Reweighting Cannot Resolve a Continuum: Key Placement and Prompt Complexity for In-Context Estimation

Anonymous Authors

Abstract

When the latent task variable behind a prompt is continuous, no two prompt examples share a task, so the familiar "retrieve examples from the same class and average" account of in-context learning is unavailable almost surely. We compare the two mechanisms that remain. *Reweighting* treats the prompt as a sample: softmax dictionary attention over prompt examples is self-normalized importance sampling targeting the query posterior. *Interpolation* treats the prompt as a regression design: it estimates the smooth map from latents to observation laws and evaluates it wherever the query requires. For an informative query of m observations, the posterior over a p-dimensional latent has width $m^{-1/2}$, while the nearest of n prior-drawn prompt examples sits at distance $(n\bar{p}(x_0))^{-1/p}$; below $n \asymp m^{p/2}$, *every* reweighting of the prompt—not just softmax weights—misses the posterior at its own scale, with the exact exponential threshold given by a Clarke–Barron expansion. We show this exponent is a *search* cost, not a representation cost: optimally *designed* query-independent keys pay the same $m^{p/2}$, and we prove a matching necessity bound, while keys placed *after* localizing the query cost only polylogarithmically. Interpolation escapes the exponential barrier entirely, requiring $n \gtrsim (m \log m)^{1+p/(2\beta)}$ examples for β-smooth context maps, and we give an end-to-end finite-key pipeline achieving this. The resulting crossover $p > 2\beta/(\beta-1)$ identifies when interpolation provably dominates; we are explicit that the lower bound constrains distributional resolution, and that point-functional readouts of convex reweightings can escape at polynomially smaller prompt lengths.

1 Introduction

The standard mechanism offered for in-context learning is retrieval and averaging: locate prompt examples generated by the same latent task as the query, then pool their sufficient statistics. With a continuous task prior $\bar{\pi}$, this mechanism dies before asymptotics begin—$\mathbb{P}(x_s = x_t) = 0$ for $s \neq t$, so every class count is zero or one almost surely. Two mechanisms survive. The prompt can be used as a *sample*: approximate the query posterior by reweighting the prompt latents, which is exactly what dictionary softmax attention computes. Or it can be used as a *design*: estimate the context map by regression over latent positions and evaluate it at arbitrary candidates, including ones no prompt example visited. This paper shows that the two mechanisms obey different prompt-complexity laws, and locates the obstruction to the first not in the randomness of the prompt but in the timing of key placement.

Why block queries. Throughout, the query is a block of m observations from a single latent x_0—multiple query tokens from one task instance, a long query segment summarized by its sufficient statistic, or test-time aggregation over repeated measurements of the same instance. The block length m is the resolution knob: at $m = 1$ the posterior is as wide as the prior and $O(1)$ prompt length suffices, so every phenomenon below is trivial; the interesting regime, and the one relevant to informative queries, is $m \gg 1$, where the posterior concentrates at scale $m^{-1/2}$ and the prompt must keep up. All statements are asymptotic in m with $n = n(m)$.

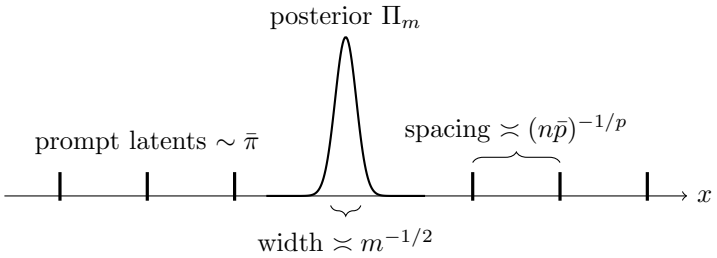


Figure 1: Width versus spacing. Below $n\bar{p}(x_0) \asymp m^{p/2}$ the bulk of the posterior contains no prompt atom, so no reweighting supported on the atoms can resolve it (Theorem 3.3).

Key set	Count (up to logs/constants)	What is covered
Sampled from $\bar{\pi}$	$m^{p/2}\sqrt{\det G(x_0)}/\bar{p}(x_0)$ (Thm. 3.3)	hitting the bulk at all
Designed, query-independent	$\varepsilon^{-p} m^{p/2} V_G$ (Props. 4.1, 4.2)	all of $\mathcal{X}$ at posterior resolution
Adaptive, post-localization	$(\Lambda_m R_\varepsilon/\varepsilon)^p = \mathrm{polylog}(m)$ (Cor. 4.3)	a located ball of radius $\Lambda_m/\sqrt{m}$

Table 1: The key-placement trichotomy. Here $V_G = \int \sqrt{\det G}\, d\nu$ and $\Lambda_m/\sqrt{m}$ is the radius delivered by a localization mechanism.

The width-versus-spacing argument. Figure 1 contains the core of the lower bound. An m-observation query concentrates the posterior into a ball of Fisher width $m^{-1/2}$ around the maximum-likelihood latent. Prompt latents are drawn from the broad prior, so local prior thickness places the nearest one at distance $\Theta_P\{(n\bar{p}(x_0))^{-1/p}\}$. A measure supported on the prompt atoms can resolve the posterior only if some atom lands inside the bulk: $(n\bar{p}(x_0))^{-1/p} \lesssim m^{-1/2}$, i.e. $n\bar{p}(x_0) \gtrsim m^{p/2}$. No weighting scheme can repair an empty bulk—one cannot fix a dartboard with no darts near the bullseye by re-scoring the darts. A Clarke–Barron expansion sharpens the threshold to $n \asymp e^{L_m}$ with

$$L_m := \mathrm{KL}(\Pi_m \| \bar{\pi}) = \underbrace{\tfrac{p}{2}\log\tfrac{m}{2\pi e}}_{\text{resolution}} + \underbrace{\tfrac{1}{2}\log\det G(x_0)}_{\text{geometry}} + \underbrace{\log\tfrac{1}{\bar{p}(x_0)}}_{\text{typicality}} + o_P(1),$$

where G is the pullback Fisher metric. Theorem 3.3 proves failure below this threshold for *arbitrary* weights on the prompt support, in Wasserstein distance standardized by the posterior's own scale (the total-variation version being vacuous for atomic approximants); Proposition 3.4 shows exact softmax weights succeed just above it, so the threshold is two-sided at exponential scale.

A search cost, not a representation cost. The exponent $m^{p/2}$ does not reflect any difficulty of *representing* the posterior—it reflects the difficulty of *finding* it with keys committed before the query. Table 1 summarizes the trichotomy. A query-independent designed net pays the same m-exponent as prior-sampled keys because it must cover all of $\mathcal{X}$ at posterior resolution; we prove this necessity, not only sufficiency (Propositions 4.1–4.2). Randomness is therefore not the obstruction; query-independence is. Only keys placed *after* another mechanism localizes the query escape, at polylogarithmic cost (Corollary 4.3). A single attention layer scores a key set chosen before the query and so pays the search cost; localization is a sequential mechanism, which is why depth (or a two-round coarse-to-fine pass) is the natural—though not information-theoretically necessary—implementation of interpolation.

Scope of the lower bound: distributions, not point readouts. We are explicit about what the barrier does and does not forbid. Theorem 3.3 constrains the *measure* produced by reweighting: any distribution supported on the prompt atoms is far from the posterior at posterior scale, so any downstream use requiring distributional resolution—uncertainty quantification, sampling, tail functionals—inherits the $m^{p/2}$ cost, and Proposition 3.5 shows the exact softmax *point* output collapses onto the nearest atom and fails too. But a convex reweighting with *non-Bayes* logits can produce accurate point summaries below the threshold: tempering the softmax widens the effective kernel and turns the attention mean into a Nadaraya–Watson estimate of the posterior mode, accurate once $n\bar{p}(x_0) \gg m^{(p+2)/4}$ (Remark 5.5)—polynomially below $m^{p/2}$ for every $p > 2$. Such widened-bandwidth readouts are interpolation of the target functional performed inside the convex hull of the atoms, and we classify them with the interpolation mechanism. The honest statement of our dichotomy is therefore: *Bayes-weighted retrieval cannot resolve a continuous posterior as a distribution below $m^{p/2}$; mechanisms that smooth—whether by explicit regression or by widening the attention kernel—escape polynomially.*

Contributions.

(i) *Reweighting lower bound, arbitrary weights.* Dictionary attention is self-normalized importance sampling (SNIS); we prove a resolution lower bound at threshold e^{L_m} for arbitrary measures on the prompt support in posterior-standardized W_1, compute the exponent via a Clarke–Barron expansion, and match it from above at exponential scale (Theorem 3.3, Proposition 3.4). The SNIS point output fails at the same threshold (Proposition 3.5).

(ii) *Key-placement trichotomy with matching necessity.* Query-independent designed nets require $\Theta(m^{p/2})$ keys—sufficiency *and* necessity—while post-localization keys cost polylogarithmically (Propositions 4.1–4.2, Corollary 4.3).

(iii) *Interpolation upper bound and an end-to-end pipeline.* A smoothing-plus-scoring construction has block-query plug-in error $\sqrt{m \log m}\, r_n + m r_n^2$, giving sufficient prompt length $(m \log m)^{1+p/(2\beta)}$; a finite-key, estimated-context, two-round coarse-to-fine pipeline achieves the same error with polylog(m) keys (Theorems 5.2–5.3, Corollary 5.4). The crossover $p > 2\beta/(\beta - 1)$ identifies when interpolation provably dominates; we state explicitly that the regime below the crossover is left open by our one-sided comparison (Section 5.1).

Related work. In-context learning as implicit Bayesian inference over a latent task variable originates with Xie et al. [12]; our setting is its continuum limit. The finite-latent-class theory of Rubacher [8] is recovered exactly as the atomic-prior limit of our lower bound (Remark 3.7). A complementary line studies transformers performing in-context *regression*: empirically [7], via constructions implementing gradient descent or least squares in the forward pass [11, 2], and via statistical guarantees for transformer-implementable algorithms [3]. That literature instantiates our interpolation mechanism; our contribution is the matching impossibility result for the retrieval mechanism and the resulting crossover. Closest in spirit is the retrieval-versus-learning dichotomy of **?**] for finite mixtures; the continuum prior is precisely the regime where their retrieval mode has measure zero, and our results quantify its cost as the prior refines. Prior-fitted networks [**?**] amortize the posterior into weights at training time and thus sit outside both prompt-driven mechanisms; they pay the search cost during pre-training rather than at inference. Softmax attention as kernel smoothing is due to Tsai et al. [9], and the sample-size law $n \asymp e^{\mathrm{KL}}$ for importance sampling is classical [4, 1]; our contribution is not that law but (i) the computation of its exponent for the continuum-latent query problem via a Clarke–Barron expansion [5, 6],

(ii) its extension from SNIS weights to *arbitrary* measures on the prompt support, and (iii) the posterior-standardized Wasserstein formulation. Random-feature approximations of attention [13] sample over feature variables at fixed attention output; we sample over key locations with a Bayes posterior as target. The Bernstein–von Mises theorem we use is standard [**?**]; sup-norm kernel regression rates are from Tsybakov [10].

2 Setup

Let $(\mathcal{X}, \mathcal{A})$ be a latent space with prior probability measure $\bar{\pi}$; when $\mathcal{X}$ carries a volume measure ν, write $\bar{p} = d\bar{\pi}/d\nu$. A context θ specifies a natural-parameter field $\eta_\theta : \mathcal{X} \to \mathbb{R}^k$ in a regular minimal exponential family

$$p_\theta(y \mid x) = h(y) \exp\{\eta_\theta(x)^\top T(y) - A(\eta_\theta(x))\}. \tag{1}$$

The mean map is $\mu = \nabla A$, the Fisher information in natural coordinates is $F = \nabla^2 A$, the pullback Fisher metric is $G(x) = J(x)^\top F(\eta_\theta(x)) J(x)$ with $J(x) = \partial \eta_\theta(x)/\partial x$, and $D_A(\eta', \eta) = A(\eta') - A(\eta) - \nabla A(\eta)^\top(\eta' - \eta)$.

A *prompt* is a sequence $(x_t, y_t)_{t \le n}$ with latents drawn i.i.d. from $\bar{\pi}$ and responses from (1); the latents are observed—in the in-context-regression reading, the "latent" is the input x_t and the hidden object is the context field η_θ, so our setting matches Garg et al. [7] rather than the hidden-task setting of Xie et al. [12], whose continuum limit it nevertheless analyzes. The *query* is a block of m i.i.d. observations at latent x_0, summarized by $\bar{T}_m$.

Continuum attention is the integral-operator limit of dictionary softmax attention: a head with keys k, query q, and bias b produces

$$\alpha(dx \mid y) = \frac{\exp\{k(x)^\top q(y) + b(x)\}}{\int_{\mathcal{X}} \exp\{k(u)^\top q(y) + b(u)\}\, \bar{\pi}(du)}\, \bar{\pi}(dx), \tag{2}$$

and a finite-key head with key set $\mathcal{D}_N = \{x_1, \dots, x_N\} \subset \mathcal{X}$ produces an atomic measure supported on $\mathcal{D}_N$, with quadrature masses q_i entering as logit biases $\log q_i$. Bayes' rule corresponds to the oracle choice $k = \eta_\theta$, $q = T$, $b = -A \circ \eta_\theta$, giving the posterior logit field $s_\theta(x; y) = \eta_\theta(x)^\top T(y) - A(\eta_\theta(x))$ up to y-only terms; normalizability and the block Bregman form $\pi_\theta(dx \mid y_{1:m}) \propto \exp\{-m D_{A^*}(\bar{T}_m, \mu_\theta(x))\}\bar{\pi}(dx)$ are recorded in Appendix B. We write $\Pi_m := \pi_\theta(\cdot \mid y_{1:m})$ for the query posterior throughout.

Let d_G be the Riemannian distance of the pullback metric, $\hat{x}_m$ the block maximum-likelihood latent, and

$$B_m(R) := \{x : d_G(x, \hat{x}_m) \le R/\sqrt{m}\}$$

the tunable posterior bulk; its posterior tail is $O(e^{-cR^2}) + e^{-cm}$. Every "bulk" below is an instance of $B_m(R)$ with R chosen per result ($R = \sqrt{K \log m}$ for the expansions and stability results, $R \asymp \sqrt{s}$ in the lower bound, accuracy-driven R_ε for key discretization). Chart conventions, the equivalence of d_G with fixed chart metrics under compactness, and the interchangeability of balls centered at x_0 versus $\hat{x}_m$ are routine and deferred to Appendix A. Total variation is $\|P - Q\|_{\mathrm{TV}} = \sup_A |P(A) - Q(A)|$; W_1 is Wasserstein-1 for a fixed chart metric d; $\kappa\{G\} := \lambda_{\max}(G)/\lambda_{\min}(G)$. Here m is query-block length, n prompt length, p latent dimension, and $\mathcal{H}^\beta$ the Hölder–Zygmund class.

2.1 Running systems

(S1) Gaussian location. $\mathcal{X} \subset \mathbb{R}^p$ compact, $Y \mid x \sim \mathcal{N}(x, \sigma^2 I)$: $\eta_\theta(x) = x/\sigma^2$, $G \equiv \sigma^{-2} I$, $V_G = \sigma^{-p}\nu(\mathcal{X})$. This is the flat testbed for the key-placement results; all constants of Sections 3–4 are made explicit for it in Example 3.2, and it is the system of the diagnostics in Section 6.

(S2) Gaussian unknown gain. As (S1) with $p = 1$ and $Y \mid x \sim \mathcal{N}(hx, \sigma^2)$, the context being the scalar gain $\theta = h$; the context map is finite-dimensional, enabling the one-layer method-of-moments readout of Section 5.2.

A curved example (warped von Mises on S^1), with its Clarke–Barron constant instantiated, appears in Appendix H.1.

2.2 Bernstein–von Mises scale for attention

Assumption 2.1. *$\mathcal{X}$ is a compact set in $\mathbb{R}^p$ with $x_0 \in \operatorname{int}\mathcal{X}$, or a boundaryless compact C^2 manifold of dimension p read through a chart around x_0; $\eta_\theta \in C^2$ with Lipschitz Hessian, injective, with full-rank Jacobian; the family (1) is regular and minimal; the prior density $\bar{p}$ is continuous and positive at x_0.*

Theorem 2.2 (BvM for attention weights)**.** *Under Assumption 2.1, with the query as above and $\alpha_\theta(\cdot \mid \bar{T}_m)$ the exact continuum attention measure, in any fixed local C^2 chart ψ with $\hat{z}_m = \psi(\hat{x}_m)$,*

$$\big\|\psi_\# \alpha_\theta(\cdot \mid \bar{T}_m) - \mathcal{N}\big(\hat{z}_m, \tfrac{1}{m} G_\psi(x_0)^{-1}\big)\big\|_{\mathrm{TV}} \xrightarrow{P} 0,$$

where $G_\psi(x_0) = J_\psi(x_0)^\top F(\eta_\theta(x_0)) J_\psi(x_0)$; the same holds with $G_\psi(\hat{x}_m)$.

The proof (Appendix B) is a Laplace argument organized by the logit Hessian decomposition

$$-\nabla_x^2 s_\theta(x; \bar{T}_m) = \underbrace{J(x)^\top F J(x)}_{\text{pullback Fisher}} - \underbrace{\textstyle\sum_j \partial_x^2 \eta_{\theta,j}(x)\big(\bar{T}_m - \mu_\theta(x)\big)_j}_{O_P(m^{-1/2}) \text{ at the mode}}, \tag{3}$$

so extrinsic curvature contributes only a fluctuation and the pullback Fisher metric fixes the posterior scale used throughout.

3 Reweighting: the Lower Bound

Suppose first the context is known and the prompt supplies only latents used as keys; nothing is estimated, so everything in this section is a *key-placement* statement, and it applies a fortiori when reweighting must also estimate the context. Dictionary attention computes

$$\hat{g}_n(y) = \frac{\sum_{t \le n} g(x_t)\, w(x_t)}{\sum_{t \le n} w(x_t)}, \qquad w := \frac{d\,\pi(\cdot \mid y)}{d\,\bar{\pi}},$$

up to cancelling y-only factors: exactly SNIS with proposal $\bar{\pi}$ and target the posterior. Writing $w_m = d\Pi_m / d\bar{\pi}$, the SNIS effective sample size is

$$n_{\mathrm{eff}} = \frac{n}{1 + \chi^2(\Pi_m \| \bar{\pi})} \asymp \frac{n \bar{p}(x_0) m^{-p/2}}{\sqrt{\det G(x_0)}},$$

so the threshold is $n_{\mathrm{eff}} \gtrsim 1$. The classical Clarke–Barron expansion refines the scale:

Lemma 3.1 (Clarke–Barron expansion)**.** *Under Assumption 2.1 with $\eta_\theta \in C^3$,*

$$L_m := \mathrm{KL}(\Pi_m \| \bar{\pi}) = \frac{p}{2} \log \frac{m}{2\pi e} + \frac{1}{2} \log \det G(x_0) + \log \frac{1}{\bar{p}(x_0)} + o_P(1), \qquad \mathrm{Var}_{\Pi_m}(\log w_m) = \frac{p}{2} + o_P(1).$$

The verification (Appendix D) exploits exponential-family structure to obtain uniformity from the single statistic $\bar{T}_m$. We measure W_1 in the posterior's standard-deviation units, making the criterion scale-free across m; we standardize by the scalar $\sqrt{m}$ rather than $G(x_0)^{1/2}$, which leaves the condition number $\kappa\{G(x_0)\}$ in constants but not in exponents.

Example 3.2 (System (S1) instantiated)**.** In (S1), $G \equiv \sigma^{-2} I$ and $\det G = \sigma^{-2p}$, so Lemma 3.1 reads $L_m = \frac{p}{2} \log \frac{m}{2\pi e \sigma^2} + \log \frac{1}{\bar{p}(x_0)}$ and the sampled-key threshold is

$$n^* \asymp \frac{1}{\bar{p}(x_0)} \Big(\frac{\sqrt{m}}{\sigma} \Big)^p :$$

one prompt example per cell of side $\sigma/\sqrt{m}$ (the posterior width) in a unit of prior mass—the dartboard count of Figure 1 made exact. For the uniform prior $\bar{p} \equiv 1/\nu(\mathcal{X})$ this is $\nu(\mathcal{X})(\sqrt{m}/\sigma)^p \asymp m^{p/2} V_G$, i.e. the sampled and designed rows of Table 1 coincide up to constants: with nothing to be atypical relative to, the entire cost is search. The adaptive row is $(\sqrt{\log m}/\varepsilon)^p$. The mean map of (S1) is affine, so $\beta = 2$ binds only through the kernel order and the crossover of Section 5.1 reads: interpolation provably dominates for $p > 4$.

Theorem 3.3 (Resolution lower bound, arbitrary weights)**.** *In the setting of Lemma 3.1 there exist a constant $K = K(p)$, query events E_m with $\mathbb{P}(E_m) \to 1$ on which the Clarke–Barron localization holds on $B_m(\sqrt{K \log m})$, and constants $c, C > 0$, $s_0 = s_0(p, \kappa\{G(x_0)\})$, such that for every query realization in E_m, every $s_0 \le s \le L_m/4$, and every prompt length $n \le e^{L_m - 2s}$: with conditional probability at least $1 - e^{-s}$ over the prompt latents,* every *probability measure $\hat{\alpha}$ supported on $\{x_1, \dots, x_n\}$—in particular dictionary attention with arbitrary weights; the bound is deterministic on the coverage event—satisfies*

$$W_1(\hat{\alpha}, \Pi_m) \ge c \sqrt{\frac{s}{m \, \lambda_{\max}\{G(x_0)\}}}.$$

Since the posterior spread obeys $\mathbb{E}_{\Pi_m} d(X, \hat{x}_m) \le C\sqrt{p/(m\lambda_{\min}\{G(x_0)\})}$, the error exceeds the spread by the factor $c\sqrt{s/p}\,\kappa\{G(x_0)\}^{-1/2} \gg 1$ once $s \gg p\,\kappa\{G(x_0)\}$. Taking $s = \epsilon \log m$ for $0 < \epsilon < p/8$: resolution fails for all $n \le c\, m^{p/2 - 2\epsilon}$ with a failure factor diverging like $\sqrt{\epsilon \log m}$. Consequently, at exponential scale,

$$n \gtrsim e^{L_m} \asymp m^{p/2} \frac{\sqrt{\det G(x_0)}}{\bar{p}(x_0)}.$$

The confidence and the range trade off through s: near the full threshold e^{L_m} the statement holds only at constant s, hence constant failure probability and excess factor, while the polynomial-range display above gives divergence; the two-sided sharpness claimed is at exponential scale only.

Proof sketch. Normalize $\mathbb{E}_{\bar{\pi}} w_m = 1$. Markov plus a union bound over $n \le e^{L_m - 2s}$ atoms shows that with probability $\ge 1 - e^{-s}$ every prompt latent lies outside the superlevel set $\{\log w_m \ge L_m - s\}$. On E_m the local expansion of Lemma 3.1 shows this set contains the Fisher bulk $B_m(c_1\sqrt{s})$, hence a chart ball $B(\hat{x}_m, r_s)$, $r_s \asymp \sqrt{s/(m\lambda_{\max})}$, while Theorem 2.2 puts posterior mass $\ge \frac{1}{2}$ on $B(\hat{x}_m, r_s/2)$ for $s \ge s_0$. The 1-Lipschitz witness $f(x) = \min\{d(x, \{x_1, \dots, x_n\}), r_s/2\}$ integrates to zero under any $\hat{\alpha}$ on the atoms and to $\ge r_s/4$ under Π_m; Kantorovich duality concludes. Full proof: Appendix E. □

Proposition 3.4 (Two-sidedness at exponential scale)**.** *In the setting of Lemma 3.1, let $\hat{\alpha}_n$ be the SNIS dictionary measure with exact weights and $\Gamma_m := \mathbb{E}_{\bar{\pi}}[w_m^2] = 1 + \chi^2(\Pi_m \| \bar{\pi})$. If $\Gamma_m/n \xrightarrow{P} 0$ then $\sqrt{m}\, W_1(\hat{\alpha}_n, \Pi_m) \xrightarrow{P} 0$; and the Laplace computation gives $\log \Gamma_m = L_m + \frac{p}{2} \log \frac{e}{2} + o_P(1)$, so the condition follows from $\log n - L_m \to \infty$. Hence $n^* = e^{L_m(1+o(1))}$ from both sides; no matching of constants is claimed.*

Proposition 3.5 (Point-output failure of SNIS below threshold)**.** *Under local prior thickness ($c_\pi r^p \le \bar{\pi}\{u : d_G(u,x) \le r\} \le C_\pi r^p$ near x_0), if $n\bar{p}(x_0) \le m^{p/2-\epsilon}$ for some $\epsilon > 0$, then the nearest prompt atom $x_{(1)}$ to $\hat{x}_m$ satisfies $d_G(x_{(1)}, \hat{x}_m) = \Theta_P\{(n\bar{p}(x_0))^{-1/p}\}$, the exact-weight SNIS output collapses, $\hat{g}_n = x_{(1)} + o_P(m^{-1/2})$, and $d_G(\hat{g}_n, \hat{x}_m)/m^{-1/2} \to \infty$ in probability. (Proof: Appendix E.)*

Remark 3.6 (Bridging metrics). The lower bound is stated in rescaled W_1 (necessarily: TV between an atomic measure and a continuous posterior is 1), while the interpolation upper bounds of Section 5 are in TV. The comparison is coherent because both posteriors localize: if $\|\Pi - \Pi_m\|_{\mathrm{TV}} \le \tau$ and both put mass $1 - m^{-c}$ on $B_m(\sqrt{K \log m})$, then $\sqrt{m}\, W_1(\Pi, \Pi_m) \le C\sqrt{\log m}\, \tau + O(m^{-c'})$ (Appendix A). Thus TV control implies rescaled-W_1 control up to a $\sqrt{\log m}$ factor, and the crossover of Section 5.1 compares like with like.

Remark 3.7 (The atomic limit is the finite-class theory). Positivity and continuity of $\bar{p}$ at x_0 are load-bearing. If $\bar{\pi}$ has an atom at x_0, then L_m stays bounded—the posterior concentrates on the atom at finite KL cost $\log(1/\bar{\pi}\{x_0\})$—and dictionary attention succeeds with $O(1)$ prompt length: exactly the oracle dictionary regime of Rubacher [8]. Conversely a prior density vanishing like $|x - x_0|^a$ inflates the threshold to $m^{(p+a)/2}$ (Appendix D); thin prompt distributions near the truth are polynomially more expensive.

4 Key Placement

Designed keys separate discretization from search. Let $\{C_i\}_{i \le N}$ be cells represented by a d_G–ρ-net $\mathcal{D}_N = \{x_i\}$ with quadrature masses $q_i = \bar{\pi}(C_i)$ entering as logit biases, and for $\varphi_m(x) = \eta_\theta(x)^\top \bar{T}_m - A(\eta_\theta(x))$ set

$$\mathrm{Attn}_N(v) := \frac{\sum_i q_i v(x_i) e^{m\varphi_m(x_i)}}{\sum_i q_i e^{m\varphi_m(x_i)}}.$$

Proposition 4.1 (Designed keys suffice at $m^{p/2}$)**.** *Under Assumption 2.1, if the cells have d_G-diameter at most $\rho \le c/(R\sqrt{m})$ and v is bounded and L_v-Lipschitz, then on a query event of probability $1 - o(1)$,*

$$|\mathrm{Attn}_N(v) - \mathrm{Attn}(v)| \le L_v \rho + 2\|v\|_\infty \Big[\tanh\Big\{ \tfrac{C}{4}(R\sqrt{m}\,\rho + m\rho^2) \Big\} + Ce^{-cR^2} + e^{-cm} \Big].$$

With $V_G := \int_{\mathcal{X}} \sqrt{\det G}\, d\nu$, fixed accuracy ε costs $N \asymp m^{p/2} V_G$ keys, and $\varepsilon \downarrow 0$ costs $N = \widetilde{O}(m^{p/2} \varepsilon^{-p} V_G)$. (Proof: Appendix F.)

The mechanism is a localized perturbation bound (Lemma C.2): moving a key by ρ perturbs the logit by $O_P(R\sqrt{m}\,\rho + m\rho^2)$ on the bulk, exactly the structure $\sqrt{m}\, r_n + m r_n^2$ produced by context error r_n in Section 5—discretizing the latent space and estimating the context field are the same perturbation problem.

The count in Proposition 4.1 is also necessary: a query-independent net cannot beat prior-sampled keys in the m-exponent.

Proposition 4.2 (Query-independent keys must pay $m^{p/2}$). *Under Assumption 2.1 strengthened to uniform two-sided prior thickness on $\mathcal{X}$, there are constants t_0, c, C such that for every $t \geq t_0$ and every deterministic key set $\mathcal{D}_N$ with $N \leq \delta m^{p/2}$: the set*

$$\mathcal{X}_{\mathrm{bad}}(\mathcal{D}_N, t) := \{x_0 : d_G(x_i, x_0) \geq 2t/\sqrt{m} \ \ \forall i \leq N\}$$

has $\bar{\pi}(\mathcal{X}_{\mathrm{bad}}) \geq 1 - Ct^p\delta$, and for every $x_0 \in \mathcal{X}_{\mathrm{bad}}$, on a query event of probability $1 - Ce^{-ct^2} - o(1)$, every probability measure supported on $\mathcal{D}_N$ satisfies $W_1(\hat{\alpha}, \Pi_m) \geq c\,t/\sqrt{m\lambda_{\max}\{G(x_0)\}}$, exceeding the posterior spread for $t \geq t_0$. Hence resolving typical queries at posterior scale forces $N \gtrsim m^{p/2}$. (Proof: Appendix F.)

Randomness is therefore not the obstruction: designed and sampled keys share the exponent for different reasons—sampled keys because the bulk is exponentially atypical under $\bar{\pi}$, designed keys because a pre-query net must cover $\mathcal{X}$ at posterior resolution. Only placement *after* localization escapes:

Corollary 4.3 (Adaptive keys after localization). *If another mechanism produces a ball of d_G-radius $\Lambda_m/\sqrt{m}$ containing $B_m(R_\varepsilon)$ with probability $1 - o(1)$, then restricting Proposition 4.1 to that ball gives output accuracy ε with $N = O\{(\Lambda_m R_\varepsilon/\varepsilon)^p\}$ keys. The plug-in construction of Section 5 supplies $\Lambda_m = O(\sqrt{K \log m})$ (Corollary 5.4), hence* $\mathrm{polylog}(m)$ *post-localization keys, exclusive of the prompt used to estimate the context.*

A Jeffreys-type minimax statement—the prior $\propto \sqrt{\det G}$ minimizes and equalizes the worst-case sampled-key requirement—is recorded in Appendix F.

5 Interpolation: Smoothing and Polynomial Rates

Interpolation uses the prompt as a regression design: a first stage performs Nadaraya–Watson smoothing of the mean map [9], a second maps $\hat{\mu} \mapsto \hat{\eta} = (\nabla A)^{-1}(\hat{\mu})$ and applies plug-in Bayes logits. An m-observation block amplifies context error: across the bulk the query residual varies by $O_P(Rm^{-1/2})$, so uniform context error r_n produces logit *oscillation* $O_P(R\sqrt{m}\,r_n + mr_n^2)$—total variation responds to oscillation, not mean displacement, and directions that do not move the local mode are not amplified.

Assumption 5.1 (Regularity; informal). *The context class Θ consists of fields η_θ that are uniformly C^2-bounded, injective with a margin, and Fisher-sandwiched ($g_- I \preceq J^\top \nabla^2 A\, J \preceq g_+ I$); equivalently the one-observation KL contrast is uniformly locally quadratic with a global identifiability gap. The prior has two-sided local thickness, and the query statistic is uniformly conditionally sub-Gaussian. The precise statement, including the definition of Θ, is Assumption G.1 in Appendix G; all results below are proved under it.*

Theorem 5.2 (Posterior stability under a two-scale context estimate). *Under Assumption G.1, let $\widehat{\Pi}_{n,m}$ be the continuum plug-in posterior built from an estimate $\widehat{\eta}_n$ with values in $\mathcal{K}$ satisfying (H1) local rate $\|\widehat{\eta}_n - \eta_\theta\|_{\infty, B(x_0, r_0)} = O_P(r_n)$; (H2) global consistency $\|\widehat{\eta}_n - \eta_\theta\|_{\infty, \mathcal{X}} = o_P(1)$; (H3) query scale $\sqrt{m}\,r_n \leq c$. Then for every deterministic $R \geq R_0$ and $t \geq 1$, with probability at least $1 - Ce^{-ct^2} - o(1)$, uniformly over x_0 and Θ,*

$$\|\widehat{\Pi}_{n,m} - \Pi_m\|_{\mathrm{TV}} \leq C\{(R+t)\sqrt{m}\,r_n + mr_n^2 + e^{-cR^2}\} + e^{-cm};$$

in particular $R = t = \sqrt{K \log m}$ gives $\|\widehat{\Pi}_{n,m} - \Pi_m\|_{\mathrm{TV}} = O_P(\sqrt{m \log m}\,r_n + mr_n^2 + m^{-cK})$. (Proof: Appendix G.)

Theorem 5.3 (Two-layer in-context estimation, block queries)**.** *Assume additionally that the design density $\bar{p}$ is Lipschitz and bounded, and $x \mapsto \mu_\theta(x) \in \mathcal{H}^\beta(\mathcal{X})$ with $1 \le \beta \le 2$ (a single nonnegative kernel has approximation order two, motivating the cap). The two-stage operator at oracle local bandwidth $\lambda_n \asymp (n\bar{p}(x_0)/\log n)^{2/(2\beta+p)}$ achieves, for $\sqrt{m}\, r_n(x_0) \le c$ and $cK > 1$, with probability $1 - o(1)$,*

$$\|\widehat{\Pi}_{n,m} - \Pi_m\|_{\mathrm{TV}} \le C \min\{1, \sqrt{m \log m}\, r_n(x_0) + m r_n(x_0)^2\} + m^{-cK} + e^{-cm}, \qquad r_n(x_0) \asymp \Big(\frac{\log n}{n\bar{p}(x_0)}\Big)^{\frac{\beta}{2\beta+p}}.$$

A sufficient prompt complexity for vanishing TV error is therefore

$$\frac{n\bar{p}(x_0)}{\log n} \gg (m \log m)^{1+\frac{p}{2\beta}},$$

uniformly over x_0 with $\bar{p}_{\min}$ in place of $\bar{p}(x_0)$. (Proof: Appendix G.)

Corollary 5.4 (End-to-end finite-key pipeline)**.** *In the setting of Theorem 5.3, let $\hat{x}' = \arg\max_{\mathcal{X}} \widehat{\varphi}_m$ be the plug-in mode. Then $d_G(\hat{x}', \hat{x}_m) = O_P(\sqrt{\log m / m})$, so the ball $B(\hat{x}', C\sqrt{K \log m}/\sqrt{m})$ contains the bulk with probability $1 - o(1)$; placing a quadrature net of spacing $\rho \asymp \varepsilon/(\sqrt{m} R_\varepsilon)$ in that ball uses $N = O\{(\sqrt{K \log m}\, R_\varepsilon/\varepsilon)^p\}$ keys and yields, for every bounded L_v-Lipschitz v,*

$$\Big|\mathrm{Attn}_N^{\widehat{\eta}}(v) - \mathbb{E}_{\Pi_m} v\Big| \le L_v \rho + C\|v\|_\infty \big\{\sqrt{m \log m}\, r_n + m r_n^2 + \varepsilon + m^{-cK}\big\}$$

with probability $1 - o(1)$. Thus a two-round coarse-to-fine pass—estimate the context, localize by the plug-in mode, then key the located ball—achieves interpolation accuracy with polylog(m) *keys and prompt length as in Theorem 5.3. (Proof: Appendix G.)*

5.1 Crossover, and what is left open

Reweighting pays the full latent dimension through the exponent $p/2$; interpolation pays a unit block-query lever plus smoothness-discounted dimension $p/(2\beta)$. Since logarithms are subordinate, interpolation provably wins when

$$p \;>\; \frac{2\beta}{\beta - 1}$$

($p > 4$ at $\beta = 2$). Three caveats keep this honest. First, the comparison is *one-sided*: it sets a sharp threshold (two-sided by Proposition 3.4) against a sufficient condition, and we prove no lower bound for interpolation; the regime $p < 2\beta/(\beta-1)$ is indeterminate, and the sup-norm route through (H1) may be loose—only accuracy in score-relevant directions near the bulk should matter, so the $\sqrt{m}\, r_n$ lever is plausibly improvable. Second, as $\beta \downarrow 1$ the crossover diverges, and at $\beta = 1$ (allowed by the assumption) interpolation never wins *under these bounds*. Third, the metrics differ but bridge with a $\sqrt{\log m}$ loss (Remark 3.6). On homogeneous latent spaces this is the whole comparison; global structural estimators can avoid local-density dependence.

5.2 Mechanisms, not depth

Mechanism class. Theorem 3.3 constrains distributions supported on prompt latents; it does not constrain every one-layer architecture, nor every use of prompt points, nor every readout of a convex reweighting. Three escapes delimit the claim.

Widened kernels (reweighting that interpolates).

Remark 5.5 (Tempered softmax escapes at the point level). A tempered head with logits $\beta_T m\varphi_m$, $\beta_T < 1$, is still a measure on the atoms—so Theorem 3.3 correctly says its rescaled W_1 to Π_m is large—yet its *mean* is a Nadaraya–Watson estimate of the posterior mode at bandwidth $h \asymp (\beta_T m)^{-1/2}$, with error of heuristic order $h^2 + h/\sqrt{n\bar{p}(x_0)h^p}$; optimizing h shows the mode is located to $o(m^{-1/2})$ once $n\bar{p}(x_0) \gg m^{(p+2)/4}$, polynomially below $m^{p/2}$ for every $p > 2$ (calculation sketch: Appendix H). Widening the kernel is smoothing over prompt positions, i.e. interpolation of the target functional inside the convex hull; the barrier binds distributional resolution, not point readouts.

Off-atom evaluation. A local quadratic fit to Bayes logits sampled at spacing $(n\bar{p})^{-1/p}$ locates their vertex with error $O\{(n\bar{p})^{-2/p}\}$, below posterior width once $n\bar{p}(x_0) \gg m^{p/4}$; extrapolating off the atoms is again interpolation.

One-layer counterexample. A uniform-attention head can output empirical moments: in (S2), values $(x_t y_t, x_t^2)$ followed by a ratio recover $\hat{h}_n = \sum_t x_t y_t / \sum_t x_t^2$ exactly, so one layer can estimate the context and beat the dictionary threshold. Depth is therefore the natural implementation of interpolation—localize, then key—not an information-theoretic necessity.

Restatement. The dichotomy is Bayes-weighted retrieval versus smoothing. A dictionary token must serve as datum and key simultaneously, although data follow $\bar{\pi}$ while useful keys lie in the posterior bulk; every escape above works by decoupling those roles.

6 Synthetic Diagnostics

The theory prescribes a minimal, fully specified diagnostic in system (S1) (and the curved system of Appendix H.1). *Panel A (phase transition):* for fixed m and growing n, plot the normalized dictionary error $\sqrt{m}\, W_1(\hat{\alpha}_n, \Pi_m)$ against $\log n$; the transition is predicted at $\log n^* = L_m$, with the explicit (S1) value of Example 3.2, and sweeping x_0 and σ tests the typicality ($\bar{p}(x_0)$) and geometry ($\det G$) terms of Lemma 3.1 through the predicted horizontal shifts. *Panel B (trichotomy):* at matched key counts, overlay sampled keys, a designed d_G-net (Proposition 4.1), and post-localization adaptive keys (Corollary 5.4); the sampled–designed gap should vanish in the exponent and the adaptive curve should sit at polylog(m). *Panel C (interpolation law):* the two-layer error against n at several m, testing the collapse predicted by $\sqrt{m \log m}\, r_n + m r_n^2$.

7 Discussion

The positive message is architectural: the $m^{p/2}$ barrier is a search cost, dissolved entirely by spending the prompt on estimating the context field and the keys on a located ball. The limitations are the ones flagged above: the lower bound governs distributional resolution and the exact-Bayes point readout, not smoothed point readouts (Remark 5.5); the crossover is one-sided, so the small-p/low-smoothness regime is open; the block-query model is the informative-query regime and degenerates at $m = O(1)$; and the analysis is of idealized operators, not trained transformers—whether training discovers the coarse-to-fine pipeline of Corollary 5.4 is an empirical question that Section 6's Panel B is designed to make visible.

A Chart Conventions, Bulk Hygiene, and the Metric Bridge

Charts and metric equivalence. On a manifold all local statements are read through a fixed C^2 chart around x_0 and chart coordinates are relabelled x; on S^1, the arclength chart. Under Assumption 2.1, compactness gives $g_- I \preceq G \preceq g_+ I$ with $g_- := \inf_x \lambda_{\min}\{G(x)\} > 0$, $g_+ := \sup_x \lambda_{\max}\{G(x)\} < \infty$, so d_G and any fixed chart metric are equivalent up to constants; all constants below absorb this equivalence.

Centering. $d_G(\hat{x}_m, x_0) = O_P(m^{-1/2})$, hence bulks centered at x_0 and $\hat{x}_m$ differ only by replacing R with $R + O_P(1)$; we center at $\hat{x}_m$ throughout and use this interchange silently.

The bridge inequality (Remark 3.6). Suppose $\|\Pi - \Pi_m\|_{\mathrm{TV}} \le \tau$ and $\Pi(B^c) + \Pi_m(B^c) \le m^{-c}$ for $B = B_m(\sqrt{K \log m})$, and that both measures have first moments bounded by the diameter of the compact $\mathcal{X}$. Couple Π and Π_m maximally: with probability $1 - \tau$ the coupled pair is equal, contributing zero transport; on the discrepancy event, transport within B costs at most $\mathrm{diam}(B) \le C\sqrt{\log m / m}$, and the tail mass costs at most $\mathrm{diam}(\mathcal{X})\, m^{-c}$. Hence

$$W_1(\Pi, \Pi_m) \le C\sqrt{\tfrac{\log m}{m}}\,\tau + C' m^{-c}, \qquad \sqrt{m}\, W_1(\Pi, \Pi_m) \le C\sqrt{\log m}\,\tau + C' m^{-c+1/2},$$

which is the display of Remark 3.6 for $c > 1/2$ (guaranteed by choosing K large).

B Bernstein–von Mises Proof

Normalizability and block form. For model-generated queries the attention normalizer is finite almost surely on the support of the marginal law; for external queries this is an explicit condition. Compactness and continuity of η_θ with $\eta_\theta(\mathcal{X}) \subset \mathrm{dom}\, A$ suffice. For averaged statistics in $\mathrm{int}\, \mathcal{M}$, $\pi_\theta(dx \mid y_{1:m}) \propto \exp\{-m D_{A^*}(\bar{T}_m, \mu_\theta(x))\}\bar{\pi}(dx)$.

Proof of Theorem 2.2. By sufficiency the attention measure depends on the query block through $\bar{T}_m$, and its log-density with respect to $\bar{\pi}$ is $m\varphi_m(x) - \log \int e^{m\varphi_m}\, d\bar{\pi}$, with $\varphi_m(x) = \eta_\theta(x)^\top \bar{T}_m - A(\eta_\theta(x))$ and population version $\varphi(x) = \varphi(x_0) - \mathrm{KL}(x_0 \| x)$, where $\mathrm{KL}(x_0\|x)$ is the divergence between the corresponding observation laws.

Step 1: uniform LLN, MLE consistency, localization. $\|\varphi_m - \varphi\|_\infty \le \sup_x \|\eta_\theta(x)\| \cdot \|\bar{T}_m - \mu_0\| = O_P(m^{-1/2})$: uniformity in x is inherited from the convergence of the single vector $\bar{T}_m$, with no entropy conditions. Injectivity and compactness give the identifiability gap $\sup_{d(x,x_0)\ge\varepsilon} \varphi \le \varphi(x_0) - c(\varepsilon)$, yielding fixed-neighborhood localization by the two-sided normalizer sandwich. Locally, $\nabla\varphi_m(x_0) = J(x_0)^\top(\bar{T}_m - \mu_0) = O_P(m^{-1/2})$ and $-\nabla^2\varphi_m(x_0) = G(x_0) + O_P(m^{-1/2})$, and uniform local Hessian continuity gives $-\nabla^2\varphi_m \succeq G(x_0)/2$ on a fixed neighborhood with high probability. The local maximizer is unique with $\hat{x}_m - x_0 = G(x_0)^{-1} J(x_0)^\top(\bar{T}_m - \mu_0) + o_P(m^{-1/2})$, and strong concavity upgrades to the shrinking tail bound: for fixed sufficiently large K, $\alpha_\theta(B_m(\sqrt{K\log m})^c) \le Cm^{-cK} + e^{-cm}$ with high probability; the fixed far region, including secondary modes, stays e^{-cm}-negligible.

Step 2: local expansion via (3). On $B_m(\sqrt{K\log m})$,

$$-\nabla_x^2 \varphi_m(x) = J(x)^\top \nabla^2 A(\eta_\theta(x)) J(x) - \sum_j \partial_x^2 \eta_{\theta,j}(x)\big(\bar{T}_m - \mu_\theta(x)\big)_j = G(x_0) + O_P\Big(\sqrt{\tfrac{K\log m}{m}}\Big),$$

using the Lipschitz Hessian of η_θ for the drift of the first term over the shrinking ball and $\|\bar{T}_m - \mu_\theta(x)\| = O_P(\sqrt{K\log m/m})$ on it for the extrinsic term. Taylor expansion at $\hat{x}_m$ gives accumulated

remainder $O_P\{(K\log m)^{3/2}/\sqrt{m}\} = o_P(1)$, so $m\varphi_m(x) = m\varphi_m(\hat{x}_m) - \frac{m}{2}(x-\hat{x}_m)^\top G(x_0)(x-\hat{x}_m) + o_P(1)$ uniformly on the ball.

Step 3: assembly. The normalizer localizes to the shrinking ball, where the Gaussian integral applies with $\bar{p}$ continuous at x_0 contributing $\bar{p}(x_0)(1+o(1))$; the ratio of the local chart density to $\mathcal{N}(\hat{z}_m, (mG_\psi(x_0))^{-1})$ converges to one uniformly on the ball, the tail is uniformly integrable, and local L^1 convergence yields TV convergence. Replacing $G_\psi(x_0)$ by $G_\psi(\hat{x}_m)$ costs $O_P(\sqrt{p}\, m^{-1/2})$ in TV between the two Gaussians. □

C Softmax Stability Lemmas

Lemma C.1 (Softmax stability)**.** *Let q be a probability density with respect to a σ-finite reference measure and $\tilde{q} \propto q\, e^{\delta}$ for measurable δ with $\Omega := \operatorname{ess\,sup}_q \delta - \operatorname{ess\,inf}_q \delta < \infty$. Then $\|q - \tilde{q}\|_{\mathrm{TV}} \le \tanh(\Omega/4)$. If instead δ is σ^2-sub-Gaussian under q, then $\mathrm{KL}(q\|\tilde{q}) \le \sigma^2/2$ and $\|q-\tilde{q}\|_{\mathrm{TV}} \le \sigma/2$.*

Proof. The first inequality is the extremal two-point bound for exponential tilts with likelihood-ratio range e^{Ω}. The second follows from $\mathrm{KL}(q\|\tilde{q}) = \log \mathbb{E}_q e^{\delta} - \mathbb{E}_q\delta$, the sub-Gaussian moment bound, and Pinsker. □

Lemma C.2 (Localized exponent perturbation)**.** *Let $\Pi \propto e^{m\varphi}\bar{\pi}$ and $\Pi' \propto e^{m\varphi+\psi}\bar{\pi}$ be probability measures and B measurable with $\operatorname{osc}_B \psi \le \varepsilon$ and $\Pi(B^c) + \Pi'(B^c) \le \zeta$. Then $\|\Pi - \Pi'\|_{\mathrm{TV}} \le \tanh(\varepsilon/4) + \zeta$.*

Proof. Conditioned on B the measures differ by the tilt $\psi|_B$ of range at most ε; Lemma C.1 and the conditioning inequality $\|\Pi - \Pi'\|_{\mathrm{TV}} \le \|\Pi_B - \Pi'_B\|_{\mathrm{TV}} + \Pi(B^c) + \Pi'(B^c)$ conclude. □

D Clarke–Barron Expansion

Proof of Lemma 3.1. With $w = d\Pi_m/d\bar{\pi}$, carrier terms cancel and $\log w(x) = m\varphi_m(x) - \log \int e^{m\varphi_m} d\bar{\pi}$, the landscape of Appendix B; all randomness enters through $\bar{T}_m$.

Step 1 (localization). As in Appendix B: uniform LLN at rate $O_P(m^{-1/2})$, identifiability gap, posterior mass outside $B(x_0,\varepsilon)$ at most $e^{-mc(\varepsilon)/2}$ with high probability. Since φ_m is bounded on the compactum, $|\log w| \le m\,O(1) + |\log Z|$ there, so the far region contributes $m\,O(1)e^{-cm} \to 0$ to every moment below; compactness, boundedness, and the exponential tail give uniform integrability.

Step 2 (local Laplace). On $B_m(\sqrt{K\log m})$, with $K = K(p)$ chosen large enough both for the moment bounds and for the full range $s \le L_m/4$ in Theorem 3.3, and with $\eta_\theta \in C^3$ controlling the cubic remainder,

$$m\varphi_m(x) = m\varphi_m(\hat{x}_m) - \tfrac{m}{2}(x-\hat{x}_m)^\top G(x_0)(x-\hat{x}_m)(1+O_P(\epsilon_m)), \qquad \epsilon_m := \sqrt{\tfrac{K\log m}{m}}.$$

Step 3 (normalizer). $\log \int e^{m\varphi_m} d\bar{\pi} = m\varphi_m(\hat{x}_m) + \frac{p}{2}\log\frac{2\pi}{m} - \frac{1}{2}\log\det G(x_0) + \log\bar{p}(x_0) + o_P(1)$, by the Gaussian integral over the inner ball (radius $\to\infty$ in standardized units; $\bar{p}$ continuous and positive at x_0) and exponential negligibility of the complement.

Step 4 (moments of $\frac{1}{2}\chi^2_p$). Combining, on the ball

$$\log w(X) = \tfrac{p}{2}\log\tfrac{m}{2\pi} + \tfrac{1}{2}\log\det G(x_0) + \log\tfrac{1}{\bar{p}(x_0)} - \tfrac{m}{2}(X-\hat{x}_m)^\top G(X-\hat{x}_m)(1+O_P(\epsilon_m)) + o_P(1),$$

and under Π_m the quadratic form is a $(1+O_P(\epsilon_m))$-perturbed $\frac{1}{2}\|Z\|^2$, $Z\sim\mathcal{N}(0,I_p)$, by Theorem 2.2; dominated convergence over the ball (Step 1 killing the complement) gives $\mathbb{E}_{\Pi_m}[\frac{m}{2}(X-\hat{x})^\top G(X-$

$\hat{x})] \to p/2$ and variance $\to p/2$. Assembling yields the stated expansion via $\log \frac{m}{2\pi e} = \log \frac{m}{2\pi} - 1$ and $\mathrm{Var}_{\Pi_m}(\log w) = p/2 + o_P(1)$. The Step-4 display is also the superlevel-set geometry used in Theorem 3.3: on the localization event, $\{\log w \geq L_m - s\}$ contains the Fisher bulk $B_m(c_1\sqrt{s})$, hence a chart ball of radius $c_1'\sqrt{s/(m\lambda_{\max}(G))}$, for $s \geq s_0$; the choice of K guarantees this throughout $s \leq L_m/4$. □

Remark D.1 (Vanishing prior density changes the exponent). If $\bar{p}(x) \asymp |x - x_0|^a$ near the truth, Step 3's prior factor is no longer $\bar{p}(x_0)(1 + o(1))$; integrating $|x - \hat{x}|^a$ against the local Gaussian gives $\mathbb{E}_{\Pi_m} \log(1/\bar{p}(X)) \asymp \frac{a}{2} \log m$ and the threshold inflates to $m^{(p+a)/2}$.

E Proofs for Section 3

Proof of Theorem 3.3. Normalize $\mathbb{E}_{\bar{\pi}} w = 1$ and define E_m as the intersection of (i) the Clarke–Barron localization event of Appendix D on $B_m(\sqrt{K \log m})$ and (ii) the event that the constant c_1 of Step 4 there satisfies $c_1\sqrt{L_m/4} \leq \sqrt{K \log m}$, which holds for all large m by the choice $K = K(p)$ since $L_m = O_P(\log m)$; $\mathbb{P}(E_m) \to 1$.

Coverage. By Markov, $\mathbb{P}_{\bar{\pi}}(\log w \geq L_m - s) \leq e^{-(L_m - s)}$; a union bound over $n \leq e^{L_m - 2s}$ prompt latents shows that with probability $\geq 1 - e^{-s}$ every prompt latent lies outside $U_s := \{\log w \geq L_m - s\}$.

The superlevel set contains a bulk ball. By Step 4 of Appendix D, on E_m, U_s contains the Fisher bulk $B_m(c_1\sqrt{s})$ for all $s_0 \leq s \leq L_m/4$, hence a chart ball $B(\hat{x}_m, r_s)$ with $r_s := c_2\sqrt{s/(m\lambda_{\max}\{G(x_0)\})}$; every prompt latent is therefore at distance $\geq r_s$ from $\hat{x}_m$.

The posterior lives in the bulk. By Theorem 2.2 and a Gaussian ball computation, $\Pi_m(B(\hat{x}_m, r_s/2)) \geq 1 - Ce^{-c_3 s/\kappa\{G(x_0)\}} - o_m(1) \geq \frac{1}{2}$ for $s \geq s_0(p, \kappa\{G(x_0)\})$.

Wasserstein via a Lipschitz witness. $f(x) := \min\{d(x, \{x_1, \ldots, x_n\}), r_s/2\}$ is 1-Lipschitz and bounded; $\int f \, d\hat{\alpha} = 0$ for every $\hat{\alpha}$ on the atoms, while every point of $B(\hat{x}_m, r_s/2)$ is at distance $\geq r_s/2$ from every atom, so $\int f \, d\Pi_m \geq (r_s/2)\Pi_m(B(\hat{x}_m, r_s/2)) \geq r_s/4$. Kantorovich duality gives $W_1(\hat{\alpha}, \Pi_m) \geq r_s/4$. The spread bound is the Gaussian first moment under Theorem 2.2, and the displayed consequences follow by substitution. □

Proof of Proposition 3.4. Rényi computation. $\Gamma_m = \mathbb{E}_{\bar{\pi}} w_m^2 = \mathbb{E}_{\Pi_m} w_m$. The Step-3 normalizer of Appendix D and a second Gaussian integral give

$$\mathbb{E}_{\Pi_m} w_m = \frac{\int e^{2m\varphi_m} d\bar{\pi}}{\{\int e^{m\varphi_m} d\bar{\pi}\}^2} = \frac{m^{p/2}\sqrt{\det G(x_0)}}{2^p \pi^{p/2} \bar{p}(x_0)}\{1 + o_P(1)\},$$

and subtracting the expansion of L_m yields $\log \Gamma_m - L_m = \frac{p}{2}\log(e/2) + o_P(1)$.

Wasserstein upper bound. Let $d_m(x, x') = \sqrt{m}\, d_G(x, x')$; compact metric equivalence changes only constants relative to a fixed chart metric. With $S_n = \frac{1}{n}\sum_i w_m(X_i)$: $\mathbb{E}S_n = 1$, $\mathrm{Var}(S_n) \leq \Gamma_m/n$, so $\mathbb{P}(S_n < 1/2) \leq 4\Gamma_m/n = o_P(1)$. Fix $R > 1$ and partition $B_m(R)$ into $N \leq C(R/\varepsilon)^p$ cells of d_m-diameter at most ε. For the unnormalized weighted empirical measure $\nu_n = \frac{1}{n}\sum_i w_m(X_i)\delta_{X_i}$, each cell has $\mathbb{E}\nu_n(C_j) = \Pi_m(C_j)$ and $\mathrm{Var}\{\nu_n(C_j)\} \leq \frac{1}{n}\mathbb{E}_{\bar{\pi}}[w_m^2 \mathbf{1}_{C_j}]$; Cauchy–Schwarz gives $\mathbb{E}\sum_j |\nu_n(C_j) - \Pi_m(C_j)| \leq \sqrt{N\Gamma_m/n}$. On $\{S_n \geq 1/2\}$ normalization adds $O_P(\sqrt{\Gamma_m/n})$; collapsing cells to representatives costs 2ε; cell-mass discrepancy costs at most CR times the total variation; and the mass outside $B_m(R)$ contributes the truncated first moment $t_m(R) := \int d_m(x, \hat{x}_m)\mathbf{1}\{d_m > R\}\,\Pi_m(dx) \to 0$ as $R \to \infty$ by the local quadratic posterior bounds, with the empirical tail $O_P(t_m(R))$ by Markov on $\{S_n \geq 1/2\}$. Therefore

$$\sqrt{m}\, W_1(\hat{\alpha}_n, \Pi_m) = O_P\Big[\varepsilon + R\sqrt{(R/\varepsilon)^p\, \Gamma_m/n} + t_m(R)\Big];$$

taking $\varepsilon = R(\Gamma_m/n)^{1/(p+2)}$ and then $R = R_m \to \infty$ slowly proves the claim. □

Proof of Proposition 3.5. Condition on a query realization in the localization event. If n stays bounded along a subsequence, the nearest prompt point is unique a.s., $D_1 = \Theta_P(1)$, and the finite set of likelihood ratios concentrates exponentially on it; so assume $n \to \infty$. In Fisher normal coordinates about $\hat{x}_m$, local prior thickness and continuity give the point-process convergence $\sum_i \delta_{\{n\bar{p}(x_0)\}^{1/p} G(x_0)^{1/2}(x_i - \hat{x}_m)} \Rightarrow \mathcal{P}$, a homogeneous Poisson process; hence $D_1 := d_G(x_{(1)}, \hat{x}_m) = \Theta_P(\{n\bar{p}(x_0)\}^{-1/p})$ and $\{n\bar{p}(x_0)\}^{2/p}(D_2^2 - D_1^2) \Rightarrow R_2^2 - R_1^2$ with $0 < R_1 < R_2$ a.s. Under $n\bar{p}(x_0) \le m^{p/2-\epsilon}$, $m(D_2^2 - D_1^2)/\log n \xrightarrow{P} \infty$ since $m\{n\bar{p}(x_0)\}^{-2/p}/\log n \ge m^{2\epsilon/p}/\log n \to \infty$ and the limiting gap is strictly positive. The local expansion of Lemma 3.1, uniformly over every prompt point on the fixed localization neighborhood (where strong concavity holds uniformly), gives $w(x_{(j)})/w(x_{(1)}) = \exp[-\frac{m}{2}(D_j^2 - D_1^2)\{1 + o_P(1)\}]$, so

$$\sum_{j \ge 2} \frac{w(x_{(j)})}{w(x_{(1)})} \le n \exp\{-cm(D_2^2 - D_1^2)\} = o_P(m^{-A})$$

for every fixed A, points outside the neighborhood carrying an extra e^{-cm} penalty. The nearest atom thus has normalized weight $1 - o_P(m^{-1/2})$, compactness gives $\hat{g}_n = x_{(1)} + o_P(m^{-1/2})$, and $D_1/m^{-1/2} \ge \Theta_P(m^{\epsilon/p})$ concludes. □

F Proofs for Section 4

Proof of Proposition 4.1 and Corollary 4.3. Let $Q_\rho(x) = x_i$ for $x \in C_i$ and define the cell-lifted posterior $\widetilde{\alpha} \propto \exp\{m\varphi_m(Q_\rho(x))\}\bar{\pi}(dx)$. The quadrature weights are exact cell masses, so $\mathrm{Attn}_N(v) = \int v(Q_\rho(x))\,\widetilde{\alpha}(dx)$ and the atomic collapse costs only value discretization, $|\mathrm{Attn}_N(v) - \int v\,d\widetilde{\alpha}| \le L_v\rho$. For the comparison of $\widetilde{\alpha}$ with the continuum posterior α, set $e_\rho(x) := m\{\varphi_m(Q_\rho(x)) - \varphi_m(x)\}$. On the localization event, the score of φ_m vanishes at $\hat{x}_m$ and its Hessian is uniformly bounded in Fisher normal coordinates, so Taylor's theorem gives $|e_\rho(x)| \le C\{m\rho\, d_G(x, \hat{x}_m) + m\rho^2\}$ on a fixed neighborhood, hence $\mathrm{osc}_{B_m(R)} e_\rho \le C(R\sqrt{m}\,\rho + m\rho^2)$. Lemma C.1 bounds the conditional TV by $\tanh\{C(R\sqrt{m}\,\rho + m\rho^2)/4\}$. For the tails: Lipschitz continuity gives $|\varphi_m \circ Q_\rho - \varphi_m| \le L\rho$, so cell lifting reduces the fixed far-region gap by at most $2L\rho = o(1)$ and the lifted far-region mass is e^{-cm}; inside the fixed neighborhood the quadratic localization of Theorem 2.2, with completion of squares for the perturbed exponent, gives $\alpha(B_m(R)^c) + \widetilde{\alpha}(B_m(R)^c) \le Ce^{-cR^2} + e^{-cm}$ under $R\sqrt{m}\,\rho \le c$. Lemma C.2 assembles the bound, and $|\int v\,d\alpha - \int v\,d\widetilde{\alpha}| \le 2\|v\|_\infty \|\alpha - \widetilde{\alpha}\|_{\mathrm{TV}}$ gives the display. Riemannian volume comparison gives $N(\mathcal{X}, d_G, \rho) \asymp \rho^{-p} \int \sqrt{\det G}\,d\nu$; for fixed ε take R large and $\rho = c_\varepsilon/\sqrt{m}$; for $\varepsilon \downarrow 0$ take $R_\varepsilon \asymp \sqrt{\log(1/\varepsilon)}$ and $\rho \asymp \varepsilon/(\sqrt{m}R_\varepsilon)$. A d_G-ball of radius $\Lambda_m/\sqrt{m}$ has covering number $O\{(\Lambda_m/(\sqrt{m}\rho))^p\}$, giving Corollary 4.3. □

Proof of Proposition 4.2. Uniform two-sided thickness and the metric sandwich give, for each key x_i, $\bar{\pi}\{x_0 : d_G(x_i, x_0) < 2t/\sqrt{m}\} \le C(t/\sqrt{m})^p$; a union bound over $N \le \delta m^{p/2}$ keys bounds the complement of $\mathcal{X}_{\mathrm{bad}}(\mathcal{D}_N, t)$ by $Ct^p\delta$. Fix $x_0 \in \mathcal{X}_{\mathrm{bad}}$. On the event $\{Z_m \le t\}$ with $Z_m := \sqrt{m}\,d_G(\hat{x}_m, x_0)$—probability $\ge 1 - Ce^{-ct^2}$ by the sub-Gaussian query statistic and the linearization of $\hat{x}_m$ (Appendix B)—every key satisfies $d_G(x_i, \hat{x}_m) \ge d_G(x_i, x_0) - d_G(x_0, \hat{x}_m) \ge t/\sqrt{m}$. Passing to the chart metric, every key is at distance $\ge r_t := ct/\sqrt{m\lambda_{\max}\{G(x_0)\}}$ from $\hat{x}_m$, while Theorem 2.2 gives $\Pi_m(B(\hat{x}_m, r_t/2)) \ge \frac{1}{2}$ for $t \ge t_0(p, \kappa\{G(x_0)\})$. The witness argument of Theorem 3.3—deterministic given key positions— yields $W_1(\hat{\alpha}, \Pi_m) \ge r_t/4$ for every $\hat{\alpha}$ supported on $\mathcal{D}_N$. □

Remark F.1 (Jeffreys minimax key placement). Let $j(x) = \sqrt{\det G(x)}/V_G$. Then $\inf_{\bar{p}} \operatorname{ess\,sup}_x \sqrt{\det G(x)}/\bar{p}(x) = V_G$, attained at $\bar{p} = j$: Jeffreys minimizes and equalizes the worst-case sampled-key requirement of Theorem 3.3, paralleling its asymptotic least-favourability under entropy risk [6].

G Proofs for Section 5

Assumption G.1 (Uniform regularity; full statement)**.** *Let $\mathcal{X}$ be a compact boundaryless Riemannian manifold. The context class is*

$$\Theta := \Big\{\theta :\ \eta_\theta \in C^2(\mathcal{X};\mathbb{R}^k),\ \|\eta_\theta\|_{C^2} \le B,\ \eta_\theta \text{ injective},\ \eta_\theta(\mathcal{X}) \subset \mathcal{K},\ g_- I \preceq J_\theta^\top \nabla^2 A(\eta_\theta) J_\theta \preceq g_+ I\Big\}$$

for fixed B, compact $\mathcal{K} \subset \operatorname{dom} A$, and $0 < g_- \le g_+ < \infty$. There exist $r_0, c_0, C_0, \kappa_0 > 0$, uniform over Θ and $x_0 \in \mathcal{X}$, such that the one-observation contrast $K_{x_0}(x) := D_A(\eta_\theta(x), \eta_\theta(x_0))$ satisfies $c_0 d(x,x_0)^2 \le K_{x_0}(x) \le C_0 d(x,x_0)^2$ for $d(x,x_0) \le r_0$ and $\inf_{d(x,x_0)\ge r_0} K_{x_0}(x) \ge \kappa_0$. (A Taylor expansion of D_A shows the Fisher sandwich plus uniform C^2 bounds imply these local quadratic bounds after decreasing r_0.) Moreover η_θ and μ_θ are uniformly Lipschitz; $\underline{\lambda}_F I \preceq \nabla^2 A \preceq \overline{\lambda}_F I$ on $\mathcal{K}$, so D_A is jointly Lipschitz there with constant L_A; the prior has local thickness ($\bar{p}$ continuous and positive at x_0 for pointwise statements; $c_\pi r^p \le \bar{\pi}(B(x,r)) \le C_\pi r^p$ for uniform statements); and the query statistic is uniformly conditionally sub-Gaussian.

Proof of Theorem 5.2. Write $\ell_m(x) = \eta_\theta(x)^\top \bar{T}_m - A(\eta_\theta(x))$, $\widehat{\ell}_m$ likewise with $\widehat{\eta}_n$, $\Delta = \widehat{\eta}_n - \eta_\theta$, $\xi_m = \bar{T}_m - \mu_\theta(x_0)$, $Z_m = \sqrt{m}\|\xi_m\| = O_P(1)$.

Step 0 (far region; uses (H2)). Let $F = \{x : d(x,x_0) \ge r_0\}$. For the true posterior, $K_{x_0} \ge \kappa_0$ on F and compactness bounds the stochastic linear term, so on $\{Z_m \le t\}$ with $t \le c\sqrt{m}$, $m\{\ell_m(x) - \ell_m(x_0)\} \le -m\kappa_0 + Ct\sqrt{m} \le -\frac{1}{2}m\kappa_0$. For the plug-in exponent, joint Lipschitz continuity and (H2) give, on its high-probability event, $D_A\{\widehat{\eta}_n(x), \widehat{\eta}_n(x_0)\} \ge K_{x_0}(x) - 2L_A\|\Delta\|_{\infty,\mathcal{X}} \ge \frac{1}{2}\kappa_0$ on F, and the identity

$$\widehat{\ell}_m(x) - \widehat{\ell}_m(x_0) = -D_A\{\widehat{\eta}_n(x), \widehat{\eta}_n(x_0)\} + \{\widehat{\eta}_n(x) - \widehat{\eta}_n(x_0)\}^\top \{\bar{T}_m - \nabla A(\widehat{\eta}_n(x_0))\}$$

with (H1)+(H3) shows the remaining term is $O(t/\sqrt{m} + r_n)$ per observation and cannot close the fixed gap. Both normalizers are bounded below by their contribution from $B(x_0, m^{-1/2})$: prior thickness gives mass $\ge c_\pi m^{-p/2}$, the true relative exponent there is $O(1+t)$, and the plug-in oscillation differs by at most $C\{(1+t)\sqrt{m}\, r_n + m r_n^2\}$. Hence $\Pi_m(F) + \widehat{\Pi}_{n,m}(F) \le C m^{p/2} e^{-m\kappa_0/4 + C(1+t)} \le e^{-cm}$; for $t > c\sqrt{m}$, intersecting with $\{Z_m \le m^{1/4}\}$ adds $o(1)$ to the failure probability.

Step 1 (local shells; uses (H1), (H3)). Locally, $m\{\ell_m(x) - \ell_m(x_0)\} \le -c_0 m d(x,x_0)^2 + C L_\eta Z_m \sqrt{m}\, d(x,x_0)$. By softmax gauge invariance compare the perturbation through differences:

$$\begin{aligned}\delta_m(x) - \delta_m(x_0) = {} & m\{\Delta(x) - \Delta(x_0)\}^\top \xi_m + m\Delta(x)^\top\{\mu_\theta(x_0) - \mu_\theta(x)\} \\ & - m\big[D_A(\eta_\theta(x) + \Delta(x), \eta_\theta(x)) - D_A(\eta_\theta(x_0) + \Delta(x_0), \eta_\theta(x_0))\big],\end{aligned}$$

so uniform Lipschitz and Fisher bounds give, for $x \in B(x_0, r_0)$, $|\delta_m(x) - \delta_m(x_0)| \le C[Z_m\sqrt{m}\, r_n + m r_n d(x,x_0) + m r_n^2]$. The shell-controlling term $-c_0 m d^2 + C m r_n d \le -\frac{c_0}{2} m d^2 + C' m r_n^2$ shows both posteriors are root-m localized whenever $\sqrt{m}\, r_n$ is small. On the query event, $d_G(\hat{x}_m, x_0) \le C Z_m/\sqrt{m}$, so with $B_R := B_m(R)$, $\operatorname{osc}_{B_R} \delta_m \le C[(Z_m + R)\sqrt{m}\, r_n + m r_n^2]$. On $\{Z_m \le t\}$, Lemma C.2 on B_R gives the displayed bound; sub-Gaussianity gives $\mathbb{P}(Z_m > t) \le C e^{-ct^2}$, and the consequences follow with $R = t = \sqrt{K \log m}$. □

Proof of Theorem 5.3. Layer 1 (sup-norm regression, boundaryless design). With design density Lipschitz and locally positive at x_0, response $T(Y)$ conditionally sub-Gaussian, and a positive kernel at bandwidth $h_n = \lambda_n^{-1/2}$: uniform Bernstein over a local net with kernel L^2 norms $\asymp h_n^{-p/2}$ and denominators of order $nh_n^p\bar{p}(x_0)$ gives, for $1 \le \beta \le 2$ [10],

$$r_n(x_0) = \|\hat{\mu}_\theta - \mu_\theta\|_{\infty, B(x_0,r_0)} = O_P\Big(h_n^\beta + \sqrt{\tfrac{\log n}{nh_n^p\bar{p}(x_0)}}\Big) = O_P\Big(\big(\tfrac{\log n}{n\bar{p}(x_0)}\big)^{\frac{\beta}{2\beta+p}}\Big)$$

at the stated bandwidth. Projection onto $\nabla A(\mathcal{K})$ and Lipschitz inversion of $(\nabla A)^{-1}$ give the same local rate for $\hat{\eta} - \eta_\theta$. Thus (H1) holds with $r_n(x_0)$, crude global consistency gives (H2) for large n, and (H3) is the stated condition; Theorem 5.2 applies. The prompt-complexity display inverts $\sqrt{m \log m}\, r_n(x_0) \to 0$ (which is binding: $mr_n^2 \to 0$ alone needs only $n\bar{p}(x_0)/\log n \gg m^{1+p/(2\beta)}$); uniformity replaces $\bar{p}(x_0)$ by $\bar{p}_{\min}$. □

Proof of Corollary 5.4. Localization by the plug-in mode. On the events of Theorem 5.2 with $t \asymp \sqrt{\log m}$, the true exponent satisfies $m\{\varphi_m(\hat{x}_m) - \varphi_m(x)\} \ge cmd_G(x, \hat{x}_m)^2$ locally with a fixed global gap, and the plug-in perturbation obeys the Step-1 bound $|\delta_m(x) - \delta_m(x_0)| \le C[t\sqrt{m}\, r_n + mr_n d + mr_n^2]$. At the maximizer $\hat{x}'$ of $\widehat{\varphi}_m$, $cmd_G(\hat{x}', \hat{x}_m)^2 \le C[t\sqrt{m}\, r_n + mr_n d_G(\hat{x}', \hat{x}_m) + mr_n^2]$, whence $d_G(\hat{x}', \hat{x}_m) \le C\{r_n + t^{1/2}m^{-1/4}r_n^{1/2}\}$. Under $\sqrt{m}\, r_n \le c$ this is $\le C\sqrt{\log m/m}$ (both terms, using $r_n \le c/\sqrt{m}$), so $B(\hat{x}', C\sqrt{K \log m}/\sqrt{m}) \supseteq B_m(\sqrt{K \log m})$ with probability $1 - o(1)$.

Combined perturbation. Let $\widetilde{\alpha} \propto e^{m\widehat{\varphi}_m \circ Q_\rho}\bar{\pi}$ be the cell-lifted plug-in posterior on the located ball, so that as in Appendix F the finite-key output is $\mathrm{Attn}_N^{\widehat{\eta}}(v) = \int v \circ Q_\rho\, d\widetilde{\alpha}$ up to the $L_v\rho$ value-discretization cost. Compare $\widetilde{\alpha}$ with Π_m directly via Lemma C.2 with $\psi = m(\varphi_m \circ Q_\rho - \varphi_m) + \delta_m \circ Q_\rho$: on $B = B_m(R)$, the first summand oscillates by $C(R\sqrt{m}\,\rho + m\rho^2)$ by the C^2 bound on the *true* exponent (Appendix F), and the second by $C[(t + R + \sqrt{m}\rho)\sqrt{m}\, r_n + mr_n^2]$ by Step 1 of Theorem 5.2 evaluated at cell representatives, the extra $\sqrt{m}\rho$ being absorbed for $\rho \le c/(R\sqrt{m})$. The tail bounds of both proofs tolerate the added $L\rho$ and r_n perturbations (Step 0 and the completion-of-squares argument, respectively), giving $\Pi_m(B^c) + \widetilde{\alpha}(B^c) \le Ce^{-cR^2} + m^{-cK} + e^{-cm}$. With $R = \sqrt{K \log m}$, $\rho \asymp \varepsilon/(\sqrt{m}R_\varepsilon)$, and $t \asymp \sqrt{\log m}$, Lemma C.2 and $|\int v\, d\widetilde{\alpha} - \mathbb{E}_{\Pi_m} v| \le 2\|v\|_\infty \|\widetilde{\alpha} - \Pi_m\|_{\mathrm{TV}}$ assemble the display; the key count is the covering number of the located ball from Corollary 4.3. □

H Curved Example and Heuristic Calculations

H.1 (S3) Warped von Mises on S^1

$\mathcal{X} = S^1$, $Y \mid x \sim \mathrm{vM}(g(x), \varkappa)$, $\eta_\theta(x) = \varkappa(\cos g(x), \sin g(x))$, so $G(x) = \varkappa A_1(\varkappa) g'(x)^2$ with $A_1 = I_1/I_0$; $\eta_\theta(\mathcal{X})$ is a circle in $\mathbb{R}^2$, giving nonzero extrinsic curvature in (3). Here $p = 1$ and Lemma 3.1 reads

$$L_m = \tfrac{1}{2}\log\tfrac{m}{2\pi e} + \tfrac{1}{2}\log\{\varkappa A_1(\varkappa) g'(x_0)^2\} + \log\tfrac{1}{\bar{p}(x_0)} + o_P(1),$$

so steepening the warp ($|g'(x_0)| \uparrow$) or thinning the prompt density at x_0 shifts the predicted transition of Section 6 rightward by exactly the corresponding additive term—a directly testable prediction.

H.2 Tempered softmax (Remark 5.5): calculation sketch

With logits $\beta_T m\varphi_m$ the weights are proportional to a kernel of d_G-bandwidth $h \asymp (\beta_T m)^{-1/2}$ centered near $\hat{x}_m$. The attention mean is then a Nadaraya–Watson estimate of location: its bias against $\hat{x}_m$ is $O(h^2)$ (prior-gradient and curvature terms), and its standard deviation is $O(h/\sqrt{n\bar{p}(x_0)h^p})$

(values within a bandwidth of the mode have spread $O(h)$; the effective count is $n\bar{p}(x_0)h^p$). Requiring both $\ll m^{-1/2}$ forces $h \ll m^{-1/4}$ and $n\bar{p}(x_0) \gg mh^{2-p}$; at $h \asymp m^{-1/4}$ this is $n\bar{p}(x_0) \gg m^{(p+2)/4}$. This is a heuristic bandwidth calculation, not a theorem; it is recorded to delimit the scope of Theorem 3.3, which the tempered measure does not contradict (its rescaled W_1 to Π_m remains bounded away from zero, since its width $h \gg m^{-1/2}$).

References

[1] Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3): 405–431, 2017.

[2] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023.

[3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, 2023.

[4] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[5] Bertrand S. Clarke and Andrew R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[6] Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994. doi: 10.1016/0378-3758(94)90153-8.

[7] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

[8] Benjamin Rubacher. Attention implements the Fisher geometry of exponential families. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

[9] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of EMNLP-IJCNLP*, 2019.

[10] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[11] Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023.

[12] Sang Michael Xie, Anirudh Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

[13] Lin Zheng, Chong Wang, and Lingpeng Kong. Linear complexity randomized self-attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27011–27041, 2022.