Nonparametric In-Context Learning under Growing Geometric Complexity: Minimax Optimality and Local Geometry-Adaptivity of Transformers

Jaehee Seo
Department of Statistics
Seoul National University, Korea
seojaehee02@snu.ac.kr

Jisu Kim*
Department of Statistics
Seoul National University, Korea
jkim82133@snu.ac.kr

Abstract

Transformers have become a central architecture for *in-context learning* (ICL), particularly through their state-of-the-art performance in large language models. This success motivates understanding how transformers exploit intrinsic geometry of real-world data. However, existing nonparametric ICL theory has largely focused on simplified settings, such as Euclidean domains or single manifold models, failing to fully explain its phenomenal empirical success in complex tasks. To address this gap, we study the prediction problem under unknown heterogeneous geometry, modeled by sample size-dependent mixtures of manifolds with heterogeneous dimensions, smoothness, and sampling masses. Under the growing geometric complexity model, we establish a minimax lower bound capturing the aggregate difficulty of the components and construct an oracle tangent local polynomial estimator with a matching upper bound. This estimator is connected to a structure-informed, two-stage softmax transformer with a geometric preconditioner and local polynomial solvers. The transformer achieves negligible approximation error relative to the minimax rate with logarithmic depth and polynomial size. Finally, we derive an in-context generalization bound for near empirical risk minimizers (ERM) over this class. Together, these results identify conditions under which the resulting predictor exploits local geometry and attains minimax rate.

1 Introduction

Transformers were introduced as attention-based architectures that process a sequence by extracting and recombining task-relevant information across tokens [33]. Unlike architectures that rely on a fixed sequential state, self-attention allows each token to interact directly with all other tokens in a data-dependent way. This mechanism is particularly well-suited to prompt-based prediction, where instructions, demonstrations, contextual information, and a query are presented in a single sequence. It has become a cornerstone of large language models, whose strong performance in few-shot and ICL suggests that a model can adapt its prediction from examples already contained in the prompt, without task-specific parameter updates [9]. In this sense, the prompt is not merely an input string but also a source of task information from which the model must infer how to predict at the query point.

Recent works have studied a statistical explanation of this phenomenon by viewing transformers as prompt-dependent statistical estimators implementing learning procedures inside their forward pass [4, 6, 16, 34]. In this paper, we focus on the nonparametric ICL problem: given a prompt $\mathfrak{s} = ((X_i, Y_i)_{i=1}^n; X_{n+1})$ with $X_i = X_i^\star + \xi_i$ and $Y_i = f(X_i^\star) + \epsilon_i$, the predictor estimates $f(X_{n+1}^\star)$. Here $X_i^\star$ is the latent covariate, ξ_i and ϵ_i are observation perturbations, and f is an

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

unknown task function. The query response is not an input, and prediction uses the examples in the prompt without task-specific parameter updates. Existing nonparametric ICL results have primarily been developed for Euclidean domains or a single manifold [11, 21, 28]. Such settings provide a useful starting point. However, a central question for transformers, demonstrating the state-of-the-art performances in many complex tasks, still remains unanswered:

Can transformers still attain minimax optimality for nonparametric in-context learning under unknown heterogeneous geometry?

Data for contemporary machine learning, such as images and language representations, strongly motivate this question. Image data have been linked to low intrinsic dimension and a union-of-manifolds structure [8, 25]; a related account uses CW complexes [35]. Region-dependent intrinsic dimension [5], stratified language-model embeddings [22], and intrinsic-dimension-based text analysis [32] further motivate local heterogeneity. Transformer representations also exhibit training-dependent variation in intrinsic dimension [26]. An embedding layer therefore need not remove low-dimensional or heterogeneous geometry.

To reflect this geometric heterogeneity, we construct the statistical model through a sample-size-dependent mixture of manifolds. Components can have different dimensions, smoothness levels, and masses, while their number, reach, separation, and perturbation scales may vary with n. The geometry remains locally resolvable at the regression scale. We ask what risk is unavoidable, which estimator attains it, and whether that estimator can be realized by a transformer that generalizes across tasks.

1.1 Contribution

- **Growing Geometric setup.** We formulate a growing heterogeneous manifold-mixture model using sample size-dependent mixture numbers and weights, and specify its locally resolvable regression regime.
- **Aggregate minimax lower bound.** We establish a minimax lower bound of order $\mathfrak{R}_n$ in (2), retaining each component's dimension, smoothness, and effective sample size (Theorem 1).
- **Oracle tangent local polynomial estimator and upper bound.** We fit a polynomial graph and then regress in the estimated tangent coordinates (Algorithm 1). The oracle supplies dimension, smoothness, and bandwidth, not the tangent projector or task function. The estimator attains the matching upper bound $\mathfrak{R}_n$ (Theorem 2).
- **Transformer approximation.** We construct a structure-informed transformer using softmax attention and ReLU FFNs that performs geometric preconditioning and chartwise reduced regression in one forward pass (Figure 2). Stable charts and ReLU-to-transformer compilation give mean squared oracle-approximation error $O(n^{-A_0})$ for each fixed $A_0 > 0$, with logarithmic depth and polynomial embedding dimension, feed-forward width, and parameter bounds (Theorem 3).
- **Generalization error bound.** We separate near-ERM risk into in-context, approximation, finite-task, and optimization terms, yielding fresh-task risk $O(\mathfrak{R}_n)$ with sufficient pretraining and small optimization error (Theorem 4). A matching $\mathfrak{R}_n$ lower bound holds for every pretraining budget, worst-case over task distributions.

Relation to prior geometric results. Shen et al. [27] analyze noisy and task-level manifolds, and Shen et al. [28] connect structured-manifold ICL to kernel prediction. Their noisy-manifold regression and structured-manifold ICL results use a single manifold and Hölder exponents in $(0, 1]$. Higher-order local polynomial based ICL is already available in Euclidean domains [11]; the issue here is to control regression in estimated coordinates uniformly across a growing mixture. The perturbation models and structural information also differ, so our results do not uniformly extend all prior guarantees.

Extended related works are collected in Appendix B. Appendix C reports empirical performance gains with increasing pretraining meta-sample size and comparisons between the transformer and a local polynomial estimator using the oracle tangent.

2 Problem setup

Conventions. For scalars, $\wedge$ and $\vee$ denote minimum and maximum, respectively. Write $[m] = \{1, \dots, m\}$ for $m \in \mathbb{N}$ and $[m]_0 = \{0, \dots, m\}$ for integers $m \geq 0$. For $a \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer not less than a. Let $\Pi_{T_x\mathcal{M}}$ be the orthogonal projector onto the tangent space. The reach of a closed manifold is the largest tubular radius with unique nearest-point projection [15]. For $\tau > 0$, $\mathcal{C}^{\beta}_{d,L_{\mathcal{M}},\tau}$ consists of closed, connected, embedded d-manifolds without boundary satisfying $\mathrm{reach}(\mathcal{M}) \geq \tau$, with uniformly regular tangent-normal charts of fixed radius independent of τ. The Hölder ball $\mathcal{H}^{\alpha}(\mathcal{M}; L_{\mathcal{F}})$ is defined by uniform chart pullbacks; at integer α we use the $C^{\alpha-1,1}$ convention. Precise definitions are in Appendix D; the transformer class is defined in Section 4.

Mixture of manifolds. Independently of n, fix $D \geq 2$, $d_{\max} \in [D-1]$, $0 < \alpha_{\min} \leq \alpha_{\max}$, $B_X, L_{\mathcal{M}} \geq 1$, $c_\mu \in (0, 1]$, and $L_{\mathcal{F}}, B_\epsilon, \sigma_Y > 0$. Set $\beta_{\max} = \lceil \alpha_{\max} \rceil + 1$. Component geometry and smoothness may depend on n; for convenience, we abbreviate $(\mathcal{M}_{k,n}, d_{k,n}, \alpha_{k,n}, \beta_{k,n})$ by $(\mathcal{M}_k, d_k, \alpha_k, \beta_k)$. Reach, separation, masses, and perturbation bounds retain their n subscripts.

Assumption 1 (Growing heterogeneous geometry)**.** *For each n, let $K_n \in \mathbb{N}$ and $\tau_{0,n}, \delta_{0,n} > 0$. There are K_n manifolds $\mathscr{M}_n = (\mathcal{M}_k)_{k=1}^{K_n}$, with $\mathcal{M}_k \in \mathcal{C}^{\beta_k}_{d_k, L_{\mathcal{M}}, \tau_{0,n}}$, $\mathcal{M}_k \subset [-B_X, B_X]^D$, $d_k \in [d_{\max}]$, and $2 \leq \beta_k \leq \beta_{\max}$, such that*

$$\min_{k \neq \ell} d(\mathcal{M}_k, \mathcal{M}_\ell) \geq \delta_{0,n} > 0.$$

Here $d(A, B) = \inf_{a \in A, b \in B} \|a - b\|$; the separation condition is vacuous when $K_n = 1$.

The local regularity constant stays fixed, whereas $\tau_{0,n}$ and $\delta_{0,n}$ may shrink. The local chart derivatives remain uniformly controlled along the component sequence. Figure 1 illustrates these distinct roles of local regularity, component count, and separation.

Assumption 2 (Latent design and bounded perturbations)**.** *The label $Z \in [K_n]$ has masses $\pi_{k,n} > 0$ with $\sum_{k \in [K_n]} \pi_{k,n} = 1$. For each $k \in [K_n]$, $X^\star \mid (Z = k)$ has law $\mu_{k,n}(dx) = \varrho_{k,n}(x) d\,\mathrm{vol}_{\mathcal{M}_k}(x)$ with $c_\mu \leq \varrho_{k,n} \leq c_\mu^{-1}$ $\mathrm{vol}_{\mathcal{M}_k}$-a.e. For $\sigma_{k,n} \geq 0$,*

$$\mathbb{P}(\|\xi\| \leq \sigma_{k,n} \mid X^\star, Z = k) = 1.$$

The response noise satisfies, almost surely,

$$\mathbb{E}[\epsilon \mid X^\star, \xi, Z] = 0, \qquad |\epsilon| \leq B_\epsilon, \qquad \mathbb{E}[\epsilon^2 \mid X^\star, \xi, Z] \leq \sigma_Y^2. \tag{1}$$

The perturbation magnitude is restricted at the regression resolution below.

Tasks and observations. Write $\mathscr{M} = \mathscr{M}_n$, $\boldsymbol{\pi} = (\pi_{k,n})_k$, and $\boldsymbol{\sigma}_X = (\sigma_{k,n})_k$. For $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ assume $\beta_k \geq \lceil \alpha_k \rceil + 1$, and set

$$\mathcal{H} = \mathcal{H}(\boldsymbol{\alpha}, \mathscr{M}; L_{\mathcal{F}}) := \left\{ f : \bigsqcup_k \mathcal{M}_k \to \mathbb{R} : f|_{\mathcal{M}_k} \in \mathcal{H}^{\alpha_k}(\mathcal{M}_k; L_{\mathcal{F}}) \right\}.$$

Let $\mathcal{P}_\star = \mathcal{P}_\star(\mathscr{M}, \boldsymbol{\pi}, \boldsymbol{\sigma}_X; \sigma_Y^2, B_\epsilon, c_\mu)$ be the nonempty class of laws satisfying Assumption 2. For $P \in \mathcal{P}_\star$ and $f \in \mathcal{H}$, draw i.i.d. latent/noise tuples and observe $X_i = X_i^\star + \xi_i$, $Y_i = f(X_i^\star) + \epsilon_i$, for $i \leq n$, together with X_{n+1}. The target is $f(X_{n+1}^\star)$. In particular, $|Y_i| \leq B_Y := L_{\mathcal{F}} + B_\epsilon$. For a prompt predictor g, abbreviate its risk by

$$\mathcal{E}_{P,f}(g) := \mathbb{E}_{P,f}\left[\{ g(\mathfrak{s}) - f(X_{n+1}^\star) \}^2 \right].$$

When g is written as an estimate at X_{n+1}, its dependence on the in-context sample is implicit.

Assumption 3 (Feasible local regime)**.** *For a fixed $\kappa \in (0, 1)$, put $N_k = n\pi_{k,n}$, $h_k = h_{k,n} = N_k^{-1/(2\alpha_k + d_k)}$, and $r_n = \max_k h_k$. For all sufficiently large n, uniformly over k,*

$$N_k \geq n^\kappa, \qquad \tau_{0,n} \wedge \delta_{0,n} \geq a_{\mathrm{sc}}^{-1} r_n, \qquad \sigma_{k,n} \leq a_{\mathrm{sc}} h_k^{\alpha_k \vee 1},$$

where $a_{\mathrm{sc}} > 0$ is a sufficiently small constant specified in Appendix E.2.

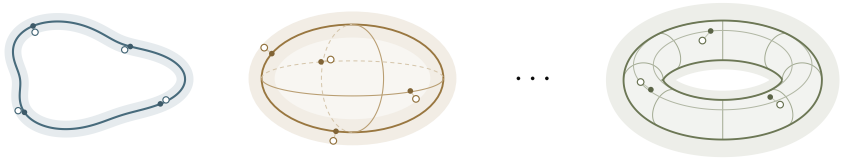


Figure 1: A heterogeneous manifold mixture with bounded covariate noise. The closed curve, ellipsoidal surface, and torus represent components of intrinsic dimensions one, two, and two, respectively; the ellipsis denotes omitted components. Filled dots mark latent covariates $X_i^\star$ on the manifolds, and open circles mark observed covariates $X_i = X_i^\star + \xi_i$. The connecting segments indicate the perturbations ξ_i, while the pale shaded neighborhoods schematically depict component-specific noise bounds $\|\xi_i\| \le \sigma_{k,n}$. In the growing-complexity setting, the number of components K_n may increase while the reach and separation lower bounds $\tau_{0,n}$ and $\delta_{0,n}$ may decrease with n.

The three requirements in Assumption 3 provide enough local samples, prevent different components from entering the same regression window, and keep covariate perturbations below the target bias scale. They imply $K_n \le n^{1-\kappa}$ and polynomially increasing local counts $N_k h_k^{d_k}$. Labels index the mixture components; dimensions and smoothness levels may coincide across components. $X^\star$ is the generating latent point; it can differ from the nearest-point projection of X. The risk and tangent comparison refer to this same generating point. We treat $X \in \mathbb{R}^D$ as a numerical covariate, possibly obtained from a fixed learned representation map. This map is distinct from Emb_n, which packs examples into tokens; learning the representation is outside our analysis.

As [13] pointed out, the tubular noise model may not be identifiable, since there are several admissible couples $(X^\star, \xi)$ such that $X = X^\star + \xi$ follows the same distribution. Our model is specified by the full latent law $P \in \mathcal{P}_\star$; the target and risk are defined under this law, and our bounds hold uniformly over admissible P.

Reading the scales. The number of observations relevant to a query is smaller than N_k: within a bandwidth-h_k neighborhood it is of order $N_k h_k^{d_k}$. The local volume and density bounds give this scale; the chosen bandwidth balances the squared approximation bias and response-noise variance:

$$h_k^{2\alpha_k} = \frac{1}{N_k h_k^{d_k}}, \qquad N_k h_k^{d_k} = N_k^{2\alpha_k/(2\alpha_k+d_k)}.$$

The effective-sample condition makes this local count grow polynomially, so the logarithmic cost of comparing tangent candidates remains smaller. The geometric condition has a different role: reach permits a stable local graph, while separation prevents observations from another component from entering that graph fit. Finally, the perturbation condition ensures $\sigma_{k,n} \lesssim h_k^{\alpha_k}$ when $\alpha_k > 1$ and $\sigma_{k,n} \lesssim h_k$ otherwise. These bounds keep the perturbation at the regression resolution.

Uniformity of constants. The asymptotic notations $\lesssim, \gtrsim$ hide constants depending only on the fixed parameters above, κ, and fixed construction margins. They are uniform in n, K_n, k, P, f and in admissible component sequences. Shrinking masses, reaches, and separations enter through the displayed scales. We distinguish statistical and implementation constants below.

3 Minimax rate and oracle estimation

3.1 Minimax lower bound

If the query belongs to component k, its intrinsic dimension is d_k, its smoothness is α_k, and its effective sample size is $N_k = n\pi_{k,n}$. Accordingly, define the target minimax rate as

$$\mathfrak{R}_n = \mathfrak{R}_n(\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\pi}) := \sum_{k \in [K_n]} \pi_{k,n} (n\pi_{k,n})^{-2\alpha_k/(2\alpha_k+d_k)}. \tag{2}$$

This averages the component-specific statistical difficulties. In the balanced homogeneous case it equals $(n/K_n)^{-2\alpha/(2\alpha+d)}$; for $K_n = 1$ it is the classical nonparametric rate [7, 18, 29, 31].

The aggregate also makes the effect of growing mixture complexity explicit. For any admissible balanced sequence with $K_n \asymp n^\eta$ and common (d, α), it becomes $n^{-(1-\eta)2\alpha/(2\alpha+d)}$; the effective sample condition requires $\eta \le 1 - \kappa$. This calculation applies to sequences satisfying the geometric assumptions.

More generally, writing $a_k = 2\alpha_k/(2\alpha_k + d_k)$, component k contributes $n^{-a_k}\pi_{k,n}^{1-a_k}$. Small mass worsens the conditional estimation problem through $n\pi_{k,n}$, but simultaneously reduces how often that difficulty is encountered at the query. The benchmark keeps both effects. When masses or component counts vary with n, replacing this average by a single largest dimension can therefore discard the relevant distribution of difficulty.

Theorem 1 (Minimax lower bound)**.** *Under Assumptions 1–3, for all sufficiently large n,*

$$\inf_{g} \sup_{f \in \mathcal{H},\, P \in \mathcal{P}_\star} \mathcal{E}_{P,f}(g) \gtrsim \mathfrak{R}_n,$$

where the infimum ranges over measurable predictors g of the observed prompt $\mathfrak{s}$.

Here, the hidden lower bound constant depends only on fixed statistical model parameters and scale margins. Its proof constructs one Assouad family with perturbations on every component at its own resolution. A single testing argument therefore makes all component contributions simultaneous. A bounded, smooth response-noise submodel and a Hellinger translation bound respect (1). Formal details and the proof are given in Appendix F.

3.2 Oracle tangent local polynomial estimator

Now we construct the statistical estimator that the transformer will approximate in the next section. For a given query $x = X_{n+1}$ from component k_x, we suppose the oracle knows $d_x = d_{k_x}$, $\alpha_x = \alpha_{k_x}$, $p_x = \lceil \alpha_x \rceil - 1$, $s_x = p_x + 1$, and $h_x = h_{k_x}$. Hence the inputs are the observed prompt and these structural quantities. The tangent projector and local regression coefficients are fitted from that prompt: a polynomial graph fit estimates the tangent, followed by regression in the projected coordinates. Higher-order graph fitting controls geometric bias at the regression resolution [1, 10].

Why fit the geometry before regression? At the latent query $x^\star$, a tangent-normal chart writes a nearby point as $x^\star + u + G_{x^\star}(u)$, with $G_{x^\star}(0) = DG_{x^\star}(0) = 0$. The manifold regularity supplies a s_x-order Taylor expansion of G. Substituting $u = h_x t$ and dividing ambient displacements by h_x, its remainder becomes $O(h_x^{s_x})$. This explains the choice $s_x = \lceil \alpha_x \rceil$ for geometry and $p_x = s_x - 1$ for regression: the two polynomial fits approximate different objects. Moreover, the observed query can be off the manifold. In the proof, the candidate offset $a^\circ = (I - \Pi^\star)\xi_{n+1}/h_x$ compensates its normal displacement, where $\Pi^\star = \Pi_{T_{x^\star}\mathcal{M}_{k_x}}$. The remaining coordinate perturbation is controlled at order $\sigma_{k_x,n}/h_x$. These latent quantities define the comparison candidate used in the proof. A nearby grid point provides the reference graph loss against which the observed-data fit is analyzed.

Finite grid and its cost. A candidate $q = (a, \Pi, T_2, \ldots, T_s)$ has $a \in \mathbb{R}^D$ with $\|a\| \le 1$, a rank-d orthogonal projector Π, and symmetric multilinear maps $T_\ell : (\mathbb{R}^D)^\ell \to \mathbb{R}^D$ with $\|T_\ell\|_{\mathrm{op}} \le 2L_{\mathcal{M}}$. Let $\Theta_{d,s}$ be this compact parameter space. Use the product n^{-1}-net $\mathcal{Q}_{d,s,n}$ of Lemma G.1, under the sum of the offset norm, projector Frobenius norm, and tensor operator norms, and fix $\Delta_{\tan,n} = n^{-3}$. This grid satisfies

$$|\mathcal{Q}_{d,s,n}| \le C n^{m_\Theta(d,s,D)}, \qquad m_\Theta(d,s,D) = D + d(D-d) + D\sum_{\ell=2}^{s} \binom{D+\ell-1}{\ell}.$$

Its logarithmic cardinality enters tangent concentration, while its cardinality determines computational cost, with an exponent depending on D, s. The deterministic grid and its loss coefficients are encoded in the comparator weights (Appendix H.3). The polynomial bound treats D, s as fixed; its exponent records the ambient-dimensional cost of the geometric search in Algorithm 1.

Localized graph fit. For the scaled ambient displacement $z_i = (X_i - x)/h_x$, use the fixed cutoff $\mathsf{A} : [0,\infty) \to [0,1]$, where $\mathsf{A}(t) = 1$ for $0 \le t \le 1/2$, $\mathsf{A}(t) = 2 - 2t$ for $1/2 < t < 1$, and $\mathsf{A}(t) = 0$ for $t \ge 1$. Define

$$R_q(z) = a + (I - \Pi)z - \sum_{\ell=2}^{s_x} T_\ell\{(\Pi z)^{\otimes \ell}\}, \qquad L_q(x) = \frac{h_x^2}{n}\sum_{i \in [n]} \mathsf{A}(\|z_i\|_1)\|R_q(z_i)\|^2. \quad (3)$$

The tensor sum is empty for $s_x = 1$. Loss-thresholded weights average $\Delta_{\tan,n}$-near-minimizing projectors, with normalizing denominator at least $\Delta_{\tan,n} > 0$. This convex average can have fractional eigenvalues. We spectrally round it to $\widehat{\Pi}_x = \Pi_{(d_x)}(\bar{P}_x)$, where $\Pi_{(d)}$ selects the top-d eigenspace with fixed measurable tie-breaking rule. With high probability, a fixed spectral gap makes this step stable (Lemma G.5).

Ambient localization and projected regression weights. Let $\widehat{v}_i = \widehat{\Pi}_x z_i \in \mathbb{R}^D$, the scaled displacement projected onto the estimated tangent subspace. For the fixed kernel $\mathsf{K}(v) = (1 - \|v\|_1)_+$, where $(t)_+ = t \vee 0$, the regression weight is $\widehat{W}_i = h_x^{-d_x} \mathsf{A}(\|z_i\|_1)\mathsf{K}(\widehat{v}_i)$. The cutoff excludes observations with $\|X_i - x\|_1 \geq h_x$, yielding a single-component window under the feasible local regime. Within that window, the kernel tapers weights according to $\|\widehat{v}_i\|_1$ and is zero when $\|\widehat{v}_i\|_1 \geq 1$. Projection can hide a large ambient displacement, which motivates the ambient cutoff. The resulting nonnegative weights define the local least-squares objective. Its fitted solution is invariant to the common scale $h_x^{-d_x}$. When the weight sum is positive, dividing by it also preserves the fitted coefficient (Remark H.1). Let $\widehat{\phi}_i = \Phi_{D,p_x}(\widehat{v}_i)$, where $\Phi_{D,p}(v) = (v^\nu/\nu!)_{|\nu|\leq p}$ and e_0 selects the constant coordinate. Algorithm 1 summarizes the estimator.

Algorithm 1 Oracle tangent local polynomial estimation

Input: $(X_i, Y_i)_{i=1}^n$, query x, d_x, p_x, s_x, h_x, $\mathcal{Q}_{d_x,s_x,n}$, $\Delta_{\tan,n}$, and $L_{\mathcal{F}}$.
Set $z_i = (X_i - x)/h_x$. Evaluate $L_q(x)$ in (3) for every $q \in \mathcal{Q}_{d_x,s_x,n}$.
Set $L_{\min} = \min_q L_q(x)$ and $t_q = \{\Delta_{\tan,n} - (L_q(x) - L_{\min})\}_+$. Normalize $\omega_q = t_q / \sum_{q'} t_{q'}$.
Form $\bar{P}_x = \sum_q \omega_q \Pi_q$ and $\widehat{\Pi}_x = \Pi_{(d_x)}(\bar{P}_x)$ with fixed measurable tie-breaking.
For each $i \in [n]$, set $\widehat{v}_i = \widehat{\Pi}_x z_i$, $\widehat{\phi}_i = \Phi_{D,p_x}(\widehat{v}_i)$, and $\widehat{W}_i = h_x^{-d_x} \mathsf{A}(\|z_i\|_1)\mathsf{K}(\widehat{v}_i)$.
Form $\widehat{G}_x^{\Pi} = n^{-1} \sum_{i\in[n]} \widehat{W}_i \widehat{\phi}_i \widehat{\phi}_i^\top$ and $\widehat{b}_x^{\Pi} = n^{-1} \sum_{i\in[n]} \widehat{W}_i \widehat{\phi}_i Y_i$.
Compute $\widehat{w}_x = (\widehat{G}_x^{\Pi})^\dagger \widehat{b}_x^{\Pi}$.
Return: $\widehat{f}_{\text{plug}}^{\Pi}(x) = (-L_{\mathcal{F}}) \vee e_0^\top \widehat{w}_x \wedge L_{\mathcal{F}}$.

The final prediction is

$$\widehat{f}_{\text{plug}}^{\Pi}(x) = (-L_{\mathcal{F}}) \vee e_0^\top (\widehat{G}_x^{\Pi})^\dagger \widehat{b}_x^{\Pi} \wedge L_{\mathcal{F}}. \tag{4}$$

Ambient monomials become redundant after projection, so $\widehat{G}_x^{\Pi}$ may be singular even on a good design. The pseudoinverse acts on the identifiable polynomial subspace, where stability follows from

$$\lambda_{\min}^+(\widehat{G}_x^{\Pi}) \geq c\pi_{k_x,n}, \qquad e_0^\top (\widehat{G}_x^{\Pi})^\dagger \widehat{G}_x^{\Pi} = e_0^\top$$

on the regression good event. These properties identify the fitted intercept.

Indeed, $\Phi_{D,p}(Uu) = B_{U,p}\Phi_{d,p}(u)$ for an orthonormal frame U and a full-column-rank $B_{U,p}$. A positive definite reduced Gram H therefore induces the ambient Gram $B_{U,p}HB_{U,p}^\top$, whose zero eigenvalues reflect redundant coordinates. Since $e_{0,D,p} = B_{U,p}e_{0,d,p}$, the intercept remains identifiable. The transformer solves in the reduced dictionary; Lemma H.8 proves that, on that event, all positive-weight charts return the same intercept as (4).

Local averaging as a special case. When $0 < \alpha_x \leq 1$, the regression degree is $p_x = 0$, so its dictionary contains only the constant. On a nonempty weighted window,

$$\widehat{f}_{\text{plug}}^{\Pi}(x) = (-L_{\mathcal{F}}) \vee \frac{\sum_{i\in[n]} \widehat{W}_i Y_i}{\sum_{i\in[n]} \widehat{W}_i} \wedge L_{\mathcal{F}}.$$

For $p_x \geq 1$, higher-order terms fit local variation while the intercept estimates the value at the query origin. Thus the same matrix formulation extends local averaging to exploit higher smoothness.

Theorem 2 (Oracle minimax upper bound)**.** *Under Assumptions 1–3, Algorithm 1 satisfies, for all sufficiently large n,*

$$\sup_{f\in\mathcal{H},\, P\in\mathcal{P}_\star} \mathcal{E}_{P,f}(\widehat{f}_{\text{plug}}^{\Pi}) \lesssim \mathfrak{R}_n.$$

The constant may additionally depend on the fixed kernel, cutoff, grid resolution exponent, and near-minimizer rule, uniformly in n, K_n, k and the growing grid cardinality. Together with Theorem 1, this attains the benchmark $\mathfrak{R}_n$.

Geometric error and the regression rate. Condition on the query variables with component k, and put $s_k = \lceil \alpha_k \rceil$ and $\delta_x = \|\widehat{\Pi}_x - \Pi_{T_{x^\star}\mathcal{M}_k}\|_{\mathrm{op}}$. Appendix G.5 gives

$$\mathbb{E}[\delta_x^2 \mid k, x^\star, \xi_{n+1}] \lesssim h_k^{2s_k} + \sigma_{k,n}^2/h_k^2 + \frac{\log(en)}{N_k h_k^{d_k}} + h_k^{2\alpha_k},$$

$$\mathbb{E}[(\widehat{f}^{\Pi}_{\mathrm{plug}}(x) - f(x^\star))^2 \mid k, x^\star, \xi_{n+1}] \lesssim h_k^{2\alpha_k} + h_k^2 \mathbb{E}[\delta_x^2 \mid k, x^\star, \xi_{n+1}] + \sigma_{k,n}^2 + \frac{1}{N_k h_k^{d_k}}.$$

The tail and bad-design terms are absorbed at order $h_k^{2\alpha_k}$. The comparison polynomial is $O(h_k)$-Lipschitz in normalized coordinates; thus coordinate error $\delta_x + \sigma_{k,n}/h_k$ produces regression error $O(h_k\delta_x + \sigma_{k,n})$. Consequently the tangent stochastic term is multiplied by h_k^2, and $h_k^2 \log(en)$ stays bounded. Since $(N_k h_k^{d_k})^{-1} = h_k^{2\alpha_k}$, the component-wise bias and variance balance. The tangent estimator uses only covariates, so the response-noise variance calculation remains valid conditional on the estimated geometry.

Two uniformity steps are essential in this argument. The localized covariate window must contain only observations from the query component, even when the separation scale changes with n. Next, the regression design event must hold for every rank-d_k projector in a neighborhood of the true projector. This uniform event accommodates the dependence of $\widehat{\Pi}_x$ on the regression covariates. The positive-spectrum bound and the weight-sum bound then give variance $O((N_k h_k^{d_k})^{-1})$ after conditioning on all covariates. Finally, the component-wise failure probabilities and risk bounds are uniform, so averaging over the query label preserves the component-wise constants.

4 Transformer realization

Transformers. For N tokens $Z \in \mathbb{R}^{N \times d_E}$ of dimension d_E and M heads with $Q_m, K_m, V_m \in \mathbb{R}^{d_E \times d_E}$, define

$$\mathrm{MHA}(Z) = \sum_{m \in [M]} \mathrm{softmax}_{\mathrm{row}}\left(\frac{ZQ_m(ZK_m)^\top}{\sqrt{d_E}}\right) ZV_m.$$

The row-wise FFN of width d_{FFN} is $\mathrm{FFN}(Z) = \mathrm{ReLU}(ZW_1 + \mathbf{1}_N b_1^\top)W_2 + \mathbf{1}_N b_2^\top$, where $W_1, W_2^\top \in \mathbb{R}^{d_E \times d_{\mathrm{FFN}}}$, $b_1 \in \mathbb{R}^{d_{\mathrm{FFN}}}$, $b_2 \in \mathbb{R}^{d_E}$, and $\mathbf{1}_N$ is the all-ones vector. Let $\theta^{(\ell)}$ collect these matrices and biases in block $\ell \in [L]$, and put $\theta = (\theta^{(1)}, \ldots, \theta^{(L)})$. Define

$$H = Z + \mathrm{MHA}_{\theta^{(\ell)}}(Z), \quad \mathrm{Block}_{\theta^{(\ell)}}(Z) = H + \mathrm{FFN}_{\theta^{(\ell)}}(H), \quad \mathrm{TF}_\theta = \mathrm{Block}_{\theta^{(L)}} \circ \cdots \circ \mathrm{Block}_{\theta^{(1)}}.$$

For $B > 0$, $\mathcal{T}_M(N, d_E, d_{\mathrm{FFN}}, L, B)$ comprises these maps $\mathbb{R}^{N \times d_E} \to \mathbb{R}^{N \times d_E}$ with every parameter entry bounded in absolute value by B; write $\mathcal{T}(d_E, d_{\mathrm{FFN}}, L, B)$ when M, N are fixed.

Input and output interface. The scalar predictor is

$$\widehat{f}^{\mathrm{TF}}_{\theta_n}(\mathfrak{s}) = \mathrm{Read}_n \circ \mathrm{TF}_{\theta_n} \circ \mathrm{Emb}_n(\mathfrak{s}). \tag{5}$$

The fixed affine embedding packs each (X_i, Y_i) into one sample row and $(X_{n+1}, 0)$ into one query row, giving $N_n^{\mathrm{end}} = n + 1$. The construction preserves raw data and row-type markers; extra coordinates store intermediate geometric and regression computations. The fixed readout uses ReLU multiplication to approximate the chart-weighted sum and clips the resulting scalar to $[-L_{\mathcal{F}}, L_{\mathcal{F}}]$ (Appendix H.2). The query response is used in the pretraining loss.

From geometric summaries to an intrinsic regression system. A selector encoded in the comparator parameters evaluates $x = X_{n+1}$ to obtain d_x, p_x, s_x, h_x and $\pi_{k_x,n}^{-1/2}$. Its network is fixed by the deterministic structural model and n; the query determines the component state (Appendix H.3). Using this state, the geometric block aggregates localized polynomial moments of the observed covariates, approximates the finite-grid projector fit, and produces tangent and chart states. The regression block combines the resulting coordinates and local weights with the observed responses to form chartwise normal equations. A globally defined bounded ReLU decoder approximates

their inverse on well-conditioned inputs. These are internal stages of one forward pass (Figure 2, Appendix A). The reduced regression dictionary has size

$$q_{\mathrm{reg,max}} = \binom{d_{\max} + p_{\max}}{p_{\max}}, \qquad p_{\max} = \lceil \alpha_{\max} \rceil - 1,$$

compared with the ambient dictionary size $\binom{D+p_{\max}}{p_{\max}}$. Geometric calculations and network-size exponents retain their dependence on D.

Fixed-dimensional summaries. With $s_{\max} = \lceil \alpha_{\max} \rceil$, the geometric loss depends on the in-context covariates through

$$m_{\tan}(x) = \frac{1}{n} \sum_{i \in [n]} \mathsf{A}(\|z_i\|_1) \Phi_{D,2s_{\max}}(z_i).$$

Indeed, each $\|R_q(z)\|^2$ is a polynomial of degree at most $2s_{\max}$, so $L_q(x) = h_x^2 c_q^\top m_{\tan}(x)$ for a deterministic coefficient c_q. All tangent candidates can therefore be evaluated using the same summary vector. After charting, the analogous regression summaries are

$$G_j = \frac{1}{n} \sum_{i \in [n]} \pi_{k_x,n}^{-1} W_{i,j} \phi_{i,j} \phi_{i,j}^\top, \qquad g_j = \frac{1}{n} \sum_{i \in [n]} \pi_{k_x,n}^{-1} W_{i,j} \phi_{i,j} Y_i.$$

The factor $\pi_{k_x,n}^{-1}$ keeps the good-event Gram spectrum bounded below by a fixed positive constant, and cancels in $G_j^{-1} g_j$. Row-wise FFNs compute local weighted contributions in value coordinates. Uniform softmax with binary marker gates and deterministic value scaling performs exact query broadcast and sample averaging (Lemma H.2). Compact localization therefore acts through values. The summary dimensions are fixed with respect to n, although their numerical ranges and approximation widths can grow. Each arithmetic approximator acts on a single row or a fixed-dimensional summary, yielding polynomial resource bounds as the prompt grows.

Why several charts represent one estimator. A rank-d projector specifies a subspace; continuous frame choices are local to coordinate charts. Using all charts avoids discontinuities from choosing one frame globally. Let P_T be an exact rank-d tangent projector, let $\mathfrak{C}_d$ be the collection of all size-d subsets of $[D]$, and let E_J select the coordinates in J. Define

$$A_J = E_J^\top P_T E_J, \qquad \sum_{J \in \mathfrak{C}_d} \det A_J = 1.$$

The second identity is Cauchy–Binet: at least one coordinate choice sees all tangent directions. The chart rule assigns positive weight only when $\det A_J > \tau_d$, with $\tau_d = (4|\mathfrak{C}_d|)^{-1}$. Since the eigenvalues of A_J lie in $[0, 1]$, positivity gives $\lambda_{\min}(A_J) > \tau_d$. On such a chart,

$$U_J = P_T E_J A_J^{-1/2}, \qquad U_J^\top U_J = I_d, \qquad U_J U_J^\top = P_T.$$

Thus the chart supplies stable orthonormal coordinates, and $U_J U_J^\top (X_i - x)/h_x$ is the same projected covariate for every positive-weight chart. The reduced polynomial dictionaries are different coordinate descriptions of the same polynomial space. On the regression good event, their Gram matrices are positive definite, so their fitted intercepts agree with the ambient pseudoinverse intercept. Their weighted average therefore represents the same estimator on that event. Safe chart maps extend the coordinate formulas to other inputs.

Only positive-weight charts require inverse approximation. A zero-weight chart may have a singular Gram matrix, but its globally defined decoder still returns a clipped value. If approximate weights place small mass on such a chart, the resulting error is bounded by $L_{\mathcal{F}}$ times that weight error. The decoder is evaluated on every branch, including those with zero weight. Appendix H.7 gives the precise comparison and controls failures of the spectral and regression good events.

Theorem 3 (Transformer approximation)**.** *Fix $A_0 > 0$ and a deterministic structural model satisfying Assumptions 1–3. For all sufficiently large n, there are deterministic dimensions and a parameter vector $\widehat{\theta}_n$ defining a transformer*

$$\mathrm{TF}_{\widehat{\theta}_n} \in \mathcal{T}_{M_0}(n+1, d_{E,n}, d_{\mathrm{FFN},n}, L_n, B_n)$$

whose fixed-interface predictor satisfies

$$\sup_{f\in\mathcal{H},\, P\in\mathcal{P}_\star} \mathbb{E}_{P,f}\left[\{\widehat{f}^{\mathrm{TF}}_{\widehat{\theta}_n}(\mathfrak{s}) - \widehat{f}^{\Pi}_{\mathrm{plug}}(X_{n+1})\}^2\right] \lesssim n^{-A_0}.$$

Here $M_0 = O(1)$, $L_n \le C_{\mathrm{depth}} \log(en)$, and $d_{E,n} + d_{\mathrm{FFN},n} + B_n \le n^{C_{\mathrm{size}}}$, with $C_{\mathrm{depth}}, C_{\mathrm{size}}$ and implicit constants depending only on A_0 and fixed model and construction bounds (Appendix H.9). The parameter vector is fixed by the structural model and n, and the same vector applies uniformly over $f \in \mathcal{H}$ and $P \in \mathcal{P}_\star$.

Combining the two comparisons by the squared triangle inequality gives

$$\sup_{f\in\mathcal{H},\, P\in\mathcal{P}_\star} \mathcal{E}_{P,f}(\widehat{f}^{\mathrm{TF}}_{\widehat{\theta}_n}) \lesssim \mathfrak{R}_n + n^{-A_0}.$$

Thus a sufficiently accurate implementation preserves the statistical rate of the oracle target. This establishes a good comparator in the transformer class before any pretraining analysis. Choosing a predictor from that class using finitely many tasks is a separate statistical step, addressed next.

Accuracy, workspace, and structural information. The proof compiles ReLU modules into FFNs with residual connections using two polynomial-width scratch banks, with raw observations and markers protected (Lemma H.3). These banks store the intermediate module states. One communication head copies or averages an entire stored vector, so the head count stays fixed. Corollary H.1 makes the accuracy dependence explicit:

$$L_n \le (c_{L,0} + c_{L,1} A_0)\log(en), \qquad d_{E,n} + d_{\mathrm{FFN},n} + B_n \le n^{c_{S,0}+c_{S,1}A_0}, \tag{6}$$

with coefficients fixed by the structural bounds and module choices; the threshold in n can depend on A_0. The grid and selector are encoded in the comparator's weights. Their construction uses the specified geometry and component parameters, and is fixed before sampling any prompt. The same construction covers the admissible within-component density and perturbation laws. Thus the comparator is shared across tasks within that structural environment, while its encoded construction may vary with the environment. Its resource bounds are given in (6), and its guarantee is in mean square, with clipping controlling rare spectral or design failures.

5 In-context generalization

Training and target tasks. Fix $P \in \mathcal{P}_\star$ and a distribution ρ_f on $\mathcal{H}$. Each of Γ independent training tasks draws $f^{(\gamma)} \sim \rho_f$ and generates $n+1$ independent latent/noise tuples from P as in Section 2. Training observes $\mathcal{D}^{\mathrm{tr}}_\Gamma = \{(\mathfrak{s}^{(\gamma)}, Y^{(\gamma)}_{n+1})\}_{\gamma=1}^{\Gamma}$. An independent target task is generated in the same way from a fresh $f \sim \rho_f$. Here n measures information within a task; Γ measures the number of training tasks. Tasks share the structural model and sampling law.

Let $\mathcal{G}^{\mathrm{end}}_n$ be the class of all scalar predictors (5) over the deterministic architecture envelope in Theorem 3, with fixed embedding and readout. All parameters, including the query and key matrices, vary freely within the prescribed bound. Define

$$R^\star_{P,\rho_f}(g) = \mathbb{E}_{f\sim\rho_f}\mathcal{E}_{P,f}(g), \qquad \widehat{R}_\Gamma(g) = \frac{1}{\Gamma}\sum_{\gamma\in[\Gamma]}\{Y^{(\gamma)}_{n+1} - g(\mathfrak{s}^{(\gamma)})\}^2.$$

Fix a deterministic tolerance $\varepsilon_{\mathrm{opt},\Gamma} \ge 0$. A measurable estimator $\widehat{g}_\Gamma \in \mathcal{G}^{\mathrm{end}}_n$ is a near ERM if, almost surely,

$$\widehat{R}_\Gamma(\widehat{g}_\Gamma) \le \inf_{g\in\mathcal{G}^{\mathrm{end}}_n} \widehat{R}_\Gamma(g) + \varepsilon_{\mathrm{opt},\Gamma}. \tag{7}$$

The tolerance bounds the achieved empirical-risk gap and enters the generalization bound explicitly, quantifying how imperfect optimization affects the final statistical guarantee.

Theorem 4 (Near-ERM generalization)**.** *Under Assumptions 1–3, fix $A_0 > 0$ and the architecture of Theorem 3. Every measurable near ERM in (7) satisfies*

$$\mathbb{E}_{\mathcal{D}^{\mathrm{tr}}_\Gamma} R^\star_{P,\rho_f}(\widehat{g}_\Gamma) \lesssim \mathfrak{R}_n + n^{-A_0} + \frac{\mathfrak{H}^{\mathrm{end}}_{n,\Gamma} + 1}{\Gamma} + \varepsilon_{\mathrm{opt},\Gamma},$$

where the entropy proxy defined in Appendix I.2 obeys $\mathfrak{H}^{\mathrm{end}}_{n,\Gamma} \le C_{\mathrm{ent}} n^{c_{\mathrm{ent}}}\{1 + \log(e\Gamma)\}$. Constants are uniform over P, ρ_f and admissible component sequences; they may depend on A_0 and fixed statistical and architectural bounds.

The four terms separate intrinsic in-context error, implementation error, finite-task complexity, and optimization tolerance. The proof uses a bounded square-loss oracle inequality and parameter covering of the full softmax class. Row-stochastic attention controls hidden-state growth, and the global Lipschitz bound for softmax controls parameter sensitivity. Discretizing the parameter cube then yields the entropy term, including the polynomial embedding dimension (Lemma I.1). Conditional mean-zero query noise allows training against Y_{n+1} while evaluating against the latent target. See Appendix I.

In particular, for a fixed predictor g the noisy and latent risks satisfy

$$\mathbb{E}\{Y_{n+1} - g(\mathfrak{s})\}^2 = R^\star_{P,\rho_f}(g) + \mathbb{E}\epsilon^2_{n+1}.$$

The same identity holds for the learned predictor after conditioning on the independent training tasks. The empirical-process argument is centered at $g_0(\mathfrak{s}) = \mathbb{E}[Y_{n+1} \mid \mathfrak{s}]$, with the oracle inequality allowing a misspecified predictor class. The good comparator provided by Theorem 3 controls approximation to the latent regression value, while the covering term controls choosing a predictor from the entire parameter-bounded class. A training procedure may select any admissible parameter vector, so its generalization cost depends on the capacity of that whole class. Following the covering-based risk analysis of Ching et al. [11], we use a finite sup-norm net to reduce uniform control to finitely many predictors. Since the variance of squared-loss differences relative to g_0 is controlled by excess risk, Bernstein's inequality and a union bound over a net at resolution $\asymp \Gamma^{-1}$ convert its log-cardinality into the $(\mathfrak{H}^{\mathrm{end}}_{n,\Gamma} + 1)/\Gamma$ near-ERM remainder (Appendix I.3).

Corollary 1 (Sufficient meta-sample regime)**.** *Choose* $A_0 \geq 2\alpha_{\max}/(2\alpha_{\max} + 1)$. *If*

$$\Gamma \gtrsim n^{c_{\mathrm{ent}}} \log(en)\mathfrak{R}_n^{-1}, \qquad \varepsilon_{\mathrm{opt},\Gamma} \lesssim \mathfrak{R}_n,$$

then $\mathbb{E}_{\mathcal{D}^{\mathrm{tr}}_\Gamma} R^\star_{P,\rho_f}(\widehat{g}_\Gamma) \lesssim \mathfrak{R}_n$.

This gives a sufficient pretraining sample-complexity bound. It also separates two uses of data. More independent training tasks can reduce the cost of selecting an in-context procedure, whereas the new function is still observed through only n examples in its own prompt. Corollary F.1 formalizes the distinction: it takes the uniform prior on the heterogeneous Assouad family, draws a fresh function index for each task, and conditions on all pretraining observations. The target index and its prompt retain their original joint law, so the same prior-average lower bound applies to every learned prediction rule. The assertion allows knowledge of the hard task law while retaining uncertainty about the new function. Optimality is therefore worst-case over task distributions. This quantifier matters for interpreting the benchmark: a fixed task distribution may be easier, as illustrated by a point mass at the zero function, for which the identically zero predictor has zero latent risk.

6 Discussion

The benchmark in (2) connects heterogeneous local difficulty to one geometry-first in-context estimator. A higher-order tangent fit, stable reduced regression, and an architecture-compatible realization attain this benchmark through a single forward pass. The resulting near-ERM guarantee separates representation from statistical learning.

Our statistical formulation motivates extending transformer analysis to other inference problems. Martinez et al. [23] establish convergence guarantees for generative learning on heterogeneous stratified spaces. A further direction is to derive sharp minimax rates for growing mixtures under suitable distributional losses, through matching aggregate upper and lower bounds that retain component dimensions, smoothness, masses, and effective sample sizes, and to investigate whether diffusion transformers attain these rates.

Acknowledgments and Disclosure of Funding

Jaehee Seo was supported by the Next Generation Scholarship for Basic Studies (Type C) from Seoul National University. This work was also supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)).

References

[1] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.

[2] Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019.

[3] Eddie Aamari, Clément Berenfeld, and Clément Levrard. Optimal reach estimation and metric learning. *The Annals of Statistics*, 51(3):1086–1108, 2023.

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

[5] Michele Allegra, Elena Facco, Francesco Denti, Alessandro Laio, and Antonietta Mira. Data segmentation based on the local intrinsic dimension. *Scientific reports*, 10(1):16449, 2020.

[6] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

[7] Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pages 177–186, 2007.

[8] Bradley CA Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Ming-Yen Cheng and Hau-tieng Wu. Local linear regression on manifolds and its geometric interpretation. *Journal of the American Statistical Association*, 108(504):1421–1434, 2013.

[11] Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. In *Forty-third International Conference on Machine Learning*, 2026.

[12] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[13] Vincent Divol. Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, 183(1):581–647, 2022.

[14] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[15] Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93 (3):418–491, 1959.

[16] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

[17] Gene H. Golub and Charles F. Van Loan. *Matrix Computations - 4th Edition*. Johns Hopkins University Press, Philadelphia, PA, 4th edition, 2013. doi: 10.1137/1.9781421407944.

[18] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

[19] Alex Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. *Advances in Neural Information Processing Systems*, 37:42162–42210, 2024.

[20] Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

[21] Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. *Advances in Neural Information Processing Systems*, 37:106667–106713, 2024.

[22] Xin Li and Anand Sarwate. Unraveling the localized latents: Learning stratified manifold structures in llm embedding space with sparse mixture-of-experts. *arXiv preprint arXiv:2502.13577*, 2025.

[23] Randy Martinez, Rong Tang, and Lizhen Lin. A deep generative approach to stratified learning. *arXiv preprint arXiv:2604.10650*, 2026.

[24] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

[25] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[26] Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874, 2024.

[27] Zhaiming Shen, Alex Havrilla, Rongjie Lai, Alexander Cloninger, and Wenjing Liao. Transformers for learning on noisy and task-level manifolds: Approximation and generalization insights. *arXiv preprint arXiv:2505.03205*, 2025.

[28] Zhaiming Shen, Alexander Hsu, Rongjie Lai, and Wenjing Liao. Understanding in-context learning on structured manifolds: Bridging attention to kernel methods. In *International Conference on Learning Representations*, volume 2026, pages 42067–42103, 2026.

[29] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[30] Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2024.

[31] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

[32] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36:39257–39276, 2023.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[35] Yi Wang and Zhiren Wang. CW complex hypothesis for image data. In *Forty-first International Conference on Machine Learning*, 2024.

[36] Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090. PMLR, 2021.

[37] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks*, 94:103–114, 2017.

[38] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Appendix

A Construction overview

Figure 2 separates the fixed input/output interface from computations inside the transformer. Its notation follows Appendices H.2–H.5. Displayed summaries are the targets of the approximate modules at query $x = X_{n+1}$. Uniform softmax attention copies and averages numerical registers; row-wise ReLU networks with residual connections compute the statistical modules.

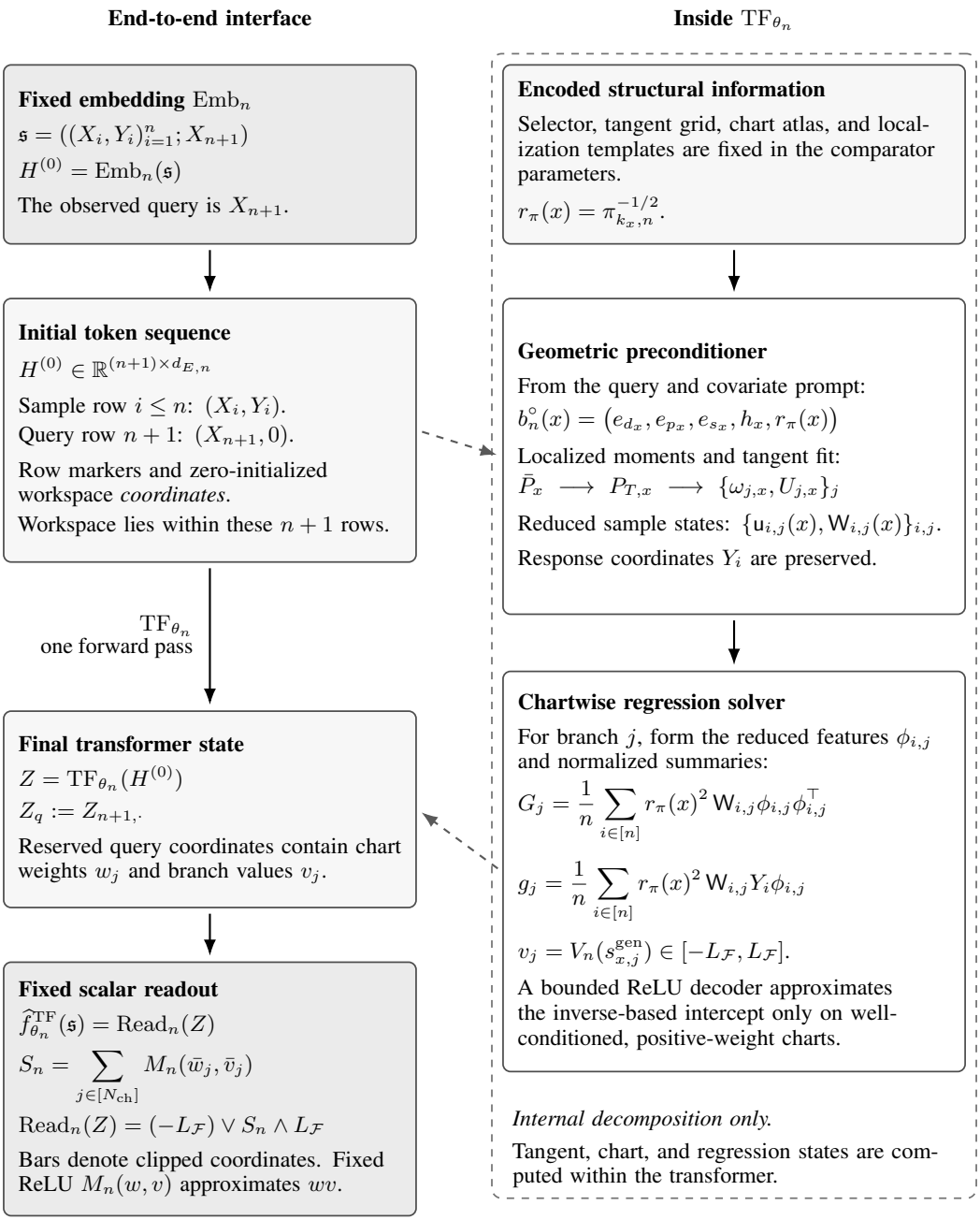


Figure 2: End-to-end interface and internal computation of the predictor in Theorem 3. Solid arrows show successive transformations; dashed arrows associate the initial and final token states with the internal modules. The sequence has $n+1$ rows and polynomially many workspace coordinates. The readout is a clipped scalar.

B Extended related work

In-context statistical procedures. Viewing the prompt as a sample from a new task separates a transformer's capacity to represent a statistical procedure from whether training finds that procedure. Empirical work on simple function classes [16], investigations of linear-model learning algorithms [4], and constructive analyses of gradient-descent-like updates and algorithm selection [6, 34] develop this perspective. Our theorems use it for nonparametric tasks, with the optimization tolerance kept separate from approximation and statistical error.

Euclidean and manifold nonparametric ICL. Kim et al. [21] analyze nonparametric in-context estimation through learned basis representations and distinguish pretraining from in-context generalization. Ching et al. [11] show that efficient transformer implementations of local-polynomial regression can exploit arbitrary positive Hölder smoothness in Euclidean covariates. Our analysis supplies usable intrinsic coordinates in an admissible growing geometric environment and controls the interaction between coordinate estimation and higher-order regression.

For related supervised regression on noisy manifolds, Shen et al. [27] analyze targets of the form $g \circ \pi_{\mathcal{M}}$ with a single manifold and Hölder-continuous g with exponent in $(0, 1]$. Their discussion identifies projection approximation as an obstacle to exploiting higher regularity. The structured-manifold ICL model of Shen et al. [28] assumes a fixed manifold, a uniform covariate distribution, and Hölder exponents in $(0, 1]$, and connects attention to kernel prediction. In contrast, our benchmark explicitly retains $(d_k, \alpha_k, n\pi_{k,n})$ for each component. A tangent fit through degree $\lceil \alpha_k \rceil$ and a regression polynomial of degree $\lceil \alpha_k \rceil - 1$ yield the geometric propagation term $h_k^2 \mathbb{E}\delta_x^2$. These comparisons concern the stated regression settings, each with its own perturbation model and structural assumptions. For Euclidean complexity, Ching et al. [11] provide sharper guarantees than our conservative polynomial resource bounds.

Nonparametric regression and geometric inference. Local polynomial regression on manifold-supported data achieves rates driven by intrinsic dimension [7]; local linear regression connects this principle to tangent-space estimation [10]. Deep estimators have also been analyzed on approximate manifolds [20]. Higher-order tangent and curvature estimation [1], and estimation of the reach [2, 3], motivate the separation between local graph regularity and the global localization scale in our model. The present construction uses these geometric ideas as part of a prompt-dependent estimator.

Representation geometry and the modeling assumption. Image-dimension studies [25] and evidence for a union-of-manifolds viewpoint [8] motivate heterogeneous low-dimensional structure. Analyses of transformer embeddings document changes in anisotropy and intrinsic dimension during training [26]. The mixture represents this local heterogeneity. Bounded tubular perturbations describe observations concentrated near an idealized latent manifold, connecting the model to the approximate-manifold viewpoint [20, 27]. A fixed feature map producing X and the affine token-packing map Emb_n have different roles: only the latter is part of the formal input interface. The covariate dimension D is fixed, whereas the workspace dimension $d_{E,n}$ of the constructed transformer may grow.

Quantitative implementation and generalization. ReLU approximation theory provides general tools for constructive networks [24, 37]; program-style attention analyses explain selection and aggregation [36]. We give the required Lipschitz-preserving approximation and residual-block compilation directly in Appendix H.1, using polynomial workspace for intermediate ReLU states. Existing covering analyses [14, 19, 30] provide relevant context, but our entropy calculation is for the specific bounded-parameter softmax class and polynomial workspace used here. Lemma H.2 implements exact copying and averaging with uniform softmax weights and bounded value gates.

C Numerical illustration

We compare a trained softmax workspace predictor with a geometry-informed local-polynomial benchmark at fixed context size and heterogeneous mixture geometry, varying the number of pretraining prompts $\Gamma \in \{100{,}000, 200{,}000, 500{,}000, 1{,}000{,}000\}$. The experiment combines heterogeneity in dimension, smoothness, sampling mass, shape, and design density. We specify the data generator, empirical architecture, optimization, bandwidth tuning, and evaluation protocol below.

C.1 Data and task generation

Heterogeneous mixture and geometry. We fix ambient dimension $D = 64$, context size $n = 40{,}000$, and $K = 12$ components, one for each pair $(d_k, \alpha_k) \in \{2, 4, 6, 8\} \times \{0.75, 1.5, 3\}$. The component probabilities and expected context counts are

	$\alpha = 0.75$	$\alpha = 1.5$	$\alpha = 3$
$d = 2$	0.02 (800)	0.03 (1,200)	0.05 (2,000)
$d = 4$	0.04 (1,600)	0.06 (2,400)	0.10 (4,000)
$d = 6$	0.06 (2,400)	0.09 (3,600)	0.15 (6,000)
$d = 8$	0.08 (3,200)	0.12 (4,800)	0.20 (8,000).

For indices $i = 0, \dots, 3$ and $j = 0, 1, 2$, write $k = 3i + j + 1$, and assign anisotropy and density tilt by $\eta_k = (0, 0.15, 0.30)_{(i+j) \bmod 3}$ and $\lambda_k = (0, 0.4, 0.7)_{(i+2j) \bmod 3}$, with tuple entries indexed from zero. Let $a_d = \mathrm{vol}(\mathbb{S}^d)$ denote the area of $\mathbb{S}^d$, and define

$$R_d = (4/a_d)^{1/d}, \qquad A_k = \mathrm{diag}\big(e^{-\eta_k + 2\eta_k \ell / d_k}\big)_{\ell=0}^{d_k}, \qquad \mathcal{M}_k = c_k + R_{d_k} Q_k A_k \mathbb{S}^{d_k}.$$

The frames $Q_k \in \mathbb{R}^{64 \times (d_k+1)}$ have orthonormal columns. With $r_{\max} = \max_k R_{d_k} e^{\eta_k}$ and $s = 2r_{\max} + 1$, the centers are the rows of $(s/\sqrt{2})(I_{12} - \mathbf{1}\mathbf{1}^\top/12)V^\top$, where $V \in \mathbb{R}^{64 \times 12}$ also has orthonormal columns. Thus the centers form a regular simplex of side s. Gaussian reduced QR factorizations, with signs fixed by the diagonal of the triangular factor, generate V and Q_k using NumPy `SeedSequence` inputs $(2026, 0)$ and $(2026, 1, k-1)$, respectively. This geometry is fixed across tasks and training budgets.

For component k, sample $U \in \mathbb{S}^{d_k}$ with density $1 + \lambda_k u_1$ relative to the uniform probability law, using rejection sampling from normalized Gaussian directions with envelope $1 + \lambda_k$, and set $X^\star = c_k + R_{d_k} Q_k A_k U$. Since $\det A_k = 1$ and $R_{d_k}^{d_k} a_{d_k} = 4$, the density relative to manifold volume is

$$\varrho_k(c_k + R_{d_k} Q_k A_k u) = \frac{1 + \lambda_k u_1}{4\|A_k^{-T} u\|}, \qquad 4e^{-\eta_k} \le \mathrm{vol}(\mathcal{M}_k) \le 4e^{\eta_k}.$$

Across the twelve components, $0.05556 < \varrho_k < 0.57370$, so $c_\mu = 1/20$ is a valid common density constant. The reach is $R_{d_k} e^{-3\eta_k}$, with minimum 0.22938; the conservative separation lower bound is 1.10143, and every latent point has Euclidean norm at most $3.15035 < 4$.

Task functions. Each task independently draws one function per component and shares it across all context and query points on that component. In sphere coordinates, it has the form

$$f_k(c_k + R_{d_k} Q_k A_k u) = b_k + \sum_{\ell \in [32]} a_{k\ell} \phi_{\alpha_k}\big(4(v_{k\ell}^\top u - t_{k\ell})\big) + \sum_{\ell \in [8]} c_{k\ell} \psi_{k\ell}(u), \qquad \phi_\alpha(z) = \frac{(z_+)^\alpha}{1 + (z_+)^\alpha}.$$

Here $b_k \sim \mathrm{Unif}[-0.15, 0.15]$, $t_{k\ell} \sim \mathrm{Unif}[-0.35, 0.35]$, and the ridge directions $v_{k\ell}$ are uniform on $\mathbb{S}^{d_k}$. Independently draw bump centers $w_{k\ell}$ uniformly on $\mathbb{S}^{d_k}$ and widths $q_{k\ell} \sim \mathrm{Unif}[1, 1.6]$; with $z_{k\ell}(u) = \|u - w_{k\ell}\|^2 / q_{k\ell}^2$, set

$$\psi_{k\ell}(u) = \begin{cases} \exp\{1 - (1 - z_{k\ell}(u))^{-1}\}, & z_{k\ell}(u) < 1, \\ 0, & z_{k\ell}(u) \ge 1. \end{cases}$$

The ridge and bump coefficient vectors have absolute sums 0.4 and 0.2, respectively. For each vector, the first two coefficients receive 45% and 25% of its absolute mass; the remaining 30% is

allocated in proportion to independent $\mathrm{Unif}[0.5, 1.5]$ draws. All coefficient signs are independent and equiprobable. Consequently $|f_k| \le 0.15 + 0.4 + 0.2 = 0.75$ deterministically. The ridge knots yield the smoothness classes $C^{0,0.75}$, $C^{1,0.5}$, and $C^{2,1}$ for the three values of α_k, while the bumps are smooth. On the fixed ellipsoid family these functions admit a common finite Hölder bound, independent of the task draw and training budget; the value bound 0.75 is distinct from that Hölder norm bound.

Observations and reproducibility. Each prompt draws $n + 1$ independent labels from the mixture and latent points from their corresponding component laws. Write $N_k = n\pi_{k,n}$ and set

$$h_k = N_k^{-1/(2\alpha_k + d_k)}, \qquad \sigma_{k,n} = \frac{h_k^{\alpha_k \vee 1}}{n}, \qquad X = X^\star + \xi, \qquad Y = f_k(X^\star) + \epsilon.$$

Conditionally on the component, ξ is uniform on the ambient 64-dimensional Euclidean ball of radius $\sigma_{k,n}$: $\xi = \sigma_{k,n} B^{1/64} G/\|G\|$ for independent $B \sim \mathrm{Unif}[0, 1]$ and $G \sim N(0, I_{64})$. The independent response noise is $\epsilon \sim \mathrm{Unif}[-0.1\sqrt{3}, 0.1\sqrt{3}]$, with conditional mean zero and variance 0.01. Evaluation targets the latent query value $f_k(X^\star)$.

Task randomness uses NumPy's `SeedSequence` with the ordered input

$$(23619, \text{split}, \text{task index}, \text{stream}, \text{component}).$$

Split identifiers $0, 1, 2$ denote training, validation, and testing; stream identifiers $0, 1, 2, 3, 4$ generate labels, latent points, task functions, covariate noise, and response noise, respectively. The component entry is $k - 1$ for component-specific streams and zero for the label and response streams. The 2,000 test prompts have indices 3,000,017 through 3,002,016 and are shared across methods and budgets. The fixed bounded geometry and shrinking perturbation rule reflect the asymptotic model; the full finite-sample constant conditions of Assumption 3 have not been certified at $n = 40{,}000$.

C.2 Empirical architecture and optimization

Softmax workspace predictor. The learned predictor has width 64, four workspace blocks, four attention heads per module, 32 learned workspace tokens, and FFN width 256. Each block applies workspace-to-context cross-attention, workspace self-attention, and a GELU FFN, with residual connections. Layer normalization and dropout are disabled. For the observed query $x = X_{n+1}$, the implemented preprocessing is

$$\widetilde{X}_i = X_i/6, \qquad \widetilde{x} = x/6, \qquad r_i = \widetilde{X}_i - \widetilde{x}, \qquad t_i = \|r_i\|_2^2.$$

A learned affine–GELU–affine encoder maps the 130-dimensional vector $(\widetilde{X}_i, r_i, Y_i, t_i)$ through width 64 to a context token. A separate encoder of the same form maps $\widetilde{x} \in \mathbb{R}^{64}$ to a width-64 vector added to every learned workspace token. Each attention head uses scaled dot-product softmax weights $\mathrm{softmax}(q_a^\top k_j/\sqrt{16})$. The context tokens remain fixed across the four blocks; attention from the 32 workspace tokens accesses all 40,000 context tokens through a $32 \times 40{,}000$ score matrix per head. An affine readout of the first final workspace token followed by $\tanh$ produces a prediction in $[-1, 1]$. The model has 286,401 trainable parameters. Its learned encoders and GELU blocks define the empirical model studied here. The theoretical construction uses the residual-ReLU architecture in Section 4.

Training and checkpoint selection. For each of the four budgets, we train the model from scratch with training seed 0. Training traverses the indexed task pool $\{0, \ldots, \Gamma - 1\}$ exactly once, using a permutation with seed 101. Prompts are generated on demand. Gradient accumulation over 25 microbatches of one prompt gives an effective batch size of 25; at most one additional microbatch is prefetched. The completed runs have

Γ	100,000	200,000	500,000	1,000,000
optimizer updates	4,000	8,000	20,000	40,000
final effective batch size	25	25	25	25

AdamW uses initial learning rate 3×10^{-4} and weight decay 10^{-4}. The learning rate follows cosine annealing to zero over each run's prescribed update budget, and the accumulated gradient norm is clipped at 1. Training minimizes squared error against the noisy query response Y_{n+1}, and validation

evaluates the same loss using the 96 fixed validation-split tasks with indices 0–95, evaluated after the first update, every 250 updates, and at the final update. All four runs complete the full schedule. Test evaluation uses the final `last.pt` checkpoint after all prescribed updates. Training and transformer inference use float32 on the PyTorch MPS backend. Increasing Γ therefore changes both the number of distinct training prompts and the amount of optimization, with the cosine schedule stretched accordingly.

C.3 Geometry-informed local-polynomial benchmark

LP-CV receives the query component's dimension d_k, smoothness α_k, mixture mass $\pi_{k,n}$, and true tangent basis $U_\star$ at the latent query $X^\star_{n+1}$. For a candidate multiplier c, set $b = c(n\pi_{k,n})^{-1/(2\alpha_k+d_k)}$ and $u_i = U_\star^\top(X_i - X_{n+1})/b$. The total polynomial degree is $p_k = \lceil \alpha_k \rceil - 1 \in \{0, 1, 2\}$, with $q_k = \binom{d_k+p_k}{p_k}$ features $\phi_\nu(u) = u^\nu/\nu!$, $|\nu| \le p_k$. The weighted fit uses all observed context responses, with

$$w_i = \mathsf{A}\left(\frac{\|X_i - X_{n+1}\|_1}{b}\right)(1 - \|u_i\|_1)_+, \qquad \mathsf{A}(t) = 1 \wedge \{0 \vee (2 - 2t)\}.$$

Thus both ambient and tangent-coordinate cutoffs use the ℓ_1 norm, and the ambient cutoff scale is 1. Common positive factors such as b^{-d_k}/n cancel from the normal equations. Before solving, positive weights are divided by their maximum. The implementation uses float64 normal equations and an eigendecomposition, retaining Gram eigenvalues greater than 10^{-8} times the largest eigenvalue. The resulting pseudoinverse fit supplies the intercept, clipped to $[-1, 1]$. If no point has positive weight, the fallback is the clipped response of the nearest observed context point in Euclidean distance. Context component labels supply contamination diagnostics; all context points enter the weight rule and the nearest-point fallback. This benchmark uses the true latent-query tangent while centering at the observed query; it differs from the estimated-tangent construction in Algorithm 1.

Bandwidth selection uses the common grid $\{2, 4, 6, 8, 10, 12, 16, 24\}$ separately for all 12 (d_k, α_k) strata. For stratum $j = 0, \ldots, 11$, ordered by dimension and then smoothness, 100 validation prompts have task indices $2{,}000{,}003 + 100j$ through $2{,}000{,}102 + 100j$, with the query component fixed to that stratum. These 1,200 tuning prompts are disjoint from the 96 training-validation prompts and the test split. A multiplier is admissible if its mean positive-weight count is at least $2q_k$ and its Gram matrix has full numerical rank on at least 90% of the stratum's tuning prompts. Among admissible candidates, LP-CV minimizes mean squared error against the independent noisy query response, breaking ties toward the smaller multiplier. Every stratum has an admissible candidate; the selected multipliers, fixed for all four model comparisons, are

	$\alpha = 0.75$	$\alpha = 1.5$	$\alpha = 3$
$d = 2$	16	10	6
$d = 4$	10	8	6
$d = 6$	10	12	10
$d = 8$	12	10	16

C.4 Test protocol and uncertainty

All four final checkpoints and LP-CV use the same 2,000 test-split prompts, indexed 3,000,017 through 3,002,016. Test query components follow the specified mixture probabilities. Transformer inference processes one prompt at a time. The test target is the latent regression value $f(X^\star_{n+1})$; training and tuning use the noisy query response. The saved per-prompt input/target hashes agree across all four evaluations and the LP-CV reference; the error arithmetic and intervals have been checked by recomputing them from the saved predictions.

Reported 95% intervals are nonparametric percentile-bootstrap intervals using 5,000 replicates. Each replicate samples 2,000 test-prompt indices with replacement; its mean squared error gives the bootstrap risk, and the 2.5% and 97.5% quantiles give the interval. The transformer and LP-CV risk intervals use seeds 3,000,017 and 3,000,018, respectively. For the transformer/LP-CV risk ratio, seed 3,000,019 generates paired resamples: the same indices are used for the numerator and denominator, and each replicate takes the ratio of their mean squared errors. These intervals describe test-task variation conditional on the fitted models and selected LP bandwidths; they do not measure training-seed or bandwidth-selection uncertainty.

C.5 Results and interpretation

Table 1: Latent-target test MSE in units of 10^{-3} and risk relative to geometry-informed LP-CV on the same 2,000 prompts. Brackets give 95% percentile-bootstrap intervals; risk-ratio intervals use paired resampling. LP-CV has MSE $0.690\ [0.645, 0.739]$ in the same units at every budget.

Pretraining budget Γ	Softmax workspace MSE	Transformer / LP-CV
100,000	14.455	20.948
	$[13.619, 15.342]$	$[19.177, 22.776]$
200,000	0.919	1.332
	$[0.849, 0.995]$	$[1.222, 1.461]$
500,000	0.538	0.780
	$[0.503, 0.575]$	$[0.727, 0.837]$
1,000,000	0.505	0.733
	$[0.472, 0.540]$	$[0.686, 0.782]$

The test MSE decreases across the four budgets. At $\Gamma = 500{,}000$, the transformer has 21.98% lower MSE than LP-CV and 41.45% lower MSE than the $\Gamma = 200{,}000$ model. Its paired risk-ratio interval $[0.727, 0.837]$ lies below one. At $\Gamma = 1{,}000{,}000$, the MSE is 0.505×10^{-3}, which is 26.75% lower than LP-CV and 6.10% lower than the $\Gamma = 500{,}000$ model. The paired risk ratio is 0.733 with interval $[0.686, 0.782]$, again below one. These comparisons quantify test-prompt uncertainty for the fitted seed-0 models.

Table 2: Component-wise point estimates at $\Gamma = 1{,}000{,}000$. MSEs are in units of 10^{-3}; n_{test} is the number of queries from that component among the 2,000 shared test prompts. Ratios use unrounded MSEs.

d_k	α_k	n_{test}	LP-CV MSE	Transformer MSE	Ratio
2	0.75	41	0.606	0.448	0.739
2	1.5	60	0.376	0.310	0.825
2	3	97	0.545	0.480	0.880
4	0.75	72	0.435	0.371	0.853
4	1.5	106	0.632	0.538	0.851
4	3	217	0.845	0.418	0.495
6	0.75	117	0.692	0.474	0.684
6	1.5	175	0.882	0.404	0.458
6	3	317	0.605	0.546	0.903
8	0.75	159	0.667	0.515	0.772
8	1.5	252	0.648	0.433	0.667
8	3	387	0.780	0.679	0.871

Table 2 gives the component means. At $\Gamma = 1{,}000{,}000$, the transformer has a lower MSE point estimate than LP-CV in all twelve components. These component-wise comparisons are descriptive: query counts range from 41 to 387, and no component-wise confidence intervals are reported.

This comparison measures the combined effect of increasing the number of distinct pretraining prompts and the optimizer update budget. It illustrates learned prediction at a fixed context size and heterogeneous geometry. The theoretical guarantees concern the constructed residual-ReLU architecture and its asymptotic scaling; the empirical results concern the specified workspace architecture and task distribution.

Code is available at https://github.com/seojaehee02/nonparametric_ICL.

D Mathematical preliminaries

This section fixes the notation and basic objects used throughout the paper. We collect conventions for asymptotic notation, polynomial features, matrix norms, localization, and the Hölder and manifold regularity classes. The transformer class is defined in Section 4.

D.1 Notations

For an event A, $\mathbb{I}\{A\}$, also written $\mathbb{I}_A$, denotes its indicator. The vector $\mathbf{1}_N \in \mathbb{R}^N$ has all entries equal to one. Let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Write $[m] := \{1, \dots, m\}$ for $m \in \mathbb{N}$, and $[m]_0 := \{0, \dots, m\}$ for $m \in \mathbb{N}_0$. For $a \in \mathbb{R}$, let $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the floor and ceiling of a. For $a, b \in \mathbb{R}$, write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. For non-negative sequences a_n, b_n, write $a_n \lesssim b_n$ if $a_n \le C b_n$ for a constant $C > 0$ independent of n. Write $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. For a metric space $(\mathcal{X}, d)$, $x \in \mathcal{X}$, and $r > 0$, let $B_{\mathcal{X}}(x, r) := \{x' \in \mathcal{X} : d(x, x') < r\}$. For a set $A \subset \mathcal{X}$, write $d(x, A) := \inf_{a \in A} d(x, a)$ and $d(A, B) := \inf_{x \in A} d(x, B)$. Here $d(A, B)$ denotes the separation between A and B.

Constants. Throughout the Appendix, $c, C > 0$ denote generic constants whose values may change between occurrences. Unless a different dependence is specified, they depend only on the fixed model bounds in Section 2, κ, and fixed kernel, cutoff, and construction choices. All bounds are uniform in n, K_n, k, P, f and admissible component sequences; varying masses, reach, and separation enter through their displayed scales. Additional dependence on fixed parameters of auxiliary statements or on approximation and architecture parameters is stated locally. Named thresholds and exponents retain their fixed values throughout each argument.

For a multi-index $\nu = (\nu_1, \dots, \nu_q) \in \mathbb{N}_0^q$, define $|\nu| := \sum_{j \in [q]} \nu_j$, $\nu! := \prod_{j \in [q]} \nu_j!$, and $u^\nu := \prod_{j \in [q]} u_j^{\nu_j}$. For $p \ge 0$, set $\mathcal{I}_{q,p} := \{\nu \in \mathbb{N}_0^q : |\nu| \le p\}$, $q_{q,p} := |\mathcal{I}_{q,p}|$, and $\Phi_{q,p}(u) := \left(\frac{u^\nu}{\nu!}\right)_{\nu \in \mathcal{I}_{q,p}}$. The zero-multi-index coordinate vector is denoted by $e_{0,q,p}$, or simply e_0.

For a linear map A, define $\|A\|_{\mathrm{op}} := \sup_{\|v\|=1} \|Av\|$. For a matrix A, $\|A\|_F$ denotes its Frobenius norm and $\|A\|_{\max} := \max_{i,j} |A_{ij}|$. For a symmetric positive semidefinite matrix G, let $\lambda_{\min}^+(G)$ and $\lambda_{\max}^+(G)$ denote its smallest and largest positive eigenvalues. We write $G^\dagger$ for the Moore–Penrose pseudoinverse and $\mathrm{Im}(G)$ for the image of G. For symmetric matrices, $A \preceq B$ means that $B - A$ is positive semidefinite. In an orthonormal eigendecomposition $G = \sum_j \lambda_j u_j u_j^\top$, the convention is

$$G^\dagger = \sum_{j : \lambda_j > 0} \lambda_j^{-1} u_j u_j^\top, \qquad G^\dagger G = G G^\dagger = \sum_{j : \lambda_j > 0} u_j u_j^\top.$$

Thus $G^\dagger G$ is the orthogonal projector onto $\mathrm{Im}(G)$, including when G is singular.

We use the kernel $\mathsf{K}(v) := (1 - \|v\|_1)_+$. The cutoff A is the piecewise-linear function given by $\mathsf{A}(t) = 1$ for $0 \le t < 1/2$, $\mathsf{A}(t) = 2 - 2t$ for $1/2 \le t < 1$, and $\mathsf{A}(t) = 0$ for $t \ge 1$.

Bernstein inequality used in the proofs. Let $U_1, \dots, U_m$ be independent centered real random variables with $|U_i| \le b$ almost surely and $\sum_i \mathbb{E} U_i^2 \le v$, where $b > 0$ and $v \ge 0$. For every $t > 0$, the bounded-variable Bernstein inequality gives

$$\mathbb{P}\left\{ \left| \sum_{i=1}^m U_i \right| \ge \sqrt{2vt} + \frac{2bt}{3} \right\} \le 2e^{-t}. \tag{D.B}$$

We apply the same statement conditionally on a sigma-field when the variables are conditionally independent, conditionally centered, and obey the displayed bounds conditionally. The resulting probability inequality then holds almost surely for the conditional law.

D.2 Hölder classes and manifolds

Hölder class on Euclidean domains. Let $U \subset \mathbb{R}^D$ be open. For $\gamma \in (0, 1]$, a function $g : U \to \mathbb{R}$ is called γ-*Hölder continuous* if

$$[g]_{C^{0,\gamma}(U)} := \sup_{\substack{u, v \in U \\ u \ne v}} \frac{|g(u) - g(v)|}{\|u - v\|^\gamma} < \infty.$$

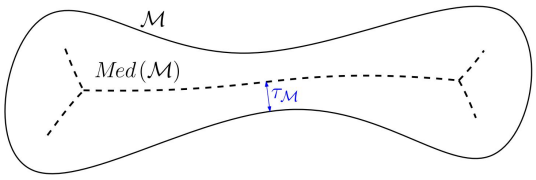


Figure 3: The medial axis $Med(\mathcal{M})$ consists of points having multiple nearest points on $\mathcal{M}$. The shortest distance from $\mathcal{M}$ to this axis is $\mathrm{reach}(\mathcal{M})$. A narrow bottleneck brings the medial axis closer to $\mathcal{M}$ and therefore reduces the reach.

Let $s \in \mathbb{N}_0$. A scalar-valued map $g : U \to \mathbb{R}$ is called $C^{s,\gamma}$ if it has continuous derivatives up to order s, all derivatives of order at most s are bounded, and every derivative of order s is γ-Hölder continuous. Equivalently, its $C^{s,\gamma}$-norm

$$\|g\|_{C^{s,\gamma}(U)} := \max_{|\mu|\le s} \|D^\mu g\|_\infty + \max_{|\mu|=s} [D^\mu g]_{C^{0,\gamma}(U)}$$

is finite, where $[D^\mu g]_{C^{0,\gamma}(U)} := \sup_{\substack{u,v\in U \\ u\neq v}} \frac{|D^\mu g(u) - D^\mu g(v)|}{\|u-v\|^\gamma}$. For vector-valued maps, the same definition is used component-wise, equivalently with Euclidean norms and multilinear operator norms. For $\alpha > 0$, put $s_\alpha := \lceil \alpha \rceil - 1$ and $\gamma_\alpha := \alpha - s_\alpha \in (0, 1]$. We define the α-Hölder class by $\mathcal{H}^\alpha(U) := C^{s_\alpha,\gamma_\alpha}(U)$ with $\|g\|_{\mathcal{H}^\alpha(U)} := \|g\|_{C^{s_\alpha,\gamma_\alpha}(U)}$.

Manifolds. A d-dimensional embedded C^{s_β,γ_β} submanifold of $\mathbb{R}^D$ is a set that can be represented locally, after a rigid change of coordinates, as the graph of a C^{s_β,γ_β} map from $\mathbb{R}^d$ to $\mathbb{R}^{D-d}$. We denote the tangent space and the normal space at $x \in \mathcal{M}$ as $T_x\mathcal{M}$ and $T_x\mathcal{M}^\perp$. $\Pi_{T_x\mathcal{M}}$ is the nearest-point projection map onto $T_x\mathcal{M}$. Also, let $d_{\mathcal{M}}$ denote the intrinsic length metric on $\mathcal{M}$. Standard terminology about reach is in the sense of [15].

We recall the reach of a closed set $\mathcal{M}$, first introduced in Federer [15]. The medial axis of $\mathcal{M}$ is defined as

$$Med(\mathcal{M}) := \left\{ y \in \mathbb{R}^D : \exists x_1 \neq x_2 \in \mathcal{M} \quad \text{s.t.} \quad \|x_1 - y\| = \|x_2 - y\| = d(y, \mathcal{M}) \right\}.$$

Thereby, the medial axis of A is the set of points whose nearest-point projection onto $\mathcal{M}$ is not uniquely defined. The *reach* of $\mathcal{M}$ is defined as

$$\mathrm{reach}(\mathcal{M}) := \tau_{\mathcal{M}} = \inf_{x\in\mathcal{M},\ y\in Med(\mathcal{M})} \|x - y\|.$$

Figure 3 illustrates the medial axis and the reach of the closed set.

We now define the class of regular manifolds.

Definition D.1 (Regular embedded manifold class)**.** Fix $D \ge 2$. For $d \in [D-1]$, $\beta \ge 2$, $L \ge 1$, and $\tau > 0$, let $r_L := 1/(4L)$. We write $\mathcal{M} \in \mathcal{C}^\beta_{d,L,\tau}(\mathbb{R}^D)$ if $\mathcal{M} \subset \mathbb{R}^D$ is a closed, connected, embedded d-dimensional C^{s_β,γ_β} submanifold without boundary satisfying

$$\mathrm{reach}(\mathcal{M}) \ge \tau,$$

such that, for every $x \in \mathcal{M}$, there is a tangent-normal chart

$$\Psi_x(v) = x + v + N_x(v), \qquad v \in B_{T_x\mathcal{M}}(0, r_L),$$

parametrizing a relatively open neighborhood of x in $\mathcal{M}$. Here

$$N_x : B_{T_x\mathcal{M}}(0, r_L) \to T_x\mathcal{M}^\perp, \qquad N_x(0) = 0, \qquad D_0 N_x = 0.$$

The graph maps have uniformly bounded higher-order regularity:

$$\max_{2\le j\le s_\beta} \sup_{v\in B_{T_x\mathcal{M}}(0,r_L)} \|D^j N_x(v)\|_{\mathrm{op}} \le L, \qquad [D^{s_\beta} N_x]_{C^{0,\gamma_\beta}(B_{T_x\mathcal{M}}(0,r_L))} \le L,$$

where the first maximum is omitted when $s_\beta = 1$. The chart radius r_L and regularity bound L do not depend on τ.

Note that for a given $\mathcal{M} \in \mathcal{C}^2_{d,L,\tau}(\mathbb{R}^D)$, $\mathrm{reach}(\mathcal{M}) = \inf_{\substack{x,x' \in \mathcal{M} \\ x \neq x'}} \frac{\|x-x'\|^2}{2d(x', x+T_x\mathcal{M})}$ holds (see [15, Theorem 4.18]). Positive reach is a standard regularity condition in geometric inference [1–3]: Classical formulations often impose a fixed reach lower bound. The following lemma characterizes the local Euclidean behavior with positive reach. The proof is given in Appendix E.1.

Lemma D.1. *Let $\mathcal{M} \in \mathcal{C}^\beta_{d,L,\tau}(\mathbb{R}^D)$, where $\tau > 0$. There exists $c_{\mathrm{geo}} \in (0,1)$, depending only on D, d, L, β, such that the following holds. For every $x \in \mathcal{M}$, set $\rho := c_{\mathrm{geo}}(\tau \wedge r_L)$. Then the following local estimates hold.*

(i) For every $r \in (0, \rho]$, $\mathrm{vol}_{\mathcal{M}}\big(\mathcal{M} \cap B_{\mathbb{R}^D}(x,r)\big) \asymp r^d$, with comparison constants depending only on d.

(ii) For all $y, z \in \mathcal{M} \cap B_{\mathbb{R}^D}(x,\rho)$, $\|y-z\| \le d_{\mathcal{M}}(y,z) \le 2\|y-z\|$.

(iii) For all $y \in \mathcal{M} \cap B_{\mathbb{R}^D}(x,\rho)$, $\|\Pi_{T_y\mathcal{M}} - \Pi_{T_x\mathcal{M}}\|_{\mathrm{op}} \le L\|y-x\|$.

Hölder class on manifolds. We next recall Hölder class defined on regular embedded manifolds by pulling them back to tangent-normal charts. Since the admissible charts have uniform C^β bounds and fixed radius, this gives the same smoothness class, up to equivalent norms, as any other atlas with comparable uniform regularity constants. Thus the definition is stable under admissible choices of uniformly regular charts.

Definition D.2 (Hölder functions on regular manifolds)**.** Let $\mathcal{M} \in \mathcal{C}^\beta_{d,L,\tau}(\mathbb{R}^D)$, and let $\Psi_x : B_{T_x\mathcal{M}}(0, r_L) \to \mathcal{M}$ be the tangent-normal chart from Definition D.1. For $\alpha > 0$ and $L_{\mathcal{F}} > 0$ with $\beta \ge \lceil \alpha \rceil + 1$, define

$$\mathcal{H}^\alpha(\mathcal{M}; L_{\mathcal{F}}) := \left\{ f : \mathcal{M} \to \mathbb{R} : \sup_{x \in \mathcal{M}} \|f \circ \Psi_x\|_{\mathcal{H}^\alpha(B_{T_x\mathcal{M}}(0,r_L))} \le L_{\mathcal{F}} \right\}.$$

Here $T_x\mathcal{M}$ is regarded as a Euclidean space with the inherited inner product.

E Auxiliary details

This appendix collects technical details for the geometric assumptions used in the main text. We first prove the local consequences of positive reach stated in Lemma D.1. We then give a constant-explicit version of Assumption 3 and record its elementary consequences for effective sample sizes, local bandwidths, and the growth of K_n.

E.1 Proof of Lemma D.1

Proof. Class membership gives $\mathrm{reach}(\mathcal{M}) \ge \tau$. Fix $x \in \mathcal{M}$, and let $r_L := 1/(4L)$. Let

$$\Psi_x(u) = x + u + N_x(u), \qquad u \in B_{T_x\mathcal{M}}(0, r_L),$$

be the tangent-normal chart from Definition D.1. Choose $c_{\mathrm{geo}} \in (0, 1/4]$ sufficiently small and put $\rho = c_{\mathrm{geo}}(\tau \wedge r_L)$. We first verify

$$\mathcal{M} \cap B_{\mathbb{R}^D}(x,\rho) \subset \Psi_x\big(B_{T_x\mathcal{M}}(0,\rho)\big).$$

For y in the left-hand side, set $u = \Pi_{T_x\mathcal{M}}(y-x)$, $q = \Psi_x(u)$, and $w = y - q$. Chart regularity gives $\|DN_x(u)\|_{\mathrm{op}} \le L\|u\| \le 1/4$ and $\|N_x(u)\| \le L\|u\|^2/2$, so $\|w\| < 3\rho < \tau$. Moreover, $w \in T_x\mathcal{M}^\perp$, whereas $T_q\mathcal{M}$ is the graph of $DN_x(u)$. Hence $d(w, T_q\mathcal{M}) \ge \|w\|/2$. The reach inequality [15, Theorem 4.18] yields

$$\frac{\|w\|}{2} \le d(w, T_q\mathcal{M}) \le \frac{\|w\|^2}{2\tau}.$$

Since $\|w\| < \tau$, this forces $w = 0$, proving the inclusion.

In the rest of the proof, write

$$G_x := N_x\big|_{B_{T_x\mathcal{M}}(0,\rho)}, \qquad F_x(u) := x + u + G_x(u).$$

Boundary values of G_x and its derivatives are inherited from N_x, which is defined on the larger ball $B_{T_x\mathcal{M}}(0, r_L)$. By the chart regularity and $DG_x(0) = 0$,

$$\|DG_x(u)\|_{\mathrm{op}} = \|DN_x(u) - DN_x(0)\|_{\mathrm{op}} \le L\|u\|, \qquad \|u\| \le \rho.$$

Since $\rho \le c_{\mathrm{geo}} r_L$ and $r_L = (4L)^{-1}$,

$$\sup_{\|u\|\le\rho} \|DG_x(u)\|_{\mathrm{op}} \le L\rho \le \frac{c_{\mathrm{geo}}}{4} < \frac{1}{4}. \tag{8}$$

We first prove the volume estimate. Write $\omega_d := \mathrm{vol}_{\mathbb{R}^d}(B_{\mathbb{R}^d}(0,1))$. The Jacobian of F_x is

$$J_x(u) = \sqrt{\det\{I + DG_x(u)^\top DG_x(u)\}}.$$

By (8),

$$1 \le J_x(u) \le \left(1 + \frac{1}{16}\right)^{d/2} = \left(\frac{17}{16}\right)^{d/2}. \tag{9}$$

If $F_x(u) \in \mathcal{M} \cap B_{\mathbb{R}^D}(x, r)$, then

$$\|u\| = \|\Pi_{T_x\mathcal{M}}(F_x(u) - x)\| \le \|F_x(u) - x\| < r.$$

Thus

$$\mathcal{M} \cap B_{\mathbb{R}^D}(x, r) \subset F_x\big(B_{T_x\mathcal{M}}(0, r)\big).$$

Using (9),

$$\mathrm{vol}_{\mathcal{M}}\big(\mathcal{M} \cap B_{\mathbb{R}^D}(x, r)\big) \le \omega_d \left(\frac{17}{16}\right)^{d/2} r^d.$$

For the lower bound, if $\|u\| < r/2$, then

$$\|G_x(u)\| \le \int_0^1 \|DG_x(tu)\|_{\mathrm{op}}\|u\|\, dt \le \frac{L}{2}\|u\|^2.$$

Hence, since $r \le \rho$ and $L\rho \le 1/4$,

$$\|F_x(u) - x\| \le \|u\| + \|G_x(u)\| \le \frac{r}{2} + \frac{L}{2}\left(\frac{r}{2}\right)^2 \le \frac{r}{2} + \frac{r}{32} < r.$$

Therefore

$$F_x\big(B_{T_x\mathcal{M}}(0, r/2)\big) \subset \mathcal{M} \cap B_{\mathbb{R}^D}(x, r).$$

Again using (9),

$$\mathrm{vol}_{\mathcal{M}}\big(\mathcal{M} \cap B_{\mathbb{R}^D}(x, r)\big) \ge \omega_d \left(\frac{r}{2}\right)^d.$$

This proves (i). Let

$$y = F_x(u), \qquad z = F_x(v), \qquad u, v \in B_{T_x\mathcal{M}}(0, \rho).$$

The ambient distance is always bounded above by the intrinsic distance on the manifold $\mathcal{M}$:

$$\|y - z\| \le d_{\mathcal{M}}(y, z).$$

For the reverse inequality, consider the lifted line segment

$$\gamma(t) := F_x((1-t)u + tv), \qquad t \in [0, 1].$$

By (8),

$$\mathrm{length}(\gamma) \le \int_0^1 \sqrt{1 + \|DG_x((1-t)u + tv)\|_{\mathrm{op}}^2}\, \|u - v\|\, dt \le \left(\frac{17}{16}\right)^{1/2} \|u - v\|.$$

Moreover, $u - v = \Pi_{T_x\mathcal{M}}(y - z)$ implies $\|u - v\| \le \|y - z\|$. Hence

$$d_{\mathcal{M}}(y, z) \le \mathrm{length}(\gamma) \le \left(\frac{17}{16}\right)^{1/2} \|y - z\| \le 2\|y - z\|.$$

This proves (ii). Write $y = F_x(u)$. Then $T_y\mathcal{M}$ is the graph of

$$DG_x(u) : T_x\mathcal{M} \to T_x\mathcal{M}^\perp.$$

For a linear map $A : T_x\mathcal{M} \to T_x\mathcal{M}^\perp$ with $\|A\|_{\mathrm{op}} \le 1$, let $\mathrm{graph}(A) := \{v + Av : v \in T_x\mathcal{M}\} \subset \mathbb{R}^D$ and $\theta_{\max}$ is the largest canonical angle between $\mathrm{graph}(A)$ and $T_x\mathcal{M}$. Then we have

$$\|\Pi_{\mathrm{graph}(A)} - \Pi_{T_x\mathcal{M}}\|_{\mathrm{op}} = \sin\theta_{\max} = \sup_{v \in T_x\mathcal{M}\setminus\{0\}} \frac{\|Av\|}{\|v + Av\|} = \frac{\|A\|_{\mathrm{op}}}{\sqrt{1 + \|A\|_{\mathrm{op}}^2}} \le \|A\|_{\mathrm{op}},$$

where the first equality comes from linear algebra; see [17]. Using this with $A = DG_x(u)$, and then using the chart regularity,

$$\|\Pi_{T_y\mathcal{M}} - \Pi_{T_x\mathcal{M}}\|_{\mathrm{op}} \le \|DG_x(u)\|_{\mathrm{op}} \le L\|u\| \le L\|y - x\|.$$

This proves (iii). □

E.2 Constant-explicit version of Assumption 3

We use the fixed-n aliases of the main text and retain explicit n indices in $\tau_{0,n}$, $\delta_{0,n}$, $\pi_{k,n}$, $\sigma_{k,n}$, $\mu_{k,n}$, and $\varrho_{k,n}$. Write $\sigma_{\max,n} := \max_{k\in[K_n]} \sigma_{k,n}$ and $R_{\mathrm{loc}} = R_{\mathrm{loc},n}$. Recall $N_k = n\pi_{k,n}$, $h_k = h_{k,n} = N_k^{-1/(2\alpha_k + d_k)}$, and $r_n = \max_k h_k$. Fix the exponent $\kappa \in (0,1)$ from Assumption 3. Set

$$r_{L_\mathcal{M}} = (4L_\mathcal{M})^{-1}, \qquad R_{\mathrm{loc}} = c_{\mathrm{geo}}(\tau_{0,n} \wedge r_{L_\mathcal{M}} \wedge 1), \qquad a_{\mathrm{sc}} = \frac{c_{\mathrm{geo}}}{64\sqrt{D}}.$$

Choose $c_{\mathrm{geo}} \in (0,1)$ sufficiently small that the local estimates of Lemma D.1 hold up to radius $2R_{\mathrm{loc}}$ and the fixed small-perturbation requirements in Appendix G hold. In particular, the requirements involving $2a_{\mathrm{sc}}$ in the regression design proof must hold. Such a choice is possible by reducing the uniform localization radius and then imposing finitely many smallness conditions, all depending only on fixed model bounds.

Assumption E.1 (Formal restatement of Assumption 3)**.** *For all sufficiently large n, uniformly over $k \in [K_n]$,*

$$N_k \ge n^\kappa, \qquad \tau_{0,n} \wedge \delta_{0,n} \ge a_{\mathrm{sc}}^{-1} r_n, \qquad \sigma_{k,n} \le a_{\mathrm{sc}} h_k^{\alpha_k \vee 1}.$$

Define

$$\rho_h := \frac{\kappa}{2\alpha_{\max} + d_{\max}}, \qquad \bar\rho_h := \frac{1}{2\alpha_{\min} + 1}, \qquad \rho_{\tan} := \frac{2\kappa\alpha_{\min}}{2\alpha_{\min} + d_{\max}}.$$

Lemma E.1. *Under Assumption E.1, for all sufficiently large n, uniformly over k,*

$$N_k \ge n^\kappa, \qquad K_n \le n^{1-\kappa}, \qquad n^{-\bar\rho_h} \le h_k \le r_n \le n^{-\rho_h},$$
$$n\pi_{k,n}h_k^{d_k} \ge n^{\rho_{\tan}}, \qquad (n\pi_{k,n}h_k^{d_k})^{-1} = h_k^{2\alpha_k},$$
$$h_k \le R_{\mathrm{loc}}, \qquad r_n + 2\sigma_{\max,n} \le \delta_{0,n}/2, \qquad \frac{2\sigma_{k,n}}{h_k} \le 2a_{\mathrm{sc}} \le \frac{1}{4\sqrt{D}}.$$

Proof. Since $\sum_{k\in[K_n]} N_k = n$ and $n^\kappa \le N_k \le n$,

$$K_n n^\kappa \le n, \qquad n^{-1/(2\alpha_{\min}+1)} \le N_k^{-1/(2\alpha_k + d_k)} \le n^{-\kappa/(2\alpha_{\max}+d_{\max})}.$$

Furthermore,

$$N_k h_k^{d_k} = N_k^{2\alpha_k/(2\alpha_k+d_k)} \ge n^{\rho_{\tan}}, \qquad (N_k h_k^{d_k})^{-1} = h_k^{2\alpha_k}.$$

For large n, $h_k \le r_n \le 1$. Hence

$$\sigma_{k,n}/h_k \le a_{\mathrm{sc}}, \qquad \sigma_{\max,n} \le a_{\mathrm{sc}} r_n, \qquad r_n + 2\sigma_{\max,n} \le (1 + 2a_{\mathrm{sc}})a_{\mathrm{sc}}\delta_{0,n} \le \delta_{0,n}/2.$$

Finally, $h_k \le r_n \le a_{\mathrm{sc}}\tau_{0,n} \le c_{\mathrm{geo}}\tau_{0,n}$, while $r_n \to 0$ ensures $r_n \le c_{\mathrm{geo}}(r_{L_\mathcal{M}} \wedge 1)$. Together these give $h_k \le R_{\mathrm{loc}}$. □

F Minimax lower bound

F.1 Submodel for lower bound

It suffices to prove the lower bound on a submodel, so we take $\xi \equiv 0$. Fix a symmetric nonnegative $r_0 \in C_c^\infty((-1,1))$ with $\int r_0^2 = 1$, and set $q_0 := r_0^2$. Then q_0 is a symmetric density supported on $(-1,1)$, and

$$\int_{\mathbb{R}} \left\{ \frac{d}{du} \sqrt{q_0(u)} \right\}^2 du = \int_{\mathbb{R}} r_0'(u)^2 du < \infty.$$

Let

$$\sigma_{\mathrm{lb}} := (B_\epsilon \wedge \sigma_Y)/2, \qquad q_{\mathrm{lb}}(u) := \sigma_{\mathrm{lb}}^{-1} q_0(u/\sigma_{\mathrm{lb}}).$$

In the submodel, ϵ has density q_{lb} and is independent of $X^\star$. Since q_0 is symmetric and supported on $(-1,1)$, we have

$$\mathbb{E}[\epsilon] = 0, \qquad |\epsilon| \le \sigma_{\mathrm{lb}} \le B_\epsilon, \text{ a.s.}, \qquad \mathbb{E}[\epsilon^2] \le \sigma_{\mathrm{lb}}^2 \le \sigma_Y^2.$$

Thus this submodel satisfies the noise conditions. Under Assumption 1 and non-empty condition of $\mathcal{P}_\star$,

$$\inf_k \operatorname{reach}(\mathcal{M}_k) \ge \tau_{0,n}, \qquad \inf_{k \ne \ell} d(\mathcal{M}_k, \mathcal{M}_\ell) \ge \delta_{0,n}.$$

Throughout this section, the constants c, C follow the convention in Section D.1, with dependence restricted to the fixed statistical model bounds and scale margins, q_0, and the fixed bump functions used below.

F.2 Auxiliary lemmas

We use the following standard L^2-form of Assouad's lemma; see [31, Lemma 2.12].

For probability laws P, Q, total variation means

$$d_{\mathrm{TV}}(P,Q) := \sup_{A \text{ measurable}} |P(A) - Q(A)| = \frac{1}{2} \int |p - q| \, d\lambda,$$

where $p = dP/d\lambda$, $q = dQ/d\lambda$, and any common dominating measure, for example $\lambda = P + Q$, may be used.

Lemma F.1. *Let J be finite, $\Theta = \{0,1\}^J$, and $f_\theta \in L^2(\nu)$. Let P_θ be the corresponding observation law and $\theta^{(r)}$ denote a flip in coordinate r. Suppose*

$$\|f_\theta - f_{\theta'}\|_{L^2(\nu)}^2 \ge \sum_{r \in J} \rho_r^2 \mathbb{I}\{\theta_r \ne \theta_r'\}. \tag{10}$$

Then

$$\begin{aligned} \inf_{\widetilde{f}} \sup_\theta \mathbb{E}_\theta \|\widetilde{f} - f_\theta\|_{L^2(\nu)}^2 &\ge \inf_{\widetilde{f}} 2^{-|J|} \sum_\theta \mathbb{E}_\theta \|\widetilde{f} - f_\theta\|_{L^2(\nu)}^2 \\ &\ge \frac{1}{8} \sum_{r \in J} \rho_r^2 \left\{ 1 - \sup_{\theta : \theta_r = 0} d_{\mathrm{TV}}(P_\theta, P_{\theta^{(r)}}) \right\}. \end{aligned}$$

The infima include randomized measurable estimators with independent auxiliary randomness.

Proof. The bound is immediate for estimators with infinite prior-average risk. For any estimator with finite prior-average risk, choose a nearest f_ϑ in the finite family, with deterministic tie-breaking, and call its index $\widehat{\theta}$. The triangle inequality and (10) give

$$\|\widetilde{f} - f_\theta\|_{L^2(\nu)}^2 \ge \frac{1}{4} \sum_r \rho_r^2 \mathbb{I}\{\widehat{\theta}_r \ne \theta_r\}.$$

For each r, pairing θ and $\theta^{(r)}$ yields

$$2^{-|J|} \sum_\theta P_\theta(\widehat{\theta}_r \ne \theta_r) \ge \frac{1}{2} \left\{ 1 - \sup_{\theta : \theta_r = 0} d_{\mathrm{TV}}(P_\theta, P_{\theta^{(r)}}) \right\}.$$

Sum these inequalities and take the infimum. Independent auxiliary randomness can be conditioned on, or appended to both laws while preserving their total variation distance. □

Next, we give a lemma bounding Hellinger distance by L^2 distance. Let

$$\mathsf{H}^2(P,Q) := \int \left(\sqrt{dP/d\lambda} - \sqrt{dQ/d\lambda}\right)^2 d\lambda$$

for two probability measures $P, Q \ll \lambda$.

The comparison with total variation follows from Cauchy–Schwarz:

$$\begin{aligned} d_{\mathrm{TV}}(P,Q) &= \frac{1}{2}\int |\sqrt{p} - \sqrt{q}|(\sqrt{p} + \sqrt{q})\, d\lambda \\ &\le \frac{1}{2}\mathsf{H}(P,Q)\left\{\int (\sqrt{p} + \sqrt{q})^2\, d\lambda\right\}^{1/2} \le \mathsf{H}(P,Q). \end{aligned}$$

Writing $\mathfrak{a}(P,Q) = \int \sqrt{pq}\, d\lambda = 1 - \mathsf{H}^2(P,Q)/2$, product integration gives

$$\mathsf{H}^2(P^{\otimes n}, Q^{\otimes n}) = 2\{1 - \mathfrak{a}(P,Q)^n\} \le 2n\{1 - \mathfrak{a}(P,Q)\} = n\mathsf{H}^2(P,Q). \tag{F.H}$$

The inequality uses $1 - a^n = (1-a)\sum_{j=0}^{n-1} a^j \le n(1-a)$ for $a \in [0,1]$.

Lemma F.2. *Let Q_a have density $q_{\mathrm{lb}}(y-a)$. Then, for all $a, b \in \mathbb{R}$,*

$$\mathsf{H}^2(Q_a, Q_b) \le C(a-b)^2.$$

Here C depends only on q_0 and σ_{lb}. Consequently, if $X \sim \nu$, $Y_f = f(X) + \epsilon$, and $Y_g = g(X) + \epsilon$ with $\epsilon \sim q_{\mathrm{lb}}$ independent of X, let $P_f^{(n)}$ and $P_g^{(n)}$ denote the laws of n i.i.d. covariate–response pairs. Then

$$\mathsf{H}^2(P_f^{(1)}, P_g^{(1)}) \le C\|f-g\|^2_{L^2(\nu)}, \qquad \mathsf{H}^2(P_f^{(n)}, P_g^{(n)}) \le Cn\|f-g\|^2_{L^2(\nu)}.$$

Proof. Write $r_{\mathrm{lb}} := \sqrt{q_{\mathrm{lb}}}$. Since Q_a has density $q_{\mathrm{lb}}(y-a)$, its square-root density is $r_{\mathrm{lb}}(y-a)$. Hence

$$\mathsf{H}^2(Q_a, Q_b) = \int_{\mathbb{R}} \{r_{\mathrm{lb}}(y-a) - r_{\mathrm{lb}}(y-b)\}^2\, dy = \|r_{\mathrm{lb}}(\cdot - a) - r_{\mathrm{lb}}(\cdot - b)\|_2^2.$$

Set $t := a - b$ and note that

$$\|r_{\mathrm{lb}}(\cdot - a) - r_{\mathrm{lb}}(\cdot - b)\|_2 = \|r_{\mathrm{lb}}(\cdot - t) - r_{\mathrm{lb}}\|_2.$$

Since r_{lb} is absolutely continuous and $r'_{\mathrm{lb}} \in L^2$,

$$r_{\mathrm{lb}}(y-t) - r_{\mathrm{lb}}(y) = -\int_0^t r'_{\mathrm{lb}}(y-s)\, ds = -\operatorname{sgn}(t)\int_0^{|t|} r'_{\mathrm{lb}}(y - \operatorname{sgn}(t)s)\, ds.$$

Minkowski's integral inequality gives

$$\|r_{\mathrm{lb}}(\cdot - t) - r_{\mathrm{lb}}(\cdot)\|_2 \le \int_0^{|t|} \|r'_{\mathrm{lb}}(\cdot - \operatorname{sgn}(t)s)\|_2\, ds = |t|\|r'_{\mathrm{lb}}\|_2,$$

so

$$\mathsf{H}^2(Q_a, Q_b) \le (a-b)^2\|r'_{\mathrm{lb}}\|_2^2.$$

Since $q_{\mathrm{lb}}(u) = \sigma_{\mathrm{lb}}^{-1} q_0(u/\sigma_{\mathrm{lb}})$, we have

$$r_{\mathrm{lb}}(u) = \sigma_{\mathrm{lb}}^{-1/2} r_0(u/\sigma_{\mathrm{lb}})$$

and

$$r'_{\mathrm{lb}}(u) = \sigma_{\mathrm{lb}}^{-3/2} r'_0(u/\sigma_{\mathrm{lb}}).$$

Hence, by the change of variables $v = u/\sigma_{\mathrm{lb}}$,

$$\|r'_{\mathrm{lb}}\|_2^2 = \int_{\mathbb{R}} \sigma_{\mathrm{lb}}^{-3}\{r'_0(u/\sigma_{\mathrm{lb}})\}^2\, du = \sigma_{\mathrm{lb}}^{-2}\int_{\mathbb{R}} \{r'_0(v)\}^2\, dv \le C.$$

This proves $\mathsf{H}^2(Q_a, Q_b) \le C(a-b)^2$.

Now consider the one-sample joint laws. Let $P_f^{(1)}$ and $P_g^{(1)}$ denote the laws of (X, Y_f) and (X, Y_g), respectively. Since $X \sim \nu$ and, conditionally on $X = x$, $Y_f = f(x) + \epsilon$ has density $q_{\mathrm{lb}}(y - f(x))$, the law $P_f^{(1)}$ is absolutely continuous with respect to $\nu(dx)\,dy$, with density

$$p_f(x, y) = q_{\mathrm{lb}}(y - f(x)).$$

Similarly,

$$p_g(x, y) = q_{\mathrm{lb}}(y - g(x)).$$

Therefore

$$\begin{aligned}\mathsf{H}^2(P_f^{(1)}, P_g^{(1)}) &= \iint \{r_{\mathrm{lb}}(y - f(x)) - r_{\mathrm{lb}}(y - g(x))\}^2\, dy\, \nu(dx) \\ &= \int \mathsf{H}^2(Q_{f(x)}, Q_{g(x)})\, \nu(dx) \le C\|f - g\|_{L^2(\nu)}^2.\end{aligned}$$

Then we have

$$\mathsf{H}^2(P_f^{(n)}, P_g^{(n)}) \le n\mathsf{H}^2(P_f^{(1)}, P_g^{(1)}) \le Cn\|f - g\|_{L^2(\nu)}^2.$$

□

Lemma F.3. *Let $\mathcal{M} \subset \mathbb{R}^D$ be compact with $\mathcal{M} \in \mathcal{C}^{\beta}_{d, L_{\mathcal{M}}, \tau}(\mathbb{R}^D)$, where $\tau > 0$. Suppose $d \in [d_{\max}]$, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, $\beta \in [2, \beta_{\max}]$, and $\beta \ge \lceil \alpha \rceil + 1$. There exists $a_{\mathrm{geo}} > 0$, depending only on $D, d_{\max}, \beta_{\max}, L_{\mathcal{M}}$, such that, whenever $h \in (0, a_{\mathrm{geo}}(\tau \wedge 1))$, there are points $\mathcal{X}_h = \{x_1, \dots, x_m\} \subset \mathcal{M}$ satisfying*

$$m \ge c\,\mathrm{vol}_{\mathcal{M}}(\mathcal{M}) h^{-d}, \tag{11}$$

and there are functions $\varphi_1, \dots, \varphi_m : \mathcal{M} \to \mathbb{R}$ with pairwise disjoint supports, i.e.,

$$\mathrm{supp}(\varphi_i) \cap \mathrm{supp}(\varphi_j) = \emptyset \qquad \text{for } i \ne j,$$

satisfying

$$\mathrm{supp}(\varphi_j) \subset \mathcal{M} \cap B_{\mathbb{R}^D}(x_j, h),$$

and further satisfy

$$\left\| \sum_{j \in [m]} \vartheta_j h^{\alpha} \varphi_j \right\|_{\mathcal{H}^{\alpha}(\mathcal{M})} \le C \quad \text{for all } \vartheta \in \{0, 1\}^m, \tag{12}$$

and

$$ch^d \le \int_{\mathcal{M}} \varphi_j^2\, d\,\mathrm{vol}_{\mathcal{M}} \le Ch^d.$$

The constants c, C depend only on $D, d_{\max}, \alpha_{\min}, \alpha_{\max}, \beta_{\max}, L_{\mathcal{M}}$ and on the fixed bump functions.

Proof. Let $c_{\mathrm{geo}} \in (0, 1)$ be the localization constant in Lemma D.1, chosen uniformly over $d \in [d_{\max}]$, $\beta \in [2, \beta_{\max}]$, and $L = L_{\mathcal{M}}$. Put

$$r_{L_{\mathcal{M}}} := (4L_{\mathcal{M}})^{-1}, \qquad a_{\mathrm{geo}} := c_{\mathrm{geo}}(r_{L_{\mathcal{M}}} \wedge 1)/16.$$

If $h \le a_{\mathrm{geo}}(\tau \wedge 1)$, then $4h \le c_{\mathrm{geo}}(\tau \wedge r_{L_{\mathcal{M}}})$ and $h \le r_{L_{\mathcal{M}}}$.

Let $\mathcal{X}_h = \{x_1, \dots, x_m\}$ be a maximal $4h$-separated subset of $\mathcal{M}$, with separation measured in ambient Euclidean distance, that is, the following is satisfied:

(a) for any $j \ne j' \in [m]$, $\|x_j - x_{j'}\| \ge 4h$.

(b) for any $x \in \mathcal{M}$, there is $j \in [m]$ such that $\|x - x_j\| < 4h$.

Note that (b) implies

$$\mathcal{M} \subset \bigcup_{j \in [m]} B_{\mathbb{R}^D}(x_j, 4h).$$

Since $4h \le c_{\mathrm{geo}}(\tau \wedge r_{L_{\mathcal{M}}})$, Lemma D.1(i) gives

$$\mathrm{vol}_{\mathcal{M}}(\mathcal{M}) \le \sum_{j \in [m]} \mathrm{vol}_{\mathcal{M}}\big(\mathcal{M} \cap B_{\mathbb{R}^D}(x_j, 4h)\big) \le Cmh^d.$$

Hence,

$$m \geq c \operatorname{vol}_{\mathcal{M}}(\mathcal{M}) h^{-d}.$$

Fix a nonzero bump $\eta_d \in C_c^\infty(B_{\mathbb{R}^d}(0, 1/4))$. For each j, let

$$\Psi_j(u) = x_j + u + N_j(u), \qquad u \in B_{T_{x_j}\mathcal{M}}(0, r_{L_{\mathcal{M}}}),$$

be the tangent-normal chart at x_j. Identifying $T_{x_j}\mathcal{M}$ with $\mathbb{R}^d$ by an arbitrary orthonormal basis, define

$$\varphi_j(\Psi_j(u)) := \eta_d(u/h) \qquad \text{for } \|u\| < h/4,$$

and set $\varphi_j = 0$ outside.

If $\varphi_j(\Psi_j(u)) \neq 0$, then $\|u\| \leq h/4$. Since $DN_j(0) = 0$ and $\|DN_j(v)\|_{\mathrm{op}} \leq L_{\mathcal{M}}\|v\|$,

$$\|N_j(u)\| \leq \int_0^1 \|DN_j(tu)\|_{\mathrm{op}} \|u\| \, dt \leq \frac{L_{\mathcal{M}}}{2}\|u\|^2.$$

Thus

$$\|\Psi_j(u) - x_j\| \leq \|u\| + \|N_j(u)\| \leq \frac{h}{4} + \frac{L_{\mathcal{M}}}{2}\left(\frac{h}{4}\right)^2 < h,$$

because $h \leq r_{L_{\mathcal{M}}}$. So

$$\operatorname{supp}(\varphi_j) \subset \mathcal{M} \cap B_{\mathbb{R}^D}(x_j, h).$$

The centers are $4h$-separated, so these supports are pairwise disjoint. Let

$$J_j(u) := \sqrt{\det\{I + DN_j(u)^\top DN_j(u)\}}$$

be the Jacobian in jth chart. On the support of φ_j, $\|DN_j(u)\|_{\mathrm{op}} \leq L_{\mathcal{M}}\|u\| \leq 1/16$, and hence

$$J_j(u) \in \left[1, \left(\frac{17}{16}\right)^{d/2}\right]$$

by (9). Therefore, by the change of variables $u = hv$,

$$\int_{\mathcal{M}} \varphi_j^2 \, d\operatorname{vol}_{\mathcal{M}} = \int \eta_d(u/h)^2 J_j(u) \, du = h^d \int \eta_d(v)^2 J_j(hv) \, dv.$$

Since η_d is fixed and nonzero,

$$ch^d \leq \int_{\mathcal{M}} \varphi_j^2 \, d\operatorname{vol}_{\mathcal{M}} \leq Ch^d.$$

Let $F_\vartheta := \sum_{j\in[m]} \vartheta_j h^\alpha \varphi_j$ for $\vartheta \in \{0,1\}^m$. In the chart centered at x_j,

$$h^\alpha \varphi_j(\Psi_j(u)) = h^\alpha \eta_d(u/h).$$

For every multi-index μ with $|\mu| = r \leq s_\alpha$, the chain rule gives

$$D^\mu\{h^\alpha \eta_d(u/h)\} = h^{\alpha - r}(D^\mu \eta_d)(u/h).$$

Hence all derivatives of order $r \leq s_\alpha$ are uniformly bounded for $0 < h \leq 1$, since $r < \alpha$. For the top Hölder seminorm, if $|\mu| = s_\alpha$, then

$$[D^\mu\{h^\alpha \eta_d(\cdot/h)\}]_{C^{0,\gamma_\alpha}} \leq h^{\alpha - s_\alpha} h^{-\gamma_\alpha} [D^\mu \eta_d]_{C^{0,\gamma_\alpha}} = [D^\mu \eta_d]_{C^{0,\gamma_\alpha}},$$

because $\alpha = s_\alpha + \gamma_\alpha$. Thus the factor h^α exactly compensates for the derivative scaling h^{-s_α} and the Hölder scaling $h^{-\gamma_\alpha}$.

In any admissible chart Ψ_x, the transition to the j-th chart on their overlap is

$$\Psi_j^{-1}(\Psi_x(v)) = \Pi_{T_{x_j}\mathcal{M}}(\Psi_x(v) - x_j).$$

Its derivatives through order $s_\alpha + 1$ are uniformly bounded because $\beta \geq s_\alpha + 2$ and the chart radius and derivative bounds are fixed independently of τ. Each bump vanishes near the boundary of its defining chart, so its extension by zero has the same regularity. The chain rule therefore gives

derivative bounds $Ch^{\alpha-r}$ for each pullback $h^{\alpha}\varphi_j \circ \Psi_x$, for $0 \le r \le s_\alpha + 1$. Since the bump supports are disjoint, these bounds also hold for $g_x := F_\vartheta \circ \Psi_x$:

$$\|D^r g_x\|_\infty \le Ch^{\alpha-r}, \qquad 0 \le r \le s_\alpha + 1.$$

The chart domain is convex. Thus the bounds for orders s_α and $s_\alpha + 1$ imply, for any two points u, v in that domain,

$$\|D^{s_\alpha} g_x(u) - D^{s_\alpha} g_x(v)\| \le C \min\{h^{\gamma_\alpha}, h^{\gamma_\alpha - 1}\|u - v\|\} \le C\|u - v\|^{\gamma_\alpha}.$$

Together with $h \le 1$, this proves

$$\|F_\vartheta\|_{\mathcal{H}^\alpha(\mathcal{M})} \le C$$

uniformly over $\vartheta \in \{0,1\}^m$, all admissible charts, and h, with constants independent of τ. □

F.3 Proof of Theorem 1

Proof. It suffices to prove the lower bound on a submodel fixed in Section F.1. Let

$$\nu^\circ := \sum_{k \in [K_n]} \pi_{k,n} \mu^\circ_{k,n},$$

where $d\mu^\circ_{k,n} = \varrho^\circ_{k,n}\, d\,\mathrm{vol}_{\mathcal{M}_k}$ and $c_\mu \le \varrho^\circ_{k,n} \le c_\mu^{-1}$. Since

$$1 = \int_{\mathcal{M}_k} \varrho^\circ_{k,n}\, d\,\mathrm{vol}_{\mathcal{M}_k} \le c_\mu^{-1}\,\mathrm{vol}_{\mathcal{M}_k}(\mathcal{M}_k),$$

we have $\mathrm{vol}_{\mathcal{M}_k}(\mathcal{M}_k) \ge c_\mu$. Recall

$$N_k := n\pi_{k,n}, \qquad h_{k,n} := N_k^{-1/(2\alpha_k + d_k)}, \qquad r_n := \max_{k \in [K_n]} h_{k,n}.$$

By Lemma E.1, $r_n \to 0$, and by Assumption E.1,

$$r_n \le a_{\mathrm{sc}}(\tau_{0,n} \wedge \delta_{0,n}).$$

Let a_{geo} be the constant in Lemma F.3. Define

$$a_0 := 1 \wedge a_{\mathrm{geo}}, \qquad \bar{h}_k := a_0 h_{k,n}.$$

Since $a_{\mathrm{sc}} \le 1$, for all sufficiently large n,

$$\bar{h}_k \le a_0 r_n \le a_0 a_{\mathrm{sc}} \tau_{0,n} \le a_{\mathrm{geo}} \tau_{0,n}.$$

Also, since $r_n \to 0$, $\bar{h}_k \le a_{\mathrm{geo}}$ for all sufficiently large n. Hence

$$\bar{h}_k \le a_{\mathrm{geo}}(\tau_{0,n} \wedge 1) \le a_{\mathrm{geo}}\{\mathrm{reach}(\mathcal{M}_k) \wedge 1\}.$$

Thus Lemma F.3 applies to every component $\mathcal{M}_k \in \mathcal{C}^{\beta_k}_{d_k, L_{\mathcal{M}}, \tau_{0,n}}(\mathbb{R}^D)$ with $\tau = \tau_{0,n}$ and $h = \bar{h}_k$.

For each k, let

$$\{(x_{k,j}, \varphi_{k,j}) : j \in [m_k]\}$$

be the corresponding bump packing. Since $\mathrm{vol}_{\mathcal{M}_k}(\mathcal{M}_k) \ge c_\mu$, the packing bound in (11) gives

$$m_k \ge c\,\mathrm{vol}_{\mathcal{M}_k}(\mathcal{M}_k)\bar{h}_k^{-d_k} \ge c\bar{h}_k^{-d_k}. \tag{13}$$

The supports are disjoint within each component by construction. Across different components, they are also disjoint because

$$d(\mathcal{M}_k, \mathcal{M}_\ell) \ge \delta_{0,n} > 0 \qquad \text{for } k \ne \ell.$$

Choose a fixed $\lambda_{\mathrm{lb}} \in (0,1]$, depending only on the fixed model constants, q_0, and the bump functions, sufficiently small for the Hölder and testing bounds below; their constants are independent of λ_{lb}. For

$$\mathcal{J}_n := \{(k,j) : k \in [K_n],\, j \in [m_k]\}$$

and $\theta \in \{0,1\}^{|\mathcal{J}_n|}$, define, for $x \in \mathcal{M}_k$,

$$f_\theta(x) := \sum_{j \in [m_k]} \theta_{k,j} \lambda_{\mathrm{lb}} \bar{h}_k^{\alpha_k} \varphi_{k,j}(x).$$

By (12) in Lemma F.3,

$$\left\| \sum_{j\in[m_k]} \theta_{k,j}\bar{h}_k^{\alpha_k}\varphi_{k,j} \right\|_{\mathcal{H}^{\alpha_k}(\mathcal{M}_k)} \le C.$$

The first requirement, $C\lambda_{\mathrm{lb}} \le L_{\mathcal{F}}$, therefore ensures $f_\theta \in \mathcal{H}(\boldsymbol{\alpha}, \mathscr{M}; L_{\mathcal{F}})$ for every $\theta \in \{0,1\}^{|\mathcal{J}_n|}$.

Let $P_\theta^{(n)}$ be the distribution of the training sample under

$$X_i^\star \overset{\text{i.i.d.}}{\sim} \nu^\circ, \qquad X_i = X_i^\star, \qquad Y_i = f_\theta(X_i^\star) + \epsilon_i.$$

For $\theta, \theta' \in \{0,1\}^{|\mathcal{J}_n|}$, the pairwise disjointness of the supports implies that, pointwise on $\mathcal{M}_k$,

$$\begin{aligned}
|f_\theta - f_{\theta'}|^2 &= \lambda_{\mathrm{lb}}^2 \bar{h}_k^{2\alpha_k} \left| \sum_{j\in[m_k]} (\theta_{k,j} - \theta'_{k,j})\varphi_{k,j} \right|^2 \\
&= \lambda_{\mathrm{lb}}^2 \bar{h}_k^{2\alpha_k} \sum_{j\in[m_k]} (\theta_{k,j} - \theta'_{k,j})^2 \varphi_{k,j}^2 \\
&= \lambda_{\mathrm{lb}}^2 \bar{h}_k^{2\alpha_k} \sum_{j\in[m_k]} \mathbb{I}\{\theta_{k,j} \ne \theta'_{k,j}\} \varphi_{k,j}^2.
\end{aligned}$$

Moreover, the density bounds $c_\mu \le \varrho_{k,n}^\circ \le c_\mu^{-1}$ and Lemma F.3 give

$$c\bar{h}_k^{d_k} \le \int_{\mathcal{M}_k} \varphi_{k,j}(x)^2 \varrho_{k,n}^\circ(x)\, d\,\mathrm{vol}_{\mathcal{M}_k}(x) \le C\bar{h}_k^{d_k}.$$

Consequently,

$$\begin{aligned}
\|f_\theta - f_{\theta'}\|_{L^2(\nu^\circ)}^2 &= \sum_{k\in[K_n]} \pi_{k,n} \int_{\mathcal{M}_k} |f_\theta - f_{\theta'}|^2 \varrho_{k,n}^\circ \, d\,\mathrm{vol}_{\mathcal{M}_k} \\
&= \sum_{k\in[K_n]} \sum_{j\in[m_k]} \lambda_{\mathrm{lb}}^2 \pi_{k,n} \bar{h}_k^{2\alpha_k} \mathbb{I}\{\theta_{k,j} \ne \theta'_{k,j}\} \int_{\mathcal{M}_k} \varphi_{k,j}^2 \varrho_{k,n}^\circ \, d\,\mathrm{vol}_{\mathcal{M}_k} \\
&\ge \sum_{(k,j)\in\mathcal{J}_n} c\lambda_{\mathrm{lb}}^2 \pi_{k,n} \bar{h}_k^{2\alpha_k+d_k} \mathbb{I}\{\theta_{k,j} \ne \theta'_{k,j}\}.
\end{aligned}$$

Fix a uniform lower constant $c > 0$ in the last bound. Then Lemma F.1 applies with

$$\rho_{k,j}^2 := c\lambda_{\mathrm{lb}}^2 \pi_{k,n} \bar{h}_k^{2\alpha_k+d_k}.$$

Let $\theta^{(k,j)}$ be obtained from θ by flipping only the coordinate (k,j). Then

$$f_\theta - f_{\theta^{(k,j)}} = \pm\lambda_{\mathrm{lb}} \bar{h}_k^{\alpha_k} \varphi_{k,j}$$

on the corresponding bump support, and the difference is zero elsewhere. By Lemma F.2 and the density upper bound,

$$\begin{aligned}
\mathsf{H}^2(P_\theta^{(n)}, P_{\theta^{(k,j)}}^{(n)}) &\le Cn\|f_\theta - f_{\theta^{(k,j)}}\|_{L^2(\nu^\circ)}^2 \\
&\le Cn\lambda_{\mathrm{lb}}^2 \pi_{k,n} \bar{h}_k^{2\alpha_k+d_k} \\
&= C\lambda_{\mathrm{lb}}^2 a_0^{2\alpha_k+d_k}.
\end{aligned}$$

Since $a_0 \le 1$, the second requirement, $C\lambda_{\mathrm{lb}}^2 \le 1/16$, gives

$$\mathsf{H}^2(P_\theta^{(n)}, P_{\theta^{(k,j)}}^{(n)}) \le \frac{1}{16}.$$

Using $d_{\mathrm{TV}}(P,Q) \le \mathsf{H}(P,Q)$, we get

$$d_{\mathrm{TV}}(P_\theta^{(n)}, P_{\theta^{(k,j)}}^{(n)}) \le \frac{1}{4}.$$

Therefore Lemma F.1 gives

$$\begin{aligned}\inf_{\widetilde{f}} \sup_{\theta\in\{0,1\}^{|\mathcal{J}_n|}} \mathbb{E}_\theta \|\widetilde{f} - f_\theta\|^2_{L^2(\nu^\circ)} &\geq \frac{1}{8} \sum_{(k,j)\in\mathcal{J}_n} \rho^2_{k,j} \left(1 - \frac{1}{4}\right) \\ &= \frac{3}{32} \sum_{(k,j)\in\mathcal{J}_n} \rho^2_{k,j} \\ &= \frac{3}{32} c\lambda^2_{\mathrm{lb}} \sum_{k\in[K_n]} m_k \pi_{k,n} \bar{h}_k^{2\alpha_k+d_k}.\end{aligned}$$

By (13),

$$m_k \geq c\bar{h}_k^{-d_k},$$

and hence

$$m_k \pi_{k,n} \bar{h}_k^{2\alpha_k+d_k} \geq c\pi_{k,n} \bar{h}_k^{-d_k} \bar{h}_k^{2\alpha_k+d_k} = c\pi_{k,n} \bar{h}_k^{2\alpha_k}.$$

It follows that

$$\inf_{\widetilde{f}} \sup_{\theta\in\{0,1\}^{|\mathcal{J}_n|}} \mathbb{E}_\theta \|\widetilde{f} - f_\theta\|^2_{L^2(\nu^\circ)} \geq c \sum_{k\in[K_n]} \pi_{k,n} \bar{h}_k^{2\alpha_k}.$$

Since $\bar{h}_k = a_0 N_k^{-1/(2\alpha_k+d_k)}$,

$$\begin{aligned}\sum_{k\in[K_n]} \pi_{k,n} \bar{h}_k^{2\alpha_k} &= \sum_{k\in[K_n]} \pi_{k,n} a_0^{2\alpha_k} N_k^{-2\alpha_k/(2\alpha_k+d_k)} \\ &\geq a_0^{2\alpha_{\max}} \sum_{k\in[K_n]} \pi_{k,n} (n\pi_{k,n})^{-2\alpha_k/(2\alpha_k+d_k)} \\ &= a_0^{2\alpha_{\max}} \mathfrak{R}_n(\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\pi}).\end{aligned}$$

The same calculation uses the middle (uniform-prior) bound in Lemma F.1 and therefore also proves

$$\inf_g 2^{-|\mathcal{J}_n|} \sum_\theta \mathbb{E}_{P^\circ, f_\theta} \{g(\mathfrak{s}) - f_\theta(X^\star_{n+1})\}^2 \geq c\mathfrak{R}_n. \tag{14}$$

Here P° is the fixed latent design and bounded-noise law of the submodel.

Finally, on this submodel,

$$X_{n+1} = X^\star_{n+1} \sim \nu^\circ$$

independently of the training sample. Therefore, conditional on $(X_i, Y_i)_{i\in[n]}$, the prediction risk against $f_\theta(X^\star_{n+1})$ is exactly the $L^2(\nu^\circ)$-risk of the function

$$x \mapsto \widehat{f}_n((X_i, Y_i)_{i\in[n]}, x).$$

Thus the Assouad lower bound above applies to the original prediction problem. Since the finite family

$$\{f_\theta : \theta \in \{0,1\}^{|\mathcal{J}_n|}\},$$

together with the fixed design and noise law constructed above, is a submodel of the full model class, the proof is complete. □

F.4 Lower bound with independent pretraining tasks

Corollary F.1. *Under the hypotheses of Theorem 1, for every sufficiently large n there are $P^\circ \in \mathcal{P}_\star$ and a finite-support task distribution $\rho^\circ_{f,n}$ such that, for every number of independent training tasks $\Gamma \geq 0$,*

$$\inf_{\mathcal{A}} \mathbb{E}_{P^\circ, \rho^\circ_{f,n}} \left[\left\{\mathcal{A}(\mathcal{D}^{\mathrm{tr}}_\Gamma, \mathfrak{s}) - f(X^\star_{n+1})\right\}^2\right] \geq c\mathfrak{R}_n.$$

The tasks are generated as in Section 5; the infimum ranges over all measurable learning rules, also allowing independent randomization. The rule may be given P° and $\rho^\circ_{f,n}$. Consequently the same lower bound holds for $\inf_{\mathcal{A}} \sup_{P,\rho_f}$ in this pretraining experiment.

Proof. Use the submodel and finite family from the proof above and set

$$\rho_{f,n}^{\circ} = 2^{-|\mathcal{J}_n|} \sum_{\theta \in \{0,1\}^{\mathcal{J}_n}} \delta_{f_\theta}.$$

The index θ_0 of the target task is independent of all training-task indices and observations. Thus $\mathcal{D}_\Gamma^{\mathrm{tr}}$ is independent of the pair $(\theta_0, \mathfrak{s})$. Conditional on almost every realized training dataset D, the map $g_D(s) = \mathcal{A}(D, s)$ is a measurable prompt predictor, and the conditional target law is unchanged. Applying (14) gives

$$\mathbb{E}[\{g_D(\mathfrak{s}) - f_{\theta_0}(X_{n+1}^\star)\}^2 \mid \mathcal{D}_\Gamma^{\mathrm{tr}} = D] \geq c\mathfrak{R}_n.$$

Average over D and any independent randomization. The hard prior is allowed to depend on n and is fixed independently of the realized training data. □

This is a worst-case statement over task distributions. For a fixed ρ_f, the lower bound depends on that distribution; for example a point mass at the zero function has zero latent prediction risk.

G Tangent local polynomial estimator

The upper-bound proof proceeds from local geometry to regression risk. Local purity restricts each window to one component; polynomial coercivity and concentration over the finite candidate grid then control the tangent fit. Lemma G.5 converts this control into a rounded projector estimate. Lemma G.6 establishes a single Gram event uniformly over nearby projectors, so it also applies to the estimated projector. The final bias–variance argument combines these two events to obtain the componentwise and aggregate rates.

G.1 Standing notation and scale consequences

We use the notation from the main text and Appendix E.2. All spectral projectors use a fixed measurable tie-breaking rule, including outside the good events. Let $\Pi_{(d)}(B)$ denote the projector onto the top d-dimensional eigenspace of a symmetric matrix B. For $k \in [K_n]$, recall

$$p_k := \lceil \alpha_k \rceil - 1, \qquad s_k := \lceil \alpha_k \rceil, \qquad h_k := h_{k,n} = N_k^{-1/(2\alpha_k + d_k)},$$
$$p_{\max} := \lceil \alpha_{\max} \rceil - 1, \qquad s_{\max} := \lceil \alpha_{\max} \rceil.$$

The regression polynomial degree is p_k, whereas the tangent graph fit uses one additional degree $s_k = p_k + 1$. The degree s_k tangent fit is used because the $C^{s_k,1}$ regularity supplied by $\beta_k \geq s_k + 1$ yields, after the rescaling $u = h_k t$ and division by h_k, a graph Taylor remainder of order $h_k^{s_k}$ on bounded t-sets.

For $d \in [d_{\max}]$ and $s \in [s_{\max}]$, let $\mathscr{P}_d$ be the set of rank-d orthogonal projectors in $\mathbb{R}^D$. The tensors below are understood as bounded symmetric multilinear maps $T_\ell : (\mathbb{R}^D)^\ell \to \mathbb{R}^D$. Define

$$\Theta_{d,s} := \{(a, \Pi, T_2, \ldots, T_s) : \|a\| \leq 1,\ \Pi \in \mathscr{P}_d,\ \|T_\ell\|_{\mathrm{op}} \leq 2L_{\mathcal{M}},\ 2 \leq \ell \leq s\}.$$

Here, $a \in \mathbb{R}^D$. The tensors are symmetric in their arguments. This restriction preserves the diagonal evaluations $T_\ell(v, \ldots, v)$ entering R_q and the operator-norm bound, and the Taylor tensors used below are symmetric. If $s = 1$, the tensor part is absent. For $q = (a, \Pi, T_2, \ldots, T_s) \in \Theta_{d,s}$, recall

$$R_q(z) := a + (I - \Pi)z - \sum_{\ell=2}^{s} T_\ell\{(\Pi z)^{\otimes \ell}\}.$$

The metric on $\Theta_{d,s}$ is

$$d_\Theta(q, q') := \|a - a'\| + \|\Pi - \Pi'\|_F + \sum_{\ell=2}^{s} \|T_\ell - T_\ell'\|_{\mathrm{op}}.$$

Lemma G.1 (Product nets for tangent candidates)**.** *For $d \in [D-1]$, $s \in \mathbb{N}$, $L_{\mathcal{M}} \geq 1$, and $\eta \in (0, 1]$, the space $(\Theta_{d,s}, d_\Theta)$ has a finite η-net $\mathcal{Q}_{d,s}(\eta) \subset \Theta_{d,s}$ such that*

$$|\mathcal{Q}_{d,s}(\eta)| \leq C\eta^{-m_\Theta(d,s,D)},$$
$$m_\Theta(d, s, D) := D + d(D - d) + D \sum_{\ell=2}^{s} \binom{D + \ell - 1}{\ell}. \tag{15}$$

The constant C depends only on $D, d, s, L_{\mathcal{M}}$. All net centers belong to their respective candidate factors, and the nets can be fixed deterministically.

Proof. We first record a covering bound for a ball of radius $R > 0$ in an m-dimensional normed vector space. A set of centers in that ball whose pairwise distances exceed $t > 0$ has disjoint open balls of radius $t/2$. These balls lie in the ball of radius $R + t/2$. Comparing Lebesgue volumes in any linear coordinates bounds the number of centers by $(1 + 2R/t)^m$. A greedy selection therefore terminates in a maximal such set, which is a t-net with centers in the original ball. In particular, the offset ball $\{a \in \mathbb{R}^D : \|a\| \le 1\}$ has a t-net of size at most $(1 + 2/t)^D$.

For the projector factor, put $M_{D,d} := \binom{D}{d}$. Given $\Pi \in \mathscr{P}_d$, choose $U \in \mathbb{R}^{D\times d}$ with $U^\top U = I_d$ and $\Pi = UU^\top$. For each size-d subset $J \subset [D]$, let U_J be the square matrix of rows indexed by J, in increasing order. Cauchy–Binet gives

$$\sum_{|J|=d} \det(U_J)^2 = \det(U^\top U) = 1.$$

Thus some J satisfies $|\det U_J| \ge M_{D,d}^{-1/2}$. All singular values of U_J are at most one, so $\sigma_{\min}(U_J) \ge |\det U_J|$ and $\|U_J^{-1}\|_{\mathrm{op}} \le M_{D,d}^{1/2}$. Define

$$A := U_{J^c} U_J^{-1} \in \mathbb{R}^{(D-d)\times d}, \qquad R_{D,d} := \sqrt{dM_{D,d}}.$$

Since $\|U_{J^c}\|_{\mathrm{op}} \le 1$, we have $\|A\|_F \le \sqrt{d}\|A\|_{\mathrm{op}} \le R_{D,d}$. Let S_J be the permutation matrix restoring the original coordinate order from the order (J, J^c), and set

$$X_A := \begin{pmatrix} I_d \\ A \end{pmatrix}, \qquad H_A := (I_d + A^\top A)^{-1}, \qquad P_J(A) := S_J X_A H_A X_A^\top S_J^\top.$$

The columns of U and $S_J X_A$ have the same span, so $\Pi = P_J(A)$. For every matrix A, $P_J(A)$ is itself a rank-d orthogonal projector.

On the Frobenius ball $\|A\|_F, \|B\|_F \le R := R_{D,d}$, put $S := \sqrt{1 + R^2}$. Then $\|X_A\|_{\mathrm{op}}, \|X_B\|_{\mathrm{op}} \le S$, $\|H_A\|_{\mathrm{op}}, \|H_B\|_{\mathrm{op}} \le 1$, and the inverse identity gives

$$\|H_A - H_B\|_F = \|H_A(B^\top B - A^\top A)H_B\|_F \le 2R\|A - B\|_F.$$

Expanding $X_A H_A X_A^\top - X_B H_B X_B^\top$ by changing one factor at a time therefore yields

$$\|P_J(A) - P_J(B)\|_F \le \Lambda_{D,d}\|A - B\|_F, \qquad \Lambda_{D,d} := 2S + 2RS^2.$$

Apply the ball-covering bound to this $d(D-d)$-dimensional Frobenius ball at radius $t/\Lambda_{D,d}$, and map its centers through each P_J. Taking the union over the $M_{D,d}$ coordinate choices gives a t-net in $\mathscr{P}_d$, with centers in $\mathscr{P}_d$, of size at most

$$M_{D,d}\left(1 + \frac{2R_{D,d}\Lambda_{D,d}}{t}\right)^{d(D-d)}.$$

For each $\ell \ge 2$, a symmetric multilinear map $T : (\mathbb{R}^D)^\ell \to \mathbb{R}^D$ is determined by its values on unordered ℓ-tuples of coordinate vectors. There are $\binom{D+\ell-1}{\ell}$ such tuples and D output coordinates, so this vector space has dimension $m_\ell := D\binom{D+\ell-1}{\ell}$. The multilinear operator norm is a norm on this space: its vanishing forces all these coordinate values to vanish. Its ball $\{T : \|T\|_{\mathrm{op}} \le 2L_{\mathcal{M}}\}$ therefore has a t-net, in that same norm and with centers in the ball, of size at most $(1 + 4L_{\mathcal{M}}/t)^{m_\ell}$ by the preceding volume argument.

There are $s+1$ factors: the offset, the projector, and $s-1$ tensor factors. Choose the factor nets above at radius $t := \eta/(s+1)$ and take their Cartesian product. The sum metric gives total approximation error at most $(s+1)t = \eta$, and all product centers belong to $\Theta_{d,s}$. Its cardinality is at most

$$(1 + 2/t)^D\, M_{D,d}\left(1 + \frac{2R_{D,d}\Lambda_{D,d}}{t}\right)^{d(D-d)} \times \prod_{\ell=2}^{s}\left(1 + \frac{4L_{\mathcal{M}}}{t}\right)^{m_\ell}.$$

Substituting $t = \eta/(s+1)$ and using $\eta \le 1$ proves (15). For $s = 1$, the tensor product is empty and the same argument uses the two remaining factors. Fixing the factor nets once for each prescribed radius makes the construction deterministic. □

We fix the tangent grid and near-minimizer tolerance by

$$\eta_{\mathrm{grid},n} := n^{-1}, \qquad \Delta_{\tan,n} := n^{-3}, \qquad \mathcal{Q}_{d,s,n} := \mathcal{Q}_{d,s}(n^{-1}),$$

where the deterministic product net is the one fixed in Lemma G.1. Consequently, every $q \in \Theta_{d,s}$ has $q^\sharp \in \mathcal{Q}_{d,s,n}$ with $d_\Theta(q, q^\sharp) \le \eta_{\mathrm{grid},n}$, and

$$N_{\mathrm{grid},d,s,n} := |\mathcal{Q}_{d,s,n}| \le C n^{m_\Theta(d,s,D)}. \tag{16}$$

For $q \in \mathcal{Q}_{d,s,n}$, let Π_q denote its projector component.

Let $c_{\mathrm{Ber}} \in (0, 1]$ be the fixed Bernstein exponent used in the concentration bounds below. Define

$$\bar\rho_h := \frac{1}{2\alpha_{\min} + 1}, \qquad a^{\mathrm{or}}_{\tan,n} := \frac{2\alpha_{\max}\bar\rho_h + 3}{c_{\mathrm{Ber}}} \log(en).$$

Unless explicitly stated otherwise, Appendix G uses the proof-level tail parameter

$$a_{\tan,n} := a^{\mathrm{or}}_{\tan,n}.$$

This quantity is used only in concentration bounds.

Also recall

$$r_{L_\mathcal{M}} := (4L_\mathcal{M})^{-1}, \qquad R_{\mathrm{loc}} := c_{\mathrm{geo}}(\tau_{0,n} \wedge r_{L_\mathcal{M}} \wedge 1),$$

where c_{geo} is the constant fixed in Appendix E.2. By Lemma E.1, uniformly over $k \in [K_n]$, for all sufficiently large n,

$$n\pi_{k,n} h_k^{d_k} \ge c n^{\rho_{\tan}}, \qquad h_k \le R_{\mathrm{loc}}, \qquad \sup_{\ell \in [K_n]} h_\ell + 2\sigma_{\max,n} \le \frac{\delta_{0,n}}{2}, \tag{17}$$

and

$$\sigma_{k,n} \le a_{\mathrm{sc}} h_k^{\alpha_k \vee 1}, \qquad \{n\pi_{k,n} h_k^{d_k}\}^{-1} = h_k^{2\alpha_k}. \tag{18}$$

We also use, when converting polynomial or exponential remainders into powers of h_k, the bounds

$$n^{-\bar\rho_h} \le h_k \le r_n \le n^{-\rho_h}.$$

Local notation after conditioning on the query. Whenever we condition on

$$(Z_{n+1}, X^\star_{n+1}, \xi_{n+1}) = (k, x^\star, \xi_{\mathrm{qry}}),$$

we write $\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}$, $\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}$, and use the query-level notation

$$x := x^\star + \xi_{\mathrm{qry}}, \quad d_x := d_k, \quad p_x := p_k, \quad s_x := s_k, \quad h_x := h_k, \quad \Pi^\star := \Pi_{T_{x^\star}\mathcal{M}_k}.$$

Let $U_\star \in \mathbb{R}^{D \times d_x}$ be a fixed measurable orthonormal basis of $T_{x^\star}\mathcal{M}_k$, so that

$$U_\star^\top U_\star = I_{d_x}, \qquad U_\star U_\star^\top = \Pi^\star.$$

Let $N_{x^\star}$ be the tangent-normal graph map at $x^\star$, and define its coordinate version by

$$G_{x^\star}(u) := N_{x^\star}(U_\star u), \qquad u \in B_{\mathbb{R}^{d_x}}(0, r_{L_\mathcal{M}}).$$

Put $N_{\mathrm{grid},k,n} := N_{\mathrm{grid},d_k,s_k,n}$. For $i \in [n]$, define

$$z_i := \frac{X_i - x}{h_x}, \qquad \mathsf{A}_i := \mathsf{A}\left(\frac{\|X_i - x\|_1}{h_x}\right).$$

Also define

$$b_{\tan} := h_x^{s_x} + \frac{\sigma_{k,n}}{h_x}, \qquad \zeta_{\tan} := \left(\frac{\log N_{\mathrm{grid},k,n} + a_{\tan,n}}{n\pi_{k,n} h_x^{d_x}}\right)^{1/2}, \qquad \epsilon_{\tan} := b_{\tan} + \zeta_{\tan}. \tag{19}$$

The symbol $\zeta_{\tan}$ is used for the stochastic tangent fluctuation scale, to avoid confusion with the integer polynomial degree s_x.

For $q \in \Theta_{d_x,s_x}$, define the population and empirical tangent losses

$$L(q; x) := \mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}\left[h_x^2 \mathsf{A}_i \|R_q(z_i)\|^2\right], \qquad L_n(q; x) := \frac{h_x^2}{n} \sum_{i \in [n]} \mathsf{A}_i \|R_q(z_i)\|^2.$$

For $q \in \mathcal{Q}_{d_x,s_x,n}$, set

$$L_q(x) := L_n(q;x), \qquad L_{\min}(x) := \min_{q\in\mathcal{Q}_{d_x,s_x,n}} L_q(x),$$

$$\omega_q(x) := \frac{\{\Delta_{\tan,n} - (L_q(x) - L_{\min}(x))\}_+}{\sum_{q'\in\mathcal{Q}_{d_x,s_x,n}}\{\Delta_{\tan,n} - (L_{q'}(x) - L_{\min}(x))\}_+}.$$

The denominator is at least $\Delta_{\tan,n}$, because any minimizer of $L_q(x)$ over the finite grid contributes exactly $\Delta_{\tan,n}$. The corresponding averaged projector and tangent estimator are

$$\bar{P}_x := \sum_{q\in\mathcal{Q}_{d_x,s_x,n}} \omega_q(x)\Pi_q, \qquad \widehat{\Pi}_x := \Pi_{(d_x)}(\bar{P}_x).$$

For a rank-d_x projector Π, define the regression coordinates, weights, features, and Gram matrix by

$$v_i^\Pi := \Pi z_i, \quad W_i^\Pi := h_x^{-d_x}\mathsf{A}_i\mathsf{K}(v_i^\Pi), \quad \phi_i^\Pi := \Phi_{D,p_x}(v_i^\Pi), \quad G^\Pi := \frac{1}{n}\sum_{i\in[n]} W_i^\Pi \phi_i^\Pi(\phi_i^\Pi)^\top.$$

Write

$$v_i^\circ := v_i^{\Pi^\star}, \qquad W_i^\circ := W_i^{\Pi^\star}, \qquad e_0 := e_{0,D,p_x}.$$

When an orthonormal basis U of a rank-d subspace is used and $v = Uu$, the kernel remains the ambient kernel $\mathsf{K}(v)$. Thus, for example, $\mathsf{K}(v_i^\circ) = \mathsf{K}(U_\star U_\star^\top z_i)$.

We use the generic-constant convention in Section D.1. Any additional dependence on fixed local parameters is stated in the auxiliary lemmas below.

G.2 Auxiliary lemmas

We use the following convention. A probability kernel $\{\Lambda_t : t \in T\}$ from a measurable space T to $\mathbb{R}^m$ means that Λ_t is a probability measure on $\mathbb{R}^m$ for every $t \in T$, and that $t \mapsto \Lambda_t(A)$ is measurable for every Borel set $A \subset \mathbb{R}^m$. If $B \subset \mathbb{R}^m$ is Borel and we say that the kernel is supported on B, we mean $\Lambda_t(B) = 1$ for every $t \in T$. Conditional laws are chosen in versions supported on their stipulated closed perturbation balls for every conditioning value: an arbitrary version may be replaced by a point mass at zero on the measurable null set where the support restriction fails. Integrals against the design law are unchanged. Pushforward kernels below inherit the stated pointwise support bounds.

Lemma G.2. *Uniformly over $d \in [d_{\max}]$ and $p \in \{0,\dots,p_{\max}\}$, the following hold.*

(i) If $U \in \mathbb{R}^{D\times d}$ has orthonormal columns, then there exists $B_{U,p} \in \mathbb{R}^{q_{D,p}\times q_{d,p}}$ such that

$$\Phi_{D,p}(Uu) = B_{U,p}\Phi_{d,p}(u), \qquad e_{0,D,p} = B_{U,p}e_{0,d,p}.$$

Moreover, all singular values of $B_{U,p}$ are bounded above and below away from zero by fixed constants. Let

$$\psi_i := \Phi_{d,p}(u_i), \qquad H := \frac{1}{n}\sum_{i\in[n]} W_i\psi_i\psi_i^\top.$$

If H satisfies $c\pi I \preceq H \preceq c^{-1}\pi I$ for some $\pi > 0$ and fixed $c \in (0,1)$, and if $\phi_i = \Phi_{D,p}(Uu_i)$, then

$$G := \frac{1}{n}\sum_{i\in[n]} W_i\phi_i\phi_i^\top = B_{U,p}HB_{U,p}^\top$$

satisfies

$$\lambda_{\min}^+(G) \ge c'\pi, \qquad \|G\|_{\mathrm{op}} \le \frac{1}{c'}\pi, \qquad e_{0,D,p} \in \mathrm{Im}(G), \qquad e_{0,D,p}^\top G^\dagger G = e_{0,D,p}^\top$$

for some $c' \in (0,1)$ depending only on $c, D, d_{\max}, p_{\max}$.

(ii) Fix $R > 0$. There exist constants $c, C, \eta_{\mathrm{poly}} > 0$, depending only on $R, d_{\max}, p_{\max}$, such that for every measurable map $e : \overline{B_{\mathbb{R}^d}(0,R)} \to \mathbb{R}^d$ satisfying

$$\|e\|_\infty := \sup_{\|t\| \le R} \|e(t)\| \le \eta_{\mathrm{poly}},$$

and every $a \in \mathbb{R}^{q_{d,p}}$,

$$c\|a\|^2 \le \int_{\overline{B_{\mathbb{R}^d}(0,R)}} \{a^\top \Phi_{d,p}(t + e(t))\}^2 dt \le C\|a\|^2.$$

The same bounds hold in the following averaged form. Let

$$\Lambda : \overline{B_{\mathbb{R}^d}(0,R)} \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$$

be a probability kernel satisfying

$$\Lambda_t\Big(\overline{B_{\mathbb{R}^d}(0,\eta_{\mathrm{poly}})}\Big) = 1 \qquad \text{for every } t \in \overline{B_{\mathbb{R}^d}(0,R)}.$$

Then, for every $a \in \mathbb{R}^{q_{d,p}}$,

$$c\|a\|^2 \le \int_{\overline{B_{\mathbb{R}^d}(0,R)}} \int_{\mathbb{R}^d} \{a^\top \Phi_{d,p}(t+u)\}^2 \, d\Lambda_t(u)\, dt \le C\|a\|^2.$$

(iii) Fix $c > 0$. If $U \in \mathbb{R}^{D\times d}$ has orthonormal columns, $m \in \mathcal{H}^\alpha(B_{\mathbb{R}^d}(0,ch); L_{\mathcal{F}})$, $p = \lceil \alpha \rceil - 1$, and $0 < h \le 1$, identify m with its unique continuous extension to $\overline{B_{\mathbb{R}^d}(0,ch)}$. Then there is $w_h \in \mathbb{R}^{q_{D,p}}$ such that, for $\|u\| \le ch$,

$$|m(u) - \Phi_{D,p}(Uu/h)^\top w_h| \le Ch^\alpha, \qquad e_{0,D,p}^\top w_h = m(0).$$

Moreover, for every fixed $R_0 < \infty$, $z \mapsto \Phi_{D,p}(z)^\top w_h$ is Ch-Lipschitz on $B_{\mathbb{R}^D}(0,R_0)$. The constants in this part may additionally depend on c, and the Lipschitz constant may also depend on R_0.

Proof. For part (i), restricting each ambient monomial $z^\nu/\nu!$ to the subspace $z = Uu$ gives a polynomial in u of degree at most p. Collecting its coefficients in the intrinsic monomial basis defines $B_{U,p}$ and gives $\Phi_{D,p}(Uu) = B_{U,p}\Phi_{d,p}(u)$. Also note that $e_{0,D,p} = B_{U,p}e_{0,d,p}$.

The map $U \mapsto B_{U,p}$ is continuous on the compact Stiefel manifold $\{A \in \mathbb{R}^{D\times d} : A^\top A = I_d\}$. Conversely, since $u = U^\top(Uu)$, each intrinsic coordinate u_j is a linear function of the ambient coordinate $z = Uu$. Hence every intrinsic monomial $u^\mu/\mu!$ can be written as a linear combination of ambient monomials $z^\nu/\nu!$, $|\nu| \le p$, restricted to $z = Uu$. Thus there exists $C_{U,p}$ such that

$$\Phi_{d,p}(u) = C_{U,p}\Phi_{D,p}(Uu).$$

Consequently $C_{U,p}B_{U,p} = I_{q_{d,p}}$, so $B_{U,p}$ has full column rank. Since only finitely many (d,p) are allowed, and by the compactness of Stiefel manifolds, all singular values of $B_{U,p}$ are uniformly bounded above and below.

Now let $\phi_i = B_{U,p}\psi_i$, so

$$G = B_{U,p}HB_{U,p}^\top.$$

The bounds $\lambda_{\min}^+(G) \ge c'\pi$ and $\|G\|_{\mathrm{op}} \le \frac{1}{c'}\pi$ directly follow from the spectral bounds on H and the uniform singular-value bounds for $B_{U,p}$.

Since H is positive definite, $\mathrm{Im}(G) = \mathrm{Im}(B_{U,p})$, and the identity $e_{0,D,p} = B_{U,p}e_{0,d,p}$ gives $e_{0,D,p} \in \mathrm{Im}(G)$. For a symmetric positive semidefinite matrix G, $GG^\dagger$ is the orthogonal projector onto $\mathrm{Im}(G)$, so $e_{0,D,p}^\top G^\dagger G = e_{0,D,p}^\top$. By the symmetric property of the Moore-Penrose inverse, we obtain the desired result.

For part (ii), fix (d,p) and set

$$G_{0,d,p} := \int_{B_{\mathbb{R}^d}(0,R)} \Phi_{d,p}(t)\Phi_{d,p}(t)^\top \, dt.$$

For $a \neq 0$, $t \mapsto a^\top \Phi_{d,p}(t)$ is a nonzero polynomial and cannot vanish on an open ball. Hence $G_{0,d,p}$ is positive definite. Since (d,p) ranges over a finite set, its smallest eigenvalue is uniformly bounded below, and its largest eigenvalue is uniformly bounded above.

If $\|e\|_\infty \le 1$, then for $t \in B_{\mathbb{R}^d}(0,R)$, $t + e(t) \in B_{\mathbb{R}^d}(0,R+1)$. On this fixed ball, the entries of $\Phi_{d,p}\Phi_{d,p}^\top$ are uniformly Lipschitz over the allowed finite collection of (d,p). Thus

$$\left\| \int \Phi_{d,p}(t+e(t))\Phi_{d,p}(t+e(t))^\top dt - G_{0,d,p} \right\|_{\mathrm{op}} \le C\|e\|_\infty.$$

Choosing η_{poly} small enough gives the lower bound; the upper bound follows from boundedness of the dictionary on $B_{\mathbb{R}^d}(0,R+1)$.

For the averaged assertion, the same perturbation estimate gives

$$\left\| \int_{\overline{B_{\mathbb{R}^d}(0,R)}} \int_{\mathbb{R}^d} \left\{ \Phi_{d,p}(t+u)\Phi_{d,p}(t+u)^\top - \Phi_{d,p}(t)\Phi_{d,p}(t)^\top \right\} d\Lambda_t(u)\, dt \right\|_{\mathrm{op}} \le C\eta_{\mathrm{poly}},$$

because Λ_t is supported on $\overline{B_{\mathbb{R}^d}(0,\eta_{\mathrm{poly}})}$. After decreasing η_{poly} if necessary, the same lower and upper eigenvalue bounds follow.

For part (iii), the Hölder bound makes m uniformly continuous on its open ball, giving the asserted unique continuous extension. Let T_m be the Taylor polynomial of m at zero of degree $p = \lceil \alpha \rceil - 1$. The Hölder remainder bound, extended to the boundary by continuity, gives

$$|m(u) - T_m(u)| \le Ch^\alpha, \qquad \|u\| \le ch.$$

Define the ambient polynomial $P_h(z) := T_m(hU^\top z)$. Then $P_h(z) = \Phi_{D,p}(z)^\top w_h$ for some w_h. For $z = Uu/h$,

$$\Phi_{D,p}(Uu/h)^\top w_h = T_m(u).$$

Also $e_{0,D,p}^\top w_h = P_h(0) = m(0)$. Finally, on any fixed ball $\|z\| \le R_0$, each nonconstant term in P_h contains a factor $h^{|\mu|}$ with $|\mu| \ge 1$, and the derivatives of m are bounded by $L_{\mathcal{F}}$. Hence $\|\nabla_z P_h(z)\| \lesssim h$, with the constant depending on R_0, proving the Lipschitz claim. □

The next lemma identifies the tangent projector from a polynomial graph loss, even after a small deterministic shift and a small random perturbation.

Lemma G.3 (Projector identifiability)**.** *Fix $R > 0$ and $B_\Gamma > 0$. There are constants $\eta_{\mathrm{id}}, C > 0$, depending on $D, d_{\max}, s_{\max}, R, B_\Gamma, L_{\mathcal{M}}$, such that the following holds uniformly over $d \in [d_{\max}]$ and $s \in [s_{\max}]$. Let $\Pi^\circ = U_\circ U_\circ^\top$, where $U_\circ \in \mathbb{R}^{D\times d}$ has orthonormal columns, and let*

$$\Gamma_0(t) := U_\circ t + \sum_{\ell=2}^{s} V_\ell\{t^{\otimes \ell}\}, \qquad \|V_\ell\|_{\mathrm{op}} \le B_\Gamma,$$

with the sum interpreted as absent when $s = 1$. For $\eta \in [0, \eta_{\mathrm{id}}]$ and $\|b\| \le \eta$, let

$$\Lambda : B_{\mathbb{R}^d}(0,R) \times \mathcal{B}(\mathbb{R}^D) \to [0,1]$$

be a probability kernel such that

$$\Lambda_t\big(\{e \in \mathbb{R}^D : \|e\| \le \eta\}\big) = 1 \qquad \text{for every } t \in B_{\mathbb{R}^d}(0,R).$$

Then, for all $q, q^\circ \in \Theta_{d,s}$ whose q°-projector is Π°,

$$\|\Pi - \Pi^\circ\|_{\mathrm{op}} \le C\{\mathcal{D}(q,q^\circ) + \eta\},$$
$$\mathcal{D}(q,q^\circ)^2 := \int_{B_{\mathbb{R}^d}(0,R)} \int_{\mathbb{R}^D} \|R_q(\Gamma_0(t)+b+e) - R_{q^\circ}(\Gamma_0(t)+b+e)\|^2 \, d\Lambda_t(e)\, dt.$$

where Π denotes the projector component of q.

Proof. Write

$$F_0(t) := R_q(\Gamma_0(t)) - R_{q^\circ}(\Gamma_0(t)), \qquad t \in B_{\mathbb{R}^d}(0, R).$$

The map F_0 is a $\mathbb{R}^D$-valued polynomial in t. Its degree is bounded by $s_{\max}^2$. Hence all such maps F_0 belong to a fixed finite-dimensional polynomial space.

Coefficient extraction is a bounded linear map on this fixed-dimensional polynomial space. Consequently,

$$\|D_0 F_0\|_{\mathrm{op}} \le C \left\{ \int_{B_{\mathbb{R}^d}(0,R)} \|F_0(t)\|^2 \, dt \right\}^{1/2}. \tag{20}$$

Since $\Gamma_0(t) = U_\circ t + O(\|t\|^2)$, the terms $T_\ell\{(\Pi z)^{\otimes \ell}\}$, $\ell \ge 2$, do not contribute to the linear coefficient. Thus the degree-one coefficient of $R_q(\Gamma_0(t))$ is $(I - \Pi)U_\circ$, while that of $R_{q^\circ}(\Gamma_0(t))$ is $(I - \Pi^\circ)U_\circ = 0$. Therefore

$$D_0 F_0 = (I - \Pi)U_\circ = (\Pi^\circ - \Pi)U_\circ.$$

For rank-d orthogonal projectors, the principal-angle identity gives

$$\|(\Pi^\circ - \Pi)U_\circ\|_{\mathrm{op}} = \|\Pi - \Pi^\circ\|_{\mathrm{op}}.$$

Combining this with (20) yields

$$\|\Pi - \Pi^\circ\|_{\mathrm{op}} \le C \left\{ \int_{B_{\mathbb{R}^d}(0,R)} \|F_0(t)\|^2 \, dt \right\}^{1/2}. \tag{21}$$

Define

$$F_{b,e}(t) := R_q(\Gamma_0(t) + b + e) - R_{q^\circ}(\Gamma_0(t) + b + e).$$

For all $q, q^\circ \in \Theta_{d,s}$, all $\|t\| \le R$, and all $\|b\|, \|e\| \le \eta_{\mathrm{id}}$,

$$\|F_{b,e}(t) - F_0(t)\| \le C(\|b\| + \|e\|). \tag{22}$$

Indeed, the relevant arguments $\Gamma_0(t) + u$ remain in a fixed Euclidean ball depending only on $R, B_\Gamma, s_{\max}$, and the coefficients of q, q° range over the compact parameter set $\Theta_{d,s}$. The Jacobian in z of $R_q(z) - R_{q^\circ}(z)$ is therefore uniformly bounded on this ball.

Using (22), Minkowski's inequality, and the fact that each Λ_t is a probability measure, we get

$$\left\{ \int_{B_{\mathbb{R}^d}(0,R)} \|F_0(t)\|^2 dt \right\}^{1/2} \le \left\{ \int_{B_{\mathbb{R}^d}(0,R)} \int_{\mathbb{R}^D} \|F_{b,e}(t)\|^2 d\Lambda_t(e) dt \right\}^{1/2} + C\eta.$$

Substitution into (21) proves the lemma. □

G.3 Tangent projector estimation

In this paper, we suppose n is sufficiently large and use the results in Lemma E.1: the tangent-projection argument uses the window conditions

$$h_x \le 1 \wedge R_{\mathrm{loc}}, \quad h_x + \sigma_{k,n} + \sigma_{\max,n} < \delta_{0,n}, \quad \frac{\sigma_{k,n}}{h_x} \le \frac{1}{2\sqrt{D}} \le \frac{1}{2} \tag{23}$$

and the grid conditions

$$\log N_{\mathrm{grid},k,n} + a_{\tan,n} \le cn\pi_{k,n} h_x^{d_x}, \quad \epsilon_{\tan} \le c, \quad \eta_{\mathrm{grid},n} \le c\epsilon_{\tan}^2, \quad \Delta_{\tan,n} \le c\pi_{k,n} h_x^{d_x+2} \epsilon_{\tan}^2. \tag{24}$$

Here $N_{\mathrm{grid},k,n}$ is defined in (16), $\epsilon_{\tan}$ is defined in (19), and $c > 0$ is a fixed small constant. These conditions are verified uniformly over components in the proof of the oracle upper bound below.

Lemma G.4. *For sufficiently large n, if $\|X_i - x\|_1 \le h_x$, then $Z_i = k$. Moreover, on this event,*

$$\|X_i^\star - x^\star\| \le h_x + 2\sigma_{k,n} \le 2R_{\mathrm{loc}}.$$

Proof. If $Z_i = \ell \neq k$ and $\|X_i - x\|_1 \le h_x$, then

$$d(\mathcal{M}_\ell, \mathcal{M}_k) \le \|X_i^\star - x^\star\| \le h_x + \|\xi_i\| + \|\xi_{\mathrm{qry}}\| \le h_x + \sigma_{\max,n} + \sigma_{k,n} < \delta_{0,n},$$

contradicting Assumption 1. Hence $Z_i = k$. The second claim then follows from

$$\|X_i^\star - x^\star\| \le h_x + \|\xi_i\| + \|\xi_{\mathrm{qry}}\| \le h_x + 2\sigma_{k,n} \le 2h_x \le 2R_{\mathrm{loc}}.$$

□

Lemma G.5 (Tangent estimation event)**.** *Under (23)–(24), there is an event $\mathcal{E}_{k,n}^{\tan}(x)$ such that*

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{(\mathcal{E}_{k,n}^{\tan}(x))^c\} \le C\exp(-c_{\mathrm{Ber}}a_{\tan,n}),$$

and on $\mathcal{E}_{k,n}^{\tan}(x)$,

$$\|\bar{P}_x - \Pi^\star\|_{\mathrm{op}} \le C\epsilon_{\tan}.$$

Moreover, if the small constant c in (24) is chosen sufficiently small, then on the same event $\mathcal{E}_{k,n}^{\tan}(x)$,

$$\lambda_{d_x}(\bar{P}_x) - \lambda_{d_x+1}(\bar{P}_x) \ge \frac{3}{4}, \qquad \|\widehat{\Pi}_x - \Pi^\star\|_{\mathrm{op}} \le C\epsilon_{\tan}.$$

Consequently,

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}\|\widehat{\Pi}_x - \Pi^\star\|_{\mathrm{op}}^2 \le C\left\{h_x^{2s_x} + \frac{\sigma_{k,n}^2}{h_x^2} + \frac{\log N_{\mathrm{grid},k,n} + a_{\tan,n}}{n\pi_{k,n}h_x^{d_x}} + \exp(-c_{\mathrm{Ber}}a_{\tan,n})\right\}.$$

Proof. By Lemma G.4, the tangent window contains only observations from component k, and all latent points in the window lie in the tangent-normal chart at $x^\star$.

For such observations, write

$$X_i^\star = x^\star + U_\star u_i + G_{x^\star}(u_i), \qquad t_i := u_i/h_x.$$

On $\{\mathsf{A}_i > 0\}$, $\|X_i - x\|_1 \le h_x$. By Lemma G.4,

$$\|X_i^\star - x^\star\| \le h_x + 2\sigma_{k,n} \le Ch_x.$$

Since $G_{x^\star}(u_i) \in T_{x^\star}\mathcal{M}_k^\perp$,

$$\|u_i\| = \|U_\star^\top(X_i^\star - x^\star)\| \le \|X_i^\star - x^\star\| \le Ch_x.$$

Hence $\|t_i\| \le R_{\tan}$ for a constant $R_{\tan} < \infty$ depending only on the fixed model constants. The constant c_{geo} in R_{loc} is chosen at the outset so that $h_x R_{\tan} \le r_{L_\mathcal{M}}$ whenever $h_x \le R_{\mathrm{loc}}$. Thus the following Taylor expansion is taken inside the tangent-normal chart domain.

For $2 \le \ell \le s_x$, regard $D^\ell G_{x^\star}(0)$ as a symmetric ℓ-linear map from $(\mathbb{R}^{d_x})^\ell$ to $\mathbb{R}^D$, and define

$$V_{\ell,h_x} := \frac{h_x^{\ell-1}}{\ell!}D^\ell G_{x^\star}(0).$$

Since $G_{x^\star}(u) = N_{x^\star}(U_\star u)$, $U_\star$ is an isometry from $\mathbb{R}^{d_x}$ onto $T_{x^\star}\mathcal{M}_k$, and the tangent-normal graph satisfies

$$\|D^\ell N_{x^\star}(0)\|_{\mathrm{op}} \le L_\mathcal{M}, \qquad 2 \le \ell \le s_x,$$

we have $\|D^\ell G_{x^\star}(0)\|_{\mathrm{op}} \le L_\mathcal{M}$, and therefore

$$\|V_{\ell,h_x}\|_{\mathrm{op}} \le \frac{L_\mathcal{M}}{\ell!}h_x^{\ell-1} \le L_\mathcal{M}.$$

The $C^{s_x,1}$ bound on $G_{x^\star}$ gives, uniformly over $\|t\| \le R_{\tan}$,

$$h_x^{-1}G_{x^\star}(h_x t) = \sum_{\ell=2}^{s_x} V_{\ell,h_x}\{t^{\otimes\ell}\} + \mathcal{R}_{h_x}(t), \qquad \|\mathcal{R}_{h_x}(t)\| \le Ch_x^{s_x}. \tag{25}$$

Set $\Gamma_{h_x}(t) := U_\star t + \sum_{\ell=2}^{s_x} V_{\ell,h_x}\{t^{\otimes \ell}\}$. Then

$$z_i = \Gamma_{h_x}(t_i) - \xi_{\mathrm{qry}}/h_x + e_{h_x}(t_i, \xi_i), \qquad \|e_{h_x}(t_i,\xi_i)\| \le C\left(h_x^{s_x} + \frac{\sigma_{k,n}}{h_x}\right).$$

Here $e_{h_x}(t,\xi) \in \mathbb{R}^D$ denotes the sum of the graph Taylor remainder and the rescaled sample noise ξ/h_x. Also $\|\xi_{\mathrm{qry}}/h_x\| \le 1/2$ by (23).

Define multilinear maps $T_\ell^\circ : (\mathbb{R}^D)^\ell \to \mathbb{R}^D$ by

$$T_\ell^\circ(w_1,\dots,w_\ell) := V_{\ell,h_x}(U_\star^\top w_1, \dots, U_\star^\top w_\ell).$$

Since $\|U_\star^\top\|_{\mathrm{op}} = 1$, $\|T_\ell^\circ\|_{\mathrm{op}} \le \|V_{\ell,h_x}\|_{\mathrm{op}} \le L_{\mathcal{M}} \le 2L_{\mathcal{M}}$. Moreover, $T_\ell^\circ\{(U_\star t)^{\otimes\ell}\} = V_{\ell,h_x}\{t^{\otimes\ell}\}$. Put

$$q^\circ = (a^\circ, \Pi^\star, T_2^\circ, \dots, T_{s_x}^\circ), \qquad a^\circ = \frac{(I-\Pi^\star)\xi_{\mathrm{qry}}}{h_x}.$$

Since $\|a^\circ\| \le \sigma_{k,n}/h_x \le 1/2$, this belongs to Θ_{d_x,s_x}. On $\{\mathsf{A}_i > 0\}$,

$$\Pi^\star z_i = U_\star t_i + \frac{\Pi^\star(\xi_i - \xi_{\mathrm{qry}})}{h_x}, \qquad \left\|\frac{\Pi^\star(\xi_i-\xi_{\mathrm{qry}})}{h_x}\right\| \le \frac{2\sigma_{k,n}}{h_x}.$$

The normal query shift in $(I-\Pi^\star)z_i$ is cancelled by a°. The sample perturbation contributes at most $\sigma_{k,n}/h_x$, and the polynomial arguments differ from $U_\star t_i$ by at most $2\sigma_{k,n}/h_x$. Uniform Lipschitz bounds for the fixed-degree polynomial terms on the localized window and (25) yield

$$\mathsf{A}_i > 0 \quad\Longrightarrow\quad \|R_{q^\circ}(z_i)\| \le C\left(h_x^{s_x} + \frac{\sigma_{k,n}}{h_x}\right) = C b_{\tan}. \tag{26}$$

Given this q°, define the loss-induced discrepancy

$$d_{h_x}(q,q^\circ)^2 := \frac{1}{\pi_{k,n} h_x^{d_x}} \mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}\left[\mathbb{I}\{Z_i = k\}\mathsf{A}_i \|R_q(z_i) - R_{q^\circ}(z_i)\|^2\right].$$

Fix a small constant $r_{\mathrm{id}} > 0$, depending only on the fixed regularity constants, and set

$$y_h(t) := x^\star + U_\star h_x t + G_{x^\star}(h_x t), \qquad \|t\| \le r_{\mathrm{id}}.$$

For $\|t\| \le r_{\mathrm{id}}$ and $\|\xi\|, \|\xi_{\mathrm{qry}}\| \le \sigma_{k,n}$,

$$\frac{\|y_h(t) + \xi - (x^\star + \xi_{\mathrm{qry}})\|_1}{h_x} \le C\left(r_{\mathrm{id}} + h_x r_{\mathrm{id}}^2 + \frac{\sigma_{k,n}}{h_x}\right).$$

The constants r_{id} and a_{sc} are fixed at the outset so that the right-hand side is at most a number strictly smaller than one. Since A is bounded below on a fixed compact subinterval of $[0,1)$,

$$\mathsf{A}\left(\frac{\|y_h(t) + \xi - (x^\star + \xi_{\mathrm{qry}})\|_1}{h_x}\right) \ge c.$$

Define the pushforward measure ν_h on $B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{id}})$ by

$$\nu_h(A) := \mathbb{P}\{X_i^\star \in y_h(A) \mid Z_i = k\}$$

for Borel sets $A \subset B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{id}})$. The change of variables $u = h_x t$ in the tangent-normal chart gives

$$\nu_h(A) = \int_A \varrho_{k,n}(y_h(t)) J_{x^\star}(h_x t) h_x^{d_x}\, dt,$$

where

$$J_{x^\star}(u) := \sqrt{\det\{I + DG_{x^\star}(u)^\top DG_{x^\star}(u)\}}.$$

The density lower and upper bounds and the local Jacobian bounds imply that there are constants $0 < c < C < \infty$, independent of $n, k, x^\star, \xi_{\mathrm{qry}}$, such that for every such Borel set A,

$$c h_x^{d_x}|A| \le \nu_h(A) \le C h_x^{d_x}|A|. \tag{27}$$

Let $b = -\xi_{\mathrm{qry}}/h_x$. Since the underlying spaces are standard Borel, choose a regular conditional distribution of ξ_i given $(X_i^\star, Z_i)$, supported on $\{u \in \mathbb{R}^D : \|u\| \le \sigma_{k,n}\}$ at every (y,k) with $y \in \mathcal{M}_k$. Such a version is obtained by assigning the point mass at zero on the $\mu_{k,n}$-null set where the support condition fails. For each $t \in B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{id}})$, let Λ_t be the pushforward of this conditional distribution at $(y_h(t), k)$ under the map $\xi \mapsto e_{h_x}(t, \xi)$. Choose $C_1 \ge 1$ large enough so that

$$\|b\| \le C_1 b_{\tan}, \qquad \Lambda_t\big(\overline{B_{\mathbb{R}^D}(0, C_1 b_{\tan})}\big) = 1 \quad \text{for every } t.$$

Then $\Lambda : B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{id}}) \times \mathcal{B}(\mathbb{R}^D) \to [0,1]$ is a probability kernel. Therefore, using $\mathsf{A}_i \ge c$, $\mathbb{P}(Z_i = k) = \pi_{k,n}$, and (27),

$$\begin{aligned} d_{h_x}(q, q^\circ)^2 &\ge \frac{c}{h_x^{d_x}} \int_{B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{id}})} \int_{\mathbb{R}^D} \|R_q(\Gamma_{h_x}(t) + b + e) - R_{q^\circ}(\Gamma_{h_x}(t) + b + e)\|^2 \, d\Lambda_t(e) \, d\nu_h(t) \\ &\ge c \int_{B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{id}})} \int_{\mathbb{R}^D} \|R_q(\Gamma_{h_x}(t) + b + e) - R_{q^\circ}(\Gamma_{h_x}(t) + b + e)\|^2 \, d\Lambda_t(e) \, dt. \end{aligned} \tag{28}$$

Set $\eta_0 := C_1 b_{\tan}$. Since $b_{\tan} \le \epsilon_{\tan}$, the smallness condition in (24) is chosen so that $\eta_0 \le \eta_{\mathrm{id}}$. Since $\|V_{\ell,h_x}\|_{\mathrm{op}} \le L_{\mathcal{M}}$, Lemma G.3 applies with

$$R = r_{\mathrm{id}}, \qquad B_\Gamma = L_{\mathcal{M}}, \qquad \Gamma_0 = \Gamma_{h_x}, \qquad \eta = \eta_0.$$

Together with (28), it gives

$$\|\Pi - \Pi^\star\|_{\mathrm{op}} \le C\{d_{h_x}(q, q^\circ) + b_{\tan}\}. \tag{29}$$

Local purity, the density upper bound, and the local volume bound in Lemma D.1 give

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}} \mathsf{A}_i \le \mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{\mathsf{A}_i > 0\} \le C\pi_{k,n} h_x^{d_x}.$$

The population quadratic loss is coercive in the same discrepancy. Expanding squares and using this mass bound, (26), Cauchy's inequality, and Young's inequality,

$$L(q; x) - L(q^\circ; x) \ge c\pi_{k,n} h_x^{d_x+2} d_{h_x}(q, q^\circ)^2 - C\pi_{k,n} h_x^{d_x+2} b_{\tan}^2. \tag{30}$$

Conditioning on the query, q° is fixed. For each grid point q, the summands in $L_n(q; x) - L_n(q^\circ; x)$ are uniformly bounded by Ch_x^2. Moreover, on $\{\mathsf{A}_i > 0\}$, both $R_q(z_i)$ and $R_{q^\circ}(z_i)$ are uniformly bounded, and

$$\big|\|R_q(z_i)\|^2 - \|R_{q^\circ}(z_i)\|^2\big| \le C\|R_q(z_i) - R_{q^\circ}(z_i)\|.$$

Together with the definition of d_{h_x} and (26), this gives the variance bound

$$\mathrm{Var}_{k,x^\star,\xi_{\mathrm{qry}}}\left[h_x^2 \mathsf{A}_i \{\|R_q(z_i)\|^2 - \|R_{q^\circ}(z_i)\|^2\}\right] \le C\pi_{k,n} h_x^{d_x+4} \{d_{h_x}(q, q^\circ) + b_{\tan}\}^2.$$

Apply Bernstein's inequality with $u = \log N_{\mathrm{grid},k,n} + a_{\tan,n}$, enlarging the fixed threshold constant as needed. A union bound over the $N_{\mathrm{grid},k,n}$ candidates uses $N_{\mathrm{grid},k,n} e^{-u} = e^{-a_{\tan,n}}$ and gives an event on which, uniformly over $q \in \mathcal{Q}_{d_x,s_x,n}$,

$$|\{L_n(q; x) - L_n(q^\circ; x)\} - \{L(q; x) - L(q^\circ; x)\}| \le C\pi_{k,n} h_x^{d_x+2} \{(d_{h_x}(q, q^\circ) + b_{\tan})\zeta_{\tan} + \zeta_{\tan}^2\}. \tag{31}$$

The same Bernstein argument gives

$$\frac{1}{n} \sum_{i \in [n]} \mathsf{A}_i \le C\pi_{k,n} h_x^{d_x}. \tag{32}$$

Let $\mathcal{E}_{k,n}^{\tan}(x)$ be the intersection of the events in (31) and (32). Under (24),

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{(\mathcal{E}_{k,n}^{\tan}(x))^c\} \le C \exp(-c_{\mathrm{Ber}} a_{\tan,n}).$$

Let $q^\sharp$ be a grid point within $\eta_{\mathrm{grid},n}$ of q°. Since $q \mapsto R_q(z)$ is Lipschitz uniformly on the fixed tangent window,

$$\|R_{q^\sharp}(z_i) - R_{q^\circ}(z_i)\| \le C\eta_{\mathrm{grid},n} \qquad \text{whenever } \mathsf{A}_i > 0.$$

Using (26) and (32),

$$L_n(q^\sharp; x) - L_n(q^\circ; x) \le Ch_x^2 \frac{1}{n} \sum_{i\in[n]} \mathsf{A}_i\{b_{\tan}\eta_{\mathrm{grid},n} + \eta_{\mathrm{grid},n}^2\}$$
$$\le C\pi_{k,n} h_x^{d_x+2}\{b_{\tan}\eta_{\mathrm{grid},n} + \eta_{\mathrm{grid},n}^2\} \le C\pi_{k,n} h_x^{d_x+2}\epsilon_{\tan}^2,$$

where the last inequality uses $\eta_{\mathrm{grid},n} \le c\epsilon_{\tan}^2$ and $\epsilon_{\tan} \le c$.

If q is active, $\omega_q(x) > 0$, then $L_n(q; x) \le L_{\min}(x) + \Delta_{\tan,n}$. Since $L_{\min}(x) \le L_n(q^\sharp; x)$, the preceding display and the condition on $\Delta_{\tan,n}$ imply

$$L_n(q; x) - L_n(q^\circ; x) \le C\pi_{k,n} h_x^{d_x+2}\epsilon_{\tan}^2. \tag{33}$$

Put $D_q := d_{h_x}(q, q^\circ)$ and $\mathfrak{m}_x := \pi_{k,n} h_x^{d_x+2}$. Combining (33), (31), and (30) gives

$$\mathfrak{m}_x D_q^2 \lesssim \mathfrak{m}_x\{b_{\tan}^2 + \epsilon_{\tan}^2 + (D_q + b_{\tan})\zeta_{\tan} + \zeta_{\tan}^2\}.$$

Since $b_{\tan} \le \epsilon_{\tan}$ and $\zeta_{\tan} \le \epsilon_{\tan}$, Young's inequality yields $D_q \le C\epsilon_{\tan}$. Hence, by (29),

$$\|\Pi_q - \Pi^\star\|_{\mathrm{op}} \le C\epsilon_{\tan}$$

for every active grid point q. Averaging over the nonnegative weights gives, on $\mathcal{E}_{k,n}^{\tan}(x)$,

$$\|\bar{P}_x - \Pi^\star\|_{\mathrm{op}} \le C_0\epsilon_{\tan}.$$

Since $\Pi^\star$ is a rank-d_x projector, Weyl's inequality gives

$$\lambda_{d_x}(\bar{P}_x) \ge 1 - C_0\epsilon_{\tan}, \qquad \lambda_{d_x+1}(\bar{P}_x) \le C_0\epsilon_{\tan}.$$

Choosing the small constant in (24) so that $C_0\epsilon_{\tan} \le 1/8$, we get the spectral gap lower bound $3/4$. The Davis–Kahan sin-Θ theorem [38, Theorem 1], originally due to Davis and Kahan [12], then gives

$$\|\widehat{\Pi}_x - \Pi^\star\|_{\mathrm{op}} \le C\|\bar{P}_x - \Pi^\star\|_{\mathrm{op}} \le C\epsilon_{\tan}.$$

For the second moment, on $\mathcal{E}_{k,n}^{\tan}(x)$ the squared error is at most $C\epsilon_{\tan}^2$, while on the complement it is at most one. The displayed probability bound therefore gives

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}\|\widehat{\Pi}_x - \Pi^\star\|_{\mathrm{op}}^2 \lesssim \epsilon_{\tan}^2 + \exp(-c_{\mathrm{Ber}} a_{\tan,n}),$$

which is the asserted bound after expanding $\epsilon_{\tan}$. □

G.4 Regression stability

The regression design argument also uses the window condition in (23).

Lemma G.6 (Uniform regression stability)**.** *Under the standing assumptions, including Assumption 3, and (23), there exist $\delta_\star > 0$ and an event $\mathcal{E}_{k,n}^{\mathrm{reg}}(x)$ such that*

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{(\mathcal{E}_{k,n}^{\mathrm{reg}}(x))^c\} \le C\exp(-cn\pi_{k,n}h_x^{d_x}),$$

and on $\mathcal{E}_{k,n}^{\mathrm{reg}}(x)$, for every rank-$d_x$ projector Π with $\|\Pi - \Pi^\star\|_{\mathrm{op}} \le \delta_\star$,

$$\lambda_{\min}^+(G^\Pi) \ge c\pi_{k,n}, \qquad \|G^\Pi\|_{\mathrm{op}} \le C\pi_{k,n}, \qquad \|(G^\Pi)^\dagger\|_{\mathrm{op}} \le C\pi_{k,n}^{-1},$$
$$e_0 \in \mathrm{Im}(G^\Pi), \qquad e_0^\top (G^\Pi)^\dagger G^\Pi = e_0^\top,$$

and

$$\frac{1}{n}\sum_{i\in[n]} W_i^\Pi \le C\pi_{k,n}, \qquad \frac{1}{n}\sum_{i\in[n]} (W_i^\Pi)^2 \le C\pi_{k,n}h_x^{-d_x}.$$

Moreover, on $\{\|X_i - x\|_1 \le h_x\}$,

$$\|v_i^\Pi - v_i^\circ\| \le C\|\Pi - \Pi^\star\|_{\mathrm{op}}, \qquad |W_i^\Pi - W_i^\circ| \le Ch_x^{-d_x}\|\Pi - \Pi^\star\|_{\mathrm{op}}.$$

Proof. Here and below in this proof, $\mathbb{E}$ denotes $\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}$. By Lemma G.4, the regression window contains only component k and is contained in the local chart at $x^\star$. Fix $r_{\mathrm{in}} > 0$ sufficiently small, depending only on the fixed model constants. With this choice, let η_{poly} be the constant in Lemma G.2(ii). The scale constant a_{sc} is chosen at the outset so that $2a_{\mathrm{sc}} \le \eta_{\mathrm{poly}}$ and so that the cutoff and kernel lower bounds below hold.

For $t \in B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{in}})$, write

$$y_h(t) := x^\star + U_\star h_x t + G_{x^\star}(h_x t).$$

The tangent-normal change of variables gives, for every measurable $A \subset B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{in}})$,

$$\mathbb{P}\{X_i^\star \in y_h(A) \mid Z_i = k\} = \int_A \varrho_{k,n}(y_h(t)) J_{x^\star}(h_x t) h_x^{d_x}\, dt,$$

where $J_{x^\star}(u) := \sqrt{\det\{I + DG_{x^\star}(u)^\top DG_{x^\star}(u)\}}$. The density lower and upper bounds and the local Jacobian bounds imply that there are constants $0 < c < C < \infty$, independent of $n, k, x^\star, \xi_{\mathrm{qry}}$, such that for every Borel $A \subset B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{in}})$,

$$c h_x^{d_x} |A| \le \mathbb{P}\{X_i^\star \in y_h(A) \mid Z_i = k\} \le C h_x^{d_x} |A|.$$

For $X_i^\star = y_h(t)$, define $u_i^\circ := U_\star^\top z_i$. Since $G_{x^\star}(h_x t)$ is normal to $T_{x^\star}\mathcal{M}_k$,

$$u_i^\circ = t + e_h(t, \xi_i), \qquad e_h(t, \xi_i) := \frac{U_\star^\top(\xi_i - \xi_{\mathrm{qry}})}{h_x}.$$

Moreover

$$\|e_h(t, \xi_i)\| \le \frac{2\sigma_{k,n}}{h_x} \le 2a_{\mathrm{sc}} h_x^{(\alpha_k \vee 1) - 1} \le 2a_{\mathrm{sc}}.$$

For $\|t\| \le r_{\mathrm{in}}$ and $\|\xi\|, \|\xi_{\mathrm{qry}}\| \le \sigma_{k,n}$,

$$\frac{\|y_h(t) + \xi - (x^\star + \xi_{\mathrm{qry}})\|_1}{h_x} \le C\left(r_{\mathrm{in}} + h_x r_{\mathrm{in}}^2 + \frac{\sigma_{k,n}}{h_x}\right),$$

and, using $u_i^\circ = t + U_\star^\top(\xi - \xi_{\mathrm{qry}})/h_x$,

$$\|U_\star u_i^\circ\|_1 \le C\left(r_{\mathrm{in}} + \frac{\sigma_{k,n}}{h_x}\right).$$

The constants r_{in} and a_{sc} are fixed so that both right-hand sides are bounded by constants on which A and K are bounded below. Hence

$$\mathsf{A}_i \ge c, \qquad \mathsf{K}(v_i^\circ) = \mathsf{K}(U_\star u_i^\circ) \ge c.$$

Set $\psi_i^\circ := \Phi_{d_x,p_x}(u_i^\circ)$, and define the reduced true-tangent Gram matrix

$$H^\circ := \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{Z_i = k\} h_x^{-d_x} \mathsf{A}_i \mathsf{K}(U_\star u_i^\circ) \psi_i^\circ (\psi_i^\circ)^\top.$$

Use the supported version of the regular conditional distribution of ξ_i given $(X_i^\star, Z_i)$ chosen in the tangent-projector proof. For each $t \in B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{in}})$, let Λ_t be the pushforward of this conditional distribution at $(y_h(t), k)$ under the map $\xi \mapsto \frac{U_\star^\top(\xi - \xi_{\mathrm{qry}})}{h_x}$. Then $\Lambda : B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{in}}) \times \mathcal{B}(\mathbb{R}^{d_x}) \to [0, 1]$ is a probability kernel. Moreover,

$$\Lambda_t\left(\overline{B_{\mathbb{R}^{d_x}}(0, \eta_{\mathrm{poly}})}\right) = 1 \qquad \text{for every } t \in B_{\mathbb{R}^{d_x}}(0, r_{\mathrm{in}}),$$

because

$$\left\|\frac{U_\star^\top(\xi - \xi_{\mathrm{qry}})}{h_x}\right\| \le \frac{\|\xi\| + \|\xi_{\mathrm{qry}}\|}{h_x} \le \frac{2\sigma_{k,n}}{h_x} \le 2a_{\mathrm{sc}} \le \eta_{\mathrm{poly}}.$$

Using the lower local mass bound, the lower bounds on A_i and K, and Lemma G.2(ii) applied to the kernel Λ, for every $a \in \mathbb{R}^{q_{d_x,p_x}}$,

$$a^\top \mathbb{E} H^\circ a \ge c\pi_{k,n} \int_{B_{\mathbb{R}^{d_x}}(0,r_{\mathrm{in}})} \int_{\mathbb{R}^{d_x}} \{a^\top \Phi_{d_x,p_x}(t+u)\}^2 d\Lambda_t(u) dt \ge c\pi_{k,n}\|a\|^2.$$

The upper bound is obtained on the full regression window. If $\mathsf{A}_i > 0$, then $\|X_i - x\|_1 \le h_x$. By Lemma G.4, $Z_i = k$ on this window and $\|X_i^\star - x^\star\| \le Ch_x$. Hence, in the local chart, $\|u_i^\circ\| \le C$, uniformly over all observations contributing to H°. Since A, K, and Φ_{d_x,p_x} are uniformly bounded on this fixed coordinate window, and since the local mass upper bound gives

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{\mathsf{A}_i > 0,\ Z_i = k\} \le C\pi_{k,n}h_x^{d_x},$$

we have $a^\top \mathbb{E}H^\circ a \le C\pi_{k,n}\|a\|^2$. Thus

$$c\pi_{k,n}I \preceq \mathbb{E}H^\circ \preceq C\pi_{k,n}I.$$

Define $C_i^\circ := \mathbb{I}\{Z_i = k\}h_x^{-d_x}\mathsf{A}_i$. Let $\mathcal{E}_{k,n}^{\mathrm{reg}}(x)$ be the event on which

$$c\pi_{k,n}I \preceq H^\circ \preceq C\pi_{k,n}I, \qquad \frac{1}{n}\sum_{i\in[n]} C_i^\circ \le C\pi_{k,n}, \qquad \frac{1}{n}\sum_{i\in[n]} (C_i^\circ)^2 \le C\pi_{k,n}h_x^{-d_x}. \tag{34}$$

Since q_{d_x,p_x} is uniformly bounded, entrywise Bernstein bounds imply operator-norm concentration and give

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{(\mathcal{E}_{k,n}^{\mathrm{reg}}(x))^c\} \le C\exp(-cn\pi_{k,n}h_x^{d_x}).$$

On this event we extend the bounds deterministically from $\Pi^\star$ to all nearby rank-d_x projectors. If $\|\Pi - \Pi^\star\|_{\mathrm{op}} \le \delta_\star$, then on $\{\|X_i - x\|_1 \le h_x\}$,

$$\|v_i^\Pi - v_i^\circ\| = \|(\Pi - \Pi^\star)z_i\| \le C\|\Pi - \Pi^\star\|_{\mathrm{op}},$$

and, since K is Lipschitz,

$$|W_i^\Pi - W_i^\circ| \le Ch_x^{-d_x}\|\Pi - \Pi^\star\|_{\mathrm{op}}.$$

Choose an orthonormal basis U_Π of $\mathrm{Im}(\Pi)$ aligned with $U_\star$. There exists an orthogonal matrix $R_\Pi \in \mathbb{R}^{d_x\times d_x}$ such that

$$\|U_\Pi - U_\star R_\Pi\|_{\mathrm{op}} \le C\|\Pi - \Pi^\star\|_{\mathrm{op}}.$$

Put $u_i^\Pi := U_\Pi^\top z_i$, and define

$$H^\Pi := \frac{1}{n}\sum_{i\in[n]} W_i^\Pi \Phi_{d_x,p_x}(u_i^\Pi)\Phi_{d_x,p_x}(u_i^\Pi)^\top.$$

Let S_{R_Π,p_x} be the fixed finite-dimensional rotation matrix satisfying

$$\Phi_{d_x,p_x}(R_\Pi^\top u) = S_{R_\Pi,p_x}\Phi_{d_x,p_x}(u).$$

Its singular values are uniformly bounded above and below over all orthogonal R_Π. Therefore, with $\widetilde{H}_\Pi^\circ := S_{R_\Pi,p_x}H^\circ S_{R_\Pi,p_x}^\top$, (34) implies $c\pi_{k,n}I \preceq \widetilde{H}_\Pi^\circ \preceq C\pi_{k,n}I$.

Set $\widetilde{\psi}_i^\circ := \Phi_{d_x,p_x}(R_\Pi^\top u_i^\circ)$. By local purity, $\mathbb{I}\{Z_i = k\}W_i^\circ = W_i^\circ$. Hence

$$\widetilde{H}_\Pi^\circ = \frac{1}{n}\sum_{i\in[n]} W_i^\circ \widetilde{\psi}_i^\circ(\widetilde{\psi}_i^\circ)^\top.$$

On the support of the weights,

$$\|u_i^\Pi - R_\Pi^\top u_i^\circ\| = \|(U_\Pi - U_\star R_\Pi)^\top z_i\| \le C\|\Pi - \Pi^\star\|_{\mathrm{op}}.$$

Since

$$W_i^\Pi - W_i^\circ = h_x^{-d_x}\mathsf{A}_i\{\mathsf{K}(v_i^\Pi) - \mathsf{K}(v_i^\circ)\},$$

the Lipschitz property of K, local purity on $\{\mathsf{A}_i > 0\}$, and the bound $\|v_i^\Pi - v_i^\circ\| \le C\|\Pi - \Pi^\star\|_{\mathrm{op}}$ imply

$$|W_i^\Pi - W_i^\circ| \le CC_i^\circ\|\Pi - \Pi^\star\|_{\mathrm{op}}.$$

Using the boundedness and Lipschitzness of the polynomial dictionary,

$$\left\|W_i^\Pi \Phi_{d_x,p_x}(u_i^\Pi)\Phi_{d_x,p_x}(u_i^\Pi)^\top - W_i^\circ\widetilde{\psi}_i^\circ(\widetilde{\psi}_i^\circ)^\top\right\|_{\mathrm{op}} \le CC_i^\circ\|\Pi - \Pi^\star\|_{\mathrm{op}}.$$

Averaging over i and using $n^{-1}\sum_{i\in[n]} C_i^\circ \le C\pi_{k,n}$, we obtain

$$\|H^\Pi - \widetilde{H}_\Pi^\circ\|_{\mathrm{op}} \le C\pi_{k,n}\|\Pi - \Pi^\star\|_{\mathrm{op}}. \tag{35}$$

Taking $\delta_\star > 0$ sufficiently small and using Weyl's inequality yields

$$c\pi_{k,n} I \preceq H^\Pi \preceq C\pi_{k,n} I.$$

Since $v_i^\Pi = U_\Pi u_i^\Pi$ and $\phi_i^\Pi = \Phi_{D,p_x}(U_\Pi u_i^\Pi) = B_{U_\Pi,p_x}\Phi_{d_x,p_x}(u_i^\Pi)$, Lemma G.2(i) gives the spectral, pseudoinverse, and image identities for G^Π.

Finally, $W_i^\Pi \le Ch_x^{-d_x}\mathbb{I}\{\|X_i - x\|_1 \le h_x\}$. By local purity, $\mathsf{A}_i = \mathbb{I}\{Z_i = k\}\mathsf{A}_i$ on the support of the weights. Hence

$$W_i^\Pi \le CC_i^\circ, \qquad (W_i^\Pi)^2 \le Ch_x^{-d_x}C_i^\circ.$$

The displayed bounds for C_i° therefore give

$$\frac{1}{n}\sum_{i\in[n]} W_i^\Pi \le C\pi_{k,n}, \qquad \frac{1}{n}\sum_{i\in[n]}(W_i^\Pi)^2 \le C\pi_{k,n}h_x^{-d_x}.$$

□

G.5 Proof of Theorem 2

Proof. The scale consequences (17)–(18), together with Lemma E.1, verify the window hypotheses (23). We next verify the grid conditions (24). The selected product grids satisfy (16). Since $d \in [d_{\max}]$ and $s \in [s_{\max}]$, their exponents and prefactors are bounded in terms of fixed model parameters, and hence

$$\log N_{\mathrm{grid},k,n} \le C\log(en).$$

Also $a_{\tan,n} \lesssim \log(en)$, while

$$n\pi_{k,n}h_x^{d_x} \ge cn^{\rho_{\tan}}.$$

Hence

$$\log N_{\mathrm{grid},k,n} + a_{\tan,n} \le cn\pi_{k,n}h_x^{d_x}$$

for all sufficiently large n. The condition $\epsilon_{\tan} \le c$ follows from

$$\sigma_{k,n}/h_x \le a_{\mathrm{sc}}h_x^{(\alpha_k\vee 1)-1},$$

the small fixed choice of a_{sc}, and the uniform convergence of $\zeta_{\tan}$ to zero.

Next,

$$\zeta_{\tan}^2 = \frac{\log N_{\mathrm{grid},k,n} + a_{\tan,n}}{n\pi_{k,n}h_x^{d_x}} \ge c\frac{\log(en)}{n},$$

because $n\pi_{k,n}h_x^{d_x} \le n$ and $a_{\tan,n} \asymp \log(en)$. Therefore

$$\eta_{\mathrm{grid},n} = n^{-1} \le c\epsilon_{\tan}^2$$

for all sufficiently large n. Moreover,

$$\pi_{k,n}h_x^{d_x+2}\epsilon_{\tan}^2 \ge \pi_{k,n}h_x^{d_x+2}\zeta_{\tan}^2 = \frac{h_x^2\{\log N_{\mathrm{grid},k,n} + a_{\tan,n}\}}{n} \ge cn^{-1-2\bar{\rho}_h}\log(en).$$

Since $\bar{\rho}_h < 1$,

$$\Delta_{\tan,n} = n^{-3} \le c\pi_{k,n}h_x^{d_x+2}\epsilon_{\tan}^2$$

for all sufficiently large n. Thus (24) holds uniformly over k.

Fix P, f, and condition on

$$(Z_{n+1}, X_{n+1}^\star, \xi_{n+1}) = (k, x^\star, \xi_{\mathrm{qry}}).$$

Use the local notation from Subsection G.1, and set

$$\delta_x := \|\widehat{\Pi}_x - \Pi^\star\|_{\mathrm{op}}.$$

After reducing the small constant in (24), Lemma G.5 implies

$$\mathcal{E}^{\tan}_{k,n}(x) \subset \{\delta_x \le \delta_\star\}.$$

Define

$$\mathcal{E}_{k,n}(x) := \mathcal{E}^{\mathrm{reg}}_{k,n}(x) \cap \{\delta_x \le \delta_\star\}.$$

The preceding inclusion implies $\{\delta_x > \delta_\star\} \subset (\mathcal{E}^{\tan}_{k,n}(x))^c$. Therefore, by Lemmas G.5 and G.6,

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{\mathcal{E}_{k,n}(x)^c\} \lesssim \exp(-cn\pi_{k,n}h_x^{d_x}) + \exp(-c_{\mathrm{Ber}}a_{\tan,n}).$$

For every fixed $M > 0$, with an implicit constant that may depend on M,

$$\exp(-cn\pi_{k,n}h_x^{d_x}) = \exp(-ch_x^{-2\alpha_k}) \lesssim h_x^M.$$

Taking $M = 2\alpha_{\max} + 1$ and using $h_x \le r_n \to 0$ gives

$$\exp(-cn\pi_{k,n}h_x^{d_x}) \le Ch_x^{2\alpha_k+1} = o(h_x^{2\alpha_k}).$$

Moreover, by the definition of $a_{\tan,n}$ and $h_x \ge n^{-\bar\rho_h}$,

$$\exp(-c_{\mathrm{Ber}}a_{\tan,n}) \le n^{-2\alpha_{\max}\bar\rho_h - 3} \le n^{-3}h_x^{2\alpha_k} = o(h_x^{2\alpha_k}).$$

Both little-o bounds are uniform over components, admissible models, and conditioned query values. Thus

$$\mathbb{P}_{k,x^\star,\xi_{\mathrm{qry}}}\{\mathcal{E}_{k,n}(x)^c\} = o(h_x^{2\alpha_k}). \tag{36}$$

On $\mathcal{E}_{k,n}(x)$, write

$$\widehat{v}_i := v_i^{\widehat{\Pi}_x}, \qquad \widehat{W}_i := W_i^{\widehat{\Pi}_x}, \qquad \widehat{\phi}_i := \phi_i^{\widehat{\Pi}_x}, \qquad \widehat{G}_x := G^{\widehat{\Pi}_x}.$$

Define

$$\widehat{b}_x := \frac{1}{n}\sum_{i\in[n]} \widehat{W}_i\widehat{\phi}_i Y_i, \qquad \widehat{w}^{\Pi}_{\mathrm{plug}}(x) := \widehat{G}_x^{\dagger}\widehat{b}_x.$$

The truncated plug-in estimator is

$$\widehat{f}^{\Pi}_{\mathrm{plug}}(x) := (-L_{\mathcal{F}}) \vee \{e_0^\top \widehat{w}^{\Pi}_{\mathrm{plug}}(x)\} \wedge L_{\mathcal{F}}.$$

Lemma G.6 gives

$$\lambda^{+}_{\min}(\widehat{G}_x) \ge c\pi_{k,n}, \qquad \|(\widehat{G}_x)^{\dagger}\|_{\mathrm{op}} \le C\pi_{k,n}^{-1}, \qquad e_0^\top(\widehat{G}_x)^{\dagger}\widehat{G}_x = e_0^\top,$$

as well as the required weight-sum bounds.

Since the truncation map $t \mapsto (-L_{\mathcal{F}}) \vee t \wedge L_{\mathcal{F}}$ is 1-Lipschitz and $f(x^\star) \in [-L_{\mathcal{F}}, L_{\mathcal{F}}]$, it suffices on $\mathcal{E}_{k,n}(x)$ to control the untruncated intercept error.

For observations with $\widehat{W}_i > 0$, Lemma G.4 gives $Z_i = k$, and we may write

$$X_i^\star = x^\star + U_\star u_i + G_{x^\star}(u_i).$$

On the regression support, $\|u_i\| \le Ch_x$ and $\|G_{x^\star}(u_i)\| \le Ch_x^2$. The pullback

$$m_{x^\star}(u) := f(x^\star + U_\star u + G_{x^\star}(u))$$

is uniformly α_k-Hölder on a ball of radius Ch_x. By Lemma G.2(iii), there is $w_x^0 \in \mathbb{R}^{q_{D,p_x}}$ such that

$$e_0^\top w_x^0 = f(x^\star), \qquad \left|m_{x^\star}(u_i) - \Phi_{D,p_x}(U_\star u_i/h_x)^\top w_x^0\right| \le Ch_x^{\alpha_k}.$$

Moreover,

$$\widehat{v}_i - \frac{U_\star u_i}{h_x} = \frac{(\widehat{\Pi}_x - \Pi^\star)(X_i - x)}{h_x} + \frac{\Pi^\star\xi_i}{h_x} - \frac{\Pi^\star\xi_{\mathrm{qry}}}{h_x}.$$

For arbitrary bounded covariate perturbations, the localized window gives

$$\left\|\widehat{v}_i - \frac{U_\star u_i}{h_x}\right\| \le C\left(\delta_x + \frac{\sigma_{k,n}}{h_x}\right).$$

The comparison polynomial in Lemma G.2(iii) is Ch_x-Lipschitz in the normalized coordinates; when $p_x = 0$ it is constant. Thus

$$f(X_i^\star) = \widehat{\phi}_i^\top w_x^0 + r_i, \qquad |r_i| \le C(h_x^{\alpha_k} + h_x\delta_x + \sigma_{k,n}).$$

The small-perturbation condition implies $\sigma_{k,n} \lesssim h_x^{\alpha_k \vee 1} \le h_x^{\alpha_k}$.

Write

$$\widehat{b}_x = \widehat{G}_x w_x^0 + R_x + E_x,$$

where

$$R_x := \frac{1}{n}\sum_{i\in[n]} \widehat{W}_i \widehat{\phi}_i r_i, \qquad E_x := \frac{1}{n}\sum_{i\in[n]} \widehat{W}_i \widehat{\phi}_i \epsilon_i.$$

The range identity yields

$$e_0^\top \widehat{w}_{\mathrm{plug}}^{\Pi}(x) - f(x^\star) = e_0^\top (\widehat{G}_x)^\dagger R_x + e_0^\top (\widehat{G}_x)^\dagger E_x.$$

Using $\|(\widehat{G}_x)^\dagger\|_{\mathrm{op}} \le C\pi_{k,n}^{-1}$, boundedness of the features, and $n^{-1}\sum_{i\in[n]} \widehat{W}_i \le C\pi_{k,n}$,

$$|e_0^\top (\widehat{G}_x)^\dagger R_x| \le C(h_x^{\alpha_k} + h_x\delta_x + \sigma_{k,n}).$$

For the noise term, condition on the sigma-field generated by $(Z_i, X_i^\star, \xi_i)_{i\in[n]}$ and the query variables. The event $\mathcal{E}_{k,n}(x)$, the weights, and the features are measurable with respect to this sigma-field. The response noises are conditionally mean-zero and independent.

If $\widehat{W}_i > 0$, then $\mathsf{A}_i > 0$, and hence $\|X_i - x\|_1 \le h_x$. Therefore $\|z_i\| \le 1$, and since $\widehat{\Pi}_x$ is an orthogonal projector, $\|\widehat{v}_i\| \le 1$. Because $p_x \le p_{\max}$, the polynomial dictionary is bounded on the unit ball, and therefore $\|\widehat{\phi}_i\| \le C$. Therefore,

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}\left[|e_0^\top (\widehat{G}_x)^\dagger E_x|^2 \mathbb{I}_{\mathcal{E}_{k,n}(x)}\right] \le \frac{C}{n\pi_{k,n}h_x^{d_x}}.$$

Indeed, on $\mathcal{E}_{k,n}(x)$, $\|(\widehat{G}_x)^\dagger\|_{\mathrm{op}} \le C\pi_{k,n}^{-1}$, the features are uniformly bounded, and

$$\frac{1}{n}\sum_{i\in[n]} \widehat{W}_i^2 \le C\pi_{k,n}h_x^{-d_x}.$$

More explicitly, with the covariate sigma-field just specified, put

$$a_i = \frac{1}{n} e_0^\top (\widehat{G}_x)^\dagger \widehat{W}_i \widehat{\phi}_i.$$

On $\mathcal{E}_{k,n}(x)$, the feature and pseudoinverse bounds imply

$$\sum_i a_i^2 \le \frac{C}{n^2\pi_{k,n}^2}\sum_i \widehat{W}_i^2 \le \frac{C}{n\pi_{k,n}h_x^{d_x}}.$$

Conditional independence and conditional mean zero eliminate the cross terms, so

$$\mathbb{E}\left[\left(\sum_i a_i\epsilon_i\right)^2 \middle| \text{covariate sigma-field}\right] = \sum_i a_i^2 \mathbb{E}[\epsilon_i^2 \mid \text{covariate sigma-field}] \le \sigma_Y^2 \sum_i a_i^2.$$

Multiplying by the good-event indicator and taking expectations bounds the second moment of $e_0^\top (\widehat{G}_x)^\dagger E_x$ on $\mathcal{E}_{k,n}(x)$ by $C/(n\pi_{k,n}h_x^{d_x})$.

By Lemma G.5,

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}}\delta_x^2 \le C\left\{h_x^{2s_x} + \frac{\sigma_{k,n}^2}{h_x^2} + \frac{\log N_{\mathrm{grid},k,n} + a_{\tan,n}}{n\pi_{k,n}h_x^{d_x}} + e^{-c_{\mathrm{Ber}}a_{\tan,n}}\right\}.$$

Using

$$\sigma_{k,n} \lesssim h_x^{\alpha_k \vee 1}, \qquad (n\pi_{k,n}h_x^{d_x})^{-1} = h_x^{2\alpha_k}, \qquad \log N_{\mathrm{grid},k,n} + a_{\tan,n} \lesssim \log(en),$$

and $h_x \le r_n \le n^{-\rho_h}$, we get

$$h_x^2 \mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}} \delta_x^2 \le C h_x^{2\alpha_k}.$$

The bound $h_x^2 \log(en) \lesssim 1$ absorbs the stochastic tangent factor, and the exponential term is $o(h_x^{2\alpha_k})$ by the choice of $a_{\tan,n}$. Also $\sigma_{k,n}^2 \lesssim h_x^{2\alpha_k}$.

Combining the deterministic and stochastic terms gives

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}} \left[\{\widehat{f}^{\Pi}_{\mathrm{plug}}(x) - f(x^\star)\}^2 \mathbb{I}_{\mathcal{E}_{k,n}(x)} \right] \lesssim h_x^{2\alpha_k} + \frac{1}{n\pi_{k,n} h_x^{d_x}}.$$

On $\mathcal{E}_{k,n}(x)^c$, both $\widehat{f}^{\Pi}_{\mathrm{plug}}(x)$ and $f(x^\star)$ are bounded by $L_{\mathcal{F}}$, and (36) gives an additional $o(h_x^{2\alpha_k})$ contribution. Hence

$$\mathbb{E}_{k,x^\star,\xi_{\mathrm{qry}}} \{\widehat{f}^{\Pi}_{\mathrm{plug}}(x) - f(x^\star)\}^2 \lesssim h_x^{2\alpha_k} + \frac{1}{n\pi_{k,n} h_x^{d_x}} \lesssim (n\pi_{k,n})^{-2\alpha_k/(2\alpha_k+d_k)}.$$

Averaging over Z_{n+1} gives

$$\sup_{f \in \mathcal{H}(\boldsymbol{\alpha},\mathscr{M};L_{\mathcal{F}})} \sup_{P \in \mathcal{P}_\star} \mathbb{E}_{P,f} \left[\{\widehat{f}^{\Pi}_{\mathrm{plug}}(X_{n+1}) - f(X^\star_{n+1})\}^2 \right] \lesssim \mathfrak{R}_n(\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\pi}).$$

The assertion is asymptotic; all constants above are uniform once the feasible-scale assumptions hold. □

H Transformer approximation

We prove Theorem 3 by an explicit compilation of row-wise ReLU networks and attention-based copying and averaging. The embedding dimension is allowed to grow polynomially with n. In particular, the intermediate states of a wide ReLU comparison tree are stored in polynomially many transformer coordinates.

Throughout this appendix the manifolds, component parameters, selector nets, and tangent grids are deterministic structural information encoded in the constructed parameters, independently of the task function f and the particular prompt. The only prompt inputs are $\mathfrak{s} = ((X_i, Y_i)_{i=1}^n; X_{n+1})$. We abbreviate $\mathcal{H} = \mathcal{H}(\boldsymbol{\alpha}, \mathscr{M}; L_{\mathcal{F}})$.

H.1 Architecture-compatible approximation primitives

The following elementary construction supplies all arithmetic approximants, including globally defined versions of the spectral and inverse decoders. It also gives a Lipschitz bound for the approximating networks themselves.

Lemma H.1 (ReLU approximation on a fixed-dimensional compact set)**.** *Fix input and output dimensions m, r. Suppose $\mathcal{K}_n \subset [-n^B, n^B]^m$ is a nonempty compact set and $F_n : \mathcal{K}_n \to [-n^B, n^B]^r$ is L_n-Lipschitz in maximum norm, with $1 \le L_n \le n^B$. For any fixed $E > 0$, there is a globally defined ReLU network $\widetilde{F}_n : \mathbb{R}^m \to [-n^B, n^B]^r$ such that*

$$\sup_{z \in \mathcal{K}_n} \|\widetilde{F}_n(z) - F_n(z)\|_\infty \le n^{-E}.$$

Its depth is at most $C(B+E+1)\log(en)$, its width and parameter magnitudes are at most $n^{C(B+E+1)}$, and its global Lipschitz constant in maximum norm is at most L_n. Here C depends only on m, r.

Proof. Choose an η-net $\{z_1, \dots, z_M\} \subset \mathcal{K}_n$ in maximum norm, where $\eta = n^{-E}/(4L_n)$. A maximal separated set can be chosen with

$$M \lesssim (1 + n^B/\eta)^m \lesssim n^{m(E+2B+1)}.$$

For each output coordinate $a \in [r]$, set

$$\widetilde{F}_{n,a}(z) = -n^B \vee \max_{j \in [M]} \{F_{n,a}(z_j) - L_n \|z - z_j\|_\infty\} \wedge n^B.$$

For $z \in \mathcal{K}_n$, Lipschitz continuity bounds each term by $F_{n,a}(z)$. A net point within η gives a term at least $F_{n,a}(z) - 2L_n\eta$. Clipping preserves this error bound. Absolute values, maxima, minima, and clipping are exact ReLU operations:

$$|t| = \mathrm{ReLU}(t) + \mathrm{ReLU}(-t), \qquad a \vee b = a + \mathrm{ReLU}(b - a).$$

Compute the M cones in parallel and their maximum with a balanced comparison tree. All intermediate widths are $O(mrM)$, the depth is $O(1 + \log M)$, and all scalar coefficients and biases are polynomially bounded. Each cone is globally L_n-Lipschitz; taking maxima and clipping preserves that bound coordinatewise. □

The lemma applies to multiplication, fixed-degree polynomials, reciprocals on polynomially separated positive intervals, and any fixed-dimensional map with polynomial Lipschitz and range bounds. It approximates a map on $\mathcal{K}_n$ and defines a bounded network on all inputs; inverse and spectral targets are evaluated only on their specified domains.

Lemma H.2 (Exact copying and averaging with uniform softmax)**.** *Let $N = n + 1$ tokens contain protected markers $q_i = \mathbb{I}\{i = N\}$ and $s_i = \mathbb{I}\{i \le n\}$. Suppose an r-dimensional vector register v_i satisfies $\max_i \|v_i\|_\infty \le R$ on the input domain, with $R \ge 1$. The architecture in Section 4 can implement either (i) a broadcast of v_N to all rows or (ii) the average $n^{-1}\sum_{j\in[n]} v_j$ stored only in the query row, exactly on that domain. Each operation uses one active attention head in one layer and at most three transformer blocks, at most $2r$ additional workspace coordinates, FFN width at most Cr, and parameter magnitudes at most $C(1 + N + R)$, for a numerical constant C. Other logical registers and markers are preserved.*

Proof. For $|t| \le R$ and $b \in \{0, 1\}$, the binary value gate is

$$\mathsf{Gate}_R(t, b) := \mathrm{ReLU}(t) - \mathrm{ReLU}(t - Rb) - \mathrm{ReLU}(-t) + \mathrm{ReLU}(-t - Rb) = bt.$$

Allocate source and destination banks of width r, disjoint from all protected coordinates. In the update $Z \mapsto Z + \mathrm{FFN}(Z)$, a destination coordinate w can be cleared by adding $-\mathrm{ReLU}(w) + \mathrm{ReLU}(-w) = -w$. Adding the displayed gate at the same time overwrites it by the gated value. Thus one FFN residual update prepares the gated source and clears the destination, using $O(r)$ hidden units. Set its attention value matrices to zero.

For the communication sublayer set $Q = K = 0$. The attention matrix is exactly $N^{-1}\mathbf{1}_N\mathbf{1}_N^\top$, independently of all token values. To broadcast the query, write $q_j v_j$ into the source bank and choose V to route N times that bank into the cleared destination. Every destination row then equals

$$\frac{1}{N}\sum_{j\in[N]} N q_j v_j = v_N.$$

For averaging, prepare $s_j v_j$ and route N/n times the source bank. Every destination row then equals

$$\frac{1}{N}\sum_{j\in[N]} \frac{N}{n} s_j v_j = \frac{1}{n}\sum_{j\in[n]} v_j.$$

This vector has maximum norm at most R. In the following FFN residual update, overwrite the destination by $\mathsf{Gate}_R(w, q_i)$, using the same residual cancellation. The average remains only in the query row. The temporary source bank can be cleared in that update or in one further block with zero value matrices.

All vector coordinates are processed in parallel. Update columns of protected registers are zero, and additional heads have zero value matrices. The two scratch banks and the displayed gates give the stated width bounds; their coefficients are bounded by $C(1 + N + R)$. □

Lemma H.3 (Compilation into residual transformer blocks)**.** *Let a row-wise ReLU network have depth T, maximum hidden or output width W, and scalar parameters bounded by $B \ge 1$. Its input is stored in selected coordinates of each token. Any fixed set of other coordinates can be preserved. The network can be implemented exactly by $T + O(1)$ transformer blocks of the form in Section 4, with all attention value matrices zero, embedding dimension $O(W + r_0)$, feed-forward dimension $O(W + r_0)$, and parameter bound $C(B + 1)$. Here r_0 is the number of input and protected coordinates.*

For $N = n + 1$ tokens with protected sample/query markers, a fixed number J of broadcasts or sample averages of vectors of dimension at most r, bounded by $R \geq 1$ when communicated, can be inserted with at most $3J$ additional blocks by Lemma H.2. The resulting embedding and FFN dimensions are $O(W + r_0 + Jr)$, and parameters are bounded by $C(B + N + R + 1)$. One active head suffices for each communication, independently of n, W, r.

Proof. Allocate two disjoint scratch banks of width W, and alternate their roles as source and destination. Suppose the source bank is v and the current destination bank is w. A hidden layer $v \mapsto \mathrm{ReLU}(Av + b)$ is implemented by setting the FFN output in the destination coordinates to

$$\mathrm{ReLU}(Av + b) - \mathrm{ReLU}(w) + \mathrm{ReLU}(-w).$$

The block residual adds w, so the new destination equals $\mathrm{ReLU}(Av + b)$ exactly. An affine output is handled by $t = \mathrm{ReLU}(t) - \mathrm{ReLU}(-t)$ and the same cancellation of the old destination. Set all update columns of protected coordinates to zero. These are one-hidden-layer FFN updates, with at most a fixed multiple of $W + r_0$ hidden units. Setting all value matrices to zero gives $\mathrm{MHA} \equiv 0$, so $Z + \mathrm{MHA}(Z) = Z$.

For communication, apply Lemma H.2 with separate banks, retaining the network's other state as protected registers. Its gates are exact on the stated bounded domain, and its destination values equal the required broadcast or average. Induction over the row-wise layers and communication stages therefore gives the prescribed computation exactly. Allocating the indicated banks and padding unused heads with zero value matrices yields the resource bounds. □

H.2 Embedding and readout maps

For $d \in [d_{\max}]$, let

$$\mathfrak{C}_d = \{J \subset [D] : |J| = d\}, \qquad \mathfrak{C}_{\mathrm{all}} = \{(d, J) : d \in [d_{\max}], J \in \mathfrak{C}_d\}.$$

Fix an enumeration $\{(d_j, J_j) : j \in [N_{\mathrm{ch}}]\}$ of this list. The numbers

$$q_{\mathrm{tan},\max} = \binom{D + 2s_{\max}}{2s_{\max}}, \qquad q_{\mathrm{reg},\max} = \binom{d_{\max} + p_{\max}}{p_{\max}}, \qquad N_{\mathrm{ch}}$$

are fixed with respect to n. Put $N_n^{\mathrm{end}} = n + 1$.

Embedding. The affine map Emb_n places (X_i, Y_i) in sample row i, places $(X_{n+1}, 0)$ in query row $n + 1$, and writes a constant coordinate $c_i = 1$ and the sample/query markers in Lemma H.2. All other coordinates are zero. The embedding dimension $d_{E,n}$ includes fixed logical registers for the raw data, scale state, tangent summaries, and chartwise regression summaries, together with polynomially many scratch coordinates. Its size is fixed deterministically by the architecture bounds below.

Readout. Reserve two query coordinates (w_j, v_j) per chart branch. Clip them to $\bar{w}_j \in [0, 1]$, $\bar{v}_j \in [-L_{\mathcal{F}}, L_{\mathcal{F}}]$. Using Lemma H.1 for the multiplication map on this fixed rectangle, choose a globally bounded ReLU map M_n with

$$\sup_{(w,v) \in [0,1] \times [-L_{\mathcal{F}}, L_{\mathcal{F}}]} |M_n(w, v) - wv| \leq N_{\mathrm{ch}}^{-1} n^{-E_{\mathrm{read}}}.$$

Define the fixed readout

$$\mathrm{Read}_n(H) = (-L_{\mathcal{F}}) \vee \sum_{j \in [N_{\mathrm{ch}}]} M_n(\bar{w}_j, \bar{v}_j) \wedge L_{\mathcal{F}}.$$

It is uniformly within $n^{-E_{\mathrm{read}}}$ of the clipped exact weighted sum. Moreover, the multiplication map on the rectangle has a fixed Lipschitz constant, so the construction in Lemma H.1 gives

$$\mathrm{Lip}_{\|\cdot\|_{\max}}(\mathrm{Read}_n) \leq C \tag{37}$$

on the entire transformer-output space, uniformly in n. The readout has logarithmic depth, polynomial size, and fixed parameters.

H.3 Structural selection and geometric summaries

For a closed set A, write $A^{\oplus r} = \{x : d(x, A) \le r\}$. The offsets $\mathcal{M}_k^{\oplus \sigma_{k,n}}$ are disjoint because $2\sigma_{\max,n} < \delta_{0,n}$. On the query support define $k_x = \kappa_n(x) = k$ when $x \in \mathcal{M}_k^{\oplus \sigma_{k,n}}$, and set

$$b_n^\circ(x) = (e_{d_x}, e_{p_x}, e_{s_x}, h_x, r_\pi(x)), \qquad r_\pi(x) = \pi_{k_x,n}^{-1/2}.$$

Here the e's are one-hot vectors of fixed lengths. Their values are structural states computed internally.

Lemma H.4 (Polynomial structural bounds and exact selection)**.** *Under Assumptions 1–3, the state $b_n^\circ(x)$ can be computed exactly on the query support by a ReLU network with $O(\log n)$ depth, polynomial width, and polynomially bounded parameters. The following quantities are polynomially bounded:*

$$K_n, \quad \max_k h_k^{-1}, \quad \max_k \pi_{k,n}^{-1/2}, \quad \delta_{0,n}^{-1}, \quad \max_{d,s} |\mathcal{Q}_{d,s,n}|, \quad \Delta_{\tan,n}^{-1}.$$

Proof. Lemma E.1 gives

$$K_n \le n^{1-\kappa}, \quad h_k^{-1} \le n^{\bar\rho_h}, \quad \pi_{k,n}^{-1/2} \le n^{(1-\kappa)/2}, \quad \delta_{0,n}^{-1} \le C n^{\bar\rho_h}.$$

The last inequality follows from $\delta_{0,n} \ge a_{\mathrm{sc}}^{-1} r_n \ge a_{\mathrm{sc}}^{-1} n^{-\bar\rho_h}$. Lemma G.1 and the selected grids in (16) give

$$\max_{d\in[d_{\max}],\, s\in[s_{\max}]} |\mathcal{Q}_{d,s,n}| \le C n^{m_\Theta}, \qquad m_\Theta := D + \max_{d\in[d_{\max}]} d(D-d) + D \sum_{\ell=2}^{s_{\max}} \binom{D+\ell-1}{\ell}.$$

The constant depends only on $D, d_{\max}, s_{\max}, L_{\mathcal{M}}$, and $\Delta_{\tan,n}^{-1} = n^3$.

Lemma E.1 also gives $\delta_{0,n} - 2\sigma_{\max,n} \ge \delta_{0,n}/2$. Thus, for $r_{\mathrm{sel}} := (\delta_{0,n} - 2\sigma_{\max,n})/(16\sqrt{D})$,

$$0 < r_{\mathrm{sel}}^{-1} \le 32\sqrt{D}\, \delta_{0,n}^{-1} \le C n^{\bar\rho_h}.$$

In each offset, fix a maximal set $\mathcal{C}_{k,n}$ whose pairwise maximum-norm distances exceed $r_{\mathrm{sel}}/4$. It is an $r_{\mathrm{sel}}/4$-net. Each offset lies in the fixed cube $[-B_X - 1, B_X + 1]^D$, so the disjoint balls of radius $r_{\mathrm{sel}}/8$ about these centers give

$$\sum_{k\in[K_n]} |\mathcal{C}_{k,n}| \le C K_n (1 + r_{\mathrm{sel}}^{-1})^D \le n^C.$$

For each center put

$$H_c(x) = \left(1 - \frac{2\|x - c\|_\infty}{r_{\mathrm{sel}}}\right)_+, \qquad \chi_k(x) = 0 \vee \left(2 \max_{c\in\mathcal{C}_{k,n}} H_c(x)\right) \wedge 1.$$

If $x \in \mathcal{M}_k^{\oplus \sigma_{k,n}}$, a center within $r_{\mathrm{sel}}/4$ gives $\chi_k(x) = 1$. Every center in a different offset has maximum-norm distance at least $(\delta_{0,n} - 2\sigma_{\max,n})/\sqrt{D} = 16 r_{\mathrm{sel}}$, giving $\chi_\ell(x) = 0$ for $\ell \ne k$. Consequently,

$$\sum_{k\in[K_n]} \chi_k(x)(e_{d_k}, e_{p_k}, e_{s_k}, h_k, \pi_{k,n}^{-1/2}) = b_n^\circ(x)$$

on the query support. The hats and finite maxima are exact ReLU maps; balanced comparison trees have logarithmic depth and polynomial width. Lemma H.3 stores their intermediate values in the polynomial-dimensional scratch banks. □

Tangent moment and loss-weighted projector. For $x = X_{n+1}$, define

$$m_{\tan}^{\mathrm{ora}}(x) = \frac{1}{n} \sum_{i\in[n]} \mathsf{A}\left(\frac{\|X_i - x\|_1}{h_x}\right) \Phi_{D,2s_{\max}}\left(\frac{X_i - x}{h_x}\right).$$

This is a fixed-dimensional vector, bounded by a fixed constant because the cutoff vanishes when $\|(X_i - x)/h_x\|_1 \ge 1$. For every grid candidate q, the polynomial $\|R_q(z)\|^2$ has degree at most

$2s_{\max}$. Write $\|R_q(z)\|^2 = c_q^\top \Phi_{D,2s_{\max}}(z)$, with coefficients uniformly bounded over the parameter grid. Set

$$L_q(m,h) = h^2 c_q^\top m, \qquad t_q(m,h) = \{\Delta_{\tan,n} - (L_q(m,h) - \min_{q'} L_{q'}(m,h))\}_+,$$

$$\bar{P}(m;d,s,h) = \frac{\sum_{q\in\mathcal{Q}_{d,s,n}} t_q(m,h)\Pi_q}{\sum_{q\in\mathcal{Q}_{d,s,n}} t_q(m,h)}.$$

The denominator is at least $\Delta_{\tan,n}$ for every (m,h). On bounded moment domains and $h \in [n^{-\bar{\rho}_h}, 1]$,

$$\mathrm{Lip}(\bar{P}) \le C|\mathcal{Q}_{d,s,n}|/\Delta_{\tan,n} \le n^C. \tag{38}$$

Indeed, finite minima are Lipschitz with the maximum candidate Lipschitz constant, each t_q is uniformly Lipschitz, and $\|\sum_q t_q \Pi_q\|_{\mathrm{op}} \le \sum_q t_q$. Define

$$\bar{P}_x^{\mathrm{ora}} = \bar{P}(m_{\tan}^{\mathrm{ora}}(x); d_x, s_x, h_x), \qquad P_{T,x}^{\mathrm{ora}} = \Pi_{(d_x)}(\bar{P}_x^{\mathrm{ora}}).$$

These are the averaged and spectral projectors in Algorithm 1.

Spectral decoding domain. For each d use the compact set of symmetric matrices

$$\mathcal{K}_d = \{B : \|B\|_{\mathrm{op}} \le 2,\ \lambda_d(B) \ge 2/3,\ \lambda_{d+1}(B) \le 1/3\}.$$

The map $B \mapsto \Pi_{(d)}(B)$ is uniformly Lipschitz on this set. For example, write the projector as a resolvent contour integral along a fixed rectangle with left edge at $1/2$, right edge at 3, and imaginary parts ± 1. Every point of the contour is a fixed positive distance from the spectrum. The resolvent identity bounds the difference of the integrals by $C\|B - B'\|_{\mathrm{op}}$. Lemma H.1 therefore supplies a globally defined, bounded and Lipschitz ReLU spectral decoder, accurate on $\mathcal{K}_d$. Its input and output are symmetrized by $A \mapsto (A + A^\top)/2$. This map is nonexpansive in maximum norm and fixes the target projector, so the decoder remains bounded and Lipschitz with the same accuracy on $\mathcal{K}_d$. Its symmetric output is passed to the safe chart maps below.

Globally defined safe chart maps. Let E_J be the coordinate embedding associated with $J \in \mathfrak{C}_d$, put $A_J(P) = E_J^\top P E_J$, and let $\tau_d = (4|\mathfrak{C}_d|)^{-1}$. Define

$$\vartheta_d(t) = 0 \vee (4|\mathfrak{C}_d| t - 1) \wedge 1.$$

Choose a smooth bounded function $\rho_d : \mathbb{R} \to \mathbb{R}$ equal to $t^{-1/2}$ for $t \ge \tau_d$, smoothly continued below τ_d with bounded derivatives. For every symmetric P in a fixed bounded matrix cube define, by spectral functional calculus,

$$U_J^{\mathrm{safe}}(P) = P E_J \rho_d(A_J(P)), \qquad \omega_J^{\mathrm{safe}}(P) = \frac{\vartheta_d(\det A_J(P))}{\max\{1/2, \sum_{J'\in\mathfrak{C}_d} \vartheta_d(\det A_{J'}(P))\}}. \tag{39}$$

These maps are bounded and Lipschitz on the cube. The determinant is polynomial, the denominator is bounded below, and for symmetric matrices A, B the Frobenius-norm identity in their two orthonormal eigenbases gives

$$\|\rho_d(A) - \rho_d(B)\|_F \le \sup|\rho_d'|\, \|A - B\|_F.$$

Thus (39) also covers perturbed, non-PSD matrices and eigenvalues outside $[0,1]$. All decoder inputs are clipped to the chosen cube before these maps are approximated.

If P is an exact rank-d orthogonal projector, Cauchy–Binet gives $\sum_J \det A_J(P) = 1$. Hence the denominator in (39) is at least one, the weights sum to one, and $\omega_J^{\mathrm{safe}}(P) > 0$ implies $\lambda_{\min}(A_J(P)) \ge \det A_J(P) > \tau_d$. For these charts only,

$$(U_J^{\mathrm{safe}}(P))^\top U_J^{\mathrm{safe}}(P) = I_d, \qquad U_J^{\mathrm{safe}}(P)(U_J^{\mathrm{safe}}(P))^\top = P. \tag{40}$$

Ideal chart states and positive-weight charts. For $d_j = d_x$, set

$$\omega_{j,x}^{\mathrm{ora}} = \omega_{J_j}^{\mathrm{safe}}(P_{T,x}^{\mathrm{ora}}), \qquad U_{j,x}^{\mathrm{ora}} = U_{J_j}^{\mathrm{safe}}(P_{T,x}^{\mathrm{ora}}),$$

with zero columns appended to $U_{j,x}^{\mathrm{ora}}$ when needed. Set both quantities to zero for $d_j \ne d_x$. Define

$$\mathcal{J}_x^+ = \{j : d_j = d_x,\ \omega_{j,x}^{\mathrm{ora}} > 0\}.$$

For every dimension-compatible branch, including those with zero weight, put

$$u_{i,j}^{\mathrm{ora}} = (U_{j,x}^{\mathrm{ora}})^\top (X_i - x)/h_x,$$

$$W_{i,j}^{\mathrm{ora}} = h_x^{-d_x} \mathsf{A}(\|X_i - x\|_1/h_x) \mathsf{K}(U_{j,x}^{\mathrm{ora}} u_{i,j}^{\mathrm{ora}}).$$

For dimension-incompatible branches set $u_{i,j}^{\mathrm{ora}} = 0$ and $W_{i,j}^{\mathrm{ora}} = 0$. Only indices in $\mathcal{J}_x^+$ are required to provide orthonormal coordinates and an invertible regression Gram matrix.

H.4 Implementation of the geometric block

Choose polynomial envelopes

$$R_{h,n} = n^{\bar{\rho}_h}, \quad R_{\pi,n} = n^{(1-\kappa)/2}, \quad R_{W,n} = R_{h,n}^{d_{\max}}, \quad R_{u,n} = C_u R_{h,n},$$

where C_u dominates the fixed safe-chart and observation bounds. All ideal coordinates and weights lie in these envelopes. When invoking an arithmetic approximant, inputs are clipped to its declared box; this is an exact nonexpansive operation that preserves the ideal inputs.

Lemma H.5. *For every $E > 0$, one transformer block stack computes $b_n^\circ(x)$ exactly on the query support and produces $\omega_{j,x}$, $\mathsf{u}_{i,j}(x)$, and $\mathsf{W}_{i,j}(x)$ such that, on the event $\bar{P}_x^{\mathrm{ora}} \in \mathcal{K}_{d_x}$,*

$$\max_j |\omega_{j,x} - \omega_{j,x}^{\mathrm{ora}}| + \max_{i,j} \|\mathsf{u}_{i,j}(x) - u_{i,j}^{\mathrm{ora}}\| + \max_{i,j} |\mathsf{W}_{i,j}(x) - W_{i,j}^{\mathrm{ora}}| \le C n^{-E}.$$

For every observed prompt, all outputs are defined and satisfy

$$0 \le \omega_{j,x} \le 1, \qquad \|\mathsf{u}_{i,j}(x)\|_\infty \le R_{u,n}, \qquad 0 \le \mathsf{W}_{i,j}(x) \le R_{W,n}.$$

Depth is $O(\log n)$; embedding dimension, feed-forward dimension, and parameter magnitudes are polynomially bounded. The head count is fixed.

Proof. First compute the exact selector in Lemma H.4 on the query row, and broadcast $(x, b_n^\circ(x))$ to all sample rows. The row map producing $\mathsf{A}(\|X_i - x\|_1/h_x)\Phi_{D,2s_{\max}}((X_i - x)/h_x)$ has fixed input/output dimensions and polynomial range and Lipschitz bounds on the declared observation and bandwidth domains. Approximate it with Lemma H.1, then average by Lemma H.2. The resulting moment error is bounded by the row-wise approximation error.

The moment-to-$\bar{P}$ map has a polynomial Lipschitz bound by (38); approximate this fixed-dimensional map by the same lemma, separately for each of the finitely many (d, s). One-hot states select the corresponding outputs. On the stated event, the spectral decoder is accurate at the ideal matrix. Its global Lipschitz bound also controls the change caused by the approximate moment and averaged projector, including perturbed inputs outside $\mathcal{K}_d$. The safe chart maps are globally specified on a fixed cube, so the same argument produces the chart states. Clip chart weights to $[0, 1]$ and broadcast the chart matrices.

Finally the row maps producing $u_{i,j}$ and $W_{i,j}$ have fixed input/output dimensions and polynomial Lipschitz constants, since $h_x^{-1} \le R_{h,n}$. Apply the arithmetic lemma and clip the resulting coordinates and weights to their envelopes. For each dimension/degree selector there are only finitely many labels. The exact binary gate in Lemma H.2 sets dimension-incompatible branch states to zero on the declared envelopes.

There are a fixed number of nonlinear module stages and attention communications. Their target maps and the chosen approximants have polynomial Lipschitz bounds on the clipped domains. If each module error is at most $n^{-E'}$, the final maximum error is at most $Cn^{C_{\mathrm{prop}}} n^{-E'}$, with C_{prop} independent of E'. Take $E' > E + C_{\mathrm{prop}} + 1$. The arithmetic lemma and the compilation lemma then give the stated sizes, including scratch space for all intermediate comparison states. Shared row computations on non-query rows are harmless: only marker-selected rows are read in broadcasts, averages, and the final readout. □

H.5 Globally defined regression decoder

A chart branch carries $((u_i, Y_i, W_i)_{i=1}^n; e_d, e_p, r)$, where $r = \pi_{k,n}^{-1/2}$, the coordinates are padded to $d_{\max}$, and $p \le p_{\max}$. Let $\phi_{d,p}(u) = \Phi_{d,p}(u_{1:d})$, and form

$$G = \frac{1}{n} \sum_{i \in [n]} r^2 W_i \phi_{d,p}(u_i) \phi_{d,p}(u_i)^\top, \qquad g = \frac{1}{n} \sum_{i \in [n]} r^2 W_i Y_i \phi_{d,p}(u_i).$$

They have dimension at most $q_{\mathrm{reg,max}}$. Let $\mathcal{B}_n^{\Pi,\mathrm{reg}}$ be the box of branch prompts satisfying

$$\max_i \|u_i\|_\infty \le R_{u,n}, \quad 0 \le W_i \le R_{W,n}, \quad |Y_i| \le B_Y, \quad 0 < r \le R_{\pi,n}.$$

Choose C_G so that $\|G\|_{\mathrm{op}} + \|g\|_2 \le n^{C_G}$ throughout this box. Choose a fixed $\lambda_{\mathrm{reg}} > 0$ so small that the ideal positive-weight chart matrices in the regression design good event have $\lambda_{\min}(G) \ge 4\lambda_{\mathrm{reg}}$.

Indeed, the reduced-Gram bound following (35) in the proof of Lemma G.6 is uniform over nearby projectors and their orthonormal coordinates; normalization by $r^2 = \pi_{k,n}^{-1}$ makes its lower bound a fixed positive constant. Define

$$\mathcal{G}_n^{\Pi,\mathrm{reg}} = \{s^{\mathrm{reg}} \in \mathcal{B}_n^{\Pi,\mathrm{reg}} : \lambda_{\min}(G) \geq \lambda_{\mathrm{reg}}\},$$

and define the exact local polynomial value only on this set by

$$f_{\mathrm{LP}}(s^{\mathrm{reg}}) = (-L_{\mathcal{F}}) \vee e_{0,d,p}^{\top} G^{-1} g \wedge L_{\mathcal{F}}.$$

Lemma H.6. *For any fixed $E > 0$, row-wise ReLU computations, one uniform-softmax averaging stage from Lemma H.2, and a globally defined ReLU decoder produce a branch value $V_n(s^{\mathrm{reg}}) \in [-L_{\mathcal{F}}, L_{\mathcal{F}}]$ on the entire box $\mathcal{B}_n^{\Pi,\mathrm{reg}}$, and*

$$\sup_{s^{\mathrm{reg}} \in \mathcal{G}_n^{\Pi,\mathrm{reg}}} |V_n(s^{\mathrm{reg}}) - f_{\mathrm{LP}}(s^{\mathrm{reg}})| \leq C n^{-E}.$$

Depth is $O(\log n)$ and all dimensions and parameters are polynomially bounded. Inverse evaluation is restricted to well-conditioned branches.

Proof. Approximate the sample-wise entries of $r^2 W_i \phi_{d,p}(u_i)\phi_{d,p}(u_i)^{\top}$ and $r^2 W_i Y_i \phi_{d,p}(u_i)$ to error $n^{-E'}$ by Lemma H.1, and aggregate their averages by Lemma H.2. These are fixed-degree maps in a fixed number of coordinates on polynomial boxes.

On the compact set

$$\mathcal{K}_{d,p,n}^{\mathrm{inv}} = \{(G, g) : G = G^{\top},\ G \succeq \lambda_{\mathrm{reg}} I,\ \|G\|_{\mathrm{op}} + \|g\|_2 \leq n^{C_G}\},$$

the map $(G, g) \mapsto (-L_{\mathcal{F}}) \vee e_0^{\top} G^{-1} g \wedge L_{\mathcal{F}}$ has polynomial Lipschitz constant. Indeed,

$$G^{-1} - G'^{-1} = G^{-1}(G' - G)G'^{-1}, \qquad \|G^{-1}\|_{\mathrm{op}}, \|(G')^{-1}\|_{\mathrm{op}} \leq \lambda_{\mathrm{reg}}^{-1}.$$

Apply Lemma H.1 to this compact-set map and clip its output to $[-L_{\mathcal{F}}, L_{\mathcal{F}}]$. The resulting network is defined on all inputs, including singular matrices. Its error at the ideal summary is small, and its global polynomial Lipschitz bound controls the approximate-summary error. Taking E' sufficiently large gives the claim. Finitely many dimension/degree cases are selected using exact one-hot gating. Compile these ReLU networks by Lemma H.3. □

Lemma H.7. *Consider two branch prompts in $\mathcal{G}_n^{\Pi,\mathrm{reg}}$ with the same (d, p, r, Y_i) but coordinates and weights (u_i, W_i) and (u_i', W_i'). Then*

$$|f_{\mathrm{LP}}(s) - f_{\mathrm{LP}}(s')| \leq C n^C \left(\max_i \|u_i - u_i'\| + \max_i |W_i - W_i'|\right).$$

Proof. The sample-wise polynomial summary maps are polynomially Lipschitz on the branch boxes. Averaging preserves their maximum-error bound. Apply the inverse identity in the preceding proof and the nonexpansiveness of clipping. □

Compact localization through value coordinates. The cutoff A and local weights enter FFN-computed value coordinates before uniform-softmax averaging. The ideal localized value is zero outside the cutoff, although every attention probability is positive. Compact localization thus acts through values, with approximate contributions controlled by the module error bounds.

Remark H.1 (Common normalization of local regression weights)**.** Let $W_i \geq 0$, $Z_W := \sum_{i\in[n]} W_i > 0$, and let ϕ_i be the local polynomial feature vectors. Define

$$G = \sum_{i\in[n]} W_i \phi_i \phi_i^{\top}, \qquad b = \sum_{i\in[n]} W_i \phi_i Y_i, \qquad a_i = \frac{W_i}{Z_W}.$$

The normalized statistics satisfy

$$\widetilde{G} = \sum_{i\in[n]} a_i \phi_i \phi_i^{\top} = Z_W^{-1} G, \qquad \widetilde{b} = \sum_{i\in[n]} a_i \phi_i Y_i = Z_W^{-1} b, \qquad \widetilde{G}^{\dagger}\widetilde{b} = G^{\dagger} b.$$

Indeed, scaling a positive semidefinite matrix by $c > 0$ scales its nonzero eigenvalues by c, so $(cG)^{\dagger} = c^{-1} G^{\dagger}$. Thus common positive normalization preserves the exact weighted least-squares coefficient and its clipped intercept, including when G is singular. In our construction, common factors such as $n^{-1} r_{\pi}^2$ enter both summaries; their averages are computed directly by Lemma H.2.

H.6 Model size and error propagation

All selectors and approximants above are finite networks, chosen independently of the particular prompt. Their row-wise inputs and outputs have fixed dimensions, but their intermediate widths can grow. The exact selector has polynomially many centers; the tangent grids and all normalization factors have polynomial bounds by Lemma H.4. The arithmetic approximants operate on polynomial boxes, with fixed input/output dimensions and polynomial Lipschitz constants.

Communication acts only on logical registers: raw observations, structural states, clipped geometric/chart states, and moment or regression summaries. The observation bounds, exact selector bounds, and declared module boxes give a deterministic envelope $R_n \le n^C$ for all these registers on every admissible observed prompt, including outside the statistical good event. This exponent is independent of the requested approximation accuracy: the arithmetic lemma clips outputs to the fixed target range. The intermediate states of wide comparison networks remain in local scratch banks. Lemma H.2 therefore applies at every communication stage with this polynomial envelope, preserving exact markers and all other logical registers.

More explicitly, for a prescribed final exponent $A_0 > 0$, choose common module accuracy n^{-E_*} with $E_* > A_0 + 2C_{\mathrm{prop}} + 10$, increasing it to cover the regression summary stability constant if necessary. There are a fixed number of module stages, so their propagated scalar error is $O(n^{-A_0/2})$. Here C_{prop} is computed from the target maps and the Lipschitz-preserving approximants. It is independent of the requested accuracy exponent. Table 3 summarizes the domains.

Table 3: Domains and stability bounds for the constructed modules.

Module	Input/range control	Stability used
Structural selector	Fixed observation cube; polynomially many centers	Exact ReLU construction
Tangent moment	$h^{-1} \le n^{\bar{\rho}_h}$; fixed moment dimension	Polynomial Lipschitz bound
Moment-to-projector	Bounded moments; denominator $\ge n^{-3}$	(38)
Spectral/chart maps	Fixed compact spectral domain and chart cube	Fixed Lipschitz bounds; global decoders
Regression summaries	Polynomial coordinate and weight boxes	Fixed-degree polynomial maps
Inverse decoder	$G \succeq \lambda_{\mathrm{reg}} I$ on comparison domain	Polynomial Lipschitz bound
Readout	Weights in $[0,1]$, values in $[-L_{\mathcal{F}}, L_{\mathcal{F}}]$	Uniform global Lipschitz bound

Let W_n be the maximum total simultaneous width of the ReLU modules, including all parallel chart branches and selector comparison states. Lemma H.1 gives $W_n \le n^{C_*}$ for a fixed C_* depending on A_0 and fixed model constants. Choose

$$d_{E,n} = C_0(W_n + r_0), \qquad d_{\mathrm{FFN},n} = C_1(W_n + r_0),$$

where r_0 is the fixed logical-register count and the constants provide two ReLU scratch banks, protected coordinates, and the communication banks in Lemmas H.2 and H.3.

The communicated dimensions and the number of stages are fixed, so these additional banks have fixed total width. Communication parameters are bounded by $C(1 + n + R_n)$. The widths can be padded to deterministic polynomial envelopes common to the whole model class.

Concatenating the fixed number of compiled modules and communication blocks gives

$$N_n^{\mathrm{end}} = n + 1, \quad M_0 = O(1), \quad L_n \le C_{\mathrm{depth}} \log(en), \quad d_{E,n} + d_{\mathrm{FFN},n} + B_n \le n^{C_{\mathrm{size}}}. \quad (41)$$

The internal regression dictionary has size $q_{\mathrm{reg,max}}$, whereas geometric registers and polynomial size exponents may depend on D. This is a constructive existence bound with polynomial workspace and resource exponents that may depend on the ambient dimension.

H.7 End-to-end approximation

Let $s_{x,j}^{\mathrm{ora}}$ and $s_{x,j}^{\mathrm{gen}}$ denote the ideal and generated branch prompts, respectively. For $j \in \mathcal{J}_x^+$ on the regression good event, define

$$f_{\mathrm{chart}}^{\mathrm{ora}}(x) = \sum_{j \in \mathcal{J}_x^+} \omega_{j,x}^{\mathrm{ora}} f_{\mathrm{LP}}(s_{x,j}^{\mathrm{ora}}).$$

This sum evaluates inverses only on positive-weight charts.

Lemma H.8 (Chart-intercept equivalence)**.** *If the ideal positive-weight branch prompts lie in $\mathcal{G}_n^{\Pi,\mathrm{reg}}$, then $f_{\mathrm{chart}}^{\mathrm{ora}}(x) = \widehat{f}_{\mathrm{plug}}^{\Pi}(x)$.*

Proof. For $j \in \mathcal{J}_x^+$, (40) implies $U_{j,x}^{\mathrm{ora}}(U_{j,x}^{\mathrm{ora}})^\top = P_{T,x}^{\mathrm{ora}}$, so the ambient projected weights are exactly the ideal branch weights. Let $B = B_{U,p}$ from Lemma G.2(i). Write the unnormalized reduced Gram and response vector as H, c. Then the ambient quantities are $G_{\mathrm{amb}} = BHB^\top$ and $b_{\mathrm{amb}} = Bc$, where $H \succ 0$ and B has full column rank. For $w = G_{\mathrm{amb}}^\dagger b_{\mathrm{amb}}$, the fact $b_{\mathrm{amb}} \in \mathrm{Im}(G_{\mathrm{amb}})$ gives

$$BHB^\top w = Bc \quad \Longrightarrow \quad B^\top w = H^{-1}c.$$

Since $e_{0,D,p} = Be_{0,d,p}$,

$$e_{0,D,p}^\top w = e_{0,d,p}^\top H^{-1} c.$$

The normalization r_π^2 cancels between the branch Gram and response vector. Thus every positive-weight branch has the same clipped intercept. Their weights sum to one, proving the assertion. □

Lemma H.9. *For every fixed $A_0 > 0$, the geometric approximations can be chosen so that there is an event $\mathcal{A}_n$ with $\mathbb{P}(\mathcal{A}_n^c) \le Cn^{-A_0}$, uniformly over P, f, on which $\bar{P}_x^{\mathrm{ora}} \in \mathcal{K}_{d_x}$, the regression design good event holds, and*

$$s_{x,j}^{\mathrm{ora}},\ s_{x,j}^{\mathrm{gen}} \in \mathcal{G}_n^{\Pi,\mathrm{reg}} \quad \textit{for every } j \in \mathcal{J}_x^+.$$

All generated branch prompts, including zero-weight charts, lie in $\mathcal{B}_n^{\Pi,\mathrm{reg}}$; the eigenvalue lower bound applies only to branches in $\mathcal{J}_x^+$.

Proof. Use the proof-level tangent concentration parameter

$$a_{\tan,n}^{\mathrm{app}} = \frac{A_0 + 2\alpha_{\max}\bar{\rho}_h + 3}{c_{\mathrm{Ber}}} \log(en).$$

The estimator's grid and tolerance remain n^{-1} and n^{-3}; the displayed tail parameter is used only in the concentration proof. The window and grid conditions in Appendix G.5 remain valid. Lemma G.5, with the fixed small scale constant chosen there, gives

$$\lambda_{d_x}(\bar{P}_x^{\mathrm{ora}}) \ge 7/8, \qquad \lambda_{d_x+1}(\bar{P}_x^{\mathrm{ora}}) \le 1/8,$$

and a projector error at most the regression-stability radius, except on an event of probability Cn^{-A_0}. The regression failure probability is at most $C\exp(-cN_k h_k^{d_k}) \le Cn^{-A_0}$, uniformly over k.

On the intersection, (40) makes each $j \in \mathcal{J}_x^+$ an orthonormal coordinate representation of the stable projected design. The reduced-Gram bound following (35) in Lemma G.6, together with the normalization $r^2 = \pi_{k,n}^{-1}$, gives minimum eigenvalue at least $4\lambda_{\mathrm{reg}}$ uniformly over these charts. The coordinate and weight error in Lemma H.5, propagated through the polynomial summary map, makes the generated Gram error at most λ_{reg} when the approximation exponent is large enough. Hence both branch prompts belong to the required good set. This argument is uniform as the positive weight approaches zero: positivity still implies the fixed determinant threshold $\det A_J > \tau_d$. The box assertions hold for every prompt by clipping. There are finitely many chart branches, and averaging the uniform conditional failure bounds over the query component yields the assertion. □

H.8 Proof of Theorem 3

Proof. Apply Lemmas H.5 and H.6 with sufficiently large accuracy exponents, and compose their transformer blocks as in (41). Their communication stages belong directly to $\mathcal{T}_{M_0}$ by Lemma H.2; all remaining stages use zero attention value matrices and ReLU FFNs with block residual connections. Before the readout arithmetic approximation, the predictor is the clipped version of

$$S_{\mathrm{gen}} = \sum_{j\in[N_{\mathrm{ch}}]} \omega_{j,x} V_n(s^{\mathrm{gen}}_{x,j}).$$

On $\mathcal{A}_n$, Lemma H.8 and an exact add-and-subtract decomposition give

$$\begin{aligned}|S_{\mathrm{gen}} - \widehat{f}^{\Pi}_{\mathrm{plug}}(x)| \le &\sum_{j\in\mathcal{J}_x^+} \omega^{\mathrm{ora}}_{j,x} |V_n(s^{\mathrm{gen}}_{x,j}) - f_{\mathrm{LP}}(s^{\mathrm{gen}}_{x,j})| \\ &+ \sum_{j\in\mathcal{J}_x^+} \omega^{\mathrm{ora}}_{j,x} |f_{\mathrm{LP}}(s^{\mathrm{gen}}_{x,j}) - f_{\mathrm{LP}}(s^{\mathrm{ora}}_{x,j})| \\ &+ L_{\mathcal{F}} \sum_{j\in[N_{\mathrm{ch}}]} |\omega_{j,x} - \omega^{\mathrm{ora}}_{j,x}|.\end{aligned}$$

The first two terms use only positive-weight charts and are controlled by Lemmas H.6 and H.7. The final sum controls all branches, including generated leakage onto an oracle-zero-weight chart; its decoder is bounded even when its Gram matrix is singular.

Choosing the module exponents and E_{read} sufficiently large makes the squared error on $\mathcal{A}_n$ at most Cn^{-A_0}. On its complement both final predictions are clipped to $[-L_{\mathcal{F}}, L_{\mathcal{F}}]$, giving at most $4L_{\mathcal{F}}^2\mathbb{P}(\mathcal{A}_n^c) \le Cn^{-A_0}$. Consequently,

$$\sup_{P\in\mathcal{P}_\star} \sup_{f\in\mathcal{H}} \mathbb{E}_{P,f}\{\mathrm{Read}_n \circ \mathrm{TF}_{\widehat{\theta}_n} \circ \mathrm{Emb}_n(\mathfrak{s}) - \widehat{f}^{\Pi}_{\mathrm{plug}}(X_{n+1})\}^2 \le Cn^{-A_0}.$$

The parameter vector depends only on the fixed structural model and n, independently of P's allowed within-component density or perturbation law, the prompt, and f. The structural selector and uniform bounds ensure the same constructed comparator works for all such P, f. □

H.9 Explicit dependence on the approximation exponent

Corollary H.1 (Accuracy–resource bound)**.** *The construction of Theorem 3 can be chosen so that, for fixed positive coefficients $c_{L,0}, c_{L,1}, c_{S,0}, c_{S,1}$ and every fixed $A_0 > 0$, all sufficiently large n satisfy*

$$L_n \le (c_{L,0} + c_{L,1}A_0)\log(en), \qquad d_{E,n} + d_{\mathrm{FFN},n} + B_n \le n^{c_{S,0}+c_{S,1}A_0}.$$

The coefficients depend only on the fixed model bounds and the finite module pattern, independently of A_0, n, K_n, P, f. The large-n threshold may depend on A_0. The head count is fixed and $N_n^{\mathrm{end}} = n+1$.

Proof. There are finitely many nonlinear module families. Their input and output dimensions, say m_j, r_j, depend only on $D, d_{\max}, \alpha_{\max}$. Choose one exponent $B_* > 0$ bounding their target domains, ranges, and Lipschitz constants by powers of n. This exponent is independent of the requested accuracy: all bandwidth inverses, component-mass normalizers, grid cardinalities, and the inverse near-minimizer tolerance are fixed by the structural scale conditions. The compact spectral domain and positive-chart conditioning constants are also independent of A_0.

The approximants in Lemma H.1 have the same Lipschitz bounds as their target maps. Therefore composition over the fixed module pattern amplifies a common absolute module error η by at most $Cn^{C_*}\eta$, where C_* is independent of the internal accuracy. This includes the polynomial stability of the regression summaries and decoder. Exact one-hot gating, coordinate clipping, and averaging have polynomial bounds independent of the approximation exponent as well. Choose the internal exponent as $E_* = A_0/2 + C_* + c_*$, with a fixed c_* large enough to cover fixed factors and all output coordinates. On the good event this makes the squared error $O(n^{-A_0})$. Choosing the proof-level tail parameter proportional to $(A_0+1)\log(en)$ makes the failure probability $O(n^{-A_0})$ while preserving the estimator, its grid, and its network modules.

For module j, the net in Lemma H.1 has at most $C_j n^{m_j(E_*+2B_*+1)}$ centers. Its depth is bounded by $C_j(B_*+E_*+1)\log(en)$, and its width and scalar parameters by $n^{C_j(B_*+E_*+1)}$, after increasing fixed coefficients. The exact structural selector has a polynomial number of centers with an exponent independent of A_0. The number of chart branches and nonlinear module stages is fixed. Consequently their maximum simultaneous width, combined depth, and parameter magnitudes have exponents or depth constants affine in A_0, after taking maxima and sums over a fixed list. The explicit compilation in Lemma H.3 allocates two scratch banks of that width. Lemma H.2 adds at most three blocks per communication and fixed-width logical-register banks. Its coefficients are bounded by $C(1+n+R_n)$, with exponent independent of A_0. Absorbing fixed multiplicative factors for sufficiently large n proves the stated bounds. □

The conclusion concerns inverse-polynomial accuracy for each fixed A_0 and all sufficiently large n, with logarithmic depth sufficient.

I In-context generalization

This appendix proves the in-context generalization result in Theorem 4. We first bound the covering number of the full bounded-parameter softmax transformer class and then apply a bounded square-loss oracle inequality to its near empirical-risk minimizers. The resulting entropy term is polynomial in n and logarithmic in the number of training tasks Γ. Covering-number arguments for attention and transformer classes also appear in [14, 19, 30].

I.1 Covering number and entropy bound

We use the softmax blocks of Section 4, with a residual connection around each attention and ReLU FFN map. Every entry of the query, key, value, and FFN matrices and biases varies freely in $[-B, B]$. For $B, R \geq 1$, put

$$\mathcal{Z}_N(R) := \{H \in \mathbb{R}^{N\times d_E} : \|H\|_{\max} \leq R\}.$$

Let Read have a Lipschitz bound $L_{\mathrm{Read}} \geq 1$, with respect to maximum norm, on the transformer-output domain. Define

$$\mathcal{F}_{\mathrm{Read}} := \{H \mapsto \mathrm{Read}(\mathrm{TF}_\theta(H)) : \mathrm{TF}_\theta \in \mathcal{T}_M(N, d_E, d_{\mathrm{FFN}}, L, B)\},$$
$$\|g\|_{\infty,R} := \sup_{H\in\mathcal{Z}_N(R)} |g(H)|.$$

We construct a parameter-sensitivity bound $\mathfrak{L}_L(R)$ such that

$$\sup_{H\in\mathcal{Z}_N(R)} \|\mathrm{TF}_\theta(H) - \mathrm{TF}_{\theta'}(H)\|_{\max} \leq \mathfrak{L}_L(R)\|\theta-\theta'\|_\infty$$

for all parameter vectors in the full cube.

Lemma I.1 (Entropy of bounded softmax transformers)**.** *For every $0 < \delta \leq 1$,*

$$\log \mathcal{N}(\mathcal{F}_{\mathrm{Read}}, \delta, \|\cdot\|_{\infty,R}) \leq P_{\mathrm{TF}} \log\left(1 + \frac{CBL_{\mathrm{Read}}\mathfrak{L}_L(R)}{\delta}\right),$$

where the number of trainable scalar parameters satisfies

$$P_{\mathrm{TF}} = L\{3Md_E^2 + 2d_E d_{\mathrm{FFN}} + d_{\mathrm{FFN}} + d_E\} \leq CL(M+1)(d_E+d_{\mathrm{FFN}})^2.$$

If $M = O(1)$, $L \leq C_L\log(en)$, $B + R + d_E + d_{\mathrm{FFN}} + L_{\mathrm{Read}} \leq n^{C_{\mathrm{arch}}}$, and $N \leq C_N n$, then the output radius and parameter-sensitivity bound can be chosen so that

$$\sup_{\theta\in[-B,B]^{P_{\mathrm{TF}}}} \sup_{H\in\mathcal{Z}_N(R)} \|\mathrm{TF}_\theta(H)\|_{\max} \leq R_L,$$
$$\log(1+R_L) \leq C\log^2(en), \qquad \log\{1+\mathfrak{L}_L(R)\} \leq C\log^3(en).$$

Consequently, for constants $C, c_{\mathrm{cov}} > 0$ depending only on $C_L, C_N, C_{\mathrm{arch}}$, the fixed head-count bound, and fixed architecture choices,

$$\log \mathcal{N}(\mathcal{F}_{\mathrm{Read}}, \delta, \|\cdot\|_{\infty,R}) \leq Cn^{c_{\mathrm{cov}}}\{1+\log(1/\delta)\}.$$

All bounds hold uniformly over the entire parameter cube.

Proof. For $z \in \mathbb{R}^N$, write $p = \mathrm{softmax}(z)$. Its derivative in direction $v \in \mathbb{R}^N$ is

$$(D\,\mathrm{softmax}(z)v)_j = p_j \left(v_j - \sum_{a\in[N]} p_a v_a \right).$$

Since $p_j \geq 0$ and $\sum_{j\in[N]} p_j = 1$,

$$\|D\,\mathrm{softmax}(z)v\|_1 \leq \sum_{j\in[N]} p_j \left(|v_j| + \left| \sum_{a\in[N]} p_a v_a \right| \right) \leq 2\|v\|_\infty.$$

Integration along the segment between any two logits gives the global bound

$$\|\,\mathrm{softmax}(z) - \mathrm{softmax}(z')\|_1 \leq 2\|z - z'\|_\infty. \tag{42}$$

The bound is independent of both N and the sizes of the logits.

We next control hidden states. Suppose the input to block $\ell + 1$ has maximum norm at most R_ℓ, with $R_0 = R$. For one head, let

$$S(H; Q, K) := \frac{HQ(HK)^\top}{\sqrt{d_E}}, \qquad P(H; Q, K) := \mathrm{softmax}_{\mathrm{row}}(S(H; Q, K)).$$

Each row of P is a probability vector, so

$$\|P(H; Q, K)HV\|_{\max} \leq \|HV\|_{\max} \leq d_E B R_\ell.$$

Thus $H + \mathrm{MHA}(H)$ has radius at most $\widetilde{R}_\ell = (1 + Md_E B)R_\ell$. For $\|Z\|_{\max} \leq S$, the state $Z + \mathrm{FFN}(Z)$ has radius at most $(1 + d_E d_{\mathrm{FFN}} B^2)S + d_{\mathrm{FFN}} B^2 + B$. We may therefore choose

$$R_{\ell+1} = (1 + d_E d_{\mathrm{FFN}} B^2)(1 + Md_E B)R_\ell + d_{\mathrm{FFN}} B^2 + B. \tag{43}$$

These bounds apply to all parameter choices and all inputs in $\mathcal{Z}_N(R)$.

For the sensitivity calculation, take two block inputs H, H' with maximum norm at most R_ℓ, and put $\Delta = \|H - H'\|_{\max}$. Let η bound the maximum difference between corresponding parameters of the two blocks. Direct matrix multiplication gives

$$\|HQ - H'Q'\|_{\max} \leq d_E(B\Delta + R_\ell\eta), \qquad \|HV - H'V'\|_{\max} \leq d_E(B\Delta + R_\ell\eta),$$

with the same bound for the key projections. Since every projected entry has absolute value at most $d_E B R_\ell$, the scaled logits satisfy

$$\|S(H; Q, K) - S(H'; Q', K')\|_{\max} \leq 2d_E^{5/2}\{B^2 R_\ell \Delta + BR_\ell^2 \eta\}.$$

For each output row, apply (42) to its logits and use the row sum of P to bound the value difference. This yields

$$\begin{aligned} &\|P(H; Q, K)HV - P(H'; Q', K')H'V'\|_{\max} \\ &\quad \leq d_E(B\Delta + R_\ell\eta) + 2d_E B R_\ell \|S(H; Q, K) - S(H'; Q', K')\|_{\max}. \end{aligned}$$

Set $\widetilde{H} = H + \mathrm{MHA}_\theta(H)$ and $\widetilde{H}' = H' + \mathrm{MHA}_{\theta'}(H')$. Summing over heads and including the residual bounds $\widetilde{\Delta} := \|\widetilde{H} - \widetilde{H}'\|_{\max}$ by

$$\begin{aligned} \widetilde{\Delta} \leq\ &\{1 + Md_E B + 4Md_E^{7/2} B^3 R_\ell^2\}\Delta \\ &+ M\{d_E R_\ell + 4d_E^{7/2} B^2 R_\ell^3\}\eta. \end{aligned} \tag{44}$$

For two FFN inputs Z, Z' of radius at most S, the one-Lipschitz property of ReLU and the block residual similarly give

$$\begin{aligned} &\|Z + \mathrm{FFN}_\theta(Z) - Z' - \mathrm{FFN}_{\theta'}(Z')\|_{\max} \\ &\quad \leq (1 + d_E d_{\mathrm{FFN}} B^2)\|Z - Z'\|_{\max} \\ &\quad + \{2d_{\mathrm{FFN}} B(d_E S + 1) + 1\}\eta. \end{aligned} \tag{45}$$

Indeed, the first affine map changes by at most $d_E B\|Z - Z'\|_{\max} + (d_E S + 1)\eta$, and its outputs have radius at most $B(d_E S + 1)$. The second affine map, its bias, and the residual increment give the displayed bound.

Set $\mathfrak{L}_0(R) = 0$. Equations (44)–(45) define a valid parameter-sensitivity bound recursively by

$$\begin{aligned}\widetilde{\mathfrak{L}}_\ell &:= \{1 + Md_E B + 4Md_E^{7/2} B^3 R_\ell^2\}\mathfrak{L}_\ell(R) \\ &\quad + M\{d_E R_\ell + 4d_E^{7/2} B^2 R_\ell^3\}, \\ \mathfrak{L}_{\ell+1}(R) &:= (1 + d_E d_{\mathrm{FFN}} B^2)\widetilde{\mathfrak{L}}_\ell + 2d_{\mathrm{FFN}} B(d_E \widetilde{R}_\ell + 1) + 1.\end{aligned}$$

Induction proves the required sensitivity inequality for identical initial inputs and arbitrary parameter vectors θ, θ' in the cube. In particular, both the state and sensitivity bounds allow freely varying query and key matrices.

Under the polynomial architecture envelope, (43) gives

$$1 + R_{\ell+1} \le n^C(1 + R_\ell), \qquad \log(1 + R_\ell) \le C(\ell + 1)\log(en).$$

Since $L \le C_L \log(en)$, this proves the stated $O(\log^2(en))$ bound at depth L. The sensitivity recursion, whose coefficients have fixed degree in the displayed dimensions, parameter radius, and state radius, gives

$$1 + \mathfrak{L}_{\ell+1}(R) \le n^C(1 + R_\ell)^C\{1 + \mathfrak{L}_\ell(R)\}.$$

Consequently,

$$\begin{aligned}\log\{1 + \mathfrak{L}_L(R)\} &\le C \sum_{\ell \in [L-1]_0} \{\log(en) + \log(1 + R_\ell)\} \\ &\le C(L + 1)^2 \log(en) \le C \log^3(en).\end{aligned}$$

Finally, cover the full parameter cube by a maximum-norm net of mesh $\eta = \delta/(L_{\mathrm{Read}}\mathfrak{L}_L(R))$. Its cardinality is at most

$$\left(1 + \frac{2BL_{\mathrm{Read}}\mathfrak{L}_L(R)}{\delta}\right)^{P_{\mathrm{TF}}}.$$

The parameter-sensitivity inequality and the readout Lipschitz bound make its image a δ-net in $\mathcal{F}_{\mathrm{Read}}$. Each block has $3Md_E^2$ attention parameters and $2d_E d_{\mathrm{FFN}} + d_{\mathrm{FFN}} + d_E$ FFN parameters, giving the stated parameter count. Under the polynomial envelope, $P_{\mathrm{TF}} \le Cn^C$, $\log(BL_{\mathrm{Read}}) \lesssim \log(en)$, and $\log\{1 + \mathfrak{L}_L(R)\} \lesssim \log^3(en)$. Substituting these estimates into the cube-covering bound and absorbing the fixed power of $\log(en)$ into a polynomial proves the final entropy bound. □

I.2 End-to-end class and entropy proxy

Fix $P \in \mathcal{P}_\star(\mathscr{M}, \boldsymbol{\pi}, \boldsymbol{\sigma}_X; \sigma_Y^2, B_\epsilon, c_\mu)$ and a task distribution ρ_f supported on $\mathcal{H}(\boldsymbol{\alpha}, \mathscr{M}; L_{\mathcal{F}})$. Let

$$N_n^{\mathrm{end}} := n + 1.$$

Take $B_n \ge 1$, enlarging the deterministic parameter envelope if needed. Motivated by the end-to-end construction in Theorem 3, define

$$\mathcal{G}_n^{\mathrm{end}} := \left\{\mathfrak{s} \mapsto \mathrm{Read}_n\big(\mathrm{TF}_\theta(\mathrm{Emb}_n(\mathfrak{s}))\big) : \mathrm{TF}_\theta \in \mathcal{T}_{M_0}(N_n^{\mathrm{end}}, d_{E,n}, d_{\mathrm{FFN},n}, L_n, B_n)\right\}.$$

Here $\mathcal{T}_{M_0}$ is the softmax class from Section 4, and Emb_n and Read_n are fixed deterministic interface maps. The readout truncates to $[-L_{\mathcal{F}}, L_{\mathcal{F}}]$, so every $g \in \mathcal{G}_n^{\mathrm{end}}$ is uniformly bounded by $L_{\mathcal{F}}$.

Under the bounded-response condition, $|Y_i| \le B_Y := L_{\mathcal{F}} + B_\epsilon$ almost surely. Since B_X is fixed and the covariate-perturbation radii are bounded for all sufficiently large n, there is a constant $R_0 \in [1, \infty)$, independent of $n, K_n, \Gamma, P, \rho_f$, such that

$$\|\mathrm{Emb}_n(\mathfrak{s})\|_{\max} \le R_0$$

almost surely. Define the common prompt domain and its uniform norm by

$$\mathcal{S}_n := \{\mathfrak{s} : \|\mathrm{Emb}_n(\mathfrak{s})\|_{\max} \le R_0\}, \qquad \|g\|_\infty := \sup_{\mathfrak{s} \in \mathcal{S}_n} |g(\mathfrak{s})|.$$

Every admissible prompt law is supported on $\mathcal{S}_n$. We view $\mathcal{G}_n^{\mathrm{end}}$ as a class on this common domain, and use this norm in its covering numbers and empirical-process bounds. Put $\bar{Q} := B_Y + L_{\mathcal{F}} + 1$.

Let P_n^{TF} be the number of trainable scalar parameters of the transformer in $\mathcal{G}_n^{\mathrm{end}}$. It satisfies

$$P_n^{\mathrm{TF}} \le CL_n\{M_0 d_{E,n}^2 + d_{E,n} d_{\mathrm{FFN},n} + d_{\mathrm{FFN},n}\}.$$

Let $L_{\mathrm{Read},n} \ge 1$ be a Lipschitz bound for the fixed map Read_n on the transformer-output domain generated by B_n-bounded transformers applied to $\mathcal{Z}_{N_n^{\mathrm{end}}}(R_0)$. By (37), the fixed readout is globally Lipschitz with a constant independent of n, including on outputs of arbitrary B_n-bounded transformers in the class.

Define the entropy proxy

$$\mathfrak{H}_{n,\Gamma}^{\mathrm{end}} := P_n^{\mathrm{TF}} \log\bigl(1 + CB_n L_{\mathrm{Read},n} \bar{Q}^2 \Gamma\, \mathfrak{L}_{L_n}(R_0)\bigr),$$

where $\mathfrak{L}_{L_n}(R_0)$ is evaluated with

$$(N_n^{\mathrm{end}}, d_{E,n}, d_{\mathrm{FFN},n}, L_n, B_n, M_0).$$

Applying the first bound in Lemma I.1 with $\delta = (\bar{Q}^2\Gamma)^{-1}$ gives

$$\log \mathcal{N}\left(\mathcal{G}_n^{\mathrm{end}}, \frac{1}{\bar{Q}^2\Gamma}, \|\cdot\|_\infty\right) \le \mathfrak{H}_{n,\Gamma}^{\mathrm{end}}.$$

Corollary I.1. *For the end-to-end architecture in Theorem 3, there exist constants $C_{\mathrm{ent}}, c_{\mathrm{ent}} > 0$ such that*

$$\mathfrak{H}_{n,\Gamma}^{\mathrm{end}} \le C_{\mathrm{ent}} n^{c_{\mathrm{ent}}}\{1 + \log(e\Gamma)\}.$$

Here $C_{\mathrm{ent}}, c_{\mathrm{ent}}$ may additionally depend on A_0, the fixed head count, and the polynomial architecture exponents. They are uniform in $n, K_n, \Gamma, P, \rho_f, f$.

Proof. Theorem 3 gives

$$N_n^{\mathrm{end}} = n + 1 \le Cn, \qquad M_0 \le C,$$

and

$$L_n \le C_{\mathrm{depth}} \log(en), \qquad d_{E,n} + d_{\mathrm{FFN},n} + B_n \le n^{C_{\mathrm{size}}}.$$

The input radius R_0, the response bound B_Y, and $\bar{Q} = B_Y + L_{\mathcal{F}} + 1$ are fixed. The readout Lipschitz constant is uniformly bounded by (37). Therefore Lemma I.1 gives

$$\log\{1 + \mathfrak{L}_{L_n}(R_0)\} \le C \log^3(en), \qquad P_n^{\mathrm{TF}} \le Cn^C.$$

Thus

$$\begin{aligned}\mathfrak{H}_{n,\Gamma}^{\mathrm{end}} &= P_n^{\mathrm{TF}} \log\bigl(1 + CB_n L_{\mathrm{Read},n} \bar{Q}^2 \Gamma\, \mathfrak{L}_{L_n}(R_0)\bigr) \\ &\le Cn^C\{1 + \log(e\Gamma)\}.\end{aligned}$$

Renaming constants proves the claim. □

I.3 Empirical process bound

For training data $\mathcal{D}_\Gamma^{\mathrm{tr}}$, let

$$\begin{aligned}\widehat{R}_\Gamma(g) &:= \frac{1}{\Gamma}\sum_{\gamma\in[\Gamma]}\{Y_{n+1}^{(\gamma)} - g(\mathfrak{s}^{(\gamma)})\}^2, \\ R_Y(g) &:= \mathbb{E}\{Y_{n+1} - g(\mathfrak{s})\}^2, \\ R_{P,\rho_f}^\star(g) &:= \mathbb{E}_{f\sim\rho_f}\mathbb{E}_{P,f}\{g(\mathfrak{s}) - f(X_{n+1}^\star)\}^2.\end{aligned}$$

Lemma I.2 (Bounded least-squares empirical process)**.** *Let $(S,Y), \{(S_\gamma, Y_\gamma)\}_{\gamma\in[\Gamma]}$ be i.i.d. with $|Y| \le B_Y$ almost surely. Let $\mathcal{G}$ be a class of measurable maps from the sample space of S to $[-B_G, B_G]$, and define*

$$R(g) := \mathbb{E}\{Y - g(S)\}^2, \qquad \widehat{R}_\Gamma(g) := \frac{1}{\Gamma}\sum_{\gamma\in[\Gamma]}\{Y_\gamma - g(S_\gamma)\}^2.$$

Let $g_0(S) := \mathbb{E}[Y \mid S]$ and $Q_{\mathcal{G}} := B_Y + B_G + 1$. Assume the displayed supremum is measurable; otherwise the same statement holds with outer probability. Then a universal constant $C > 0$ satisfies, for every $t > 0$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\{R(g) - R(g_0) - 2\widehat{R}_\Gamma(g) + 2\widehat{R}_\Gamma(g_0)\} \geq t\right) \leq C\,\mathcal{N}\left(\mathcal{G}, \frac{t}{CQ_{\mathcal{G}}^2}, \|\cdot\|_\infty\right)\exp\left(-\frac{\Gamma t}{CQ_{\mathcal{G}}^4}\right).$$

Proof. For $g \in \mathcal{G}$, set

$$\ell_g(S,Y) := (Y - g(S))^2, \qquad \ell_0(S,Y) := (Y - g_0(S))^2.$$

Since $|Y| \leq B_Y$, we have $|g_0(S)| \leq B_Y$ almost surely. Since $g_0(S) = \mathbb{E}[Y \mid S]$,

$$R(g) - R(g_0) = \mathbb{E}\{g(S) - g_0(S)\}^2.$$

Moreover,

$$|\ell_g - \ell_0| \leq CQ_{\mathcal{G}}^2, \qquad \mathbb{E}(\ell_g - \ell_0)^2 \leq CQ_{\mathcal{G}}^2\{R(g) - R(g_0)\}.$$

Write P and P_Γ for population and empirical averages, and set $\mu_g := P(\ell_g - \ell_0) = R(g) - R(g_0) \geq 0$. Then

$$\mu_g - 2P_\Gamma(\ell_g - \ell_0) \geq t \quad \Longrightarrow \quad (P - P_\Gamma)(\ell_g - \ell_0) \geq \frac{\mu_g + t}{2}.$$

Bernstein's inequality gives the bound $\exp\{-\Gamma t/(CQ_{\mathcal{G}}^2)\}$. Taking a $t/(CQ_{\mathcal{G}}^2)$-net in $\|\cdot\|_\infty$, using

$$\|\ell_g - \ell_{g'}\|_\infty \leq CQ_{\mathcal{G}}\|g - g'\|_\infty,$$

and applying a union bound proves the claim. □

Lemma I.3 (Near-ERM oracle inequality)**.** *Assume $|Y_{n+1}| \leq B_Y$ almost surely, where $B_Y < \infty$ is independent of $n, K_n, \Gamma, P, \rho_f$. Assume also that the fresh query-response noise satisfies*

$$\mathbb{E}[\epsilon_{n+1} \mid \mathfrak{s}, f, X_{n+1}^\star, \xi_{n+1}, Z_{n+1}] = 0.$$

Fix a deterministic tolerance $\varepsilon_{\mathrm{opt},\Gamma} \geq 0$. Let $\widehat{g}_\Gamma \in \mathcal{G}_n^{\mathrm{end}}$ be a measurable empirical-risk $\varepsilon_{\mathrm{opt},\Gamma}$-minimizer, that is, almost surely,

$$\widehat{R}_\Gamma(\widehat{g}_\Gamma) \leq \inf_{g\in\mathcal{G}_n^{\mathrm{end}}} \widehat{R}_\Gamma(g) + \varepsilon_{\mathrm{opt},\Gamma}.$$

Then

$$\mathbb{E}_{\mathcal{D}_\Gamma^{\mathrm{tr}}} R^\star_{P,\rho_f}(\widehat{g}_\Gamma) \leq 2\inf_{g\in\mathcal{G}_n^{\mathrm{end}}} R^\star_{P,\rho_f}(g) + C\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}} + 1}{\Gamma} + 2\varepsilon_{\mathrm{opt},\Gamma}.$$

Here C depends only on $B_Y, L_{\mathcal{F}}$ and fixed construction constants, uniformly in $n, K_n, \Gamma, P, \rho_f$.

Proof. Let $g_0(\mathfrak{s}) := \mathbb{E}[Y_{n+1} \mid \mathfrak{s}]$, and define

$$Z_\Gamma := \sup_{g\in\mathcal{G}_n^{\mathrm{end}}}\{R_Y(g) - R_Y(g_0) - 2\widehat{R}_\Gamma(g) + 2\widehat{R}_\Gamma(g_0)\}.$$

Lemma I.2, applied on $\mathcal{S}_n$ with $\mathcal{G} = \mathcal{G}_n^{\mathrm{end}}$, $B_G = L_{\mathcal{F}}$, and hence $Q_{\mathcal{G}} = \bar{Q}$, gives, for every $t > 0$,

$$\mathbb{P}(Z_\Gamma \geq t) \leq C\,\mathcal{N}\left(\mathcal{G}_n^{\mathrm{end}}, \frac{t}{C\bar{Q}^2}, \|\cdot\|_\infty\right)\exp\left(-\frac{\Gamma t}{C\bar{Q}^4}\right).$$

Choose

$$t = A\bar{Q}^4\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}} + u}{\Gamma}, \qquad u \geq 1,$$

with A sufficiently large. Since $\frac{t}{C\bar{Q}^2} \geq \frac{1}{\bar{Q}^2\Gamma}$, monotonicity of covering numbers and the definition of $\mathfrak{H}_{n,\Gamma}^{\mathrm{end}}$ imply

$$\mathbb{P}\left(Z_\Gamma \geq C\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}} + u}{\Gamma}\right) \leq Ce^{-u}.$$

For $\Xi_\Gamma := R_Y(\widehat{g}_\Gamma) - R_Y(g_0) - 2\widehat{R}_\Gamma(\widehat{g}_\Gamma) + 2\widehat{R}_\Gamma(g_0)$, we have $\Xi_\Gamma \le Z_\Gamma$. Integrating the preceding tail bound gives

$$\mathbb{E}_{\mathcal{D}_\Gamma^{\mathrm{tr}}}\Xi_\Gamma \le C\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}}+1}{\Gamma}.$$

For every $g \in \mathcal{G}_n^{\mathrm{end}}$, $\widehat{R}_\Gamma(\widehat{g}_\Gamma) \le \widehat{R}_\Gamma(g) + \varepsilon_{\mathrm{opt},\Gamma}$. Hence

$$R_Y(\widehat{g}_\Gamma) - R_Y(g_0) \le \Xi_\Gamma + 2\widehat{R}_\Gamma(g) - 2\widehat{R}_\Gamma(g_0) + 2\varepsilon_{\mathrm{opt},\Gamma}.$$

Taking expectations gives

$$\mathbb{E}_{\mathcal{D}_\Gamma^{\mathrm{tr}}} R_Y(\widehat{g}_\Gamma) \le 2R_Y(g) - R_Y(g_0) + C\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}}+1}{\Gamma} + 2\varepsilon_{\mathrm{opt},\Gamma}.$$

It remains to replace noisy query-response risk by latent regression risk. For any deterministic h, and also for $h = \widehat{g}_\Gamma$ after conditioning on $\mathcal{D}_\Gamma^{\mathrm{tr}}$, the prediction $h(\mathfrak{s})$ is a function of the fresh prompt $\mathfrak{s}$ only. The conditional mean-zero assumption gives

$$\mathbb{E}[\{h(\mathfrak{s}) - f(X_{n+1}^\star)\}\epsilon_{n+1}] = 0.$$

Thus

$$R_Y(h) = R_{P,\rho_f}^\star(h) + \sigma_{\rho,Y}^2, \qquad \sigma_{\rho,Y}^2 := \mathbb{E}\epsilon_{n+1}^2.$$

Moreover,

$$R_Y(g_0) = \mathbb{E}\{\mathrm{Var}(Y_{n+1} \mid \mathfrak{s})\} \ge \sigma_{\rho,Y}^2.$$

Indeed, by the law of total variance,

$$\begin{aligned}\mathrm{Var}(Y_{n+1} \mid \mathfrak{s}) &\ge \mathbb{E}\big[\mathrm{Var}(Y_{n+1} \mid \mathfrak{s}, f, X_{n+1}^\star, \xi_{n+1}, Z_{n+1}) \,\big|\, \mathfrak{s}\big] \\ &= \mathbb{E}\big[\mathrm{Var}(\epsilon_{n+1} \mid \mathfrak{s}, f, X_{n+1}^\star, \xi_{n+1}, Z_{n+1}) \,\big|\, \mathfrak{s}\big].\end{aligned}$$

Integrating and using the conditional mean-zero assumption gives $R_Y(g_0) \ge \sigma_{\rho,Y}^2$.

Substitution gives

$$\mathbb{E}_{\mathcal{D}_\Gamma^{\mathrm{tr}}} R_{P,\rho_f}^\star(\widehat{g}_\Gamma) \le 2R_{P,\rho_f}^\star(g) + C\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}}+1}{\Gamma} + 2\varepsilon_{\mathrm{opt},\Gamma}.$$

Taking the infimum over $g \in \mathcal{G}_n^{\mathrm{end}}$ proves the claim. □

Theorem I.1. *Fix $A_0 > 0$. Assume the hypotheses of Theorem 3, and suppose that $\mathcal{G}_n^{\mathrm{end}}$ contains the end-to-end predictor constructed there. Let $P \in \mathcal{P}_\star(\mathscr{M}, \boldsymbol{\pi}, \boldsymbol{\sigma}_X; \sigma_Y^2, B_\epsilon, c_\mu)$, and let ρ_f be supported on $\mathcal{H}(\boldsymbol{\alpha}, \mathscr{M}; L_{\mathcal{F}})$. Assume $|Y_{n+1}| \le B_Y := L_{\mathcal{F}} + B_\epsilon$ almost surely and*

$$\mathbb{E}[\epsilon_{n+1} \mid \mathfrak{s}, f, X_{n+1}^\star, \xi_{n+1}, Z_{n+1}] = 0.$$

For a deterministic tolerance $\varepsilon_{\mathrm{opt},\Gamma} \ge 0$, every measurable empirical-risk $\varepsilon_{\mathrm{opt},\Gamma}$-minimizer $\widehat{g}_\Gamma$ over $\mathcal{G}_n^{\mathrm{end}}$, as defined in Lemma I.3, satisfies

$$\mathbb{E}_{\mathcal{D}_\Gamma^{\mathrm{tr}}} R_{P,\rho_f}^\star(\widehat{g}_\Gamma) \le C\{\mathfrak{R}_n(\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\pi}) + n^{-A_0}\} + C\frac{\mathfrak{H}_{n,\Gamma}^{\mathrm{end}}+1}{\Gamma} + 2\varepsilon_{\mathrm{opt},\Gamma}.$$

Here C may additionally depend on A_0, uniformly in $n, K_n, \Gamma, P, \rho_f, f$.

Proof. By Theorem 2 and Theorem 3, the class $\mathcal{G}_n^{\mathrm{end}}$ contains a comparator g_n° such that, for every $f \in \mathcal{H}(\boldsymbol{\alpha}, \mathscr{M}; L_{\mathcal{F}})$,

$$\mathbb{E}_{P,f}\{g_n^\circ(\mathfrak{s}) - f(X_{n+1}^\star)\}^2 \le C\{\mathfrak{R}_n(\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\pi}) + n^{-A_0}\}.$$

Averaging over $f \sim \rho_f$ gives

$$\inf_{g \in \mathcal{G}_n^{\mathrm{end}}} R_{P,\rho_f}^\star(g) \le C\{\mathfrak{R}_n(\boldsymbol{\alpha}, \mathbf{d}, \boldsymbol{\pi}) + n^{-A_0}\}.$$

Substitution into Lemma I.3 proves the theorem. □

I.4 Proofs of Theorem 4 and Corollary 1

Theorem 4 follows directly from Theorem I.1 and Corollary I.1. The conditional query-noise assumption is verified below. We then prove the sufficient regime in Corollary 1.

The conditional mean-zero condition for the fresh query noise used below follows from the i.i.d. task construction in Section 5 and Assumption 2. Indeed, for the target task,

$$\mathbb{E}[\epsilon_{n+1} \mid \mathfrak{s}, f, X^\star_{n+1}, \xi_{n+1}, Z_{n+1}] = 0.$$

Proof. By Theorem I.1,

$$\mathbb{E}_{\mathcal{D}^{\mathrm{tr}}_\Gamma} R^\star_{P,\rho_f}(\widehat{g}_\Gamma) \le C\{\mathfrak{R}_n + n^{-A_0}\} + C\frac{\mathfrak{H}^{\mathrm{end}}_{n,\Gamma} + 1}{\Gamma} + 2\varepsilon_{\mathrm{opt},\Gamma}.$$

Let $a_k := \frac{2\alpha_k}{2\alpha_k + d_k}$ and $\lambda_{\max} := \frac{2\alpha_{\max}}{2\alpha_{\max}+1}$, then summing over k gives $\mathfrak{R}_n^{-1} \le n^{\lambda_{\max}}$. Thus any fixed $A_0 \ge \lambda_{\max}$ is sufficient. Assume also $\varepsilon_{\mathrm{opt},\Gamma} \lesssim \mathfrak{R}_n$. By Corollary I.1,

$$\mathfrak{H}^{\mathrm{end}}_{n,\Gamma} \le C_{\mathrm{ent}} n^{c_{\mathrm{ent}}}\{1 + \log(e\Gamma)\}.$$

Let $\Gamma_0 := Cn^{c_{\mathrm{ent}}} \log(en)\mathfrak{R}_n^{-1}$ for a sufficiently large fixed $C > 0$. Then $\Gamma_0 \le n^C \log(en)$, so $\log(e\Gamma_0) \le C \log(en)$ for all large n. Since $\Gamma \mapsto \frac{1+\log(e\Gamma)}{\Gamma}$ is decreasing for $\Gamma \ge 1$, every $\Gamma \ge \Gamma_0$ satisfies

$$\frac{\mathfrak{H}^{\mathrm{end}}_{n,\Gamma} + 1}{\Gamma} \lesssim \mathfrak{R}_n.$$

Combining the preceding bounds gives $\mathbb{E}_{\mathcal{D}^{\mathrm{tr}}_\Gamma} R^\star_{P,\rho_f}(\widehat{g}_\Gamma) \lesssim \mathfrak{R}_n$, as claimed. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the growing heterogeneous manifold-mixture setting, matching minimax bounds, and transformer approximation and generalization results. Claims of local geometry adaptivity and minimax optimality are restricted to the explicit assumptions and constructed class.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sections 4 and 5 specify the encoded structure, resource bounds, approximation, finite-task and optimization terms, and worst-case lower-bound scope. Appendix C distinguishes the empirical and theoretical architectures, states that finite-sample feasibility is uncertified, and defines the uncertainty captured by the intervals.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sections 2–5 state the assumptions and results, with conventions and geometric details in Appendices D and E. Complete proofs of the lower bound, estimator upper bound, transformer approximation, and generalization appear in Appendices F, G, H, and I, respectively.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix C specifies the generator, seeds, architecture, optimization schedules, final-checkpoint selection, LP benchmark, and shared test and bootstrap protocols for all four budgets through $\Gamma = 1{,}000{,}000$. The linked public repository includes configurations, predictions, and verification records.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The repository linked in Appendix C provides the generator, Transformer training, oracle LP baseline, evaluation code, and four-budget predictions and bootstrap summaries, with versions, installation and execution commands, configurations, and seeds. Prompts are regenerated from code; trained checkpoints and full prompt datasets are not distributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C specifies all four training budgets, data splits, random seeds, architecture, AdamW settings and cosine schedule, gradient accumulation and clipping, final-checkpoint selection, and LP bandwidth selection. The corresponding configurations are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Appendix C reports 95% percentile-bootstrap intervals from 5,000 resamples of 2,000 shared test prompts, with paired risk-ratio intervals for all four budgets checked against released predictions. They quantify test-task variation conditional on seed-0 fits and selected bandwidths; component results at $\Gamma = 1{,}000{,}000$ are descriptive point estimates.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C and the compute documentation specify the model, precision, batches, updates, Apple M2 Pro with 16 GiB memory, and per-run times totaling 52.37 training hours or 52.79 hours including recorded Transformer and shared LP evaluation. The totals cover those production phases; additional exploratory, validation, and preparation computation is disclosed separately.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and conducted this research in accordance with it.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work studies statistical guarantees for nonparametric in-context learning and includes a controlled synthetic numerical illustration. It does not develop an application-specific or deployed system, and the paper identifies no direct path from these foundational results to specific societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: It has no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Third-party software, including NumPy and PyTorch, is credited with exact versions, official project and license links, and bundled-component notices in the environment documentation and used under its stated licenses. Experimental data are generated by the supplied synthetic-data code.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The repository linked in Appendix C documents the generator, installation, training, evaluation, versions, dependency licenses, limitations, and CPU example, with four-budget predictions, summaries, verification records, and original/exported file hashes. The authors' code is licensed under MIT and the results under CC BY 4.0, as specified in the licensing documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments, human subjects, or user studies.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, participant studies, or data collection requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs assisted with wording and LaTeX editing, review of mathematical arguments, and preparation and review of experimental code and reproducibility documentation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.