$

Thus,

$$\gamma = \frac{q}{2}.$$

For finite candidate families contained in the local neighborhood, Theorem 3.1 consequently yields

$$\Delta(\widehat{C}) \lesssim_q \frac{(m_q + \rho)^2}{\lambda} \left(\frac{\log(M/\delta)}{n} \right)^{\frac{2(q-1)}{q}}. \tag{H.86}$$

At the same time, Appendix E shows that

$$P\ell_C = +\infty$$

for every bounded codebook whenever

$$\mathbb{E}\|X\|_2^2 = +\infty.$$

Thus, (H.86) genuinely operates in a regime where the ordinary population squared-distortion objective is infinite.

H.10 Tail index and limiting rate exponent

For any fixed

$$1 < q < \nu,$$

the Lipschitz–quadratic exponent is

$$\beta(q) = \frac{2(q-1)}{q}.$$

Since

$$\frac{2(q-1)}{q}$$

is increasing in q, letting

$$q \uparrow \nu$$

gives

$$\sup_{1<q<\nu} \beta(q) = \frac{2(\nu-1)}{\nu}. \tag{H.87}$$

The supremum need not be attained because the ν-th moment of a Student-t_ν distribution is not finite. Hence the correct statement is that every exponent

$$\beta < \frac{2(\nu-1)}{\nu}$$

is attainable by choosing $q < \nu$ sufficiently close to ν.

In particular, when

$$\nu = 2,$$

the variance remains infinite, while the attainable rate exponent can be made arbitrarily close to

$$1.$$

Summary of the generic Student-t mechanism. For an arbitrary finite Student-t mixture, the preceding results separate the moment and geometric requirements of the relative-risk theory. If $1 < q < \nu \le 2$, then $\mathbb{E}\|X\|_2^q < \infty$ while $\mathbb{E}\|X\|_2^2 = \infty$, so the q-moment condition required for relative-loss concentration holds genuinely below finite variance. Moreover, the weaker condition $\nu > 1$ guarantees $A_{\mathrm{V}} < \infty$, showing that the weighted Voronoi boundary functional is well defined under only a first moment. Finiteness alone, however, does not imply positive relative curvature: the Relative Pollard theorem additionally requires $16A_{\mathrm{V}} < p_{\min}$, which remains a separate geometric condition for a generic mixture. The symmetric two-component family treated above provides an explicit verification of this missing smallness condition. In that family,

$$A_{\mathrm{V}}(a) = O_{\nu,d}((s/a)^\nu)\,,$$

so increasing the separation-to-scale ratio a/s drives the boundary contribution to zero and eventually ensures $16A_{\mathrm{V}}(a) < p_{\min} = 1/2$. Thus, the symmetric model closes the full chain from weak q-moment integrability, through separation-induced boundary sparsity, to positive local relative-risk curvature in a regime where the ordinary squared population distortion is infinite.