WHEN PARAMETER BOUNDS LIMIT LEARNING: APPROXIMATION AND ESTIMATION IN MEAN-FIELD NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning theory for neural networks often imposes hard parameter bounds for tractability, but these restrictions can change what a model represents rather than merely simplify proofs. We study this in Gaussian multi-index regression, where targets depend on an unknown low-dimensional subspace, using mean-field networks and mean-field Langevin dynamics (MFLD) under compact and noncompact parameterizations. For a fixed compact parameter domain, the representable predictors form a closed class, so targets outside it incur an irreducible approximation error unaffected by sample size or training time. Even exactly represented multi-index targets inherit smoothness from the activation, showing that bounded weights restrict the regularity of admissible link functions. For these represented targets, the long-time limit of MFLD satisfies a finite-sample prediction upper bound that depends on the ambient dimension. By contrast, on a fixed infinite-dimensional family of rough multi-index targets separated from every fixed compact class, the long-time limit of MFLD over a noncompact model with unbounded input weights and biases achieves polynomially decaying prediction error, accompanied by feature learning. Thus hard constraints and soft regularization can define different learning problems, making approximation the first step in interpreting statistical and convergence guarantees.

1 INTRODUCTION

A central challenge in learning theory is to make complex models mathematically tractable without changing the phenomenon one intends to study. A common strategy is to restrict the parameter space—through bounded domains, norm constraints, or compactness—so that existence, convergence, and statistical complexity become easier to control. Such restrictions are often treated as just technical devices, although they also change the induced function class and potentially its closure, approximation power, and regularity. They raise a fundamental question: when does a convenient restriction preserve the original learning problem, and when does it quietly replace it with a different one?

Mean-field neural networks provide a clean setting in which to examine this question. In the infinite-width description, a two-layer network is parameterized by a probability measure, allowing training to be studied through variational objectives and Wasserstein dynamics (Nitanda & Suzuki, 2017; Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Sirignano & Spiliopoulos, 2020) while retaining a genuine feature-learning regime (Suzuki et al., 2023; Shang et al., 2026). Multi-index regression is a particularly revealing test case: the target depends only on an unknown $m^\star$-dimensional subspace of an ambient space $\mathbb{R}^d$, while the link function has unknown regularity s (Gaïffas & Lecué, 2007; Klock et al., 2021; Paris, 2014). A complete statistical account should explain, within a single framework, which targets can be approximated under a given parameterization, how the prediction-rate exponent for the compact model compares with the smooth multi-index reference exponent, and how these statistical properties are connected to the long-time behavior of the training dynamics.

In this paper, we investigate these questions for Gaussian multi-index regression learned by MFLD, comparing compact and noncompact parameterizations. Our analysis follows three steps. We first

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

investigate whether the target can be approximated by the function class induced by the chosen parameter domain. When approximation is possible, we study the finite-sample prediction error. Finally, we identify the resulting entropy-regularized estimator with the long-time limit of MFLD. Our main results and their implications for statistical learning theory are as follows:

- **Bounded parameters can exclude targets from the learning problem.** On the compact parameter domain, the induced neural network class is closed in $L^2(\gamma_d)$, where γ_d denotes the standard Gaussian measure on $\mathbb{R}^d$. A target outside this closed class remains at a strictly positive distance from every admissible predictor, regardless of the sample size, regularization level, or training time. Targets represented by this class are also restricted: their multi-index links inherit finite-order smoothness from the activation.
- **For represented compact targets, the prediction error rate is governed by the ambient dimension.** If the activation has 2α bounded derivatives, the predictor at the long-time limit of compact MFLD has a prediction-error upper bound of order $N^{-2\alpha/(2\alpha+d)}$. For comparison, we use the smooth multi-index reference exponent $2s/(2s+m^\star)$. The upper bound is at least as fast as $N^{-2s/(2s+m^\star)}$ when $\alpha \geq ds/m^\star$, and for smooth activations it can be arbitrarily close to the parametric rate N^{-1} on the represented compact class.
- **Noncompact MFLD can learn rough multi-index targets with feature learning.** We construct a fixed, rich, infinite-dimensional family of rough multi-index targets that is uniformly separated from each fixed compact parameter class. The long-time limit of MFLD without compact constraints satisfies a prediction-error upper bound of order $N^{-2s/(2s+m^\star)}$, uniformly over the family and the index frame. On the same family, the learned feature kernel stays a fixed positive distance from K_{init} in $L^2(\gamma_d^{\otimes 2})$ with high probability. A nonvanishing fraction of active hidden mass enters a shrinking angular neighborhood around the index subspace, and smoothed gradients of the fitted predictor recover the projector onto the index subspace.

Organization. Sec. 2 introduces the regression model, the estimator, and the compact and noncompact parameter domains. Sec. 3 studies how compactification changes approximation power. Sec. 4 derives the prediction guarantee for represented compact targets and compares it with a smooth multi-index reference exponent. Sec. 5 gives a prediction result for a fixed family of rough multi-index targets when input weights and biases are unbounded. Sec. 6 discusses the implications of these results, their connections to the existing literature, and their limitations. Proofs and experimental details are deferred to the appendix.

2 Problem Setup

Gaussian multi-index regression problem. For $D \geq 1$, let γ_D denote the standard Gaussian measure on $\mathbb{R}^D$, and let $(\mathrm{H}_\eta)_{\eta\in\mathbb{N}^D}$ be the normalized tensor Hermite basis of $L^2(\gamma_D)$. For $s \geq 0$, define

$$H^s(\gamma_D) := \left\{h \in L^2(\gamma_D) : \|h\|_{H^s(\gamma_D)} < \infty\right\}, \quad \|h\|^2_{H^s(\gamma_D)} := \sum_{\eta\in\mathbb{N}^D} (1+|\eta|)^s |\langle h, \mathrm{H}_\eta\rangle_{L^2(\gamma_D)}|^2.$$

Equivalently, $H^s(\gamma_D)$ is the domain of $(I-\mathcal{L}_D)^{s/2}$, where $\mathcal{L}_D = \Delta - z^\top\nabla$ is the Ornstein–Uhlenbeck generator. This definition uses s as the Sobolev derivative order: for every integer $r \geq 0$, $H^r(\gamma_D)$ consists of functions whose weak derivatives through order r lie in $L^2(\gamma_D)$, with equivalent norms (Bogachev, 1998).

We observe N i.i.d. samples $\mathcal{D}_N = \{(X_i, Y_i)\}_{i=1}^N$ from the following multi-index model:

$$X \sim \gamma_d, \qquad Y = f^\star(X) + \xi, \qquad f^\star(x) = h^\star(E_\star^\top x), \tag{1}$$

where $\xi \sim \mathcal{N}(0,\sigma_\xi^2)$ is independent of X, the index frame $E_\star \in \mathbb{R}^{d\times m^\star}$ has orthonormal columns, and $h^\star \in H^s(\gamma_{m^\star})$ for some $s > m^\star/2$. The column space $\mathrm{Range}(E_\star)$ is the index subspace. The intrinsic dimension $m^\star$, the index subspace, and the regularity s are unknown to the learner.

For a predictor f, let $\ell_f(x,y) := (y-f(x))^2$, and denote the empirical risk by $P_N\ell_f := N^{-1}\sum_{i=1}^N \ell_f(X_i,Y_i)$. Under (1), the population excess risk is $P\ell_f - P\ell_{f^\star} = \|f-f^\star\|^2_{L^2(\gamma_d)}$. Excess prediction risk and squared $L^2(\gamma_d)$ estimation error coincide.

Mean-field neural networks and estimator. Let $\Theta \subseteq \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ be a parameter domain, and let $\mathcal{P}(\Theta)$ be the set of Borel probability measures on Θ. For $\theta = (a, w, b)$ and $\nu \in \mathcal{P}(\Theta)$, define a mean-field neural network and an entropy-regularized empirical risk:

$$f_\nu(x) := \int_\Theta a\,\sigma(\langle w, x\rangle + b)\,\mathrm{d}\nu(a, w, b), \qquad J^\Theta_{N,\lambda}(\nu) := P_N \ell_{f_\nu} + \lambda\,\mathrm{KL}(\nu\|\nu_0),$$

where $\nu_0 \in \mathcal{P}(\Theta)$ is a reference measure and $\lambda > 0$ and $\mathrm{KL}(\nu\|\nu_0) := \int \log(\,\mathrm{d}\nu/\,\mathrm{d}\nu_0)\,\mathrm{d}\nu$ when $\nu \ll \nu_0$, and $\mathrm{KL}(\nu\|\nu_0) := +\infty$ otherwise. The associated function class is $\mathcal{F}_\Theta := \{f_\nu : \nu \in \mathcal{P}(\Theta)\}$. A discrete measure with M atoms represents an M-neuron two-layer network, so this definition includes finite-width networks (Nitanda & Suzuki, 2017; Mei et al., 2018). We refer to w as the input-weight vector.

Our estimator is $\widehat{f}^\Theta_{N,\lambda} := f_{\widehat{\nu}^\Theta_{N,\lambda}}$, where $\widehat{\nu}^\Theta_{N,\lambda}$ is the regularized empirical-risk minimizer (RERM) of $J^\Theta_{N,\lambda}$. MFLD is the 2-Wasserstein gradient flow of this objective, initialized from ν_0 (Chapter 11 in Ambrosio et al., 2008). We prove in Props. E.1 and E.2 that, for the compact and noncompact models, respectively, the above entropy-regularized empirical risk has a unique minimizer, which coincides with the long-time limit of the corresponding reflected MFLD.

The two parameter domains. The compact analysis uses

$$\Theta_{\mathrm{c}} := [-A, A] \times B_2^d \times [-B, B], \qquad \nu_{0,\mathrm{c}} := \mathrm{Unif}([-A, A]) \otimes \mathrm{Unif}(B_2^d) \otimes \mathrm{Unif}([-B, B]),$$

where $B_2^d := \{w \in \mathbb{R}^d : \|w\|_2 \le 1\}$. This parameterization bounds the output weight, the input-weight norm, and the bias.

The noncompact analysis specializes to the $\tanh$ activation and uses $\Theta_{\mathrm{nc}} := [-A, A] \times \mathbb{R}^d \times \mathbb{R}$, with reference measure $\nu_{0,\mathrm{nc}} := \mathrm{Unif}([-A, A]) \otimes \mathcal{N}(0, \tau_w^2 I_d) \otimes \mathcal{N}(0, \tau_b^2)$. This model retains the output-weight bound while allowing unbounded input weights and biases.

3 Compactification Changes the Statistical Problem

We begin by determining whether the compact model can approximate the target. The answer follows from a basic property of the map from parameter measures to prediction functions. On the compact domain Θ_{c}, probability measures form a compact space under weak convergence. Moreover, weak convergence of the parameter measure implies convergence of the corresponding neural-network predictions in $L^2(\gamma_d)$. The resulting function class is compact and, in particular, closed.

For $k \ge 0$, let $C_b^k(\mathbb{R})$ denote the set of k-times continuously differentiable functions whose derivatives up to order k are bounded, and write $C_b(\mathbb{R}) := C_b^0(\mathbb{R})$.

Proposition 3.1 (Compactification barrier)**.** *Assume $A, B > 0$ and $\sigma \in C_b(\mathbb{R})$. Then $\mathcal{F}_{\Theta_{\mathrm{c}}}$ is compact in $L^2(\gamma_d)$. Consequently, if $f^\star \notin \mathcal{F}_{\Theta_{\mathrm{c}}}$, then*

$$\inf_{f \in \mathcal{F}_{\Theta_{\mathrm{c}}}} \|f - f^\star\|^2_{L^2(\gamma_d)} > 0.$$

Prop. 3.1 shows that, within the fixed compact model, approximation to arbitrary accuracy is possible only for targets that are exactly representable. Since $\mathcal{F}_{\Theta_{\mathrm{c}}}$ already includes all probability mixtures over Θ_{c}, increasing network width cannot remove the gap for a target outside this class. The gap persists even under exact population-risk minimization, independently of sample size, regularization, or training time. In particular, taking $\lambda \downarrow 0$ weakens the entropy penalty but leaves the admissible function class unchanged. Consistent learning of such a target therefore requires enlarging the model class beyond the fixed constraint. This conclusion concerns fixed compact bounds and does not rule out growing compact sieves whose parameter bounds increase with the sample size. Figure 1 illustrates the regularization aspect of the represented-versus-misspecified dichotomy: as λ decreases, the error vanishes for represented targets but approaches a positive plateau for misspecified ones.

For multi-index targets, compactification imposes a second restriction beyond the approximation barrier above. Exact representability, which is necessary for consistency under fixed compact bounds, also constrains the regularity of the link function. Bounded output weights and input-weight norms ensure that a represented link admits a representative whose derivatives are controlled by the corresponding derivatives of the activation, uniformly over the representing measure. The next theorem quantifies this inherited-regularity constraint.

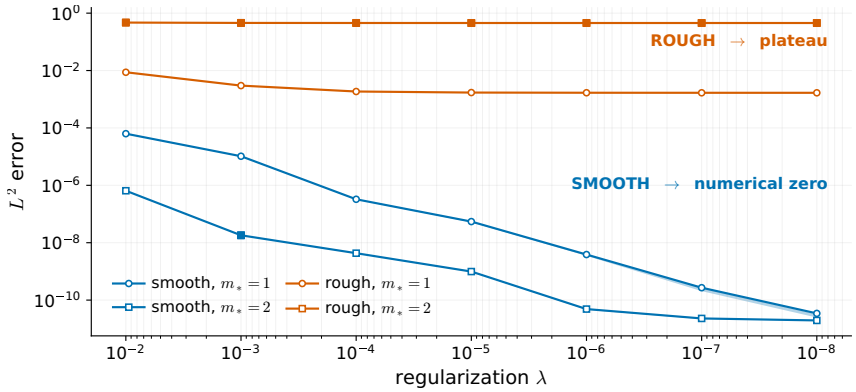


Figure 1: **Compactification distinguishes represented and misspecified targets.** For smooth represented targets, the L^2 error decreases to numerical precision as λ decreases. For rough misspecified targets, the error approaches a positive plateau for both $m^\star = 1$ and $m^\star = 2$. The plateau is consistent with the approximation barrier in Prop. 3.1. Additional numerical details are provided in Appendix F.1.

Theorem 3.2 (Finite-order regularity forced by compact representability)**.** *Let $r \in \mathbb{N}_+$. Suppose that $\sigma \in C_b^{2r}(\mathbb{R})$ and that $f^\star(x) = h^\star(E_\star^\top x) \in \mathcal{F}_{\Theta_c}$. Then, the $L^2(\gamma_{m^\star})$ equivalence class of $h^\star$ contains a representative $\widetilde{h}^\star \in C_b^{2r}(\mathbb{R}^{m^\star}) \cap H^{2r}(\gamma_{m^\star})$ with norms bounded uniformly over the representing measure. In particular, if $\sigma \in C_b^\infty(\mathbb{R})$, then $h^\star \in H^s(\gamma_{m^\star})$ for every finite $s > 0$.*

Proof idea. Let $E_{\star,\perp} \in \mathbb{R}^{d\times(d-m^\star)}$ complete $E_\star$ to an orthogonal matrix. Write $x = E_\star z + E_{\star,\perp} u$, where z and u are the coordinates along the index subspace and its orthogonal complement, respectively. Since $f^\star(x) = h^\star(E_\star^\top x)$ does not depend on u, averaging any compact representation $f^\star = f_\nu$ over $u \sim \gamma_{d-m^\star}$ gives a representative $\widetilde{h}^\star$ of the same link in $L^2(\gamma_{m^\star})$. Because $|a| \le A$ and $\|E_\star^\top w\|_2 \le 1$, differentiating this representation yields $\|\partial^\beta \widetilde{h}^\star\|_\infty \le A\|\sigma^{(|\beta|)}\|_\infty$ for all $|\beta| \le 2r$. The Gaussian Hermite–Sobolev derivative bound then gives $\widetilde{h}^\star \in H^{2r}(\gamma_{m^\star})$, with a radius independent of the particular representing measure ν. A complete proof is given in Appendix A. □

Intuitively, compact representability limits how rapidly the link function can vary. Each neuron has bounded output weight and bounded input-weight norm, so its derivatives are controlled by the corresponding derivatives of the activation. Averaging over neurons cannot create arbitrarily sharp features, because these derivative bounds hold uniformly for every neuron. Hence, if a multi-index target is represented exactly by the compact model, its link must inherit a corresponding amount of smoothness from the activation. This is a restriction on the functions that the model can represent, rather than a smoothing effect of the learning algorithm.

Takeaway: What compactification changes.

Compactification has two related effects. It leaves an irreducible error for targets outside the compact model, and it restricts well-specified targets to links that inherit the regularity of the activation.

4 Prediction Guarantees for Exactly Represented Compact Targets

We now assume that $f^\star \in \mathcal{F}_{\Theta_c}$, so the approximation barrier of Prop. 3.1 does not apply. The target depends only on $m^\star$ directions, but the estimator ranges over neural predictors on $\mathbb{R}^d$. The proof separates the prediction error into a component along the index subspace and a residual component in the ambient space. By Thm. 3.2 and the smoothness of the compact neural class, both components have the same smoothness order. The ambient component is no smaller and determines our uniform prediction rate.

Throughout this section, we set $\lambda_N = N^{-1}$.

Theorem 4.1 (Prediction bound for exactly represented compact targets)**.** *Consider the model (1). Assume $A, B > 0$, let $\alpha \in \mathbb{N}_+$ satisfy $\alpha > d/2$, and suppose $\sigma \in C_b^{2\alpha}(\mathbb{R})$ and $f^\star \in \mathcal{F}_{\Theta_c}$. Set $\widehat{f}_N := \widehat{f}^{\Theta_c}_{N,1/N}$. Then, for every fixed $c \geq 1$, there exists $C_c = C\big(c, d, \alpha, A, B, \sigma_\xi, \max_{0\leq k\leq 2\alpha} \|\sigma^{(k)}\|_\infty\big) > 0$ such that, for all $N \geq 2$, with probability at least $1 - 4(eN)^{-c}$ over the training sample,*

$$\|\widehat{f}_N - f^\star\|^2_{L^2(\gamma_d)} \leq C_c N^{-2\alpha/(2\alpha+d)}.$$

Proof idea. We consider an *information–residual decomposition* tailored to the multi-index structure. It separates the localized Rademacher complexity (e.g. Koltchinskii & Panchenko, 2000) of the neural class into an intrinsic $m^\star$-dimensional part and an ambient d-dimensional residual part. Specifically, for $g_\nu := f_\nu - f^\star$, define an information component along the true index subspace and a residual component in the ambient space

$$g_\nu^{\mathrm{info}}(X) := \mathbb{E}[g_\nu(X) \mid E_\star^\top X], \qquad g_\nu^{\mathrm{res}} := g_\nu - g_\nu^{\mathrm{info}}.$$

Then, the excess risk is decomposed as $\|g_\nu\|_2^2 = \|g_\nu^{\mathrm{info}}\|_2^2 + \|g_\nu^{\mathrm{res}}\|_2^2$. Hence, the localized error class is contained in the sum of an $m^\star$-dimensional information class and a d-dimensional residual class. The compact parameterization and $\sigma \in C_b^{2\alpha}$ give uniform $H^{2\alpha}$ bounds for both classes. Thus, the localized Rademacher complexity $\mathfrak{R}_N(r)$ of the error class $\{g_\nu : \|g_\nu\|_2 \leq r\}$ satisfies

$$\mathfrak{R}_N(r) \leq \mathfrak{R}_N^{\mathrm{info}}(r) + \mathfrak{R}_N^{\mathrm{res}}(r) \lesssim \frac{1}{\sqrt{N}} \left\{ r^{1-m^\star/(2\alpha)} + r^{1-d/(2\alpha)} \right\}.$$

The ambient residual term is dominant, and the corresponding critical-radius relation gives $r^2 \asymp N^{-2\alpha/(2\alpha+d)}$. The proof is given in Appendix B. □

Comparison with a multi-index reference exponent. For link regularity s and intrinsic dimension $m^\star$, we compare with $N^{-p_{s,m^\star}}$, where $p_{s,m^\star} := 2s/(2s + m^\star)$ is motivated by the usual smooth multi-index benchmarks (Gaïffas & Lecué, 2007; Paris, 2014; Klock et al., 2021). The compact-class upper bound in Thm. 4.1 is no slower than this rate when

$$\frac{2\alpha}{2\alpha + d} \geq \frac{2s}{2s + m^\star}, \qquad \text{equivalently} \qquad \alpha \geq \frac{sd}{m^\star}.$$

This compares upper-bound exponents for different target descriptions; it does not assert a common minimax risk for the represented compact class and the Gaussian Sobolev space above. When the condition fails, our upper bound is slower, without establishing that the ambient exponent is unavoidable. The choice $\lambda_N = N^{-1}$ is the same for every intrinsic dimension and link regularity. In particular, for any $m^\star \in \{1, \ldots, d\}$, $L > 0$, and $m^\star/2 < s \leq \alpha m^\star/d$, the upper bound is at least as fast as $N^{-2s/(2s+m^\star)}$ uniformly over represented targets $f^\star(x) = h^\star(E_\star^\top x) \in \mathcal{F}_{\Theta_c}$ with $\|h^\star\|_{H^s(\gamma_{m^\star})} \leq L$.

Near-parametric upper bounds on the represented compact class. Suppose now that $\sigma \in C_b^\infty(\mathbb{R})$. For every fixed $\varepsilon \in (0, 1)$, choose a sufficiently large but fixed integer $\alpha > d/2$ such that $\frac{2\alpha}{2\alpha+d} \geq 1 - \varepsilon$. By Thm. 4.1, for every fixed $c \geq 1$, there is a constant $C_{\varepsilon,c} > 0$ such that, uniformly over $f^\star \in \mathcal{F}_{\Theta_c}$,

$$\|\widehat{f}_N - f^\star\|^2_{L^2(\gamma_d)} \leq C_{\varepsilon,c} N^{-1+\varepsilon}$$

with probability at least $1 - 4(eN)^{-c}$. The regularity order is fixed after ε is chosen and does not increase with N. Note that the near-parametric guarantee applies only to represented targets, whose smoothness follows from compact representability.

Takeaway: What determines the compact-model rate.

For targets represented by the compact model, the prediction-error upper-bound exponent can match or exceed the smooth multi-index reference exponent $2s/(2s + m^\star)$ when the activation is sufficiently regular. In particular, smooth activations yield near-parametric upper bounds on the represented compact class.

5 BEYOND COMPACTIFICATION: LEARNING ROUGH MULTI-INDEX TARGETS

Sec. 4 gives prediction guarantees conditional on compact representability. We now return to the rough targets excluded by Sec. 3. For $\tanh$, removing fixed bounds on the input-weight norm and bias magnitude can restore learnability of a fixed, infinite-dimensional family of multi-index links, while retaining bounded output weights and the entropy-regularized squared-loss formulation.

We use the noncompact $\tanh$ model and fixed reference measure from Sec. 2, and write $\widehat{f}^{\mathrm{nc}}_{N,\lambda} := \widehat{f}^{\Theta_{\mathrm{nc}}}_{N,\lambda}$. The estimator minimizes the entropy-regularized empirical risk over all of $\mathcal{P}(\Theta_{\mathrm{nc}})$, without restriction to the true index subspace or any fixed upper bound on the input-weight norm. This corresponds to the long-time limit of MFLD (see Prop. E.2).

Fix $2 \le m^\star < d$, $L, M, B_0 > 0$, and s satisfying

$$s > \frac{m^\star}{2}, \qquad 2sd < m^\star(d+5). \tag{2}$$

All model parameters and dimensions are fixed. Write $\beta_d = (d+5)/(2d+5)$, and set

$$\lambda_N := \min\left\{\frac{1}{2}, N^{-\beta_d}\right\}. \tag{3}$$

For comparison, $\mathcal{F}_{\mathrm{c}}(R_{\mathrm{c}}, B_{\mathrm{c}})$ denotes the $\tanh$ class with $|a| \le A$, $\|w\|_2 \le R_{\mathrm{c}}$, and $|b| \le B_{\mathrm{c}}$.

Theorem 5.1 (Learning rough targets beyond fixed parameter bounds)**.** *Consider the Gaussian multi-index model (1) with $2 \le m^\star < d$, $L, M, B_0 > 0$ and (2). There exist $r_+ > s$ and a fixed $L^2(\gamma_{m^\star})$-compact family $\mathcal{G}$, independent of N, E, R_{c} and B_{c}, such that $\mathcal{G} \subset \left\{h \in H^s(\gamma_{m^\star}) : \mathrm{supp}(h) \subset [-B_0, B_0]^{m^\star},\ \|h\|_{H^s(\gamma_{m^\star})} \le L,\ \|h\|_\infty \le M\right\} \backslash H^{r_+}(\gamma_{m^\star})$. The family $\mathcal{G}$ contains nondegenerate affine boxes of every finite dimension, and every $h \in \mathcal{G}$ is nonadditive and has intrinsic dimension $m^\star$. Then, for λ_N given by (3), there exists $C > 0$ such that, for every fixed $R_{\mathrm{c}}, B_{\mathrm{c}} \in (0, \infty)$, there exists $c_{\mathrm{gap}}(R_{\mathrm{c}}, B_{\mathrm{c}}) > 0$ for which, for all $N \ge 2$,*

$$\inf_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \inf_{f \in \mathcal{F}_{\mathrm{c}}(R_{\mathrm{c}}, B_{\mathrm{c}})} \|f - h(E^\top \cdot)\|^2_{L^2(\gamma_d)} \ge c_{\mathrm{gap}}(R_{\mathrm{c}}, B_{\mathrm{c}}),$$

$$\sup_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \mathbb{E}_{h,E}\left[\|\widehat{f}^{\mathrm{nc}}_{N,\lambda_N} - h(E^\top \cdot)\|^2_{L^2(\gamma_d)}\right] \le C N^{-2s/(2s+m^\star)}. \tag{4}$$

Here, the expectation is taken over the training sample, including both the covariates $X_1, \dots, X_N$ and the independent noises $\xi_1, \dots, \xi_N$, with h and E fixed.

Proof idea. We bound the prediction error by two quantities: the approximation error of the learned feature space and the discrepancy between the fitted output layer and the best output layer under the same norm constraint:

$$\|\widehat{f}_N - f^\star\|_2^2 \le 2 \underbrace{\|f^\circ_N - f^\star\|_2^2}_{\text{feature approximation error}} + 2 \underbrace{\|\widehat{f}_N - f^\circ_N\|_2^2}_{\text{output-layer error}},$$

where f°_N is the best predictor in the learned feature space subject to the RKHS norm bound. Since $f^\star(x) = h(E^\top x)$ depends only on $E^\top x$, orthogonal projection by conditional expectation decomposes each error into a component depending on these intrinsic variables and a residual component:

$$\|f^\circ_N - f^\star\|_2^2 = \underbrace{\|\mathbb{E}[f^\circ_N(X) \mid E^\top X, \mathcal{D}_N] - h(E^\top X)\|_2^2}_{\text{intrinsic approximation error}} + \underbrace{\|f^\circ_N(X) - \mathbb{E}[f^\circ_N(X) \mid E^\top X, \mathcal{D}_N]\|_2^2}_{\text{residual approximation error}},$$

$$\|\widehat{f}_N - f^\circ_N\|_2^2 = \underbrace{\|\mathbb{E}[\widehat{f}_N(X) - f^\circ_N(X) \mid E^\top X, \mathcal{D}_N]\|_2^2}_{\text{intrinsic output-layer error}} + \underbrace{\|\widehat{f}_N(X) - f^\circ_N(X) - \mathbb{E}[\widehat{f}_N(X) - f^\circ_N(X) \mid E^\top X, \mathcal{D}_N]\|_2^2}_{\text{residual output-layer error}}.$$

The feature-space oracle bound in (5) shows that the learned feature space contains a bounded-norm predictor that approximates the target well. Because the target itself has no variation orthogonal

to the index subspace, this simultaneously controls both the intrinsic approximation error and the residual variation. The same oracle bound also shows that the fitted head is close to the best bounded-norm head in the learned feature space; projecting this discrepancy onto the intrinsic and residual components controls the remaining two terms. Thus all four errors are controlled by the same estimates for the learned feature space. □

The contrast in Thm. 5.1 is on the same fixed family $\mathcal{G}$: every fixed compact model retains a positive approximation error, whereas the noncompact estimator is uniformly consistent over $\mathcal{G}$ and the unknown index frame E. This prediction result is accompanied by feature learning: Sec. 5.1 shows how the learned feature adapts to the target and recovers its unknown index subspace.

A concrete regime. For $(m^\star, s, d) = (2, 3/2, 3)$, take $h_0(z) = c\chi_1(z_1)\chi_2(z_2)|z_1|^{11/4}$ and $\mathcal{G} = h_0 + \mathcal{V}_4(\epsilon_0)$ with fixed sufficiently small $c, \epsilon_0 > 0$. Then $\mathcal{G} \subset H^3(\gamma_2) \setminus H^4(\gamma_2) \subset H^{3/2}(\gamma_2)$. For $N \geq 4$, $\lambda_N = N^{-8/11}$ and the expected error is at most $CN^{-3/5}$, uniformly over the family and the unknown two-dimensional index subspace.

Takeaway: Beyond fixed parameter bounds.

The same rough target family is separated from every fixed compact model, yet is learned by the noncompact long-time MFLD at the stated polynomial upper bound.

5.1 Feature learning and recovery of the index subspace

We now examine what the estimator in Thm. 5.1 learns beyond accurate prediction. Write $\widehat{\nu}_N := \widehat{\nu}^{\mathrm{nc}}_{N,\lambda_N}$ and $\widehat{f}_N := f_{\widehat{\nu}_N}$. Throughout this subsection, we use the assumptions and fixed family of that theorem. All probability bounds are uniform over $h^\star \in \mathcal{G}$ and $E_\star^\top E_\star = I_{m^\star}$ under their respective data laws.

Statistical feature learning. Let $\widehat{\phi}_N$ be the (w, b)-marginal of $\widehat{\nu}_N$. The learned feature kernel is

$$K_{\mathrm{feat},N}(x, x') := \int \tanh(w^\top x + b)\tanh(w^\top x' + b)\,\mathrm{d}\widehat{\phi}_N(w, b).$$

Let K_{init} be the same kernel formed from the (w, b)-marginal of $\nu_{0,\mathrm{nc}}$, and let $\mathcal{H}_N$ and $\mathcal{H}_{\mathrm{init}}$ be the RKHS of $K_{\mathrm{feat},N}$ and that of K_{init}, respectively. Since the output weights satisfy $|a| \leq A$, the fitted predictor belongs to $\mathcal{H}_N$ with $\|\widehat{f}_N\|_{\mathcal{H}_N} \leq A$. Let f_N° be the closest function to $f^\star$ in the radius-A ball of the learned feature space $\mathcal{H}_N$; $f_N^\circ \in \arg\min_{f \in \mathcal{H}_N; \|f\|_{\mathcal{H}_N} \leq A} \|f - f^\star\|_{L^2(\gamma_d)}$.

Proposition 5.2 (Statistical feature learning)**.** *There exist constants $C, C_s, c_{\mathrm{FL}}, c_{\mathrm{init}} > 0$ such that, for all sufficiently large N, with probability at least $1 - CN^{-8}$, the following three bounds hold:*

$$\|f_N^\circ - f^\star\|^2_{L^2(\gamma_d)} + \|\widehat{f}_N - f_N^\circ\|^2_{L^2(\gamma_d)} \leq C_s N^{-2s/(2s+m^\star)}, \tag{5}$$

$\inf_{g \in \mathcal{H}_{\mathrm{init}}; \|g\|_{\mathcal{H}_{\mathrm{init}}} \leq A} \|g - f^\star\|^2_{L^2(\gamma_d)} \geq c_{\mathrm{init}}$ *and* $\|K_{\mathrm{feat},N} - K_{\mathrm{init}}\|_{L^2(\gamma_d^{\otimes 2})} \geq c_{\mathrm{FL}}$. *Moreover, the learned feature spaces $(\mathcal{H}_N)_N$ and the fitted predictors $(\widehat{f}_N)_N$ satisfy the statistical feature-learning property in Def. D.11, which specializes Shang et al. (2026, Defs. 1–2) to the present setting.*

The first term in (5) shows that the learned feature space $\mathcal{H}_N$ represents the target $f^\star$ well under the same norm bound available to the fitted network. The second shows that the fitted output layer is close to the best output layer in that space. The two lower bounds show that these properties cannot be obtained while keeping the initial feature kernel essentially unchanged.

In Appendix D.5, we state Def. D.11 and verify all five conditions, including exact factorization through the learned feature map, target approximation, spectral-tail control, and the head alignment inequality. Thus, for this multi-index problem with an unknown link function and an unknown index subspace, the result establishes that the mean-field neural network trained via MFLD exhibits statistical feature learning in the sense of Shang et al. (2026), rather than merely showing a nonvanishing change in its feature kernel (cf. Yang & Hu, 2021).

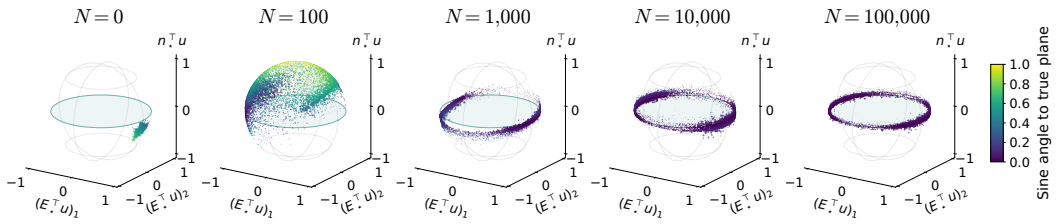


Figure 2: **Concentration towards the index subspace.** Normalized OU-smoothed predictor gradients for $(d, m^\star) = (3, 2)$. Positive-N panels use training time $T = 20$; $N = 0$ denotes initialization. The green disk marks the true plane; color indicates the sine angle to it, and marker area represents relative gradient energy. Details are in Appendix F.2.

Index-subspace recovery. Let $P_\star := E_\star E_\star^\top$. Since L^2 prediction error does not directly control gradients for rough targets, we first apply a fixed Ornstein–Uhlenbeck (OU) smoothing. For $G \sim \gamma_d$, define

$$(T_1^{(d)} f)(x) := \mathbb{E}\Big[f\Big(e^{-1}x + \sqrt{1 - e^{-2}}\, G\Big)\Big], \qquad \widehat{G}_N := \int \nabla T_1^{(d)} \widehat{f}_N(x) \nabla T_1^{(d)} \widehat{f}_N(x)^\top \, \mathrm{d}\gamma_d(x).$$

The map $f \mapsto \nabla T_1^{(d)} f$ is bounded on $L^2(\gamma_d)$, while for the target family the smoothing preserves the index subspace. Let $\widehat{P}_N$ be the spectral projector of $\widehat{G}_N$ above the deterministic threshold specified in Appendix D.7. The threshold depends only on N.

Corollary 5.3 (Index-subspace recovery)**.** *There exists $C > 0$ such that, for all sufficiently large N, with probability at least $1 - CN^{-8}$,*

$$\operatorname{rank}(\widehat{P}_N) = m^\star, \qquad \|\widehat{P}_N - P_\star\|_{\mathrm{F}}^2 \le C N^{-2s/(2s+m^\star)}.$$

For the target, the smoothed-gradient matrix has range $\mathrm{Range}(E_\star)$, and its nonzero eigenvalues are uniformly bounded away from zero over $\mathcal{G}$. The smoothed-gradient map is bounded from $L^2(\gamma_d)$ to vector-valued $L^2(\gamma_d)$. Hence the error of the gradient second-moment matrix is controlled linearly by $\|\widehat{f}_N - f^\star\|_{L^2(\gamma_d)}$. The uniform eigengap then converts this into a squared projector error of the same order as the squared prediction error. Figure 2 visualizes this predictor-level alignment in a finite-particle MFLD experiment, showing that the energy-weighted directions of OU-smoothed predictor gradients concentrate increasingly near the true index subspace as N grows.

At the parameter level, a fixed positive amount of effective active mass, weighted by $|\mathbb{E}_{\widehat{\nu}_N}[a \mid w, b]|$, lies in an angular neighborhood of $\mathrm{Range}(E_\star)$ of width $O((\log(eN))^{-1/2})$. Consequently, this neighborhood contains at least a fixed positive amount of the learned hidden marginal itself, whereas its mass under the initial Gaussian marginal tends to zero.

Takeaway: What the learned features encode.

At the noncompact MFLD limit, accurate prediction is accompanied by feature learning: the learned kernel moves a nonvanishing distance from initialization, a nonvanishing amount of active hidden mass aligns with the unknown index subspace, and that subspace, including its dimension, is consistently recoverable from the fitted predictor.

6 Discussion and Limitations

We derived prediction guarantees for the long-time limits of compact and noncompact MFLD in Gaussian multi-index regression. We constructed a fixed infinite-dimensional family of rough targets that no fixed compact class can approximate arbitrarily well, yet the long-time MFLD predictor satisfies a polynomially decaying statistical prediction bound. For this family, we also established statistical feature learning and recovery of the unknown index subspace.

Implications for learning-theoretic idealizations. Learning-theoretic idealizations should be judged by the statistical problems they preserve, not only by the proofs they simplify. Our comparison highlights that width, training time, sample size, and parameter scale are distinct resources: increasing the first three cannot compensate for a representation excluded by fixed bounds under our fixed output-weight budget. Consequently, an assumption of exact representability deserves the same scrutiny as an optimization assumption. The broader lesson is to evaluate tractability together with approximation fidelity, rather than treating model restrictions as neutral preliminaries to learning guarantees.

Implications for statistical benchmarks. Worst-case rates over broad smoothness classes cannot, by themselves, explain how a particular method learns structured targets. Our results suggest complementing smoothness and intrinsic dimension with the cost of approximation under the chosen model and prior. This model-relative perspective is closely related to the maxiset viewpoint, which fixes an estimator and a target rate and asks for the largest class of functions on which that rate is attained (Cohen et al., 2001). In the context of neural networks, a related perspective appears in Bach (2017), which connects variation-norm regularization with approximation, estimation, and adaptation to unknown low-dimensional structure for convex neural networks. More broadly, such a perspective separates favorable representation structure from smoothness alone and makes explicit which assumptions support fast learning, rather than attributing every improvement to adaptation.

Relation to prior work on mean-field learning theory. Mean-field theory formulates neural-network training as optimization over probability measures (Mei et al., 2018; Chizat & Bach, 2018). For entropy-regularized objectives, global convergence and log-Sobolev-based rates have been established for MFLD (Hu et al., 2021; Nitanda et al., 2022; Chizat, 2022). Statistical analyses establish advantages over fixed kernels (Suzuki et al., 2023) and study learned-kernel alignment in a two-timescale regime (Takakura & Suzuki, 2024). Two recent works are particularly close to ours along complementary directions. Whereas Mousavi-Hosseini et al. (2025) characterize the effective-dimension dependence of the sample complexity required for vanishing excess risk, our results provide explicit finite-sample prediction-error rates as functions of sample size, separating approximation from estimation and illuminating the effect of compactification. Shang et al. (2026) formalized the notion of statistical feature learning and developed a statistical theory for spherical MFLD, while we verify the feature-learning criterion in a noncompact setting and, beyond this, construct a fixed rough multi-index family that is separated from every fixed compact model, establish explicit excess-risk guarantees for noncompact MFLD, and recover the unknown index subspace from the fitted predictor.

Implications for regularization design and practice. Regularization design should distinguish which representations remain available from how their use is penalized. When the scales required by a target are unknown, a fixed cutoff commits to an approximation hypothesis that additional data cannot revise. Our results motivate retaining access to larger scales while controlling their cost, whether through soft penalties or suitably growing parameter domains. In practice, this suggests varying parameter cutoffs separately from regularization strength, width, and training duration when investigating error plateaus. Such experiments could help distinguish representational restrictions from optimization difficulties.

Limitations. Our statistical guarantees assume Gaussian covariates and fixed dimensions. The compact upper bound does not establish that its ambient-dimensional exponent is unavoidable, while the noncompact analysis relies on explicit representations of a fixed rough family with controlled quadratic cost. Extending these guarantees to other target classes, activations, and data distributions requires further analysis. Although convergence bounds hold at fixed sample size and temperature, they do not provide polynomial-time training or finite-particle guarantees at the statistical temperature. These extensions would connect the comparison more directly to implementations.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, design or provide feedback on experiments, implement methods, and assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks,

design or provide feedback on research methodology, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results, and no other required-disclosure tasks are applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, search for and identify relevant literature, assist with LaTeX formatting, and edit the paper to improve readability through proofreading. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Radosław Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13(34):1000–1034, 2008. doi: 10.1214/EJP.v13-521.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, Basel, second edition, 2008. doi: 10.1007/978-3-7643-8722-8.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL https://jmlr.org/papers/v18/14-546.html.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. doi: 10.1214/009053605000000282.

Vladimir I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998. doi: 10.1090/surv/062.

Lénaïc Chizat. Mean-field langevin dynamics: Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=BDqzLH1gEm.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

Albert Cohen, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Applied and Computational Harmonic Analysis*, 11(2):167–191, 2001. doi: 10.1006/acha.2000.0333.

Stéphane Gaïffas and Guillaume Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1(none):538–573, January 2007. ISSN 1935-7524. doi: 10.1214/07-EJS077. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-1/issue-none/Optimal-rates-and-adaptation-in-the-single-index-model-using/10.1214/07-EJS077.full.

Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. doi: 10.2307/2373688.

Richard Holley and Daniel Stroock. Logarithmic sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, 1987. doi: 10.1007/BF01011161.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043–2065, 2021. doi: 10.1214/20-AIHP1140.

Timo Klock, Alessandro Lanteri, and Stefano Vigogna. Estimating multi-index models with response-conditional least squares. *Electronic Journal of Statistics*, 15(1):589–629, January 2021. ISSN 1935-7524. doi: 10.1214/20-EJS1785. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-15/issue-1/Estimating-multi-index-models-with-response-conditional-least-squares/10.1214/20-EJS1785.full.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006. doi: 10.1214/009053606000001019.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In Evarist Giné, David M. Mason, and Jon A. Wellner (eds.), *High Dimensional Probability II*, pp. 443–457, Boston, MA, 2000. Birkhäuser Boston. ISBN 978-1-4612-1358-1.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Alireza Mousavi-Hosseini, Denny Wu, and Murat A. Erdogdu. Learning multi-index models with neural networks via mean-field Langevin dynamics. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/61d8577984e4ef0cba20966eb3ef2ed8-Abstract-Conference.html.

Atsushi Nitanda and Taiji Suzuki. Stochastic particle gradient descent for infinite ensembles, 2017. URL https://arxiv.org/abs/1712.05438.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9741–9757. PMLR, 2022. URL https://proceedings.mlr.press/v151/nitanda22a.html.

Quentin Paris. Minimax adaptive dimension reduction for regression. *Journal of Multivariate Analysis*, 128:186–202, July 2014. ISSN 0047259X. doi: 10.1016/j.jmva.2014.03.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0047259X14000645.

Mark S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, CA, 1964.

Grant M. Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/196f5641aa9dc87067da4ff90fd81e7b-Abstract.html.

Zong Shang, Tomoya Wakayama, Guillaume Lecué, and Taiji Suzuki. The Geometry of Statistical Feature Learning in Mean-Field Langevin Dynamics, June 2026. URL https://arxiv.org/abs/2606.31429. arXiv:2606.31429 [math.ST].

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020. doi: 10.1137/18M1192184.

Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field langevin dynamics: Classifying sparse parities and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34536–34556. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6cc321baf0a8611b1d1bdbd18822667b-Abstract-Conference.html.

Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47475–47509. PMLR, 2024. URL https://proceedings.mlr.press/v235/takakura24a.html.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, NY, 1996. doi: 10.1007/978-1-4757-2545-2.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

APPENDIX CONTENTS

A Hermite–Sobolev Tools and Compact Representability **14**

A.1 Hermite–Sobolev conventions and derivative bounds 14

A.2 Compactness of the induced neural class 14

A.3 Proof of the compactification barrier 15

A.4 Proof of inherited regularity 15

B Prediction Theory for the Compact Model **16**

B.1 Localized Hermite–Sobolev complexity 16

B.2 The information–residual local-complexity decomposition 17

B.3 Concentration and excess-risk comparison 17

B.4 Entropy-regularized approximation 19

B.5 Proof of Thm. 4.1 19

C A Fixed Rough Family and Finite-KL Approximation **19**

C.1 Variational setup and Gibbs identities 19

C.2 Dyadic ridge reconstruction and representation costs before amplitude scaling . . . 21

C.3 Construction and compactness of the rough family 23

C.4 Uniform separation from fixed compact classes 24

C.5 Exact representation with finite quadratic cost 25

C.6 Gaussian smoothing and entropy cost 25

C.7 Finite-KL approximation of the fixed family 26

C.8 Finite-KL approximation versus exact representation 26

D Feature Learning and Prediction in the Noncompact Model **27**

D.1 Empirical-process preliminaries 27

D.2 Uniform loss deviation for a fixed comparator 30

D.3 Learned RKHS geometry and the oracle 31

D.4 Feature-space oracle inequality 32

D.5 Formal feature learning and kernel non-laziness 33

D.6 Active hidden-layer mass near the index subspace 35

D.7 Smoothed gradients and index-subspace recovery 37

D.8 Intrinsic feature space and prediction transfer 39

D.9 Noncompact prediction bounds 40

E Long-Time MFLD and Projected Feature Scales **41**

E.1 Compact reflected MFLD and the RERM 41

E.2 Noncompact reflected MFLD and the RERM 43

E.3 Divergence of projected input-weight moments 45

F Experimental Details **46**

F.1 Compactification dichotomy 46

F.2 OU-gradient concentration near the index subspace 47

A Hermite–Sobolev Tools and Compact Representability

A.1 Hermite–Sobolev conventions and derivative bounds

Lemma A.1 (Gaussian Sobolev norms and weak derivatives)**.** *For every integer $r \geq 0$,*

$$h \in H^r(\gamma_D) \quad \Longleftrightarrow \quad \partial^\eta h \in L^2(\gamma_D) \quad \textit{for all } |\eta| \leq r,$$

with equivalent norms. In particular,

$$\|h\|^2_{H^r(\gamma_D)} \leq C_r \sum_{|\eta| \leq r} \|\partial^\eta h\|^2_{L^2(\gamma_D)}.$$

Moreover, for every $t \geq 0$ and every h supported on $[-B_0, B_0]^D$,

$$\|h\|_{W^{t,2}(\mathbb{R}^D)} \leq C_{D,t,B_0} \|h\|_{H^t(\gamma_D)}, \tag{6}$$

where fractional Euclidean orders are understood in the Bessel-potential sense.

Proof. Write $h_\eta = \langle h, \mathrm{H}_\eta \rangle_{L^2(\gamma_D)}$. For finite Hermite sums, the identity

$$\partial^\alpha \mathrm{H}_\eta = \sqrt{\frac{\eta!}{(\eta - \alpha)!}} \mathrm{H}_{\eta - \alpha} \quad (\eta \geq \alpha)$$

and Parseval's identity yield

$$\sum_{|\alpha| = j} \frac{j!}{\alpha!} \|\partial^\alpha h\|^2_{L^2(\gamma_D)} = \sum_\eta |\eta|^{\underline{j}} |h_\eta|^2.$$

Summing over $0 \leq j \leq r$ and using $\sum_{j=0}^r n^{\underline{j}} \asymp_r (1+n)^r$ proves norm equivalence on finite sums. Convergence of Hermite truncations then constructs all the required weak derivatives when the weighted coefficient sum is finite. Conversely, Gaussian integration by parts, with $\partial_i^* = -\partial_i + z_i$, identifies the coefficients of each weak derivative and gives the same formula. A smooth cutoff justifies integration against Hermite polynomials; its error vanishes by Cauchy–Schwarz and Gaussian polynomial integrability.

For the displayed norm inequality, use $j!/\alpha! \leq r!$ for $|\alpha| = j \leq r$ in the same coefficient identity. Its constant depends only on r.

For the last assertion, choose a fixed smooth compactly supported cutoff χ equal to one near $[-B_0, B_0]^D$. The preceding equivalence, the product rule, and the positive lower bound for the Gaussian density on $\operatorname{supp} \chi$ show that $h \mapsto \chi h$ is bounded from $H^r(\gamma_D)$ to $W^{r,2}(\mathbb{R}^D)$ at every integer $r \geq 0$. Interpolation of these Hilbert scales gives the same assertion at order t. Since $\chi h = h$, this proves (6). □

Thus the index in $H^s(\gamma_D)$ is the usual derivative order, including under the fixed-support Euclidean comparison.

A.2 Compactness of the induced neural class

Lemma A.2 (Compactness of the induced class and existence of the RERM)**.** *Let $\sigma \in C_b(\mathbb{R})$. The space $\mathcal{P}(\Theta_c)$ is compact under weak convergence, and the map $\nu \mapsto f_\nu$ is continuous from $\mathcal{P}(\Theta_c)$ to $L^2(\gamma_d)$. Its image $\mathcal{F}_{\Theta_c}$ is compact in $L^2(\gamma_d)$. Moreover, for every realized sample and every $\lambda > 0$, $J^{\Theta_c}_{N,\lambda}$ has a unique minimizer $\widehat{\nu}^{\Theta_c}_{N,\lambda}$, and $\widehat{\nu}^{\Theta_c}_{N,\lambda} \ll \nu_{0,c}$.*

Proof. The domain Θ_c is a compact metric space, so $\mathcal{P}(\Theta_c)$ is compact under weak convergence. Suppose that ν_n converges weakly to ν. For every fixed $x \in \mathbb{R}^d$, the map $(a, w, b) \mapsto a\sigma(\langle w, x\rangle + b)$ is bounded and continuous on Θ_c, and hence $f_{\nu_n}(x) \to f_\nu(x)$. Moreover, $|f_{\nu_n}(x) - f_\nu(x)| \le 2A\|\sigma\|_\infty$ uniformly in n and x. Dominated convergence gives $\|f_{\nu_n} - f_\nu\|_{L^2(\gamma_d)} \to 0$, so the image $\mathcal{F}_{\Theta_c}$ is compact.

For a fixed sample, the same pointwise argument gives $f_{\nu_n}(X_i) \to f_\nu(X_i)$ for every i, so $\nu \mapsto P_N \ell_{f_\nu}$ is weakly continuous. The variational representation

$$\mathrm{KL}(\nu\|\nu_{0,c}) = \sup_{\phi \in C(\Theta_c)} \left\{ \int \phi \, \mathrm{d}\nu - \log \int e^\phi \, \mathrm{d}\nu_{0,c} \right\}$$

shows that relative entropy is weakly lower semicontinuous. Since the empirical risk is weakly continuous, $J_{N,\lambda}^{\Theta_c}$ is weakly lower semicontinuous and finite at $\nu_{0,c}$, and compactness ensures existence of a minimizer. The empirical risk is convex in the mixing measure, while relative entropy is strictly convex on its effective domain, making the minimizer unique. Its objective is finite because it is no larger than the objective at $\nu_{0,c}$, which implies $\widehat{\nu}_{N,\lambda}^{\Theta_c} \ll \nu_{0,c}$. □

A.3 Proof of the compactification barrier

Proof of Prop. 3.1. Lem. A.2 shows that $\mathcal{F}_{\Theta_c}$ is compact and hence closed in $L^2(\gamma_d)$. A point outside a closed subset of a metric space has strictly positive distance from that subset, which proves the last claim. □

A.4 Proof of inherited regularity

Proof of Thm. 3.2. Choose $\nu \in \mathcal{P}(\Theta_c)$ such that $f_\nu = f^\star$ in $L^2(\gamma_d)$. Let $E_{\star,\perp} \in \mathbb{R}^{d\times(d-m^\star)}$ complete $E_\star$ to an orthogonal matrix. Let $Z \sim \gamma_{m^\star}$ and $U \sim \gamma_{d-m^\star}$ be independent. When $m^\star = d$, the variable U and the corresponding integral below are omitted.

Define

$$\widetilde{h}^\star(z) := \int_{\mathbb{R}^{d-m^\star}} f_\nu(E_\star z + E_{\star,\perp} u) \, \mathrm{d}\gamma_{d-m^\star}(u).$$

Since $E_\star Z + E_{\star,\perp} U \sim \gamma_d$ and $f_\nu = f^\star$ in $L^2(\gamma_d)$,

$$f_\nu(E_\star Z + E_{\star,\perp} U) = h^\star(Z) \qquad \text{almost surely.}$$

Taking conditional expectation given Z gives

$$\widetilde{h}^\star(Z) = \mathbb{E}[f_\nu(E_\star Z + E_{\star,\perp} U) \mid Z] = h^\star(Z) \qquad \text{almost surely.}$$

The functions $\widetilde{h}^\star$ and $h^\star$ are equal in $L^2(\gamma_{m^\star})$.

Using the representation of f_ν and Fubini's theorem, we may write

$$\widetilde{h}^\star(z) = \int_{\Theta_c} \int_{\mathbb{R}^{d-m^\star}} a\, \sigma(\langle w, E_\star z + E_{\star,\perp} u\rangle + b) \, \mathrm{d}\gamma_{d-m^\star}(u) \, \mathrm{d}\nu(a, w, b).$$

Let $\beta \in \mathbb{N}^{m^\star}$ satisfy $|\beta| \le 2r$. Since $\sigma^{(|\beta|)}$ is bounded, $|a| \le A$, and $\|w\|_2 \le 1$, dominated differentiation under both integrals is justified and yields

$$\partial^\beta \widetilde{h}^\star(z) = \int_{\Theta_c} \int_{\mathbb{R}^{d-m^\star}} a\, \sigma^{(|\beta|)}(\langle w, E_\star z + E_{\star,\perp} u\rangle + b) \, (E_\star^\top w)^\beta \, \mathrm{d}\gamma_{d-m^\star}(u) \, \mathrm{d}\nu(a, w, b).$$

Here

$$(E_\star^\top w)^\beta := \prod_{j=1}^{m^\star} (E_\star^\top w)_j^{\beta_j}.$$

Because $E_\star$ has orthonormal columns and $\|w\|_2 \le 1$,

$$\left|(E_\star^\top w)^\beta\right| \le 1.$$

These bounds imply

$$\left|\partial^\beta \widetilde{h}^\star(z)\right| \le A\|\sigma^{(|\beta|)}\|_\infty$$

for every $z \in \mathbb{R}^{m^\star}$. Dominated convergence also shows that $\partial^\beta \widetilde{h}^\star$ is continuous. Hence

$$\widetilde{h}^\star \in C_b^{2r}(\mathbb{R}^{m^\star}).$$

Since $\gamma_{m^\star}$ is a probability measure, the preceding bound also holds in L^2:

$$\|\partial^\beta \widetilde{h}^\star\|^2_{L^2(\gamma_{m^\star})} \le A^2\|\sigma^{(|\beta|)}\|^2_\infty.$$

Lem. A.1, applied in dimension $m^\star$, now yields

$$\begin{aligned}
\|\widetilde{h}^\star\|^2_{H^{2r}(\gamma_{m^\star})} &\le C_r \sum_{|\beta|\le 2r} \|\partial^\beta \widetilde{h}^\star\|^2_{L^2(\gamma_{m^\star})} \\
&\le C_r A^2 \sum_{k=0}^{2r} \left|\left\{\beta \in \mathbb{N}^{m^\star} : |\beta| = k\right\}\right| \|\sigma^{(k)}\|^2_\infty \\
&= C_r A^2 \sum_{k=0}^{2r} \binom{m^\star + k - 1}{k} \|\sigma^{(k)}\|^2_\infty.
\end{aligned}$$

This proves the finite-order assertion.

If $\sigma \in C_b^\infty(\mathbb{R})$, the same conditional-expectation representative can be differentiated to every finite order and belongs to $C_b^\infty(\mathbb{R}^{m^\star})$. Applying the finite-order result for every integer r and using the nesting of the Hermite–Sobolev spaces shows that $\widetilde{h}^\star \in H^s(\gamma_{m^\star})$ for every finite $s > 0$. □

B Prediction Theory for the Compact Model

B.1 Localized Hermite–Sobolev complexity

We first prove the localized Hermite–Sobolev bound used for both components of the prediction error. The subsequent fixed-point argument follows the local Rademacher complexity (Koltchinskii & Panchenko, 2000; Bartlett et al., 2005), with the spectral calculation below adapted to the Gaussian Hermite scale.

Lemma B.1 (Localized Hermite–Sobolev complexity)**.** *Let $U_1, \dots, U_N$ be independent with law γ_D, let $t > D$, and let $R \ge 1$ and $0 < r \le 1$. If $\zeta_1, \dots, \zeta_N$ are independent Rademacher or standard Gaussian variables independent of the U_i, then*

$$\mathbb{E}\left[\sup_{\substack{\|h\|_{H^t(\gamma_D)} \le R \\ \|h\|_{L^2(\gamma_D)} \le r}} \left|\frac{1}{N}\sum_{i=1}^N \zeta_i h(U_i)\right|\right] \le C_{D,t} R^{D/t} \frac{r^{1-D/t}}{\sqrt{N}}.$$

For Gaussian variables with variance σ^2, the right-hand side is multiplied by σ.

Proof. Write $h = \sum_\eta h_\eta \mathrm{H}_\eta$ and $Z_\eta = N^{-1}\sum_i \zeta_i \mathrm{H}_\eta(U_i)$, so $\mathbb{E}[Z_\eta^2] = N^{-1}$. Set $K = \lceil (R/r)^{2/t} \rceil$. Cauchy–Schwarz and $|\{\eta : |\eta| \le K\}| \le C_D K^D$ give

$$\mathbb{E}\left[\sup_h \left|\sum_{|\eta|\le K} h_\eta Z_\eta\right|\right] \le \frac{C r K^{D/2}}{\sqrt{N}}.$$

For the remaining terms, weighted Cauchy–Schwarz gives

$$\mathbb{E}\left[\sup_h \left|\sum_{|\eta|> K} h_\eta Z_\eta\right|\right] \le \frac{R}{\sqrt{N}} \left\{\sum_{|\eta|>K} (1+|\eta|)^{-t}\right\}^{1/2} \le \frac{C R K^{(D-t)/2}}{\sqrt{N}},$$

where the last sum is finite because $t > D$. Substituting the choice of K into both bounds proves the result. The Gaussian-variance statement follows by scaling. □

B.2 THE INFORMATION–RESIDUAL LOCAL-COMPLEXITY DECOMPOSITION

For $g_\nu := f_\nu - f^\star$, let $Z = E_\star^\top X$, define $g_\nu^{\mathrm{info}}(X) := \mathbb{E}[g_\nu(X) \mid Z]$, and set $g_\nu^{\mathrm{res}} := g_\nu - g_\nu^{\mathrm{info}}$. Conditional expectation is the orthogonal projection onto the closed subspace of Z-measurable functions, so

$$\|g_\nu\|_{L^2(\gamma_d)}^2 = \|g_\nu^{\mathrm{info}}\|_{L^2(\gamma_d)}^2 + \|g_\nu^{\mathrm{res}}\|_{L^2(\gamma_d)}^2. \tag{7}$$

Lemma B.2 (Information–residual local complexity)**.** *Under the assumptions of Thm. 4.1, let $q := d/(2\alpha) < 1$. There is a constant C such that, for $0 < r \le 1$ and for independent Rademacher or standard Gaussian multipliers ζ_i,*

$$\mathbb{E}\left[\sup_{\substack{\nu\in\mathcal{P}(\Theta_c)\\ \|g_\nu\|_{L^2(\gamma_d)}\le r}} \left|\frac{1}{N}\sum_{i=1}^N \zeta_i g_\nu(X_i)\right|\right] \le C\frac{r^{1-q}}{\sqrt{N}}.$$

Proof. For every multi-index β with $|\beta| \le 2\alpha$, differentiation under the compact neural integral gives

$$\partial^\beta f_\nu(x) = \int_{\Theta_c} a\,\sigma^{(|\beta|)}(\langle w, x\rangle + b) w^\beta \,\mathrm{d}\nu(a, w, b),$$

whose absolute value is uniformly bounded because $|a| \le A$, $\|w\|_2 \le 1$, and $\sigma \in C_b^{2\alpha}(\mathbb{R})$. By Lem. A.1, f_ν has a uniform $H^{2\alpha}(\gamma_d)$ bound. The information component is obtained by Gaussian averaging over directions orthogonal to $E_\star$; differentiating this average and applying the same lemma bounds it uniformly in $H^{2\alpha}(\gamma_{m^\star})$. Thm. 3.2, applied with $r = \alpha$, supplies the corresponding uniform bound for $h^\star$. These bounds imply

$$\|g_\nu^{\mathrm{info}}\|_{H^{2\alpha}(\gamma_{m^\star})} \le C.$$

In the tensor Hermite expansion, conditional expectation keeps exactly the coefficients whose indices in the orthogonal coordinates are zero. Conditional expectation is a contraction on every Hermite–Sobolev space, and $\|g_\nu^{\mathrm{res}}\|_{H^{2\alpha}(\gamma_d)} \le C$. If $\|g_\nu\|_2 \le r$, equation (7) implies $\|g_\nu^{\mathrm{info}}\|_2 \le r$ and $\|g_\nu^{\mathrm{res}}\|_2 \le r$. Applying Lem. B.1 to the two terms and using $0 < r \le 1$ yields

$$C\frac{r^{1-m^\star/(2\alpha)} + r^{1-d/(2\alpha)}}{\sqrt{N}} \le C\frac{r^{1-d/(2\alpha)}}{\sqrt{N}},$$

which proves the claim. □

For a random variable V, write $\|V\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}[\exp(|V|/t)] \le 2\}$.

B.3 CONCENTRATION AND EXCESS-RISK COMPARISON

Lemma B.3 (Adamczak concentration in normalized form)**.** *Let $W_1, \dots, W_N$ be independent and identically distributed random variables, and let $\mathcal{H}$ be a countable class of centered measurable functions. Define*

$$Z_N(\mathcal{H}) := \sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^N h(W_i)\right|, \qquad v_{\mathcal{H}}^2 := \sup_{h\in\mathcal{H}} \mathbb{E}[h(W_1)^2],$$

and let $H(w) := \sup_{h\in\mathcal{H}} |h(w)|$ and $K_N := \|\max_{1\le i\le N} H(W_i)\|_{\psi_1}$. There is an absolute constant $C_{\mathrm{A}} > 0$ such that, for every $u \ge 1$,

$$\mathbb{P}\left\{Z_N(\mathcal{H}) > C_{\mathrm{A}}\left(\mathbb{E}[Z_N(\mathcal{H})] + v_{\mathcal{H}}\sqrt{\frac{u}{N}} + K_N\frac{u}{N}\right)\right\} \le 4e^{-u}. \tag{8}$$

The same assertion holds for a separable class, or with outer probability and expectation.

Proof. Apply the unbounded empirical-process inequality of Adamczak (2008, Thm. 4) with tail parameter one to $\sup_{h\in\mathcal{H}} |\sum_{i=1}^N h(W_i)|$. Fixing its numerical parameters, absorbing the resulting constants, and dividing by N gives (8). □

Lemma B.4 (Uniform compact excess-risk comparison)**.** *There is a constant $C > 0$ such that, for every $N \geq 2$ and $x \geq 1$, with probability at least $1 - 4e^{-x}$, simultaneously for all $\nu \in \mathcal{P}(\Theta_c)$,*

$$|(P_N - P)(\ell_{f_\nu} - \ell_{f^\star})| \leq \frac{1}{4}\|f_\nu - f^\star\|^2_{L^2(\gamma_d)} + C\left\{N^{-1/(1+q)} + \kappa_N \frac{x+1}{N}\right\}, \tag{9}$$

where $\kappa_N := 1 + \sigma_\xi\sqrt{\log(eN)}$.

Proof. For $g = f_\nu - f^\star$, define $L_g := \ell_{f_\nu} - \ell_{f^\star}$. Under the regression model,

$$L_g = g^2 - 2\xi g, \qquad PL_g = \|g\|_2^2.$$

Because $f^\star \in \mathcal{F}_{\Theta_c}$, the compact model satisfies $\|g\|_\infty \leq D_c := 2A\|\sigma\|_\infty$. By symmetrization, contraction for g^2, and the Rademacher and Gaussian-multiplier versions of Lem. B.2, for $0 < r \leq 1$,

$$\mathbb{E}\left[\sup_{\|g\|_2 \leq r} |(P_N - P)L_g|\right] \leq C\frac{r^{1-q}}{\sqrt{N}}$$

and the same bound holds, after enlarging C, for $1 < r \leq D_c$ because the global compact complexity is of order $N^{-1/2}$ and $r^{1-q} \geq 1$. For the centered class $L_g - PL_g$, independence and centering of ξ imply

$$P(L_g - PL_g)^2 \leq PL_g^2 = Pg^4 + 4\sigma_\xi^2 Pg^2 \leq (D_c^2 + 4\sigma_\xi^2)r^2.$$

The weak variance in Lem. B.3 is at most Cr^2. Moreover, $|L_g - PL_g| \leq 2D_c^2 + 2D_c|\xi|$, and the Gaussian maximal estimate implies

$$\left\|\max_{1\leq i\leq N} |\xi_i|\right\|_{\psi_1} \leq C\sigma_\xi\sqrt{\log(eN)}.$$

The maximal centered-loss envelope has ψ_1 norm at most $C\kappa_N$. The loss class is separable because $\mathcal{P}(\Theta_c)$ is compact metrizable and the finite-sample loss process is continuous in the mixing measure. Lem. B.3, applied at radius r, yields with probability at least $1 - 4e^{-u}$,

$$\sup_{\|g\|_2 \leq r} |(P_N - P)L_g| \leq C\left\{\frac{r^{1-q}}{\sqrt{N}} + r\sqrt{\frac{u}{N}} + \kappa_N\frac{u}{N}\right\}.$$

Choose a numerical constant $C_\star \geq 1$ and set

$$r_0^2 := C_\star\left\{N^{-1/(1+q)} + \kappa_N\frac{x+1}{N}\right\}, \qquad r_j := 2^j r_0, \qquad u_j := x + j + 1.$$

If $r_0 \geq D_c$, apply the fixed-radius bound once at D_c and enlarge $C_\star$. Otherwise, stop at the first J such that $r_J \geq D_c$. A union bound over $0 \leq j \leq J$ gives an event of probability at least $1 - 4e^{-x}$ on which all fixed-radius bounds hold.

We now check the three terms on every radius. First,

$$\frac{r_j^{1-q}/\sqrt{N}}{r_j^2} = \frac{1}{\sqrt{N}\, r_j^{1+q}} \leq \frac{1}{\sqrt{N}\, r_0^{1+q}} \leq C_\star^{-(1+q)/2}.$$

Also $u_j \leq C(x+1)4^j$ and $r_j^2 = 4^j r_0^2$. Since $r_0^2 \geq C_\star\kappa_N(x+1)/N$ and $\kappa_N \geq 1$,

$$\frac{r_j\sqrt{u_j/N}}{r_j^2} \leq CC_\star^{-1/2}, \qquad \frac{\kappa_N u_j/N}{r_j^2} \leq CC_\star^{-1}.$$

Choosing $C_\star$ sufficiently large makes the localized complexity, weak-variance, and maximal-envelope terms at most $r_j^2/16$ in total. The innermost ball $\|g\|_2 \leq r_0$ is absorbed by the remainder in (9). On a dyadic shell $r_{j-1} < \|g\|_2 \leq r_j$, one has $r_j^2/16 = r_{j-1}^2/4 < \|g\|_2^2/4$. This proves (9). □

B.4 Entropy-regularized approximation

Lemma B.5 (Regularized approximation inside the compact model)**.** *Suppose that* $\sigma \in C_b^1(\mathbb{R})$*. If* $f^\star \in \mathcal{F}_{\Theta_c}$ *and* $0 < \lambda \le 1$*, then*

$$\inf_{\nu \in \mathcal{P}(\Theta_c)} \left\{ \|f_\nu - f^\star\|_{L^2(\gamma_d)}^2 + \lambda \,\mathrm{KL}(\nu \| \nu_{0,c}) \right\} \le C\lambda \log(e/\lambda).$$

Proof. Choose ν with $f_\nu = f^\star$. For independent $\theta \sim \nu$ and $\theta' \sim \nu_{0,c}$, let ν_λ be the law of $(1-\sqrt{\lambda})\theta + \sqrt{\lambda}\theta'$. Convexity of Θ_c keeps this measure in the parameter domain. Conditional on θ, the affine image of the uniform reference measure has Jacobian $\lambda^{(d+2)/2}$, so $\mathrm{d}\nu_\lambda / \mathrm{d}\nu_{0,c} \le \lambda^{-(d+2)/2}$ and

$$\mathrm{KL}(\nu_\lambda \| \nu_{0,c}) \le \frac{d+2}{2} \log(1/\lambda).$$

By the mean-value theorem and the bounds $|a| \le A$, $\|w\|_2 \le 1$, and $|b| \le B$,

$$|f_{\nu_\lambda}(x) - f_\nu(x)| \le 2A\sqrt{\lambda}\left\{ \|\sigma\|_\infty + \|\sigma'\|_\infty (\|x\|_2 + B) \right\}.$$

Since $X \sim \gamma_d$, integration and squaring yield $\|f_{\nu_\lambda} - f^\star\|_2^2 \le C\lambda$. Substitution of ν_λ into the infimum proves the result. □

B.5 Proof of Thm. 4.1

Proof of Thm. 4.1. Take $x = c\log(eN)$ in Lem. B.4. Because $q > 0$, $\kappa_N(x+1)/N \le C_c N^{-1/(1+q)}$ uniformly for $N \ge 2$. On this event, optimality of $\widehat{\nu}_{N,1/N}^{\Theta_c}$, equation (9), and $PL_g = \|g\|_2^2$ imply, for every comparator ν,

$$\frac{3}{4}\|\widehat{f}_N - f^\star\|_2^2 + \frac{1}{N}\,\mathrm{KL}(\widehat{\nu}_{N,1/N}^{\Theta_c} \| \nu_{0,c}) \le \frac{5}{4}\|f_\nu - f^\star\|_2^2 + \frac{1}{N}\,\mathrm{KL}(\nu \| \nu_{0,c}) + C_c N^{-1/(1+q)}.$$

Taking the infimum over ν and applying Lem. B.5 with $\lambda = N^{-1}$ yields

$$\|\widehat{f}_N - f^\star\|_2^2 \le C_c \left\{ N^{-1/(1+q)} + \frac{\log(eN)}{N} \right\} \le C_c N^{-1/(1+q)}.$$

Finally, since $q = d/(2\alpha)$,

$$\frac{1}{1+q} = \frac{2\alpha}{2\alpha + d}.$$

The probability of the event is $1 - 4(eN)^{-c}$, which proves the theorem. □

C A Fixed Rough Family and Finite-KL Approximation

C.1 Variational setup and Gibbs identities

In Appendices C and D, and in the noncompact parts of Appendix E, the activation is $\tanh$, and

$$\Theta_{\mathrm{nc}} = [-A, A] \times \mathbb{R}^d \times \mathbb{R}, \qquad \nu_0 = \mathrm{Unif}([-A, A]) \otimes \mathcal{N}(0, \tau_w^2 I_d) \otimes \mathcal{N}(0, \tau_b^2),$$

where all three parameters are fixed and positive. We abbreviate

$$\varphi_\theta(x) = a\tanh(w^\top x + b), \qquad f_\nu(x) = \int \varphi_\theta(x)\,\mathrm{d}\nu(\theta), \qquad \|g\|_2 = \|g\|_{L^2(\gamma_d)}.$$

In particular, $|f_\nu(x)| \le A$ for every probability measure ν. The data satisfy $X \sim \gamma_d$ and $Y = f^\star(X) + \xi$, where $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ is independent of X and $\|f^\star\|_\infty \le M$. Write

$$J_{N,\lambda}(\nu) = P_N \ell_{f_\nu} + \lambda\,\mathrm{KL}(\nu \| \nu_0),$$
$$\mathcal{A}_\lambda(\nu) = \|f_\nu - f^\star\|_2^2 + \lambda\,\mathrm{KL}(\nu \| \nu_0), \qquad \mathfrak{a}_\lambda(f^\star) = \inf_\nu \mathcal{A}_\lambda(\nu).$$

Their minimizers are denoted by $\widehat{\nu}_{N,\lambda}$ and ν_λ, with predictors $\widehat{f}_{N,\lambda}$ and f_λ. These are the minimizers on $\mathcal{P}(\Theta_{\mathrm{nc}})$. In particular, neither is constrained to the true index subspace.

Lemma C.1 (Existence and Gibbs identities)**.** *For every realized sample and $\lambda > 0$, the empirical and population objectives have unique minimizers of finite KL divergence. For every ν of finite KL divergence,*

$$J_{N,\lambda}(\nu) - J_{N,\lambda}(\widehat{\nu}_{N,\lambda}) = \|f_\nu - \widehat{f}_{N,\lambda}\|_N^2 + \lambda \,\mathrm{KL}(\nu\|\widehat{\nu}_{N,\lambda}), \tag{10}$$

$$\mathcal{A}_\lambda(\nu) - \mathfrak{a}_\lambda(f^\star) = \|f_\nu - f_\lambda\|_2^2 + \lambda \,\mathrm{KL}(\nu\|\nu_\lambda), \tag{11}$$

where $\|g\|_N^2 = N^{-1}\sum_{i=1}^N g(X_i)^2$.

Proof. For every Borel set D, the binary relative-entropy inequality gives

$$\mathrm{KL}(\nu\|\nu_0) \geq \nu(D)\log\frac{1}{\nu_0(D)} - \log 2.$$

Thus entropy sublevels are uniformly tight. Both objectives are finite at ν_0, and a minimizing sequence lies in an entropy sublevel. For fixed x, the map $\theta \mapsto \varphi_\theta(x)$ is bounded and continuous. Consequently, weak convergence of measures implies pointwise convergence of predictors and, by dominated convergence, convergence in $L^2(\gamma_d)$. The empirical risk and the population prediction error are therefore weakly continuous. Lower semicontinuity of relative entropy and the direct method give existence. Strict convexity of KL on its effective domain, together with convexity of the squared loss, gives uniqueness.

Here is also a direct justification of the Gibbs equation, avoiding any assumption that the minimizing density is initially positive. Let F denote either squared-loss functional and let μ be its entropy-regularized minimizer. Its first variation V_μ is bounded:

$$V_\mu(\theta) = \begin{cases} 2P_N[(f_\mu - Y)\varphi_\theta], & \text{empirical case,} \\ 2\langle f_\mu - f^\star, \varphi_\theta\rangle_{L^2(\gamma_d)}, & \text{population case.} \end{cases}$$

Define $\pi_\mu = Z_\mu^{-1}e^{-V_\mu/\lambda}\nu_0$. The Gibbs variational identity shows that π_μ uniquely minimizes $G_\mu(\eta) := \int V_\mu \,\mathrm{d}\eta + \lambda\,\mathrm{KL}(\eta\|\nu_0)$, and

$$G_\mu(\pi_\mu) - G_\mu(\mu) = -\lambda\,\mathrm{KL}(\mu\|\pi_\mu).$$

If $\mu \neq \pi_\mu$, comparison with $(1-t)\mu + t\pi_\mu$, convexity of KL, and the exact quadratic expansion of F give an objective difference bounded above by

$$-t\lambda\,\mathrm{KL}(\mu\|\pi_\mu) + t^2\|f_{\pi_\mu} - f_\mu\|_Q^2 < 0$$

for sufficiently small $t > 0$. Here Q is the empirical input measure or γ_d. This contradicts optimality. Hence

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu_0} = Z_\mu^{-1}e^{-V_\mu/\lambda}.$$

In particular, μ is equivalent to ν_0 with a bounded log-density ratio. Substitution of this formula into the quadratic expansion of $F(\nu) - F(\mu)$ gives (10)–(11). □

A useful consequence is the basic inequality. Setting $g = \widehat{f}_{N,\lambda} - f_\lambda$, empirical optimality and (11) imply

$$\|g\|_2^2 + \lambda\,\mathrm{KL}(\widehat{\nu}_{N,\lambda}\|\nu_\lambda) \leq (P - P_N)(\ell_{\widehat{f}_{N,\lambda}} - \ell_{f_\lambda}).$$

Indeed, the left-hand side equals $P(\ell_{\widehat{f}_{N,\lambda}} - \ell_{f_\lambda}) + \lambda\{\mathrm{KL}(\widehat{\nu}_{N,\lambda}\|\nu_0) - \mathrm{KL}(\nu_\lambda\|\nu_0)\}$. The process on the right contains both design and noise fluctuations, since

$$\ell_{\widehat{f}_{N,\lambda}} - \ell_{f_\lambda} = g^2 + 2(f_\lambda - f^\star)g - 2\xi g.$$

The full-predictor proof in Appendix D uses empirical optimality centered at a deterministic finite-KL comparator and controls all three terms of this loss difference.

C.2 Dyadic ridge reconstruction and representation costs before amplitude scaling

All dimensions and model constraint bounds are fixed. We first establish the representation costs of an unscaled center and of smooth perturbations, and then choose their amplitudes within the fixed output-weight bound.

Lemma C.2 (Dyadic ridge reconstruction)**.** *Let h be real-valued, integrable, square-integrable, and supported on $[-B_0, B_0]^{m^\star}$. There are real coefficients $q_j[h](u,t)$ for which, with the Fourier convention $\widehat{h}(\xi) = \int e^{-i\xi^\top z} h(z)\,\mathrm{d}z$ and inverse factor $(2\pi)^{-m^\star}$,*

$$\begin{aligned} h_j(z) &= \int_{\mathbb{S}^{m^\star-1}} \int_{\mathbb{R}} q_j[h](u,t)\tanh(T_j(u^\top z - t))\,\mathrm{d}t\,\mathrm{d}u, \\ \widehat{h}_j(\xi) &= \psi(|\xi|/T_j)\widehat{h}(\xi). \end{aligned} \tag{12}$$

Here $\mathrm{d}u$ is surface measure, $T_j = 2^j$, and ψ is a smooth dyadic partition as specified below. The sum $\sum_j h_j$ equals h in $L^2(\mathbb{R}^{m^\star})$. If $\sum_j \iint |q_j[h]| < \infty$, the neural series converges uniformly to a representative of h.

Proof. Choose a nonnegative $\psi \in C_c^\infty((1/2,2))$ with $\sum_{j\in\mathbb{Z}} \psi(2^{-j}r) = 1$ for $r > 0$. Set

$$\widehat{h}_j(\xi) = \psi(|\xi|/T_j)\widehat{h}(\xi), \qquad T_j = 2^j.$$

For $u \in \mathbb{S}^{m^\star-1}$ define the one-dimensional Fourier transform of $q_j(u,\cdot)$ by

$$\widehat{q}_j(u,r) = \frac{iT_j}{(2\pi)^{m^\star}} \sinh\left(\frac{\pi r}{2T_j}\right) |r|^{m^\star-1}\psi(|r|/T_j)\widehat{h}(ru). \tag{13}$$

This is smooth and compactly supported away from $r = 0$, so $q_j(u,\cdot)$ is Schwartz. The conjugate symmetry of (13) makes it real, and $\int q_j(u,t)\,\mathrm{d}t = 0$.

To check the reconstruction constant, the one-dimensional transform identity needed here is

$$\widehat{\tanh(T\cdot)}(r) = \frac{\pi}{iT\sinh(\pi r/(2T))}, \qquad r \neq 0.$$

For completeness, substituting $y = e^{2x}$ gives

$$\int_{\mathbb{R}} e^{-irx}\operatorname{sech}^2 x\,\mathrm{d}x = 2\int_0^\infty \frac{y^{-ir/2}}{(1+y)^2}\,\mathrm{d}y = 2\Gamma(1-ir/2)\Gamma(1+ir/2) = \frac{\pi r}{\sinh(\pi r/2)}.$$

Differentiation and scaling give the displayed transform away from zero. A distribution supported at zero does not affect (13). Consequently the transform of $\int q_j(u,t)\tanh(T_j(v-t))\,\mathrm{d}t$ is

$$\frac{\pi}{(2\pi)^{m^\star}}|r|^{m^\star-1}\psi(|r|/T_j)\widehat{h}(ru).$$

Equivalently, one may first differentiate this convolution and then use that it tends to zero at both infinities. The signed-polar identity

$$\int_{\mathbb{R}^{m^\star}} F(\xi)\,\mathrm{d}\xi = \frac{1}{2}\int_{\mathbb{S}^{m^\star-1}}\int_{\mathbb{R}} F(ru)|r|^{m^\star-1}\,\mathrm{d}r\,\mathrm{d}u$$

now proves the single-band identity in (12).

We next bound its coefficients. Cauchy–Schwarz and one-dimensional Plancherel imply, for a Schwartz function q,

$$\int_{\mathbb{R}} (1+t^2)|q(t)|\,\mathrm{d}t \le C\left(\sum_{k=0}^{3}\|\partial_r^k\widehat{q}\|_2^2\right)^{1/2}. \tag{14}$$

For example, use the two factors $(1+t^2)^{-1/2}$ and $(1+t^2)^{3/2}|q(t)|$ before applying Plancherel.

The multipliers of the partial sums $\sum_{|j|\le J} h_j$ converge to one away from zero and are bounded by one. Plancherel and dominated convergence prove convergence to h in $L^2(\mathbb{R}^{m^\star})$. Absolute summability of the coefficient masses, when available, gives uniform convergence of the neural series because $|\tanh| \le 1$. Its limit agrees almost everywhere with h. □

Lemma C.3 (An unscaled rough center)**.** *Let $m^\star \geq 2$, $s > m^\star/2$, $k = \lceil s \rceil$, and $r_+ = 2k+1$. Choose $\chi_j \in C_c^\infty((-B_0, B_0))$ equal to one near zero, and set*

$$H_0(z) = \chi_1(z_1)\,\mathrm{sgn}(z_1)|z_1|^{2k}\prod_{j=2}^{m^\star}\chi_j(z_j).$$

Then $H_0 \in H^{2k}(\gamma_{m^\star}) \setminus H^{r_+}(\gamma_{m^\star})$, with $2k \geq s$, $2k > 5/2$, and $r_+ > (m^\star+5)/2$.

Proof. Every weak derivative through order $2k$ is bounded and compactly supported, since the $2k$-th derivative of $\mathrm{sgn}(t)|t|^{2k}$ equals $(2k)!\,\mathrm{sgn}(t)$. In contrast,

$$\partial_1^{2k+1} H_0 = 2(2k)!\,\delta_0 \otimes \prod_{j=2}^{m^\star}\chi_j + r_0, \qquad r_0 \in C_c^\infty(\mathbb{R}^{m^\star}).$$

The singular measure cannot be an $L^2(\gamma_{m^\star})$ function, so Lem. A.1 proves both assertions of membership and nonmembership. Since $m^\star \geq 2$ and $s > m^\star/2$, we have $k \geq 2$. For $m^\star = 2, 3$, $r_+ \geq 5 > (m^\star+5)/2$; for $m^\star \geq 4$, $2k > m^\star$ gives $r_+ > m^\star + 1 > (m^\star+5)/2$. □

Lemma C.4 (Quadratic ridge costs before amplitude scaling)**.** *For the center and exponent in Lem. C.3, there are constants $C_{\mathrm{ctr}}, C_{\mathrm{pert}} \geq 1$ such that*

$$\sum_j \iint \{1 + T_j^2(1+t^2)\}|q_j[H_0](u,t)|\,\mathrm{d}t\,\mathrm{d}u \leq C_{\mathrm{ctr}},$$
$$\sum_j \iint \{1 + T_j^2(1+t^2)\}|q_j[g](u,t)|\,\mathrm{d}t\,\mathrm{d}u \leq C_{\mathrm{pert}}\|g\|_{H^{r_+}(\gamma_{m^\star})}.$$

The second bound holds for every $g \in H^{r_+}(\gamma_{m^\star})$ supported on the fixed box. The same conclusions hold for $m^\star = 2$, $H_0(z) = \chi_1(z_1)\chi_2(z_2)|z_1|^{11/4}$, and $r_+ = 4$. These constants do not depend on the index frame, N, or compact parameter bounds.

Proof. First consider a perturbation g. The fixed-support comparison (6) gives $\|g\|_{W^{r_+,2}(\mathbb{R}^{m^\star})} \leq C\|g\|_{H^{r_+}}$, and the same estimate holds for $z^\eta g$ with $|\eta| \leq 3$. Indeed, multiplication by a smooth compactly supported function equal to z^η near the support is bounded on this Euclidean Sobolev space. For $T = T_j \geq 1$, the ℓ-th derivative of the multiplier preceding $\widehat{g}(ru)$ in (13) is bounded by $C_\ell T^{m^\star-\ell}$ on $T/2 < |r| < 2T$. The product rule and polar coordinates give, for $0 \leq \ell \leq 3$,

$$\int_{\mathbb{S}^{m^\star-1}} \|\partial_r^\ell \widehat{q}_j[g](u,\cdot)\|_2^2\,\mathrm{d}u \leq CT^{m^\star+1}\sum_{|\eta|\leq 3}\int_{T/2<|\xi|<2T}|\partial_\xi^\eta \widehat{g}(\xi)|^2\,\mathrm{d}\xi$$
$$\leq C\|g\|_{H^{r_+}}^2 T^{m^\star+1-2r_+}.$$

Apply (14) and Cauchy–Schwarz on the sphere to obtain

$$T_j^2 \iint (1+t^2)|q_j[g](u,t)|\,\mathrm{d}t\,\mathrm{d}u \leq C\|g\|_{H^{r_+}} T_j^{(m^\star+5)/2-r_+}, \qquad T_j \geq 1.$$

This is summable because $r_+ > (m^\star+5)/2$.

For the center, its product structure gives a sharper angular estimate. Write $a = 2k$ for the general center and $a = 11/4$ for the two-dimensional example. For any sufficiently large fixed K and $|\eta| \leq 3$,

$$|\partial_\xi^\eta \widehat{H}_0(\xi)| \leq C_{K,\eta}(1+|\xi_1|)^{-a-1}\prod_{j=2}^{m^\star}(1+|\xi_j|)^{-K}. \tag{15}$$

For $a = 2k$, the $(a+1)$-st derivative of the first factor is a finite measure, and remains so after multiplication by any fixed spatial monomial. Integration by parts gives the first factor in this bound. The remaining factors are smooth and compactly supported, hence have arbitrarily fast Fourier decay. For $a = 11/4$, the third derivative has a $\mathrm{sgn}(t)|t|^{-1/4}$ singularity. Splitting its Fourier integral at

$|t| = |\xi_1|^{-1}$ bounds the inner integral by $C|\xi_1|^{-3/4}$; one integration by parts on the complement gives the same bound. Three preceding integrations by parts yield $|\xi_1|^{-15/4}$. Multiplication by a spatial monomial does not worsen this singularity, proving (15) also in this case.

On $|u_1| \geq 1/2$, the multiplier bound and its support length $O(T_j)$ imply

$$\left(\sum_{\ell=0}^{3} \|\partial_r^\ell \widehat{q}_j[H_0](u,\cdot)\|_2^2\right)^{1/2} \leq C T_j^{m^\star - a - 1/2} \prod_{j=2}^{m^\star} (1 + T_j|u_j|)^{-K}.$$

The integral of the product on this part of the sphere is $O(T_j^{-(m^\star-1)})$, by the two coordinate charts around the first axis. On $|u_1| < 1/2$, some other coordinate has magnitude at least a positive dimension-dependent constant; choosing K sufficiently large makes this part no larger. Consequently, (14) gives

$$T_j^2 \iint (1+t^2)|q_j[H_0](u,t)|\,\mathrm{d}t\,\mathrm{d}u \leq C T_j^{5/2-a}, \qquad T_j \geq 1.$$

Since $a > 5/2$, these terms are summable. For the concrete example the two high-band exponents are $-1/4$ and $-1/2$ for the center and perturbation, respectively.

It remains to include all low Fourier-frequency bands. For $T_j \leq 1$, write $\widehat{q}_j[h](u, T_j r) = T_j^{m^\star} Q_{j,u}[h](r)$. Compact support bounds every fixed spatial moment by $C\|h\|_1$, so these functions have a common compact support away from zero and derivatives through order three bounded by $C\|h\|_1$. Rescaling (14) gives

$$\iint \{1 + (T_j t)^2\}|q_j[h](u,t)|\,\mathrm{d}t\,\mathrm{d}u \leq C\|h\|_1 T_j^{m^\star}.$$

The sum over $j \leq 0$ converges. This controls both coefficient mass and quadratic bias cost, as well as the input-weight cost term since $T_j^2 \leq 1$. Combined with the bounds for the high Fourier-frequency bands, it proves both claims. □

C.3 Construction and compactness of the rough family

Fix $m^\star \geq 2$, $s > m^\star/2$, and positive L, M, B_0, A. Use H_0, $r_+ = 2\lceil s \rceil + 1$, and the constants of Lem. C.4. Choose fixed positive amplitudes satisfying

$$\begin{aligned}
0 < c_0 &\leq \min\left\{\frac{L}{2\|H_0\|_{H^s}}, \frac{M}{2\|H_0\|_\infty}, \frac{A}{8C_{\mathrm{ctr}}}\right\}, \qquad h_0 = c_0 H_0, \\
0 < \epsilon_0 &\leq \min\left\{\frac{L}{4}, \frac{M}{4}, \frac{A}{8C_{\mathrm{pert}}}\right\}, \\
\mathcal{V}_{r_+}(\epsilon_0) &:= \left\{g : \operatorname{supp} g \subset [-B_0, B_0]^{m^\star}, \quad \|g\|_{H^{r_+}(\gamma_{m^\star})} + \|g\|_\infty \leq \epsilon_0\right\}, \\
\mathcal{G} &:= h_0 + \mathcal{V}_{r_+}(\epsilon_0).
\end{aligned} \tag{16}$$

These choices are independent of N, E, and all compact parameter bounds. In the concrete $(m^\star, s, d) = (2, 3/2, 3)$ example we instead use the second center in Lem. C.4 and $r_+ = 4$, with the same amplitude rule. Its membership in $H^3 \setminus H^4$ is verified in Appendix D.9.

Lemma C.5 (Compactness and richness of the target family)**.** *The set $\mathcal{G}$ is compact in $L^2(\gamma_{m^\star})$ and satisfies*

$$\mathcal{G} \subset H^s(\gamma_{m^\star}) \setminus H^{r_+}(\gamma_{m^\star}), \qquad \sup_{h\in\mathcal{G}} \|h\|_{H^s} \leq L, \quad \sup_{h\in\mathcal{G}} \|h\|_\infty \leq M.$$

It contains nondegenerate affine boxes of every finite dimension. Every $h \in \mathcal{G}$ is nonzero and nonadditive, including after orthogonal coordinate changes. For every $E^\top E = I_{m^\star}$, the function $h(E^\top \cdot)$ has intrinsic dimension exactly $m^\star$.

Proof. The norm bounds follow from monotonicity of the Hermite weights and the triangle inequality. If $h_0 + g$ belonged to H^{r_+}, then $h_0 = (h_0 + g) - g$ would also belong to H^{r_+}, a contradiction. In particular, no element of $\mathcal{G}$ is zero.

Let $\Pi_{\le J}^{(m^\star)}$ be Hermite projection onto total degree at most J. Then

$$\sup_{g\in\mathcal{V}_{r_+}(\epsilon_0)} \|(I-\Pi_{\le J}^{(m^\star)})g\|_2^2 \le \epsilon_0^2(1+J)^{-r_+} \longrightarrow 0.$$

The projected sets are bounded in finite-dimensional spaces, so $\mathcal{V}_{r_+}(\epsilon_0)$ is totally bounded. Its support condition is closed under L^2 convergence, and both norms, and their sum, are lower semi-continuous for this convergence. For the Sobolev norm use coefficient convergence and Fatou; for the L^∞ norm use an almost-everywhere convergent subsequence. Thus it is closed and compact. Translation by h_0 preserves compactness.

For any integer $q \ge 1$, choose linearly independent smooth bumps $\eta_1, \dots, \eta_q$ with disjoint supports in $(-B_0, B_0)^{m^\star}$. Writing $b_j = \|\eta_j\|_{H^{r_+}} + \|\eta_j\|_\infty$, we have

$$|t_j| \le \frac{\epsilon_0}{q b_j} \quad (1 \le j \le q) \quad \Longrightarrow \quad \sum_{j=1}^{q} t_j\eta_j \in \mathcal{V}_{r_+}(\epsilon_0).$$

This proves the asserted infinite-dimensional richness.

A nonzero compactly supported $L^2(\mathbb{R}^{m^\star})$ function cannot be invariant under all translations in a nonzero direction: Fubini's theorem would make it constant on almost every parallel line, and compact support forces those constants to vanish. If $h(E^\top x) = \widetilde{h}(F^\top x)$ for a matrix F of rank $r < m^\star$, choose

$$0 \ne v \in \mathrm{Range}(E) \cap \ker(F^\top).$$

Such a vector exists since the intersection has dimension at least $m^\star - r$. Equivalence of Gaussian and Lebesgue null sets transfers the translation invariance of the right-hand side to h, in the nonzero direction $E^\top v$. This is impossible. Hence the intrinsic dimension is $m^\star$.

Finally, a compactly supported function of the additive form $c + \sum_{j=1}^{m^\star} h_j(z_j)$ must vanish when $m^\star \ge 2$. For each i, freeze the other coordinates with some z_j, $j \ne i$, outside a box containing the support. Fubini's theorem then forces h_i to be constant almost everywhere. Repeating for every i makes the sum constant, hence zero. An orthogonal image of a compact set is again contained in a bounded box, so the same argument applies after any orthogonal coordinate change. □

C.4 Uniform separation from fixed compact classes

Proposition C.6 (One family separated from every fixed compact class)**.** *For every fixed $R_c, B_c \in (0,\infty)$, let*

$$\Theta_c(R_c, B_c) = [-A, A] \times \overline{B}_d(0, R_c) \times [-B_c, B_c],$$

and let $\mathcal{F}_c(R_c, B_c)$ be its neural class. The family in (16) satisfies

$$\inf_{\substack{h\in\mathcal{G}\\ E^\top E = I_{m^\star}}} \inf_{f\in\mathcal{F}_c(R_c,B_c)} \|f - h(E^\top\cdot)\|_2^2 \ge c_{\mathrm{gap}}(R_c, B_c) > 0. \tag{17}$$

The family itself is independent of R_c, B_c.

Remark 1 (Quantifier order). The quantifier order is $\exists\mathcal{G} \quad \forall R_c, B_c \in (0,\infty) \quad \exists c_{\mathrm{gap}}(R_c, B_c) > 0$. No positive lower bound uniform over diverging parameter bounds is asserted. In particular, this result does not rule out growing compact sieves. For fixed bounds, however, (17) applies samplewise to every estimator taking values in the corresponding compact class, irrespective of its prior or optimization procedure.

Proof. The compactness argument of Sec. 3 applies to any fixed finite radius: probability measures on $\Theta_c(R_c, B_c)$ form a weakly compact set, and dominated convergence makes their predictor image compact in $L^2(\gamma_d)$. Moreover, if $f_\nu = h(E^\top\cdot)$, averaging over the orthogonal Gaussian coordinates gives a representative of h with

$$\|\partial^\alpha h\|_\infty \le A R_c^{|\alpha|} \|\tanh^{(|\alpha|)}\|_\infty$$

for every multi-index α. Lem. A.1 implies $h \in H^u(\gamma_{m^\star})$ for every finite u. Consequently, no element of $\mathcal{G}$ is represented by this compact class.

Fix one frame E_0. The map $h \mapsto h(E_0^\top \cdot)$ is an L^2 isometry. Its image of $\mathcal{G}$ and the compact neural class are two disjoint compact sets. Their distance

$$\delta(R_c, B_c) = \min_{\substack{h \in \mathcal{G} \\ f \in \mathcal{F}_c(R_c, B_c)}} \|f - h(E_0^\top \cdot)\|_2$$

is attained and positive. For any other frame E, choose an orthogonal matrix O with $OE_0 = E$. The isometry $f \mapsto f(O^\top \cdot)$ maps the target set for E_0 onto that for E and preserves the compact neural class through $(a, w, b) \mapsto (a, Ow, b)$. The same distance therefore holds for every E. Set $c_{\mathrm{gap}} = \delta^2$. □

C.5 Exact representation with finite quadratic cost

Proposition C.7 (Finite-quadratic-cost representation of the fixed family)**.** *For the family $\mathcal{G}$ defined in (16), there is $Q_{\mathcal{G}} < \infty$ such that, for every $h \in \mathcal{G}$ and every $E^\top E = I_{m^\star}$, there exists a finite signed Borel measure $\alpha_{h,E}$ on $\mathbb{R}^d \times \mathbb{R}$ with*

$$h(E^\top x) = \int \tanh(w^\top x + b)\, \mathrm{d}\alpha_{h,E}(w, b),$$

$$\|\alpha_{h,E}\|_{\mathrm{TV}} \le A/4, \qquad \int (1 + \|w\|^2 + b^2)\, \mathrm{d}|\alpha_{h,E}| \le Q_{\mathcal{G}}.$$

The constant is uniform in h and E and independent of N.

Proof. For $h = h_0 + g$, push the signed coefficient measure $q_j[h](u, t)\, \mathrm{d}t\, \mathrm{d}u$ forward by $(u, t) \mapsto (T_j E u, -T_j t)$ and sum over $j \in \mathbb{Z}$. Lem. C.4 gives absolute convergence in weighted total variation, since $\|Eu\| = 1$ and

$$1 + \|w\|^2 + b^2 = 1 + T_j^2(1 + t^2).$$

Denote the resulting signed measure by $\alpha_{h,E}$. By linearity of the coefficients and (16),

$$\int (1 + \|w\|^2 + b^2)\, \mathrm{d}|\alpha_{h,E}| \le c_0 C_{\mathrm{ctr}} + \epsilon_0 C_{\mathrm{pert}} \le A/4.$$

This proves the total-variation bound and a uniform finite quadratic-cost bound. Lem. C.2 identifies its uniformly convergent neural integral with $h(E^\top \cdot)$. We use these continuous representatives throughout. □

For completeness, this signed representation gives an exact probability-measure comparator with interior output weights. Put $B_h = \|\alpha_{h,E}\|_{\mathrm{TV}}$ and $s_\alpha = \mathrm{d}\alpha_{h,E} / \mathrm{d}|\alpha_{h,E}| \in \{-1, 1\}$. Then

$$\overline{\nu}_{h,E} := \frac{2}{A} \big(v \mapsto (A s_\alpha(v)/2, v)\big)_{\#} |\alpha_{h,E}| + \left(1 - \frac{2B_h}{A}\right) \delta_{(0,0,0)}. \tag{18}$$

If $B_h = 0$, omit the first term. This law is supported on $[-A/2, A/2] \times \mathbb{R}^d \times \mathbb{R}$, satisfies $f_{\overline{\nu}_{h,E}} = h(E^\top \cdot)$, and has uniformly bounded quadratic parameter moment. It may be singular relative to the Gaussian prior.

C.6 Gaussian smoothing and entropy cost

Lemma C.8 (Gaussian smoothing of an interior comparator)**.** *Suppose $\bar{\nu}$ is supported on $[-A/2, A/2] \times \mathbb{R}^d \times \mathbb{R}$ and*

$$\mathcal{M}(\bar{\nu}) := \int \left(\frac{\|w\|^2}{2\tau_w^2} + \frac{b^2}{2\tau_b^2}\right) \mathrm{d}\bar{\nu} < \infty.$$

For every $\varepsilon > 0$, there is a finite-KL measure ν_ε with

$$\|f_{\nu_\varepsilon} - f_{\bar{\nu}}\|_2^2 \le \frac{A^2}{4}(d + 1)\varepsilon^2, \tag{19}$$

$$\mathrm{KL}(\nu_\varepsilon \| \nu_0) \le \log 2 + \mathcal{M}(\bar{\nu}) + G_\varepsilon, \tag{20}$$

where

$$G_\varepsilon = \frac{d}{2}\left(\frac{\varepsilon^2}{\tau_w^2} - 1 - \log\frac{\varepsilon^2}{\tau_w^2}\right) + \frac{1}{2}\left(\frac{\varepsilon^2}{\tau_b^2} - 1 - \log\frac{\varepsilon^2}{\tau_b^2}\right).$$

Proof. Conditionally on $(a, w, b) \sim \bar{\nu}$, independently sample

$$a' \sim \mathrm{Unif}([a - A/2, a + A/2]), \qquad w' = w + \varepsilon G, \qquad b' = b + \varepsilon G_0,$$

where $G \sim \mathcal{N}(0, I_d)$ and $G_0 \sim \mathcal{N}(0, 1)$. Let ν_ε be the joint law of (a', w', b'). The output interval is inside $[-A, A]$ and has mean a. The Lipschitz property of $\tanh$ and Jensen's inequality give

$$|f_{\nu_\varepsilon}(x) - f_{\bar{\nu}}(x)|^2 \le \frac{A^2\varepsilon^2}{4}\mathbb{E}[(G^\top x + G_0)^2].$$

Integration over γ_d proves (19). Each conditional law has KL equal to $\log 2 + \|w\|^2/(2\tau_w^2) + b^2/(2\tau_b^2) + G_\varepsilon$. Convexity of KL under mixing proves (20). □

C.7 Finite-KL approximation of the fixed family

Proposition C.9 (Finite-KL approximation of the fixed family)**.** *For every $h \in \mathcal{G}$, $E^\top E = I_{m^\star}$, and $0 < \varepsilon \le 1$, there exists $\nu_{\varepsilon,h,E} \ll \nu_{0,\mathrm{nc}}$ satisfying*

$$\|f_{\nu_{\varepsilon,h,E}} - h(E^\top \cdot)\|_2^2 \le C\varepsilon^2,$$
$$\mathrm{KL}(\nu_{\varepsilon,h,E} \| \nu_{0,\mathrm{nc}}) \le C\{1 + (d+1)\log(1/\varepsilon)\}.$$

Consequently, for $0 < \lambda \le 1/2$,

$$\sup_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \mathfrak{a}_\lambda(h(E^\top \cdot)) \le C_{\mathcal{G},d}\lambda \log(e/\lambda).$$

The constants are independent of h, E, ε, and λ.

Proof. Apply Lem. C.8 to the exact comparator (18). Its quadratic moment is uniformly bounded, so (19) gives the error $C\varepsilon^2$, and (20) gives $C\{1 + (d+1)\log(1/\varepsilon)\}$ for $0 < \varepsilon \le 1$. These are bounds for the same smoothed law. Set $\varepsilon = \sqrt{\lambda}$ in the regularized approximation objective to obtain $\mathfrak{a}_\lambda \le C\lambda \log(e/\lambda)$ uniformly in h, E. □

C.8 Finite-KL approximation versus exact representation

Write

$$\mathcal{F}_{\mathrm{nc}} = \{f_\nu : \nu \in \mathcal{P}(\Theta_{\mathrm{nc}})\}, \qquad \mathcal{F}_{\mathrm{KL}} = \{f_\nu : \mathrm{KL}(\nu \| \nu_0) < \infty\}.$$

Lemma C.10 (Absolute continuity obstructs exact lower-dimensionality)**.** *Let $m^\star < d$. If $\nu \ll \nu_0$ and $f_\nu = h(E^\top \cdot)$ in $L^2(\gamma_d)$, where $E^\top E = I_{m^\star}$, then $f_\nu = 0$ and $h = 0$.*

Proof. For any probability measure ν, the predictor f_ν is continuous by bounded convergence. Fix a nonzero $v \in \mathrm{Range}(E)^\perp$. For each fixed $t \in \mathbb{R}$, equality with $h(E^\top \cdot)$ and equivalence of Gaussian and Lebesgue null sets imply $f_\nu(x + tv) = f_\nu(x)$ almost everywhere. Both sides are continuous, so this equality holds for every x.

The nondegenerate Gaussian prior on w assigns zero mass to $\{w : w^\top v = 0\}$. Absolute continuity gives the same property for ν. Therefore, dominated convergence yields, for every x,

$$\lim_{t \to +\infty} f_\nu(x + tv) = \int a \,\mathrm{sgn}(w^\top v)\, \mathrm{d}\nu =: c_v, \qquad \lim_{t \to -\infty} f_\nu(x + tv) = -c_v.$$

Translation invariance makes both limits equal to $f_\nu(x)$. Hence $f_\nu(x) = c_v = -c_v = 0$. □

Props. C.7 and C.9, together with Lem. C.10, show that every $h \in \mathcal{G}$ and $E^\top E = I_{m^\star}$, $m^\star < d$, satisfy

$$h(E^\top \cdot) \in \mathcal{F}_{\mathrm{nc}} \cap \overline{\mathcal{F}_{\mathrm{KL}}}^{L^2(\gamma_d)} \setminus \mathcal{F}_{\mathrm{KL}}.$$

The first membership follows from (18), the second by letting $\varepsilon \downarrow 0$ in the smoothing bound, and nonmembership follows from the absolute-continuity obstruction and nonzero roughness. Thus exact representation in the noncompact model is compatible with requiring approximation at every positive temperature.

D Feature Learning and Prediction in the Noncompact Model

Use the notation of Appendix C.1, and write $K(\nu) = \mathrm{KL}(\nu\|\nu_0)$, $\psi_v(x) = \tanh(w^\top x + b)$ for $v = (w, b)$, $\phi_0 = \phi_{0,\mathrm{nc}} := \mathcal{N}(0, \tau_w^2 I_d) \otimes \mathcal{N}(0, \tau_b^2)$ for the hidden-layer marginal of ν_0, and $\mathrm{St}(d, m^\star) = \{E : E^\top E = I_{m^\star}\}$. Unless a target in $\mathcal{G}$ is specified, the regression function may be any $f^\star$ with $\|f^\star\|_\infty \le M$ under the same Gaussian design and independent Gaussian noise. The empirical minimizer remains unrestricted on $\mathcal{P}(\Theta_{\mathrm{nc}})$. Conditional function norms and kernels use independent test inputs with the training sample fixed.

D.1 Empirical-process preliminaries

D.1.1 Entropy and parameter moments

Lemma D.1 (Entropy controls the quadratic parameter moment)**.** *Write $v = (w, b)$, $q(v) = 1 + \|w\|^2 + b^2$, and $M_2(\nu) = \int q(v)\,\mathrm{d}\nu(a, v)$. There is C_0, depending only on the prior variances, such that*

$$M_2(\nu) \le C_0\{\mathrm{KL}(\nu\|\nu_0) + d + 1\}$$

for every finite-KL probability measure ν.

Proof. Take $\theta_0 = (4\max\{\tau_w^2, \tau_b^2\})^{-1}$. Gaussian integration gives $\log \int e^{\theta_0(q-1)}\,\mathrm{d}\nu_0 \le (d + 1)\log 2/2$. The entropy variational inequality, first applied to a truncation of $q - 1$ and then followed by monotone convergence, yields

$$\theta_0 \int (q - 1)\,\mathrm{d}\nu \le \mathrm{KL}(\nu\|\nu_0) + (d + 1)\log 2/2.$$

This proves the result, including finiteness of the moment. □

D.1.2 Stable finite-dimensional projection

Lemma D.2 (A stable quadratic local projection)**.** *Let $R \ge 1$, $0 < \eta \le 1$, and $R\eta \le 1$. There is an orthogonal projection $Q = Q_{R,\eta}$ in $L^2(\gamma_d)$, whose range consists of functions supported on $D_R = [-R, R]^d$, such that*

$$\begin{aligned} D := \dim \operatorname{ran} Q \le C_d(R/\eta)^d, \qquad & \|Qg\|_2 \le \|g\|_2, \\ \|Qg\|_\infty \le C_d\|g\|_{L^\infty(D_R)}, & \\ \|\psi_v - Q\psi_v\|_{L^\infty(D_R)} \le C_d\eta^2\|w\|^2, \qquad & \psi_v(x) = \tanh(w^\top x + b). \end{aligned} \tag{21}$$

Moreover, $Qg(x) = \int K_Q(x, z)g(z)\,\mathrm{d}\gamma_d(z)$, where the kernel vanishes outside $D_R \times D_R$ and satisfies $\sup_x \int |K_Q(x, z)|\,\mathrm{d}\gamma_d(z) \le C_d$.

Proof. Partition each coordinate interval into equal subintervals of length between $\eta/2$ and η. On each resulting cell project, in Gaussian L^2, onto the polynomials of total degree at most two, and set the projection to zero outside D_R. These local spaces are mutually orthogonal. Their total dimension is at most $\binom{d+2}{2}(1 + 2R/\eta)^d$. Within a cell the ratio of the largest and smallest Gaussian density is at most $\exp(C_d R\eta) \le \exp(C_d)$. After rescaling to a unit cube, the L^2 and L^∞ norms of these fixed-degree polynomials are equivalent with constants depending only on d. Consequently the local orthogonal projections have uniformly bounded L^∞ operator norms. Their finite-rank kernels have the same uniform bound on the integral of their absolute values. These observations prove the first two estimates and the kernel claim. Finally, the affine Taylor polynomial of ψ_v at a cell center has uniform error at most $C_d\eta^2\|w\|^2$, because $\|\tanh''\|_\infty < \infty$. The projection reproduces that polynomial. Applying the L^∞ operator bound to its Taylor remainder proves (21). □

D.1.3 Normalized residual bounds

Lemma D.3 (An L^2 gain for the moment-normalized residual)**.** *For the quadratic projection above, set $e_v = \mathbf{1}_{D_R}(\psi_v - Q\psi_v)$ and $U_v = e_v/q(v)$. Uniformly in $v = (w, b)$,*

$$\|e_v\|_2 \le C_d\eta^{1/2}, \qquad \|U_v\|_2 \le C_d\eta^{5/2}, \qquad \|U_v\|_\infty \le C_d\eta^2.$$

Proof. Write $t = \|w\|$. For $0 < t \le \eta^{-1}$, approximate each ridge on each cell by its quadratic Taylor polynomial at the cell center. Since $t\eta \le 1$ and $|\tanh^{(3)}(z)| \le Ce^{-c|z|}$, its squared remainder is bounded by $C_d\eta^6 t^6 e^{-c'|w^\top x+b|}$. The variable $w^\top X + b$ has density bounded by C/t, uniformly in b. The best-approximation property therefore gives

$$\|e_v\|_2^2 \le C_d\eta^6 t^6 \min\{1, t^{-1}\}, \qquad \|e_v\|_2 \le C_d\eta^3 \min\{t^3, t^{5/2}\}.$$

For $t = 0$ the ridge is constant and reproduced exactly on each cell.

If $t \ge \eta^{-1}$, replacing $\tanh(w^\top x + b)$ by $\operatorname{sgn}(w^\top x + b)$ costs at most $Ct^{-1/2}$ in $L^2(\gamma_d)$: integrate $|\tanh z - \operatorname{sgn} z|^2$ against the density bound C/t. Next use the sign at each cell center. A cell on which this changes the sign intersects the hyperplane and lies in its strip of width $C_d\eta$. The Gaussian probability of that strip is at most $C_d\eta$, uniformly in the hyperplane. Piecewise constants lie in the projection space, so

$$\|e_v\|_2 \le C_d(t^{-1/2} + \eta^{1/2}) \le C_d\eta^{1/2}.$$

Together these estimates imply the first assertion. Divide by $q(v) \ge 1 + t^2$. In the ranges $t \le 1$, $1 \le t \le \eta^{-1}$, and $t \ge \eta^{-1}$, the resulting bounds are respectively $C\eta^3$, $C\eta^3 t^{1/2}$, and $C\eta^{1/2}t^{-2}$, all at most $C\eta^{5/2}$. Finally, affine Taylor approximation and L^∞ stability give $\|e_v\|_\infty \le C_d \min\{\eta^2 t^2, 1\}$. Dividing by $q(v)$ proves the last assertion. □

D.1.4 Uniform control of the full-loss remainder

We use the following variance-sensitive form of the polynomial-entropy maximal inequality; its ingredients are the VC-type expectation bound and Bousquet concentration in Koltchinskii (2006, Sec. 2).

Lemma D.4 (Variance-sensitive maximal inequality)**.** *Let $\mathcal{B}$ be a pointwise measurable class with $|f| \le B$, $\sup_{f\in\mathcal{B}} Pf^2 \le \sigma^2$, and $0 < \sigma \le B$. Suppose, for every probability measure ρ and $0 < \delta \le 1$,*

$$\log \mathcal{N}(\delta B, \mathcal{B}, L^2(\rho)) \le v \log(A_0/\delta), \qquad v \ge 1, \quad A_0 \ge e.$$

With $\mathcal{L} = v\log(A_0 B/\sigma)$, for $u \ge 1$, with probability at least $1 - 2e^{-u}$,

$$\sup_{f\in\mathcal{B}} |(P - P_N)f| \le C\left\{\sigma\sqrt{\frac{\mathcal{L}+u}{N}} + B\frac{\mathcal{L}+u}{N}\right\}.$$

Proof. The VC-type maximal inequality obtained by symmetrization and the uniform entropy integral gives, for $Z = \sup_{f\in\mathcal{B}} |(P - P_N)f|$, $\mathbb{E}[Z] \le C\{\sigma\sqrt{\mathcal{L}/N} + B\mathcal{L}/N\}$. Bousquet's inequality, applied also to the negatives of the class, adds $C\sqrt{(\sigma^2 + B\mathbb{E}[Z])u/N} + CBu/N$. Young's inequality bounds the mixed square root by $C\mathbb{E}[Z] + CBu/N$, proving the claim. This retains the variance bound rather than replacing it by the squared envelope. □

Lemma D.5 (Uniform empirical control of the remainder relative to a comparator)**.** *Let $Q = Q_{R,\eta}$ and suppose $|Y| \le B_Y$ and $X \sim \gamma_d$. For probability measures ν, μ with finite M_2, put $p_\nu = Qf_\nu$ and*

$$\widetilde{f}_{\nu,\mu} = f_\mu + Q(f_\nu - f_\mu), \qquad \mathcal{E}_{\nu,\mu}(x,y) = \mathbf{1}_{D_R}(x)\{(y - f_\nu(x))^2 - (y - \widetilde{f}_{\nu,\mu}(x))^2\}.$$

For $u \ge 1$, with probability at least $1 - 6e^{-u}$, simultaneously for all such ν, μ,

$$\begin{aligned}|(P - P_N)\mathcal{E}_{\nu,\mu}| \le C_d(A + A^2)(1 + B_Y)\left\{\eta^{5/2}\sqrt{\frac{\mathcal{L}_{R,\eta}+u}{N}} + \eta^2\frac{\mathcal{L}_{R,\eta}+u}{N}\right\} \\ \cdot \{M_2(\nu) + M_2(\mu)\}, \qquad \mathcal{L}_{R,\eta} = C_d \log\frac{C_d(R+1)}{\eta}.\end{aligned} \tag{22}$$

Proof. We establish uniform entropy for the three atomic classes $\{U_v\psi_{v'}\}$, $\{U_vQ\psi_{v'}\}$, and $\{YU_v\}$. The normalized residual satisfies

$$\|U_v\|_\infty \le C_d \min\{\eta^2, q(v)^{-1}\}, \qquad \|U_v - U_{v'}\|_\infty \le C_d(R+1)\|v - v'\|.$$

For the second inequality, differentiate $\psi_v/q(v)$ in v and use the bounded L^∞ norm of $I - Q$ on D_R. For precision $C_d\eta^2\delta$, parameters outside a ball of radius $C_d/(\eta\sqrt{\delta})$ can be approximated by zero. Cover the ball by a grid of spacing $c_d\eta^2\delta/(R+1)$ to obtain

$$\log \mathcal{N}(C_d\eta^2\delta, \{U_v\}, \|\cdot\|_\infty) \le C_d(d+1)\log\frac{C_d(R+1)}{\eta^3\delta^{3/2}}.$$

The sigmoid dictionary is a bounded VC-subgraph class: for $-1 < z < 1$, the inequality $\tanh(w^\top x + b) > z$ is equivalent to $w^\top x + b > \operatorname{arctanh}(z)$, a finite-dimensional halfspace condition. It therefore has polynomial uniform L^2 covering numbers (Thm. 2.6.7 van der Vaart & Wellner, 1996).

The projected dictionary inherits these covering numbers independently of D. For every probability measure ρ, kernel stability and Cauchy–Schwarz give

$$\|Q(f-g)\|^2_{L^2(\rho)} \le C_d \iint |K_Q(x,z)||f(z)-g(z)|^2 \,\mathrm{d}\gamma_d(z)\,\mathrm{d}\rho(x).$$

The positive measure on the right has bounded total mass; after normalization, the original dictionary's uniform entropy estimate applies. Bounded-factor product coverings now give polynomial uniform entropy for all three atomic classes, with logarithmic constant $C_d\{\log(C_d(R+1)/\eta) + \log(1/\delta)\}$. The multiplier Y is bounded, so the same residual net covers YU_v on the joint sample space.

By Lem. D.3, these classes have envelope $B = C_d(1+B_Y)\eta^2$ and second-moment bound $\sigma = C_d(1+B_Y)\eta^{5/2}$, after taking a common sufficiently large constant. Apply Lem. D.4 to each class and take a union bound. The extra logarithm $\log(B/\sigma)$ is absorbed in $\mathcal{L}_{R,\eta}$.

Finally, the remainder $\mathcal{E}_{\nu,\mu}$ factors as

$$\mathcal{E}_{\nu,\mu} = \mathbf{1}_{D_R}\{(f_\nu - p_\nu) - (f_\mu - p_\mu)\}(f_\nu + p_\nu + f_\mu - p_\mu - 2Y).$$

Moreover, $\mathbf{1}_{D_R}(f_\nu - p_\nu) = \int aq(v)U_v \,\mathrm{d}\nu(a,v)$. Expand the products as mixtures of the three atomic classes. The first factor has coefficient mass at most $A\{M_2(\nu) + M_2(\mu)\}$, and every neural term in the second factor has coefficient mass at most A. Thus only the sum of moments occurs, not its square, giving (22). Continuity in the parameters allows countable dense sets to determine the atomic suprema, ensuring measurability. □

D.1.5 Localized deviation on the projected space

Lemma D.6 (Localized deviation on the projected space)**.** *Let V be a D-dimensional space of functions supported on D_R, let f_0 be a fixed function with $\|f_0\|_\infty \le A$, and suppose $|Y| \le B_Y$. For every fixed $B > 0$, with probability at least $1 - 2e^{-u}$, simultaneously for $v \in V$ with $\|v\|_\infty \le B$,*

$$\begin{aligned}\left|(P - P_N)\left[\mathbf{1}_{D_R}\{v^2 + 2(f_0 - Y)v\}\right]\right| \le \frac{1}{8}\|v\|_2^2 \\ + C_{A,B}(1+B_Y)^2\frac{(D+1)\log(eN) + u}{N}.\end{aligned} \tag{23}$$

Proof. The L^∞ ball of radius B in V has an N^{-3}-net of size at most $(1 + 2BN^3)^D$. To see this, compare the volumes of disjoint balls of radius $N^{-3}/2$ centered at a maximal separated set with that of the ball of radius $B + N^{-3}/2$ in the finite-dimensional normed space. For $F_v = \mathbf{1}_{D_R}\{v^2 + 2(f_0 - Y)v\}$ we have

$$|F_v| \le H|v|, \qquad \mathrm{Var}(F_v) \le H^2\|v\|_2^2, \qquad H = B + 2A + 2B_Y.$$

Bernstein's inequality and a union bound over the net give

$$|(P - P_N)F_v| \le CH\|v\|_2\sqrt{t/N} + CHBt/N, \quad t = u + \log\{2(1+2BN^3)^D\}.$$

Young's inequality absorbs the first term with coefficient $1/16$ on $\|v\|_2^2$. Passing from a net point to an arbitrary v changes the empirical deviation by at most CHN^{-3} and its squared L^2 norm by at most C_BN^{-3}. Enlarging the constant yields (23). □

D.2 Uniform loss deviation for a fixed comparator

For $N \geq 2$ and $N^{-1} \leq \lambda \leq 1/2$, put

$$\ell_N = \log(eN), \qquad \Xi_{N,\lambda} = \frac{\ell_N^{3d/2+2}}{N} \max\{1, (N\lambda^2)^{-d/5}\},$$
$$\mathcal{B}_{N,\lambda}(f^\star) = \mathfrak{a}_\lambda(f^\star) + (d+1)\lambda + \Xi_{N,\lambda}.$$

Lemma D.7 (Uniform loss deviation for one deterministic comparator)**.** *Fix a probability law μ with $K(\mu) < \infty$, independently of the training sample. Under the bounded-regression assumptions of Sec. C.1, there is a measurable event $\mathcal{E}_{N,\lambda}(\mu)$ with $\mathbb{P}(\mathcal{E}_{N,\lambda}(\mu)^c) \leq CN^{-8}$ such that, simultaneously for every ν with $K(\nu) < \infty$,*

$$(P - P_N)(\ell_{f_\nu} - \ell_{f_\mu}) \leq \frac{1}{8}\|f_\nu - f_\mu\|_2^2 + \frac{\lambda}{8}\{K(\nu) + K(\mu) + 2(d+1)\} + C\Xi_{N,\lambda}. \tag{24}$$

The constants depend only on $d, A, M, \sigma_\xi, \tau_w, \tau_b$, not on μ, N, λ, or the particular bounded regression function.

Proof. The comparator is fixed before constructing any sample event. Write $t_N = N\lambda^2$ and set

$$R_N = \sqrt{24\ell_N + 2\log(2d)}, \qquad T_N = \sigma_\xi\sqrt{24\ell_N}, \qquad Y^{\mathrm{tr}} = f^\star(X) + \xi^{\mathrm{tr}},$$

where ξ^{tr} is ξ clipped to $[-T_N, T_N]$; it is zero when $\sigma_\xi = 0$. Symmetry implies $\mathbb{E}[\xi^{\mathrm{tr}} \mid X] = 0$. Thus clipping does not change any population squared-loss difference. Moreover, $|Y^{\mathrm{tr}}| \leq C\sqrt{\ell_N}$, and Gaussian tails give

$$\mathbb{P}\{\exists i : \ X_i \notin D_{R_N} \text{ or } \xi_i \neq \xi_i^{\mathrm{tr}}\} \leq CN^{-10}. \tag{25}$$

On the complementary event all empirical risks remain the original ones.

Choose a sufficiently small fixed $c > 0$ and put

$$\eta_N = c\ell_N^{-1}\min\{1, t_N^{1/5}\}, \qquad Q = Q_{R_N,\eta_N}, \qquad D = \dim \mathrm{Range}\, Q. \tag{26}$$

The constant can be chosen so that $R_N\eta_N \leq 1$ for all $N \geq 2$. Since $t_N \geq N^{-1}$, the entropy logarithm in Lem. D.5 is at most $C_c\ell_N$, with at most logarithmic dependence on $1/c$. Apply that lemma to the clipped law, with $u = 12\ell_N$. Apart from fixed constants, the two coefficients of $M_2(\nu) + M_2(\mu)$ are bounded by

$$\frac{\ell_N}{\sqrt{N}}\eta_N^{5/2} \leq c^{5/2}\ell_N^{-3/2}\lambda, \qquad \frac{\ell_N^{3/2}}{N}\eta_N^2 \leq c^2\ell_N^{-1/2}\lambda.$$

For the first estimate use $\min\{1, t_N^{1/2}\} \leq \sqrt{N}\lambda$. For the second, when $t_N \leq 1$,

$$\frac{t_N^{2/5}}{N\lambda} = N^{-3/5}\lambda^{-1/5} \leq N^{-2/5},$$

and when $t_N \geq 1$ use $(N\lambda)^{-1} \leq N^{-1/2}$. Consequently, Lem. D.1 and a sufficiently small fixed c imply, on an event of failure probability at most CN^{-10},

$$\left|(P^{\mathrm{tr}} - P_N^{\mathrm{tr}})\mathcal{E}_{\nu,\mu}\right| \leq \frac{\lambda}{8}\{K(\nu) + K(\mu) + 2(d+1)\} \tag{27}$$

for every finite-KL ν. The logarithmic dependence of the entropy bound on c is dominated by the positive powers of c above.

Set $z = Q(f_\nu - f_\mu)$. On D_{R_N}, the exact anchored identity is

$$\ell_{f_\nu} - \ell_{f_\mu} = z^2 + 2(f_\mu - Y^{\mathrm{tr}})z + \mathcal{E}_{\nu,\mu}.$$

Here $\mathcal{E}_{\nu,\mu}$ is the anchored remainder of Lem. D.5, associated with $f_\mu + Q(f_\nu - f_\mu)$. Because f_μ is deterministic, $\|z\|_\infty \leq 2C_dA$, and $\|z\|_2 \leq \|f_\nu - f_\mu\|_2$, Lem. D.6 gives a uniform bound for the projected part:

$$\frac{1}{8}\|f_\nu - f_\mu\|_2^2 + \frac{C(D+1)\ell_N^2}{N}.$$

This has failure probability at most CN^{-10}. The population contribution from $D^c_{R_N}$ is at most

$$4A(A+M)\gamma_d(D^c_{R_N}) \le CN^{-10},$$

uniformly in ν, because an independent label has conditional mean $f^\star(X)$.

Finally, Lem. D.2 and (26) give

$$D \le C_d \ell_N^{3d/2} \max\{1, (N\lambda^2)^{-d/5}\}.$$

Combine this with (25), (27), and the projected deviation estimate. The N^{-10} term is absorbed by $\Xi_{N,\lambda} \ge N^{-1}$. The union of the exceptional events has probability at most CN^{-10}, hence at most CN^{-8}. This proves (24) for all finite-KL ν on one event for the chosen deterministic μ. □

D.3 Learned RKHS geometry and the oracle

For a probability measure ϕ on $\mathbb{R}^d \times \mathbb{R}$, define

$$K_\phi(x,x') = \int \psi_v(x)\psi_v(x')\,\mathrm{d}\phi(v), \qquad \mathcal{H}_\phi = \mathcal{H}(K_\phi).$$

The feature map is $\Phi_\phi(x) = K_\phi(x,\cdot)$. Write $J_\phi : \mathcal{H}_\phi \to L^2(\gamma_d)$ for the canonical inclusion and

$$\mathbb{B}_\phi(A) = J_\phi\{g \in \mathcal{H}_\phi : \|g\|_{\mathcal{H}_\phi} \le A\}.$$

Lemma D.8 (Factorization, trace, and a well-defined oracle)**.** *For every probability law ν on Θ_{nc}, let ϕ be its hidden-layer marginal. There is $g_\nu \in \mathcal{H}_\phi$ with*

$$f_\nu(x) = \langle g_\nu, \Phi_\phi(x)\rangle_{\mathcal{H}_\phi} \quad \textit{for every } x, \qquad \|g_\nu\|_{\mathcal{H}_\phi} \le A. \tag{28}$$

Furthermore,

$$\|J_\phi\|^2_{\mathrm{HS}} = \mathrm{tr}(J_\phi^* J_\phi) = \int K_\phi(x,x)\,\mathrm{d}\gamma_d(x) \le 1. \tag{29}$$

The set $\mathbb{B}_\phi(A)$ is compact and convex in $L^2(\gamma_d)$. It therefore contains a unique metric projection f°_ϕ of $f^\star$. Choose its minimum-norm RKHS representative g°_ϕ, so that $J_\phi g^\circ_\phi = f^\circ_\phi$ and $\|g^\circ_\phi\|_{\mathcal{H}_\phi} \le A$. For every $f \in \mathbb{B}_\phi(A)$,

$$\|f^\circ_\phi - f^\star\|_2^2 + \|f - f^\circ_\phi\|_2^2 \le \|f - f^\star\|_2^2 \le 2\|f^\circ_\phi - f^\star\|_2^2 + 2\|f - f^\circ_\phi\|_2^2. \tag{30}$$

Proof. Disintegrate $\nu(\,\mathrm{d}a, \mathrm{d}v) = q(\,\mathrm{d}a \mid v)\phi(\,\mathrm{d}v)$ and put $\bar a(v) = \int a\, q(\,\mathrm{d}a \mid v)$. Then $|\bar a| \le A$ and $f_\nu = \int \bar a(v)\psi_v\,\mathrm{d}\phi(v)$. The feature-space construction identifies $\mathcal{H}_\phi$ with $L^2(\phi)$ modulo the kernel of this integral map, equipped with its minimum coefficient norm. Thus $\|g_\nu\|_{\mathcal{H}_\phi} \le \|\bar a\|_{L^2(\phi)} \le A$, and gives (28). This also establishes the minimum-norm representative assertion.

The kernel is bounded and continuous by dominated convergence. Its RKHS is separable: kernel sections at a countable dense subset of $\mathbb{R}^d$ have dense span, since $x \mapsto K_\phi(x,\cdot)$ is norm-continuous. For an orthonormal basis (e_j), Tonelli and the reproducing property give

$$\sum_j \|J_\phi e_j\|_2^2 = \int \sum_j |\langle e_j, \Phi_\phi(x)\rangle|^2\,\mathrm{d}\gamma_d(x) = \int K_\phi(x,x)\,\mathrm{d}\gamma_d(x) \le 1.$$

Hence J_ϕ is compact. A bounded sequence of RKHS representatives has a weakly convergent subsequence in the closed radius-A ball; compactness of J_ϕ sends this subsequence to a strongly convergent sequence with its limit still in $\mathbb{B}_\phi(A)$. This proves compactness of the image. Convexity is immediate.

Metric projection onto a closed convex subset of a Hilbert space gives

$$\langle f^\star - f^\circ_\phi, f - f^\circ_\phi\rangle_2 \le 0.$$

Expanding $\|f - f^\star\|_2^2$ proves the first inequality in (30). The second is the squared triangle inequality. These are deterministic geometric inequalities. □

Apply these definitions to the marginal $\widehat{\phi}_{N,\lambda}$ of the empirical minimizer. Use $\mathcal{H}_{N,\lambda}$ and $J^{\mathrm{feat}}_{N,\lambda}$ for its RKHS and inclusion, abbreviated by J below. Let $f^{\circ}_{N,\lambda} = Jg^{\circ}_{N,\lambda}$ be the metric projection of $f^{\star}$ onto $\mathbb{B}_{\widehat{\phi}_{N,\lambda}}(A)$, with $g^{\circ}_{N,\lambda}$ its minimum-norm RKHS representative, and write $\widehat{f}_{N,\lambda} = J\widehat{g}_{N,\lambda}$. Define

$$\mathcal{A}^{\mathrm{feat}}_{N,\lambda} := \|f^{\circ}_{N,\lambda} - f^{\star}\|_2^2, \qquad \mathcal{E}^{\mathrm{out}}_{N,\lambda} := \|\widehat{f}_{N,\lambda} - f^{\circ}_{N,\lambda}\|_2^2.$$

Set $\mathcal{S}^{\mathrm{feat}}_{N,\lambda} = \mathcal{A}^{\mathrm{feat}}_{N,\lambda} + \mathcal{E}^{\mathrm{out}}_{N,\lambda}$. At $\lambda = \lambda_N$, omit the temperature subscript and write $\mathcal{H}_N = \mathcal{H}_{N,\lambda_N}$, $J_N = J^{\mathrm{feat}}_{N,\lambda_N}$, $\widehat{g}_N = \widehat{g}_{N,\lambda_N}$, and $g^{\circ}_N = g^{\circ}_{N,\lambda_N}$. In particular, $f^{\circ}_N = J_N g^{\circ}_N$ is the oracle in Sec. 5.1.

Measurability of the oracle. All oracle errors above can be chosen measurable. Here is an ambient-space construction that avoids treating the random RKHS as a fixed deterministic space. Let $T = JJ^*$ act on the fixed separable space $L^2(\gamma_d)$. Its kernel is $K_{\widehat{\phi}_{N,\lambda}}$, a measurable random element of $L^2(\gamma_d^{\otimes 2})$. The polar decomposition gives $\mathbb{B}_{\widehat{\phi}_{N,\lambda}}(A) = T^{1/2}B_{L^2}(A)$. For $\varepsilon > 0$ the problem

$$\min_{\|u\|_2 \le A} \{\|T^{1/2}u - f^{\star}\|_2^2 + \varepsilon\|u\|_2^2\}$$

has a unique minimizer u_ε, measurable in T; for example, strict convexity and uniform convergence of these objectives on the bounded ball imply continuity of their minimizers as T varies in operator norm. As $\varepsilon \downarrow 0$, the images $T^{1/2}u_\varepsilon$ converge to the unique metric projection $f^{\circ}_{N,\lambda}$. Indeed their squared errors are within εA^2 of the minimum, and the projection inequality then bounds their squared distance to the projection by the same quantity. This constructs a measurable $f^{\circ}_{N,\lambda}$. Its minimum-norm RKHS representative can be obtained as the RKHS limit of $J^*(T + n^{-1}I)^{-1}f^{\circ}_{N,\lambda}$. The empirical minimizer is measurable by uniqueness and the direct method on entropy sublevels in Lem. C.1.

D.4 Feature-space oracle inequality

Proposition D.9 (Feature-space oracle inequality)**.** *Under the bounded-regression assumptions, for every $N \ge 2$ and $N^{-1} \le \lambda \le 1/2$,*

$$\mathbb{P}\left\{\mathcal{A}^{\mathrm{feat}}_{N,\lambda} + \mathcal{E}^{\mathrm{out}}_{N,\lambda} + \lambda K(\widehat{\nu}_{N,\lambda}) > C\mathcal{B}_{N,\lambda}(f^{\star})\right\} \le C_{\mathrm{fail}}N^{-8}. \tag{31}$$

The bound is uniform over $\|f^{\star}\|_\infty \le M$.

Proof. Fix the deterministic comparator $\mu = \nu_\lambda$ from Lem. C.1; it depends on the regression function and λ, but not on the observed sample. Put

$$S = \mathcal{S}^{\mathrm{feat}}_{N,\lambda}, \qquad r_\mu = \|f_\mu - f^{\star}\|_2^2, \qquad \widehat{\nu} = \widehat{\nu}_{N,\lambda}, \qquad \widehat{f} = \widehat{f}_{N,\lambda}.$$

Lem. D.8 gives the deterministic bounds

$$S \le \|\widehat{f} - f^{\star}\|_2^2 \le 2S. \tag{32}$$

Empirical optimality, the population squared-loss identity, and the first inequality of (32) imply

$$S + \lambda K(\widehat{\nu}) \le r_\mu + \lambda K(\mu) + (P - P_N)(\ell_{\widehat{f}} - \ell_{f_\mu}). \tag{33}$$

On $\mathcal{E}_{N,\lambda}(\mu)$ from Lem. D.7, (24) is valid at the random full minimizer. The second inequality of (32) gives

$$\|\widehat{f} - f_\mu\|_2^2 \le 2\|\widehat{f} - f^{\star}\|_2^2 + 2r_\mu \le 4S + 2r_\mu.$$

Substituting into (33) and collecting terms yields

$$\frac{1}{2}S + \frac{7\lambda}{8}K(\widehat{\nu}) \le \frac{5}{4}r_\mu + \frac{9\lambda}{8}K(\mu) + \frac{\lambda}{4}(d+1) + C\Xi_{N,\lambda}.$$

Since $r_\mu + \lambda K(\mu) = \mathfrak{a}_\lambda(f^{\star})$, (31) follows.

The empirical comparison uses the deterministic population minimizer ν_λ; $g^{\circ}_{N,\lambda}$ enters through the metric-projection inequalities. □

Proposition D.10 (Target-rate approximation and output-layer alignment)**.** *Assume* (2)*. For the fixed family* $\mathcal{G}$ *of* (16) *and the temperature in* (3)*, there are fixed constants* $C_s, C_{\text{fail}} > 0$ *such that, for every* $N \geq 2$*,*

$$\sup_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \mathbb{P}_{h,E}\left\{\mathcal{A}_N^{\text{feat}} + \mathcal{E}_N^{\text{out}} + \lambda_N K(\widehat{\nu}_N) > C_s N^{-p_{s,m^\star}}\right\} \leq C_{\text{fail}} N^{-8}. \tag{34}$$

Moreover,

$$\sup_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \mathbb{E}_{h,E}[\mathcal{A}_N^{\text{feat}} + \mathcal{E}_N^{\text{out}}] \leq C_s N^{-p_{s,m^\star}}. \tag{35}$$

Proof. For $N \geq 4$, the temperature choice gives $N\lambda_N^2 < 1$. Hence, by Prop. C.9, uniformly over $h \in \mathcal{G}$ and $E^\top E = I_{m^\star}$, $\mathcal{B}_{N,\lambda_N}(h(E^\top \cdot)) \lesssim N^{-p_{s,m^\star}}$. After enlarging the constant to cover $N = 2, 3$, Prop. D.9 yields (34). Moreover, by (32) and the bounds $|\widehat{f}_N| \leq A$ and $\|f^\star\|_\infty \leq M$, we have $S_N \leq \|\widehat{f}_N - f^\star\|_2^2 \leq (A+M)^2$ deterministically. Hence,

$$\mathbb{E}[S_N] \leq C_s N^{-p_{s,m^\star}} + (A+M)^2 C_{\text{fail}} N^{-8} \lesssim N^{-p_{s,m^\star}},$$

uniformly over $h \in \mathcal{G}$ and $E^\top E = I_{m^\star}$, which proves (35). □

For each fixed (h, E), define the success event

$$\Omega_{N,h,E} := \left\{S_N + \lambda_N K(\widehat{\nu}_N) \leq C_s N^{-p_{s,m^\star}}\right\}.$$

Then

$$\sup_{h,E} \mathbb{P}_{h,E}(\Omega_{N,h,E}^c) \leq C_{\text{fail}} N^{-8}.$$

Here and below, each probability is under the data law for its fixed target and frame.

D.5 Formal feature learning and kernel non-laziness

The following definition of statistical feature learning means training changes the feature space non-trivially, the learned features can represent the target with low-dimensional spectral concentration, and the fitted predictor aligns with that representation.

Definition D.11 (Statistical feature learning in the present setting)**.** Let $\mathcal{H}_N$ be the data-dependent RKHS induced by the algorithm-induced kernel $K_{\text{feat},N}$, let $\Phi_N(x) := K_{\text{feat},N}(x, \cdot)$, and let $J_N : \mathcal{H}_N \to L^2(\gamma_d)$ be the canonical inclusion. Write $(\rho_{j,N}, e_{j,N})_{j \geq 1}$ for the eigenpairs of $\Sigma_N := J_N^* J_N$, in nonincreasing order of eigenvalues. In the finite-rank case, spectral sums run over the available eigenvectors, with subsequent terms interpreted as zero.

Following Shang et al. (2026, Defs. 1–2), we say that $(\mathcal{H}_N, \widehat{f}_N)_N$ exhibits statistical feature learning if there exist an algorithm-induced estimator representative $\widehat{g}_N \in \mathcal{H}_N$, a target representative $g_N^\circ \in \mathcal{H}_N$, nondecreasing integers $k_N = o(N)$, and prescribed weights $\gamma_{j,N} \geq 0$, such that:

(FL1) $\|K_{\text{feat},N} - K_{\text{init}}\|_{L^2(\gamma_d^{\otimes 2})} = \Omega_{\mathbb{P}}(1)$;

(FL2) for every $x \in \mathbb{R}^d$, $\widehat{f}_N(x) = \langle \widehat{g}_N, \Phi_N(x) \rangle_{\mathcal{H}_N}$;

(FL3) $\|J_N g_N^\circ - f^\star\|_{L^2(\gamma_d)} = o_{\mathbb{P}}(1)$;

(FL4) $\sum_{j > k_N} \gamma_{j,N} \left|\langle g_N^\circ, e_{j,N} \rangle_{\mathcal{H}_N}\right|^2 = o_{\mathbb{P}}(1)$;

(FL5) there exist $\delta_N \downarrow 0$, nonnegative deterministic tolerances $\epsilon_N(k_N, \delta_N) = o(1)$, specified independently of the realized tail coefficients of g_N°, and admissible nondecreasing deterministic functions $\omega_N : \mathbb{R}_+ \to \mathbb{R}_+$ such that, with probability at least $1 - \delta_N$,

$$\|J_N(\widehat{g}_N - g_N^\circ)\|_{L^2(\gamma_d)}^2 \leq \epsilon_N(k_N, \delta_N) + \omega_N\left(\sum_{j > k_N} \gamma_{j,N} \left|\langle g_N^\circ, e_{j,N} \rangle_{\mathcal{H}_N}\right|^2\right).$$

Here admissibility means $\omega_N(0) = 0$ and $\lim_{\eta \downarrow 0} \limsup_{N \to \infty} \omega_N(\eta) = 0$.

The weights $(\gamma_{j,N})_{j>k_N}$ are part of the prescribed alignment criterion and are not selected post hoc from the realized tail coefficients of g_N°.

For Def. D.11, we take

$$k_N = \left\lceil N^{m^\star/(2s+m^\star)} \right\rceil, \qquad \gamma_{j,N} = \rho_{j,N}, \\ \epsilon_N(k,\delta) = C_s N^{-p_{s,m^\star}}, \qquad \omega_N \equiv 0, \qquad \delta_N = C_{\mathrm{fail}} N^{-8}. \tag{36}$$

The probability assertion uses $\delta = \delta_N \in (0,1)$ for all sufficiently large N; the deterministic tolerance is defined for every k, δ.

Since $0 < m^\star/(2s+m^\star) < 1$, the sequence k_N is nondecreasing, $k_N \to \infty$, and $k_N = o(N)$. The choice $\omega_N \equiv 0$ is admissible. We verify conditions (FL1)–(FL5) below.

Lemma D.12 (Separation from the initial RKHS ball)**.** *There exists $\eta_0 > 0$, uniform in $h \in \mathcal{G}$ and $E \in \mathrm{St}(d, m^\star)$, such that*

$$\inf_{\substack{h\in\mathcal{G} \\ E\in\mathrm{St}(d,m^\star)}} \mathrm{dist}_2\big(h(E^\top\cdot), \mathbb{B}_{\phi_0}(A)\big) \geq \eta_0.$$

Proof. The initial ball is compact by Lem. D.8. Every $f \in \mathbb{B}_{\phi_0}(A)$ has a representation

$$f(x) = \int c(v)\psi_v(x)\,\mathrm{d}\phi_0(v), \qquad \|c\|_{L^2(\phi_0)} \leq A.$$

For any multi-index α,

$$\|\partial^\alpha f\|_\infty \leq A \|\tanh^{(|\alpha|)}\|_\infty \left(\int \|w\|^{2|\alpha|}\,\mathrm{d}\phi_0\right)^{1/2} < \infty.$$

Differentiation and continuity under the integral are justified by Cauchy–Schwarz and Gaussian moments. If $f = h(E^\top\cdot)$ in $L^2(\gamma_d)$, conditioning on $E^\top X$ gives a representative of h with bounded derivatives of every finite order. By Lem. A.1, this implies $h \in H^{r+}$, contrary to Lem. C.5.

Fix $E_0 \in \mathrm{St}(d, m^\star)$. The image of $\mathcal{G}$ under $h \mapsto h(E_0^\top\cdot)$ is compact and disjoint from the initial ball, so their distance is attained and positive. For any other E, choose an orthogonal matrix O with $OE_0 = E$. The $L^2(\gamma_d)$ isometry $f \mapsto f(O^\top\cdot)$ maps the first target set to the second and preserves the initial RKHS ball: ϕ_0 is rotationally invariant in w, and the coefficient transformation preserves its $L^2(\phi_0)$ norm. The same positive distance therefore holds for every frame. □

Lemma D.13 (Kernel perturbation controls the distance between RKHS balls)**.** *For the kernels K_ϕ, K_{ϕ_0} above and every $f \in \mathbb{B}_\phi(A)$,*

$$\mathrm{dist}_2(f, \mathbb{B}_{\phi_0}(A)) \leq A\|K_\phi - K_{\phi_0}\|^{1/2}_{L^2(\gamma_d^{\otimes 2})}.$$

Proof. Let T_K denote the kernel integral operator on $L^2(\gamma_d)$. Since $T_K = J_K J_K^*$, polar decomposition gives

$$\mathbb{B}_K(A) = T_K^{1/2} B_{L^2}(A).$$

For bounded positive operators S, T,

$$\|S^{1/2} - T^{1/2}\|_{\mathrm{op}} \leq \|S - T\|^{1/2}_{\mathrm{op}}. \tag{37}$$

Indeed, set $a = \|S-T\|_{\mathrm{op}}$. Then $S \leq T + aI$ and $T \leq S + aI$. Operator monotonicity of the square root gives $S^{1/2} \leq (T+aI)^{1/2} \leq T^{1/2} + \sqrt{a}I$ and the reverse inequality with S, T interchanged. Operator monotonicity follows, for example, by applying inverse-order monotonicity to the integral formula

$$S^{1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2} S(S+tI)^{-1}\,\mathrm{d}t.$$

This proves (37) without a commutativity assumption. Now write $f = T_{K_\phi}^{1/2} u$ with $\|u\|_2 \leq A$ and compare with $T_{K_{\phi_0}}^{1/2} u \in \mathbb{B}_{\phi_0}(A)$. Use (37) and

$$\|T_{K_\phi} - T_{K_{\phi_0}}\|_{\mathrm{op}} \leq \|T_{K_\phi} - T_{K_{\phi_0}}\|_{\mathrm{HS}} = \|K_\phi - K_{\phi_0}\|_{L^2(\gamma_d^{\otimes 2})}.$$

□

Proof of Prop. 5.2. First use Prop. E.2, which is independent of all statistical bounds, to identify the long-time law at each fixed sample and positive temperature with $\widehat{\nu}_N$. The hidden-layer marginal is consequently algorithm-induced, and its kernel at initialization is K_{init}. Lem. D.12 gives the initial RKHS approximation lower bound with $c_{\mathrm{init}} = \eta_0^2$.

For exact factorization (FL2), Lem. D.8 gives

$$\widehat{f}_N(x) = \langle \widehat{g}_N, \Phi_N(x) \rangle_{\mathcal{H}_N} \quad \text{for every } x, \qquad \|\widehat{g}_N\|_{\mathcal{H}_N} \le A.$$

For the target representative take g_N°. On the event in (34),

$$\|J_N g_N^\circ - f^\star\|_2^2 \le C_s N^{-p_{s,m^\star}}, \qquad \|J_N(\widehat{g}_N - g_N^\circ)\|_2^2 \le C_s N^{-p_{s,m^\star}}.$$

Thus target approximation (FL3) holds, and the alignment inequality (FL5) holds with the deterministic tolerance in (36) and $\omega_N = 0$. This tolerance is independent of the tail coefficients of g_N°; it is specified for all k, δ, and the probability claim is applied at $\delta = \delta_N$.

For the spectral tail (FL4), (29) gives $\sum_j \rho_{j,N} \le 1$, hence $\rho_{k+1,N} \le 1/(k+1)$. Consequently, samplewise,

$$\sum_{j>k} \rho_{j,N} |\langle g_N^\circ, e_{j,N} \rangle_{\mathcal{H}_N}|^2 \le \rho_{k+1,N} \|g_N^\circ\|_{\mathcal{H}_N}^2 \le \frac{A^2}{k+1}.$$

Zero eigenvalues may be appended in the finite-rank case. Since $k_N \to \infty$ and $k_N = o(N)$, the required weighted tail vanishes. The weights are the prescribed covariance eigenvalues, not chosen from the target's tail coefficients. Measurable spectral choices can be made by resolving eigenspace ties with a fixed countable dense set in $L^2(\gamma_d)$ and transferring the positive eigenvectors to the RKHS.

For non-laziness (FL1), apply Lem. D.13 to $f_N^\circ = J_N g_N^\circ$. Lem. D.12 yields the samplewise inequality

$$\eta_0 \le \|f^\star - f_N^\circ\|_2 + \mathrm{dist}_2(f_N^\circ, \mathbb{B}_{\phi_0}(A)) \le \sqrt{\mathcal{A}_N^{\mathrm{feat}}} + A \|K_{\mathrm{feat},N} - K_{\mathrm{init}}\|_{L^2(\gamma_d^{\otimes 2})}^{1/2}. \tag{38}$$

Choose N_0 so that $C_s N^{-p_{s,m^\star}} \le \eta_0^2/16$ for every $N \ge N_0$, and set $c_{\mathrm{FL}} = \eta_0^2/(16A^2)$. On the event in (34), a kernel distance smaller than c_{FL} would make the right-hand side of (38) smaller than $\eta_0/2$, a contradiction. This proves the kernel separation asserted in the proposition and (FL1). Together with the preceding checks of (FL2)–(FL5), this verifies Def. D.11. □

D.6 Active hidden-layer mass near the index subspace

For a fixed frame E, write $S = \mathrm{Range}(E)$, $P_E = EE^\top$, $\vartheta_E(w) = \|(I - P_E)w\|/\|w\|$ for $w \ne 0$, and $\vartheta_E(0) = 0$. Let $\mathcal{C}_E(\delta) = \{(w,b) : \vartheta_E(w) \le \delta\}$ and $s_h = \|h - \mathbb{E}[h]\|_{L^2(\gamma_{m^\star})}$.

Lemma D.14 (Translation test for angular mass)**.** *Let $f = f_\nu = \int \bar{a}(v)\psi_v \,\mathrm{d}\phi(v)$ with $|\bar{a}| \le A$. Fix $h \in L^2(\gamma_{m^\star})$, $s_h = \|h - \mathbb{E}[h]\|_2 > 0$, and put $r = \|f - h(E^\top \cdot)\|_2$. For every $T > 0$ and $0 < \delta \le 1$,*

$$\int_{\mathcal{C}_E(\delta)} |\bar{a}(v)| \,\mathrm{d}\phi(v) \ge s_h - e^{T^2/2} r - \frac{C_{d-m^\star} A}{T\delta}, \tag{39}$$

$$\phi(\mathcal{C}_E(\delta)) \ge \frac{s_h}{A} - \frac{e^{T^2/2}}{A} r - \frac{C_{d-m^\star}}{T\delta}. \tag{40}$$

Proof. Let $H = h - \mathbb{E}[h]$, and let V be uniform on the unit ball of $S^\perp$, independently of $X \sim \gamma_d$. Define

$$B_T(f) = \mathbb{E}_{X,V}[H(E^\top X) f(X + TV)].$$

Translation invariance of the target along $S^\perp$ gives $B_T(f^\star) = s_h^2$. For a fixed $v \in S^\perp$ with $\|v\| \le 1$, Gaussian change of density gives

$$\begin{aligned} &\left|\mathbb{E}[H(E^\top X)(f - f^\star)(X + Tv)]\right| \\ &\quad = \left| \int H(E^\top x)(f - f^\star)(x) e^{Tv^\top x - T^2\|v\|^2/2} \,\mathrm{d}\gamma_d(x) \right| \le r s_h e^{T^2\|v\|^2/2}. \end{aligned}$$

Indeed, in Cauchy–Schwarz the squared norm of the factor multiplying $f - f^\star$ is $s_h^2 e^{T^2\|v\|^2}$, by independence of the informative and perpendicular Gaussian coordinates. Averaging over V yields

$$|B_T(f) - s_h^2| \le r s_h e^{T^2/2}. \tag{41}$$

For one feature write $u = E^\top w$, $v_\perp = (I_d - P_E)w$ and $q = d - m^\star$. After averaging over $X_\perp$ and V its informative link is

$$F_{w,b,T}(z) = \mathbb{E}[\tanh(u^\top z + v_\perp^\top X_\perp + T v_\perp^\top V + b)].$$

If $v_\perp \neq 0$, the scalar $T v_\perp^\top V$ has a density with supremum at most $c_q/(T\|v_\perp\|)$. This follows by slicing the uniform ball; for $q = 1$ the density is uniform on an interval. Convolution with the independent Gaussian $v_\perp^\top X_\perp$ does not increase that supremum. Since $\int_{\mathbb{R}} \operatorname{sech}^2(t)\,\mathrm{d}t = 2$, the scalar function of $u^\top z$ defining F is Lipschitz with constant at most $2c_q/(T\|v_\perp\|)$. For independent $Z, Z' \sim \gamma_{m^\star}$,

$$\operatorname{Var}(F(Z)) = \tfrac{1}{2}\mathbb{E}[(F(Z) - F(Z'))^2] \le \frac{4c_q^2\|u\|^2}{T^2\|v_\perp\|^2}.$$

Also $|F| \le 1$, so its variance is at most one. Hence

$$|\mathbb{E}[H(Z)F(Z)]| \le s_h \min\left\{1, \frac{2c_q\|u\|}{T\|v_\perp\|}\right\}.$$

Outside $\mathcal{C}_E(\delta)$, $\|u\|/\|v_\perp\| \le 1/\delta$. On the cone, including $v_\perp = 0$, use the bound s_h. Integrating against the signed conditional mean output weights and then taking absolute values gives

$$|B_T(f)| \le s_h \left\{\int_{\mathcal{C}_E(\delta)} |\bar{a}|\,\mathrm{d}\phi + \frac{2c_q A}{T\delta}\right\}.$$

Together with (41), division by s_h proves (39); $|\bar{a}| \le A$ gives (40). □

Proposition D.15 (A nonvanishing active fraction enters a shrinking angular neighborhood)**.** *There exist fixed $D_{\mathcal{G}} > 0$, $c_{\mathcal{G}} > 0$, and N_0 such that for all $N \ge N_0$, on $\Omega_{N,h,E}$,*

$$\int_{\mathcal{C}_E(D_{\mathcal{G}}/\sqrt{\ell_N})} |\bar{a}_N|\,\mathrm{d}\widehat{\phi}_N \ge c_{\mathcal{G}}, \qquad \widehat{\phi}_N(\mathcal{C}_E(D_{\mathcal{G}}/\sqrt{\ell_N})) \ge c_{\mathcal{G}}/A, \tag{42}$$

where $s_{\mathcal{G}} = \inf_{h\in\mathcal{G}} \|h - \mathbb{E}[h]\|_2 > 0$ and $c_{\mathcal{G}} = s_{\mathcal{G}}/2$. The failure probability is at most CN^{-8} uniformly in h, E. In contrast,

$$\phi_0(\mathcal{C}_E(D_{\mathcal{G}}/\sqrt{\ell_N})) = O(\ell_N^{-(d-m^\star)/2}) \longrightarrow 0. \tag{43}$$

Proof. The centered L^2 norm is continuous. No member of the compact set $\mathcal{G}$ is constant, since it is nonzero and compactly supported. Its minimum $s_{\mathcal{G}}$ is therefore positive. On $\Omega_{N,h,E}$, use (32) to bound $r \le \sqrt{2S_N} \le C_s N^{-p_{s,m^\star}/2}$. Set

$$T_N = \frac{1}{2}\sqrt{\ell_N}, \qquad \delta_N^{\mathrm{ang}} = D_{\mathcal{G}}/\sqrt{\ell_N}.$$

Then, because $p_{s,m^\star} > 1/2$,

$$e^{T_N^2/2} r \le C_s N^{1/8 - p_{s,m^\star}/2} \longrightarrow 0, \qquad T_N \delta_N^{\mathrm{ang}} = D_{\mathcal{G}}/2.$$

Choose $D_{\mathcal{G}}$ so that $2C_{d-m^\star}A/D_{\mathcal{G}} \le s_{\mathcal{G}}/4$, and then N_0 so that the first term is at most $s_{\mathcal{G}}/4$ and $\delta_N^{\mathrm{ang}} \le 1$. Lem. D.14 proves (42). Under the isotropic Gaussian prior, $\vartheta_E(W)^2$ has beta distribution with parameters $(d-m^\star)/2$ and $m^\star/2$, by splitting the independent sums of squares of the perpendicular and informative coordinates. Its probability below δ^2 is asymptotic to a positive constant times $\delta^{d-m^\star}$, proving (43). □

D.7 SMOOTHED GRADIENTS AND INDEX-SUBSPACE RECOVERY

For every dimension $j \geq 1$, define the time-one Ornstein–Uhlenbeck operator by

$$(T_1^{(j)} f)(x) := \mathbb{E}_{G\sim\gamma_j}\Big[f(e^{-1}x + \sqrt{1-e^{-2}}G)\Big]. \tag{44}$$

For $j = d$, this is the operator used in Sec. 5.1. We suppress the dimension when it is clear from context and set $c_{\mathrm{OU}}^2 = \sup_{k\geq 1} ke^{-2k} < \infty$. Norms of vector fields use the Euclidean norm inside the Gaussian L^2 integral.

Lemma D.16 (A bounded smoothed-gradient map)**.** *For every $g \in L^2(\gamma_j)$,*

$$\|\nabla T_1^{(j)} g\|_{L^2(\gamma_j)}^2 = \sum_{\eta} |\eta| e^{-2|\eta|} |g_\eta|^2 \leq c_{\mathrm{OU}}^2 \|g\|_2^2. \tag{45}$$

If $\mathcal{M}(g) = \int (\nabla T_1 g)(\nabla T_1 g)^\top \,\mathrm{d}\gamma_j$, then

$$\|\mathcal{M}(g) - \mathcal{M}(f)\|_{\mathrm{F}} \leq c_{\mathrm{OU}}^2 (\|g\|_2 + \|f\|_2)\|g - f\|_2. \tag{46}$$

Proof. The normalized Hermite basis diagonalizes T_1, with multiplier $e^{-|\eta|}$. The derivative identity $\partial_i \mathrm{H}_\eta = \sqrt{\eta_i}\mathrm{H}_{\eta - e_i}$ and Parseval prove (45) first for finite sums and then by closure. For vector fields $u = \nabla T_1 g$, $v = \nabla T_1 f$, use $uu^\top - vv^\top = (u-v)u^\top + v(u-v)^\top$, $\|xy^\top\|_{\mathrm{F}} = \|x\|\|y\|$, integration, and Cauchy–Schwarz. Equation (45) then proves (46). □

Lemma D.17 (A uniform identifiable gap on the fixed family)**.** *Let*

$$B(h) = \int (\nabla T_1^{(m^\star)} h)(\nabla T_1^{(m^\star)} h)^\top \,\mathrm{d}\gamma_{m^\star}.$$

Then

$$\kappa_{\mathcal{G}} := \inf_{h\in\mathcal{G}} \lambda_{\min}(B(h)) > 0, \qquad \mathcal{M}(h(E^\top \cdot)) = EB(h)E^\top. \tag{47}$$

Proof. Gaussian conditioning in (44) shows that $T_1^{(d)}[h(E^\top\cdot)] = (T_1^{(m^\star)}h)(E^\top\cdot)$. Taking its gradient proves the matrix identity.

Suppose $u^\top B(h) u = 0$ for a unit vector $u \in \mathbb{R}^{m^\star}$. Then $u^\top \nabla T_1 h = 0$ in $L^2(\gamma_{m^\star})$, and hence $T_1 h$ depends only on the projection onto $u^\perp$. Let $Q_{u^\perp}$ denote the corresponding Gaussian conditional-expectation projection. It commutes with T_1: rotate u to a coordinate vector and use the product Mehler kernel, or use the product Hermite basis. Therefore

$$T_1(h - Q_{u^\perp} h) = (I - Q_{u^\perp}) T_1 h = 0.$$

Every Hermite multiplier of T_1 is strictly positive, so it is injective on L^2, and $h = Q_{u^\perp} h$. Such a function is invariant along u. A nonzero compactly supported function cannot have this property: in coordinates $(z_\perp, r)$ it is almost everywhere equal to a function of $z_\perp$ only, while it vanishes for all sufficiently large $|r|$ for almost every fixed $z_\perp$. Fubini forces that function, and hence h, to vanish. This contradicts $h \in \mathcal{G}$. Thus each $B(h)$ is positive definite. Lem. D.16 gives continuity of $h \mapsto B(h)$ in L^2; continuity of the smallest eigenvalue and compactness of $\mathcal{G}$ give the positive minimum in (47). □

Using the fitted matrix from Sec. 5.1, write

$$v_N(x) := \nabla T_1^{(d)} \widehat{f}_N(x), \qquad \widehat{G}_N := \int v_N(x) v_N(x)^\top \,\mathrm{d}\gamma_d(x).$$

Lem. D.1 gives $\int \|w\|^2 \,\mathrm{d}\widehat{\nu}_N < \infty$, so differentiation under the neural integral is justified and gives

$$v_N(x) = e^{-1} \int aw\, \mathbb{E}_{Z\sim\gamma_1}\Big[\mathrm{sech}^2\Big(e^{-1}w^\top x + b + \sqrt{1-e^{-2}}\|w\|Z\Big)\Big] \,\mathrm{d}\widehat{\nu}_N(a,w,b). \tag{48}$$

Let $\widehat{P}_N^{(m^\star)}$ denote the top-$m^\star$ spectral projector of $\widehat{G}_N$. At an eigenvalue tie, apply Gram–Schmidt to projections of the standard coordinate vectors in their fixed order to choose a measurable eigenbasis. For Borel $B \subset \mathbb{S}^{d-1}$, define the gradient-energy-weighted direction law by

$$\mathsf{D}_N(B) = \frac{\int \|v_N(x)\|^2 \mathbf{1}_{\{v_N(x)\neq 0,\ v_N(x)/\|v_N(x)\| \in B\}}\, \mathrm{d}\gamma_d(x)}{\operatorname{tr} \widehat{G}_N}$$

when the denominator is positive, and set $\mathsf{D}_N = \delta_{e_1}$ otherwise. On the positive-trace event,

$$\int uu^\top \,\mathrm{d}\mathsf{D}_N(u) = \frac{\widehat{G}_N}{\operatorname{tr} \widehat{G}_N}. \tag{49}$$

Proposition D.18 (Effective directional concentration and subspace recovery)**.** *There exist fixed constants C_s, N_0 such that, for $N \geq N_0$, on $\Omega_{N,h,E}$,*

$$\operatorname{tr}((I_d - P_E)\widehat{G}_N) \leq C_s N^{-p_{s,m^\star}}, \tag{50}$$

$$\|\widehat{P}_N^{(m^\star)} - P_E\|_{\mathrm{F}}^2 \leq C_s N^{-p_{s,m^\star}}, \tag{51}$$

$$\int \|(I_d - P_E)u\|^2 \,\mathrm{d}\mathsf{D}_N(u) \leq C_s N^{-p_{s,m^\star}}. \tag{52}$$

Thus, for every $\delta > 0$,

$$\mathsf{D}_N\{u : \operatorname{dist}(u, S) > \delta\} \leq \frac{C_s N^{-p_{s,m^\star}}}{\delta^2}. \tag{53}$$

All statements have failure probability at most CN^{-8} uniformly in h, E. In particular, for the deterministic radius $N^{-1/8}$, the mass outside the corresponding angular neighborhood is at most $C_s N^{-p_{s,m^\star}+1/4}$ and tends to zero.

Proof. Write $G^\star = EB(h)E^\top$ and $\Delta = \widehat{G}_N - G^\star$. The target's smoothed gradient is in S. Consequently Lem. D.16 and (32) yield

$$\operatorname{tr}((I_d - P_E)\widehat{G}_N) = \|(I_d - P_E)\nabla T_1 \widehat{f}_N\|_2^2 \leq c_{\mathrm{OU}}^2 \|\widehat{f}_N - f^\star\|_2^2 \leq 2c_{\mathrm{OU}}^2 S_N, \tag{54}$$

$$\|\Delta\|_{\mathrm{F}} \leq c_{\mathrm{OU}}^2 (A + M)\|\widehat{f}_N - f^\star\|_2 \leq c_{\mathrm{OU}}^2 (A + M)\sqrt{2S_N}. \tag{55}$$

Use $S_N \leq C_s N^{-p_{s,m^\star}}$ to obtain (50). Since $N^{-p_{s,m^\star}} \to 0$ and the gap in Lem. D.17 is uniform, eventually $\|\Delta\|_{\mathrm{op}} \leq \kappa_{\mathcal{G}}/2$. The top $m^\star$ eigenvalues of $\widehat{G}_N$ are then at least $\kappa_{\mathcal{G}}/2$.

Here is a direct spectral perturbation calculation. Let $\widehat{U}$ have these $m^\star$ orthonormal eigenvectors as columns and $\widehat{\Lambda}$ contain their eigenvalues. Since $(I_d - P_E)G^\star = 0$,

$$(I_d - P_E)\widehat{U}\widehat{\Lambda} = (I_d - P_E)\Delta\widehat{U}.$$

It follows that

$$\|(I_d - P_E)\widehat{U}\|_{\mathrm{F}} \leq \frac{2\|\Delta\|_{\mathrm{F}}}{\kappa_{\mathcal{G}}}, \qquad \|\widehat{P}_N^{(m^\star)} - P_E\|_{\mathrm{F}}^2 = 2\|(I_d - P_E)\widehat{U}\|_{\mathrm{F}}^2 \leq \frac{8\|\Delta\|_{\mathrm{F}}^2}{\kappa_{\mathcal{G}}^2}.$$

This and (55) prove (51). Also $\operatorname{tr} \widehat{G}_N \geq m^\star \kappa_{\mathcal{G}}/2$. Divide (54) by this trace and use (49) to prove (52). Markov's inequality on the sphere gives (53). □

Define the thresholded projector and dimension estimate by

$$u_N := N^{-1/8}, \qquad \widehat{P}_N := \mathbf{1}_{(u_N,\infty)}(\widehat{G}_N), \qquad \widehat{m}_N := \operatorname{rank}(\widehat{P}_N), \tag{56}$$

where $\mathbf{1}_{(u_N,\infty)}$ denotes spectral projection onto eigenvalues greater than u_N.

Corollary D.19 (Unknown intrinsic dimension)**.** *For all sufficiently large N, on $\Omega_{N,h,E}$, the thresholded projector and dimension estimate in (56) satisfy*

$$\widehat{m}_N = m^\star, \qquad \widehat{P}_N = \widehat{P}_N^{(m^\star)}, \qquad \|\widehat{P}_N - P_E\|_{\mathrm{F}}^2 \leq C_s N^{-p_{s,m^\star}}.$$

The threshold depends only on N.

Proof. The nonzero target eigenvalues are at least $\kappa_{\mathcal{G}}$, and the remaining ones are zero. Equation (55) bounds every eigenvalue perturbation by $C_s N^{-p_{s,m^\star}/2}$. Since $p_{s,m^\star} > 1/2$, for all sufficiently large N,

$$C_s N^{-p_{s,m^\star}/2} < u_N = N^{-1/8} < \kappa_{\mathcal{G}}/2.$$

Exactly $m^\star$ eigenvalues exceed u_N, so the thresholded and top-$m^\star$ projectors coincide. Prop. D.18 proves the error bound. □

Proof of Cor. 5.3. Apply Prop. D.15 and Cor. D.19 on $\Omega_{N,h^\star,E_\star}$. The probability estimate and the stated rate follow from (34). □

D.8 Intrinsic feature space and prediction transfer

Fix $E \in \mathrm{St}(d, m^\star)$ and choose an orthogonal completion $E_\perp$. Define bounded linear operators

$$(U_E h)(x) = h(E^\top x),$$
$$(\Pi_E f)(z) = \int f(Ez + E_\perp u)\, \mathrm{d}\gamma_{d-m^\star}(u), \qquad Q_E = U_E \Pi_E.$$

The conditional integral is interpreted in L^2; its value is independent of the chosen orthogonal completion.

Lemma D.20 (Intrinsic and residual orthogonal decomposition)**.** *The map $U_E : L^2(\gamma_{m^\star}) \to L^2(\gamma_d)$ is an isometry, $\Pi_E = U_E^*$, and Q_E is the orthogonal projection onto the closed subspace of functions of $E^\top x$. Thus, for all $f \in L^2(\gamma_d)$ and $h \in L^2(\gamma_{m^\star})$,*

$$\|f - U_E h\|_2^2 = \|\Pi_E f - h\|_{L^2(\gamma_{m^\star})}^2 + \|(I - Q_E)f\|_2^2. \tag{57}$$

Conditionally on the training sample, set

$$C_{N,E} = \Pi_E J_N, \qquad D_{N,E} = (I - Q_E) J_N.$$

Then

$$J_N^* J_N = C_{N,E}^* C_{N,E} + D_{N,E}^* D_{N,E}. \tag{58}$$

Proof. If $Z \sim \gamma_{m^\star}$ and $U \sim \gamma_{d-m^\star}$ are independent, $EZ + E_\perp U \sim \gamma_d$. It follows that $\|U_E h\|_2 = \|h\|_{L^2(\gamma_{m^\star})}$ and, by Fubini,

$$\langle U_E h, f \rangle_{L^2(\gamma_d)} = \langle h, \Pi_E f \rangle_{L^2(\gamma_{m^\star})}.$$

Therefore $\Pi_E U_E = I$, $Q_E^* = Q_E$, and $Q_E^2 = Q_E$. The two terms in

$$f - U_E h = U_E(\Pi_E f - h) + (I - Q_E) f$$

are orthogonal. This proves (57). Apply it with $h = 0$ to $J_N g$, and polarize the resulting quadratic identity, to obtain (58). □

Proposition D.21 (Intrinsic feature space and transfer of the oracle bound)**.** *For $f^\star = U_E h$, the feature approximation and output-layer alignment errors satisfy the exact decompositions*

$$\mathcal{A}_N^{\mathrm{feat}} = \|C_{N,E} g_N^\circ - h\|_{L^2(\gamma_{m^\star})}^2 + \|D_{N,E} g_N^\circ\|_2^2, \tag{59}$$
$$\mathcal{E}_N^{\mathrm{out}} = \|C_{N,E}(\widehat{g}_N - g_N^\circ)\|_{L^2(\gamma_{m^\star})}^2 + \|D_{N,E}(\widehat{g}_N - g_N^\circ)\|_2^2. \tag{60}$$

In particular, samplewise,

$$\|\Pi_E \widehat{f}_N - h\|_{L^2(\gamma_{m^\star})}^2 + \|(I - Q_E)\widehat{f}_N\|_2^2 \le 2\mathcal{A}_N^{\mathrm{feat}} + 2\mathcal{E}_N^{\mathrm{out}}. \tag{61}$$

The range of $C_{N,E}$, equipped with its quotient RKHS norm, is the intrinsic learned RKHS $\mathcal{H}_{N,E}^{\mathrm{info}}$ with kernel

$$\bar{\psi}_v^E(z) = \mathbb{E}_G\big[\tanh\big((E^\top w)^\top z + b + \|(I - EE^\top)w\| G\big)\big], \qquad G \sim \mathcal{N}(0,1),$$
$$K_{N,E}^{\mathrm{info}}(z, z') = \int \bar{\psi}_v^E(z)\bar{\psi}_v^E(z')\, \mathrm{d}\widehat{\phi}_N(v). \tag{62}$$

Its radius-A ball satisfies

$$\mathrm{dist}_{L^2(\gamma_{m^\star})}\big(h, B_{\mathcal{H}_{N,E}^{\mathrm{info}}}(A)\big)^2 \le \|C_{N,E} g_N^\circ - h\|_{L^2(\gamma_{m^\star})}^2 \le \mathcal{A}_N^{\mathrm{feat}}. \tag{63}$$

Proof. Apply (57) to $J_N g_N^\circ - U_E h$ and (58) to $\widehat{g}_N - g_N^\circ$. This gives (59)–(60). Let $\Delta_N = \widehat{g}_N - g_N^\circ$. In the direct sum $L^2(\gamma_{m^\star}) \oplus L^2(\gamma_d)$, write

$$\begin{pmatrix} C_{N,E}\widehat{g}_N - h \\ D_{N,E}\widehat{g}_N \end{pmatrix} = \begin{pmatrix} C_{N,E}g_N^\circ - h \\ D_{N,E}g_N^\circ \end{pmatrix} + \begin{pmatrix} C_{N,E}\Delta_N \\ D_{N,E}\Delta_N \end{pmatrix}.$$

The squared triangle inequality and the two exact decompositions give (61).

For the intrinsic kernel, a coefficient $c \in L^2(\widehat{\phi}_N)$ represents $J_N g = \int c(v)\psi_v \, \mathrm{d}\widehat{\phi}_N(v)$. Gaussian conditioning and Fubini give

$$C_{N,E}g(z) = \int c(v)\bar{\psi}_v^E(z) \, \mathrm{d}\widehat{\phi}_N(v).$$

The coefficient-space feature construction therefore gives (62). Its norm is

$$\|u\|_{\mathcal{H}_{N,E}^{\mathrm{info}}} = \inf\{\|g\|_{\mathcal{H}_N} : C_{N,E}g = u\}.$$

This is a Hilbert norm because $\ker C_{N,E}$ is closed; evaluation is bounded by $\|g\|_{\mathcal{H}_N}$, so the range with this norm is indeed an RKHS. The same norm results by minimizing directly over coefficient representatives in $L^2(\widehat{\phi}_N)$, since the kernel of the original feature map is contained in that of the conditioned feature map. As $\|g_N^\circ\|_{\mathcal{H}_N} \le A$, $C_{N,E}g_N^\circ$ belongs to its radius-A ball. Equation (59) now proves (63). □

D.9 Noncompact prediction bounds

Proposition D.22 (Prediction rate on the fixed rough family)**.** *Assume* (2)*. For the fixed family $\mathcal{G}$ and temperature* (3)*, there are fixed $C_s, C_{\mathrm{fail}} > 0$ such that, for every $N \ge 2$,*

$$\sup_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \mathbb{E}_{h,E}\left[\|\widehat{f}_N - h(E^\top \cdot)\|_2^2\right] \le C_s N^{-p_{s,m^\star}}, \tag{64}$$

and

$$\sup_{\substack{h \in \mathcal{G} \\ E^\top E = I_{m^\star}}} \mathbb{P}_{h,E}\left\{\|\Pi_E \widehat{f}_N - h\|_{L^2(\gamma_{m^\star})}^2 + \|(I - Q_E)\widehat{f}_N\|_2^2 > C_s N^{-p_{s,m^\star}}\right\} \le C_{\mathrm{fail}} N^{-8}. \tag{65}$$

The sum inside the probability equals the full prediction error.

Proof. On $\Omega_{N,h,E}$, Prop. D.21 bounds the sum of intrinsic and residual errors by $2S_N \le C_s N^{-p_{s,m^\star}}$, proving (65). Integrating the same deterministic transfer and using (35) proves (64). The orthogonal identity (57) identifies both expressions with the full prediction error. □

Proposition D.23 (Prediction bound for a bounded regression function)**.** *Under the bounded-regression assumptions of Appendix C.1, for $N \ge 2$ and $N^{-1} \le \lambda \le 1/2$, the full noncompact estimator satisfies*

$$\mathbb{E}\left[\|\widehat{f}_{N,\lambda} - f^\star\|_2^2\right] \le C\mathcal{B}_{N,\lambda}(f^\star),$$
$$\mathbb{P}\{\|\widehat{f}_{N,\lambda} - f^\star\|_2^2 > C\mathcal{B}_{N,\lambda}(f^\star)\} \le C_{\mathrm{fail}} N^{-8}.$$

The constants depend only on $d, A, M, \sigma_\xi, \tau_w, \tau_b$. The expectation includes input sampling and label noise.

Proof. Lem. D.8 bounds the prediction error by twice the sum of feature approximation and output-layer alignment errors. Prop. D.9 therefore gives the probability bound. On the exceptional event the prediction error is at most $(A + M)^2$, while $\mathcal{B}_{N,\lambda} \ge N^{-1}$. Integration proves the expectation bound. □

Proof of Thm. 5.1.

Proof. Construct $\mathcal{G}$ as in Appendix C.3. Its stated structural properties follow from Lem. C.5, the uniform compact-model gap from Prop. C.6, and the prediction bound (4) from Prop. D.22. Finally, Prop. E.2 identifies the estimator with the long-time MFLD predictor at each fixed sample size and positive temperature. □

A concrete regime and the comparison range. For $(m^\star, s, d) = (2, 3/2, 3)$, choose the same smooth cutoffs and let $H_0(z_1, z_2) = \chi_1(z_1)\chi_2(z_2)|z_1|^{11/4}$. The third derivative of the first factor near zero is $\frac{231}{64}\operatorname{sgn}(z_1)|z_1|^{-1/4}$ and is locally square-integrable. Its fourth derivative on either side of zero is $-\frac{231}{256}|z_1|^{-5/4}$, which is not locally square-integrable. The cutoff terms are smooth away from zero, so Lem. A.1 gives $H_0 \in H^3(\gamma_2) \setminus H^4(\gamma_2)$. Set $r_+ = 4$ and choose $h_0 = c_0 H_0$, $\epsilon_0 > 0$ by (16), using the cost constants of Lem. C.4. Then

$$\mathcal{G} = h_0 + \mathcal{V}_4(\epsilon_0) \subset H^3(\gamma_2) \setminus H^4(\gamma_2) \subset H^{3/2}(\gamma_2) \setminus H^4(\gamma_2).$$

The quadratic costs of the high Fourier-frequency bands are summable with exponents $5/2 - 11/4 = -1/4$ for the center and $(2+5)/2 - 4 = -1/2$ for the perturbations. Thus all representation and compact-gap arguments apply to this same fixed family.

Here $\beta_3 = 8/11$, $3d/2 + 2 = 13/2$, and $p_{3/2,2} = 3/5$, with margin $8/11 - 3/5 = 7/55 > 0$. For $N \geq 4$, $\lambda_N = N^{-8/11}$, and direct substitution into the statistical budget gives $\mathcal{B}_{N,\lambda_N}(h(E^\top \cdot)) \leq CN^{-3/5}$. The smoothing error is $O(\lambda_N)$ and the penalized comparator cost is $O(\lambda_N \ell_N)$. The analysis mesh in (26) satisfies

$$N\lambda_N^2 = N^{-5/11}, \qquad \eta_N \asymp N^{-1/11}\ell_N^{-1}, \qquad D_N \lesssim N^{3/11}\ell_N^{9/2},$$

so $D_N \ell_N^2 / N \lesssim N^{-3/5}$. Thus the feature approximation, output-layer alignment, intrinsic link error, ambient residual error, and squared projector error all obey the stated $N^{-3/5}$ bound.

For every fixed $d > m^\star \geq 2$, the interval

$$\frac{m^\star}{2} < s < \frac{m^\star(d+5)}{2d}$$

has length $5m^\star/(2d) > 0$ and ensures $p_{s,m^\star} < \beta_d$. Prop. E.2 transfers the statistical guarantees to the MFLD limit, taken at each fixed sample size and positive temperature.

E Long-Time MFLD and Projected Feature Scales

Both proofs use the same proximal-Gibbs argument: a variational gap inequality, a uniform log-Sobolev inequality for the frozen Gibbs laws, and free-energy dissipation. The compact dynamics reflect on the full parameter boundary, whereas the noncompact dynamics reflect only at $a = \pm A$. For the latter, we use the notation of Appendix C.

E.1 Compact reflected MFLD and the RERM

For $\theta = (a, w, b) \in \Theta_{\mathrm{c}}$, set $\varphi_\theta(x) := a\sigma(\langle w, x\rangle + b)$ and write $J^{\mathrm{c}}_{N,\lambda}(\nu) := P_N \ell_{f_\nu} + \lambda \,\mathrm{KL}(\nu \| \nu_{0,\mathrm{c}}) = J^{\Theta_{\mathrm{c}}}_{N,\lambda}(\nu)$. For a fixed sample, the first variation of the empirical risk is

$$V^{\mathrm{c}}_\nu(\theta) := \frac{\delta P_N \ell_{f_\nu}}{\delta \nu}(\theta) = -\frac{2}{N}\sum_{i=1}^N \{Y_i - f_\nu(X_i)\}\varphi_\theta(X_i).$$

Write $\nu^{\mathrm{c}}_t = \rho_t \nu_{0,\mathrm{c}}$. The reflected compact MFLD is

$$\partial_t \rho_t = \nabla_\theta \cdot \left(\rho_t \nabla_\theta V^{\mathrm{c}}_{\nu^{\mathrm{c}}_t} + \lambda \nabla_\theta \rho_t\right) \quad \text{on } \operatorname{int}(\Theta_{\mathrm{c}}), \tag{66}$$

with $\rho_0 \equiv 1$ and the no-flux boundary condition

$$\left(\rho_t \nabla_\theta V^{\mathrm{c}}_{\nu^{\mathrm{c}}_t} + \lambda \nabla_\theta \rho_t\right) \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Theta_{\mathrm{c}}. \tag{67}$$

At the corners of the product domain, the boundary condition is understood in the weak sense. Equivalently, ν^{c}_t is the law of

$$\mathrm{d}\Theta_t = -\nabla_\theta V^{\mathrm{c}}_{\nu^{\mathrm{c}}_t}(\Theta_t)\,\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}B_t + \mathrm{d}K_t, \qquad \nu^{\mathrm{c}}_t = \mathrm{Law}(\Theta_t),$$

where K_t keeps Θ_t in Θ_{c}.

Proposition E.1 (Long-time compact MFLD and the RERM). *Assume $\sigma \in C_b^2(\mathbb{R})$ and $\lambda > 0$, and let $(\nu_t^{\mathrm{c}})_{t\geq 0}$ solve (66)–(67) from $\nu_{0,\mathrm{c}}$. Let $\rho_{\mathrm{c}} > 0$ be a log-Sobolev constant of $\nu_{0,\mathrm{c}}$ under the convention*

$$\mathrm{KL}(\mu\|\nu_{0,\mathrm{c}}) \leq \frac{1}{2\rho_{\mathrm{c}}} I(\mu\|\nu_{0,\mathrm{c}}), \qquad I(\mu\|\nu_{0,\mathrm{c}}) := \int_{\Theta_{\mathrm{c}}} \left\| \nabla_\theta \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu_{0,\mathrm{c}}}\right)\right\|_2^2 \mathrm{d}\mu. \tag{68}$$

Such a constant exists because Θ_{c} is bounded and convex. Set $M_\sigma := A\|\sigma\|_\infty$ and

$$\Omega_N^{\mathrm{c}} := \sup_{\nu\in\mathcal{P}(\Theta_{\mathrm{c}})} \mathrm{osc}_{\Theta_{\mathrm{c}}}(V_\nu^{\mathrm{c}}), \qquad \mathrm{osc}_{\Theta_{\mathrm{c}}}(V) := \sup_{\theta\in\Theta_{\mathrm{c}}} V(\theta) - \inf_{\theta\in\Theta_{\mathrm{c}}} V(\theta).$$

Then

$$\Omega_N^{\mathrm{c}} \leq \frac{4M_\sigma}{N}\sum_{i=1}^N (|Y_i| + M_\sigma) < \infty. \tag{69}$$

The following statements hold.

(i) *The objective $J_{N,\lambda}^{\mathrm{c}}$ has the unique minimizer $\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}$. This minimizer is equivalent to $\nu_{0,\mathrm{c}}$ and satisfies*

$$\frac{\mathrm{d}\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}}{\mathrm{d}\nu_{0,\mathrm{c}}}(\theta) = \frac{\exp\{-\lambda^{-1}V^{\mathrm{c}}_{\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}}(\theta)\}}{\displaystyle\int_{\Theta_{\mathrm{c}}} \exp\{-\lambda^{-1}V^{\mathrm{c}}_{\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}}(\theta')\}\,\mathrm{d}\nu_{0,\mathrm{c}}(\theta')}. \tag{70}$$

Hence $\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}$ is stationary for the reflected MFLD.

(ii) *Define $\rho_{N,\lambda}^{\mathrm{c}} := \rho_{\mathrm{c}}\exp(-\Omega_N^{\mathrm{c}}/\lambda)$ and $\kappa_{N,\lambda}^{\mathrm{c}} := 2\lambda\rho_{N,\lambda}^{\mathrm{c}}$. Then, for every $t \geq 0$,*

$$J_{N,\lambda}^{\mathrm{c}}(\nu_t^{\mathrm{c}}) - J_{N,\lambda}^{\mathrm{c}}(\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}) \leq e^{-\kappa_{N,\lambda}^{\mathrm{c}} t}\left\{J_{N,\lambda}^{\mathrm{c}}(\nu_{0,\mathrm{c}}) - J_{N,\lambda}^{\mathrm{c}}(\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}})\right\}, \tag{71}$$

$$\mathrm{KL}(\nu_t^{\mathrm{c}}\|\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}) \leq \frac{1}{\lambda}e^{-\kappa_{N,\lambda}^{\mathrm{c}} t}\left\{J_{N,\lambda}^{\mathrm{c}}(\nu_{0,\mathrm{c}}) - J_{N,\lambda}^{\mathrm{c}}(\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}})\right\}. \tag{72}$$

(iii) *Consequently, $\nu_t^{\mathrm{c}} \to \widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}$ in total variation and in W_2. Moreover, for every Borel probability measure $\mathbb{P}_X$ on $\mathbb{R}^d$,*

$$\begin{aligned}\|f_{\nu_t^{\mathrm{c}}} - f_{\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}}\|_{L^2(\mathbb{P}_X)}^2 &\leq 2M_\sigma^2\,\mathrm{KL}(\nu_t^{\mathrm{c}}\|\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}) \\ &\leq \frac{2M_\sigma^2}{\lambda}e^{-\kappa_{N,\lambda}^{\mathrm{c}} t}\left\{J_{N,\lambda}^{\mathrm{c}}(\nu_{0,\mathrm{c}}) - J_{N,\lambda}^{\mathrm{c}}(\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}})\right\}.\end{aligned} \tag{73}$$

Proof. Existence, uniqueness, and absolute continuity of $\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}$ follow from Lem. A.2. For $\nu \in \mathcal{P}(\Theta_{\mathrm{c}})$, define the frozen Gibbs measure

$$\frac{\mathrm{d}\pi_\nu^{\mathrm{c}}}{\mathrm{d}\nu_{0,\mathrm{c}}}(\theta) := \frac{\exp\{-\lambda^{-1}V_\nu^{\mathrm{c}}(\theta)\}}{\displaystyle\int_{\Theta_{\mathrm{c}}}\exp\{-\lambda^{-1}V_\nu^{\mathrm{c}}(\theta')\}\,\mathrm{d}\nu_{0,\mathrm{c}}(\theta')}.$$

For every ν, i, and θ, one has $|f_\nu(X_i)| \leq M_\sigma$ and $|\varphi_\theta(X_i)| \leq M_\sigma$. Therefore, for all $\theta, \theta' \in \Theta_{\mathrm{c}}$,

$$|V_\nu^{\mathrm{c}}(\theta) - V_\nu^{\mathrm{c}}(\theta')| \leq \frac{2}{N}\sum_{i=1}^N |Y_i - f_\nu(X_i)|\,|\varphi_\theta(X_i) - \varphi_{\theta'}(X_i)| \leq \frac{4M_\sigma}{N}\sum_{i=1}^N (|Y_i| + M_\sigma),$$

which proves (69).

The Euler–Lagrange equation at the minimizer is

$$V^{\mathrm{c}}_{\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}}(\theta) + \lambda\log\left(\frac{\mathrm{d}\widehat{\nu}_{N,\lambda}^{\Theta_{\mathrm{c}}}}{\mathrm{d}\nu_{0,\mathrm{c}}}(\theta)\right) = c_\lambda \qquad \nu_{0,\mathrm{c}}\text{-a.e.} \tag{74}$$

for a constant c_λ. The first variation is bounded, so the minimizer and $\nu_{0,\mathrm{c}}$ are equivalent. Normalization gives (70).

For every $\nu \in \mathcal{P}(\Theta_\mathrm{c})$, convexity of the empirical risk, (74), and the Gibbs variational formula give

$$\lambda \,\mathrm{KL}(\nu \| \widehat{\nu}^{\Theta_\mathrm{c}}_{N,\lambda}) \le J^{\mathrm{c}}_{N,\lambda}(\nu) - J^{\mathrm{c}}_{N,\lambda}(\widehat{\nu}^{\Theta_\mathrm{c}}_{N,\lambda}) \le \lambda \,\mathrm{KL}(\nu \| \pi^{\mathrm{c}}_\nu). \tag{75}$$

Indeed, the lower bound follows by linearizing the empirical risk at the minimizer and substituting (74). For the upper bound, linearize at ν and use

$$-\lambda \log \int_{\Theta_\mathrm{c}} e^{-V^{\mathrm{c}}_\nu/\lambda} \,\mathrm{d}\nu_{0,\mathrm{c}} = \inf_{\mu \in \mathcal{P}(\Theta_\mathrm{c})} \left\{ \int_{\Theta_\mathrm{c}} V^{\mathrm{c}}_\nu \,\mathrm{d}\mu + \lambda \,\mathrm{KL}(\mu \| \nu_{0,\mathrm{c}}) \right\}.$$

The reference measure satisfies (68). Since $\mathrm{osc}(V^{\mathrm{c}}_\nu/\lambda) \le \Omega^{\mathrm{c}}_N/\lambda$, the Holley–Stroock bounded-perturbation principle (Holley & Stroock, 1987) gives, uniformly in ν,

$$I(\mu \| \pi^{\mathrm{c}}_\nu) \ge 2\rho^{\mathrm{c}}_{N,\lambda} \,\mathrm{KL}(\mu \| \pi^{\mathrm{c}}_\nu). \tag{76}$$

Let $u_t := V^{\mathrm{c}}_{\nu^{\mathrm{c}}_t} + \lambda \log \rho_t$. Equation (66) becomes $\partial_t \rho_t = \nabla_\theta \cdot (\rho_t \nabla_\theta u_t)$. The Fokker–Planck chain rule, integration by parts, and (67) give

$$\frac{\mathrm{d}}{\mathrm{d}t} J^{\mathrm{c}}_{N,\lambda}(\nu^{\mathrm{c}}_t) = -\int_{\Theta_\mathrm{c}} \|\nabla_\theta u_t\|_2^2 \,\mathrm{d}\nu^{\mathrm{c}}_t = -\lambda^2 I(\nu^{\mathrm{c}}_t \| \pi^{\mathrm{c}}_{\nu^{\mathrm{c}}_t}). \tag{77}$$

This identity is justified first at positive times by regularization of the uniformly elliptic reflected diffusion and then at time zero by approximation, consistently with the standard Wasserstein gradient-flow chain rule (Ambrosio et al., 2008).

Set $G_t := J^{\mathrm{c}}_{N,\lambda}(\nu^{\mathrm{c}}_t) - J^{\mathrm{c}}_{N,\lambda}(\widehat{\nu}^{\Theta_\mathrm{c}}_{N,\lambda})$. Equations (75), (76), and (77) yield $\frac{\mathrm{d}}{\mathrm{d}t} G_t \le -2\lambda \rho^{\mathrm{c}}_{N,\lambda} G_t = -\kappa^{\mathrm{c}}_{N,\lambda} G_t$. Gronwall's inequality proves (71), and the lower bound in (75) gives (72).

Pinsker's inequality (Pinsker, 1964) gives total-variation convergence. Since Θ_c is compact, total-variation convergence also implies W_2 convergence. Finally,

$$|f_{\nu^{\mathrm{c}}_t}(x) - f_{\widehat{\nu}^{\Theta_\mathrm{c}}_{N,\lambda}}(x)| \le M_\sigma \int_{\Theta_\mathrm{c}} |\,\mathrm{d}\nu^{\mathrm{c}}_t - \,\mathrm{d}\widehat{\nu}^{\Theta_\mathrm{c}}_{N,\lambda}| \le M_\sigma \sqrt{2\,\mathrm{KL}(\nu^{\mathrm{c}}_t \| \widehat{\nu}^{\Theta_\mathrm{c}}_{N,\lambda})}.$$

Squaring and integrating with respect to $\mathbb{P}_X$ proves (73). □

E.2 Noncompact reflected MFLD and the RERM

Fix a realized sample and $\lambda > 0$. Define

$$U_0(a, w, b) = \frac{\|w\|^2}{2\tau_w^2} + \frac{b^2}{2\tau_b^2}, \qquad V_\nu(\theta) = \frac{2}{N} \sum_{i=1}^N (f_\nu(X_i) - Y_i)\varphi_\theta(X_i).$$

Let $\Theta_t = (a_t, w_t, b_t)$ solve

$$\mathrm{d}\Theta_t = -\nabla_\theta (V_{\nu_t} + \lambda U_0)(\Theta_t) \,\mathrm{d}t + \sqrt{2\lambda} \,\mathrm{d}B_t + \,\mathrm{d}K_t, \qquad \nu_t = \mathrm{Law}(\Theta_t), \qquad \Theta_0 \sim \nu_0. \tag{78}$$

Here B_t is Brownian motion in $\mathbb{R}^{d+2}$, independent of Θ_0. The finite-variation process K_t gives inward normal reflection only on the two faces $a = \pm A$. There is no reflection in the w- or b-coordinates.

Proposition E.2 (Long-time MFLD and RERM). *There is a unique global reflected McKean–Vlasov solution to* (78). *Put*

$$\Omega_N = 4A(A + P_N|Y|), \qquad \rho_0 = \min\left\{ \frac{\pi}{2A^2}, \tau_w^{-2}, \tau_b^{-2} \right\}, \qquad \rho_{N,\lambda} = \rho_0 e^{-\Omega_N/\lambda}.$$

Then, with $G_t = J_{N,\lambda}(\nu_t) - J_{N,\lambda}(\widehat{\nu}_{N,\lambda})$,

$$G_t \le e^{-2\lambda \rho_{N,\lambda} t} G_0, \tag{79}$$

$$\mathrm{KL}(\nu_t \| \widehat{\nu}_{N,\lambda}) \le \lambda^{-1} e^{-2\lambda \rho_{N,\lambda} t} G_0, \tag{80}$$

$$\|f_{\nu_t} - \widehat{f}_{N,\lambda}\|_2^2 \le \frac{2A^2}{\lambda} e^{-2\lambda \rho_{N,\lambda} t} G_0. \tag{81}$$

In particular, the parameter laws converge in total variation, and the predictors converge in $L^2(\gamma_d)$, *to the unique noncompact RERM. This proposition is independent of the approximation and sampling bounds above.*

Proof. Well-posedness and confinement. For each fixed X_i, $\varphi_\theta(X_i)$ has a uniformly bounded parameter gradient and Hessian on the slab. Let L_i, H_i be respective bounds. Then

$$|f_\nu(X_i) - f_\mu(X_i)| \le L_i W_2(\nu,\mu), \qquad \|\nabla V_\nu - \nabla V_\mu\|_\infty \le \frac{2}{N}\sum_{i=1}^N L_i^2 W_2(\nu,\mu).$$

Also $\|\nabla V_\nu\|_\infty$ and $\|\nabla^2 V_\nu\|_\infty$ are bounded uniformly in ν by sample-dependent constants. The drift is thus globally Lipschitz in the parameter and Lipschitz in the law in W_2.

For a prescribed law path, the reflected SDE on the slab is obtained from the one-dimensional interval Skorokhod problem in a and the usual equations in (w, b). In synchronous coupling, normal reflection on a convex set satisfies $(\Theta_t - \Theta'_t)\cdot(\,\mathrm{d}K_t - \mathrm{d}K'_t) \le 0$. The preceding Lipschitz estimates and Gronwall's inequality make the law map contractive on a sufficiently short time interval. Picard iteration and concatenation give existence and uniqueness. Writing $z_t = (w_t, b_t)$, the bounded interaction drift and the positive quadratic confinement give, by Itô's formula and Young's inequality,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\|z_t\|^2] \le -c_\lambda \mathbb{E}[\|z_t\|^2] + C_{N,\lambda}.$$

The same computation for $\|z_t\|^4$ bounds fourth moments. The output coordinate is bounded, so explosion is impossible.

Energy identity on the noncompact slab. The law solves the Fokker–Planck equation with zero normal flux on $a = \pm A$. Its Wasserstein gradient-flow structure can be checked without a compactness assumption on (w, b). For any coupling Γ of two parameter laws, interpolate $\theta_u = (1-u)\theta + u\theta'$ and let μ_u be its law. Since $|f_{\mu_u}(X_i)| \le A$, differentiation of the squared loss gives

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2} P_N \ell_{f_{\mu_u}} \ge -D_N \int \|\theta' - \theta\|^2 \,\mathrm{d}\Gamma, \qquad D_N = \frac{2}{N}\sum_{i=1}^N (A + |Y_i|)H_i.$$

Relative entropy with respect to ν_0 is displacement convex on the convex slab: it is Lebesgue entropy plus the integral of the convex quadratic U_0, up to a constant. These observations verify semiconvexity of the objective and identify its first variation. The Wasserstein subdifferential and energy-dissipation theory on convex domains therefore apply (Ambrosio et al., 2008, Chapters 10–11). Uniqueness of the reflected equation identifies this gradient flow with the law constructed above.

More explicitly, if

$$\pi_\nu = Z_\nu^{-1} e^{-V_\nu/\lambda}\nu_0,$$

the first-variation gradient at ν_t equals $\lambda\nabla\log(\,\mathrm{d}\nu_t/\,\mathrm{d}\pi_{\nu_t})$. The energy identity is consequently

$$\frac{\mathrm{d}}{\mathrm{d}t} J_{N,\lambda}(\nu_t) = -\lambda^2 I(\nu_t\|\pi_{\nu_t}) \quad \text{for almost every } t > 0, \tag{82}$$

where $I(\eta\|\pi) = \int \|\nabla\log(\,\mathrm{d}\eta/\,\mathrm{d}\pi)\|^2\,\mathrm{d}\eta$. Equivalently, this identity follows by testing the positive-time Fokker–Planck equation with its first variation. The normal-flux term vanishes; truncations in (w, b) are removed using confinement and the energy identity for the gradient flow. Continuity of the energy at the finite-energy initial law extends the integrated identity to time zero.

A uniform frozen-Gibbs log-Sobolev bound. We use the convention $\mathrm{KL}(\eta\|\mu) \le I(\eta\|\mu)/(2\rho)$ (Gross, 1975). The map $t \mapsto 2A\Phi(t) - A$ pushes a standard Gaussian onto the uniform law on $[-A, A]$ and has Lipschitz constant $2A/\sqrt{2\pi}$. Hence the latter has a valid log-Sobolev constant $\pi/(2A^2)$. Tensorization with the two Gaussian factors gives ρ_0 for ν_0. For every ν,

$$\operatorname{osc} V_\nu \le 4A(A + P_N|Y|) = \Omega_N.$$

The bounded-perturbation inequality (Holley & Stroock, 1987) therefore gives, uniformly in ν,

$$\mathrm{KL}(\eta\|\pi_\nu) \le \frac{1}{2\rho_{N,\lambda}} I(\eta\|\pi_\nu).$$

From dissipation to prediction. Convexity of the risk and the Gibbs variational identity imply

$$0 \le J_{N,\lambda}(\nu) - J_{N,\lambda}(\widehat{\nu}_{N,\lambda}) \le \lambda\,\mathrm{KL}(\nu\|\pi_\nu). \tag{83}$$

Indeed, linearizing the risk at ν bounds the gap by $G_\nu(\nu) - G_\nu(\widehat{\nu}_{N,\lambda})$, which is at most $G_\nu(\nu) - G_\nu(\pi_\nu)$. Combining (82)–(83) gives $G_t' \le -2\lambda\rho_{N,\lambda}G_t$. Gronwall's inequality proves (79); (10) gives (80). This is the proximal-Gibbs convex-analysis argument (Nitanda et al., 2022). Finally, bounded neurons and Pinsker's inequality imply

$$|f_{\nu_t}(x) - \widehat{f}_{N,\lambda}(x)| \le A \int |\,\mathrm{d}\nu_t - \,\mathrm{d}\widehat{\nu}_{N,\lambda}| \le A\sqrt{2\,\mathrm{KL}(\nu_t\|\widehat{\nu}_{N,\lambda})}.$$

Squaring and integrating proves (81). □

The convergence rate may be exponentially small as $\lambda \downarrow 0$. In the statistical application, N and λ_N are fixed before $t \to \infty$ is taken. No interchange of these limits and no polynomial-time optimization guarantee are used.

For each fixed sample and positive temperature, (80) gives total-variation convergence of the hidden marginals. By Lem. D.16, predictor L^2 convergence also gives convergence of the smoothed-gradient matrices. On $\Omega_{N,h,E}$ for sufficiently large N, the spectral gap around the threshold u_N makes the corresponding projectors continuous at the limit. Here the OU smoothing time remains one while the MFLD training time tends to infinity.

E.3 Divergence of projected input-weight moments

Corollary E.3 (Divergence of projected input-weight moments)**.** *Under the assumptions of Thm. 5.1, fix an integer $q \ge 1$ with $2q \ge r_+$. Write $\widehat{\nu}_N(\,\mathrm{d}a, \mathrm{d}w, \mathrm{d}b) = \widehat{q}_N(\,\mathrm{d}a \mid w, b)\,\widehat{\phi}_N(\mathrm{d}w, \mathrm{d}b)$ and $\overline{a}_N(w,b) := \int_{-A}^{A} a\,\widehat{q}_N(\,\mathrm{d}a \mid w, b)$ and define $\mathfrak{M}_{q,N}(E) := \int |\overline{a}_N(w,b)|\,\|E^\top w\|_2^{2q}\,\mathrm{d}\widehat{\phi}_N(w,b)$. For every fixed $0 < R < \infty$, there exist constants $C > 0$ and $N_0 = N_0(q,R)$ such that, for all $N \ge N_0$,*

$$\sup_{\substack{h\in\mathcal{G}\\ E^\top E = I_{m^\star}}} \mathbb{P}_{h,E}\{\mathfrak{M}_{q,N}(E) \le R\} \le CN^{-8}.$$

Consequently, both

$$\mathfrak{M}_{q,N}(E) \xrightarrow{\mathbb{P}} \infty \qquad \text{and} \qquad \int \|E^\top w\|_2^{2q}\,\mathrm{d}\widehat{\phi}_N(w,b) \xrightarrow{\mathbb{P}} \infty$$

uniformly over $h \in \mathcal{G}$ and $E^\top E = I_{m^\star}$.

Proof. Fix $R \ge 0$ and abbreviate $\widehat{f}_N = \widehat{f}^{\mathrm{nc}}_{N,\lambda_N}$. For an independent $X \sim \gamma_d$, let

$$\widehat{h}_{N,E}(z) := \mathbb{E}[\widehat{f}_N(X) \mid E^\top X = z, \mathcal{D}_N].$$

Gaussian conditioning gives

$$\widehat{h}_{N,E}(z) = \int \overline{a}_N(w,b)\,\mathbb{E}_G\big[\tanh\big((E^\top w)^\top z + b + \|(I_d - EE^\top)w\|_2 G\big)\big]\,\mathrm{d}\widehat{\phi}_N(w,b),$$

where $G \sim \mathcal{N}(0,1)$. Whenever $\mathfrak{M}_{q,N}(E) < \infty$, boundedness of the derivatives of $\tanh$ justifies differentiation under the integrals and yields, for $|\alpha| \le 2q$,

$$\|\partial^\alpha \widehat{h}_{N,E}\|_\infty \le C_q \int |\overline{a}_N(w,b)|\,\|E^\top w\|_2^{|\alpha|}\,\mathrm{d}\widehat{\phi}_N(w,b) \le C_q\{A + \mathfrak{M}_{q,N}(E)\}.$$

Here we used $t^j \le 1 + t^{2q}$ for $0 \le j \le 2q$ and $|\overline{a}_N| \le A$. Lem. A.1 therefore implies

$$\|\widehat{h}_{N,E}\|_{H^{2q}(\gamma_{m^\star})} \le C_{m^\star,q}\{A + \mathfrak{M}_{q,N}(E)\}. \tag{84}$$

The ball

$$\mathbb{B}_{q,R} := \big\{g \in H^{2q}(\gamma_{m^\star}) : \|g\|_{H^{2q}(\gamma_{m^\star})} \le C_{m^\star,q}(A+R)\big\}$$

is compact in $L^2(\gamma_{m^\star})$ by Hermite truncation. By Lem. C.5, $\mathcal{G}$ is also compact and contains no element of $H^{r_+}(\gamma_{m^\star})$. Since $2q \ge r_+$, these sets are disjoint, so

$$\eta_{q,R} := \mathrm{dist}_{L^2(\gamma_{m^\star})}(\mathcal{G}, \mathbb{B}_{q,R}) > 0.$$

On $\{\mathfrak{M}_{q,N}(E) \le R\}$, equation (84) and contraction of conditional expectation give

$$\eta_{q,R} \le \|\widehat{h}_{N,E} - h\|_{L^2(\gamma_{m^\star})} \le \|\widehat{f}_N - h(E^\top \cdot)\|_{L^2(\gamma_d)}.$$

Choose N_0 so that $2C_s N^{-p_{s,m^\star}} < \eta_{q,R}^2$ for $N \ge N_0$. On $\Omega_{N,h,E}$, the intrinsic transfer (61) bounds $\|\widehat{h}_{N,E} - h\|^2_{L^2(\gamma_{m^\star})}$ by $2S_N \le 2C_s N^{-p_{s,m^\star}}$. Equation (34) therefore gives the asserted probability estimate. The unweighted conclusion follows from

$$\mathfrak{M}_{q,N}(E) \le A \int \|E^\top w\|_2^{2q} \, \mathrm{d}\widehat{\phi}_N(w,b).$$

□

In the concrete regime $(m^\star, s, d) = (2, 3/2, 3)$, the family can be chosen with $r_+ = 4$, so $q = 2$ gives

$$\int |\overline{a}_N(w,b)| \, \|E_\star^\top w\|_2^4 \, \mathrm{d}\widehat{\phi}_N(w,b) \xrightarrow{\mathbb{P}} \infty.$$

Thus the projected input-weight norms of features used by the learned output weights cannot remain uniformly bounded within the two-dimensional informative subspace.

F Experimental Details

The two experiments visualize different aspects of the theory. In Fig. 1, we directly solve the entropy-regularized population problem in order to study the approximation properties of a fixed compact class while excluding finite-sample and finite-time training effects. This allows any residual error as the regularization parameter decreases to be interpreted in terms of the representational constraints of the fixed function class, rather than sampling or optimization effects. In contrast, Fig. 2 uses a finite-sample, finite-particle approximation of MFLD to visualize how the OU-smoothed gradients of the predictor align with the unknown index subspace during training.

The computations below were performed on an NVIDIA A100 GPU with 40 GB of memory using `float64` precision.

F.1 Compactification dichotomy

To isolate the represented-versus-misspecified distinction for a fixed compact class, we solve the entropy-regularized population problem directly, rather than introducing a finite training sample or a finite-time MFLD approximation. Specifically, we minimize $J_\lambda(\nu) = \|f_\nu - h\|^2_{L^2(\gamma_m)} + \lambda \, \mathrm{KL}(\nu \| \nu_0)$. This oracle population setting asks whether decreasing λ can remove the prediction error when sampling and finite-time training effects are absent. We use the compact $\tanh$ model with output-weight bound $A = 16$, input-weight radius $R = 0.25$, and bias bound $B = 4$, for $m \in \{1, 2\}$ and $\lambda \in \{10^{-2}, 10^{-3}, \ldots, 10^{-8}\}$. The reference measure has $a \sim \mathrm{Unif}[-16, 16]$, while its input-weight and bias marginals are compact pushforwards of Gaussian variables rather than uniform distributions on the constrained domains.

The represented target is generated by a bounded exponential tilt $d\nu_\eta / d\nu_0 \propto \exp\{\eta a r_m(w,b)\}$, and hence belongs to the same compact mixture class with finite relative entropy. For the rough target, we use a normalized 32-node Gauss–Jacobi approximation to a continuous OU-time mixture of product-sign functions. Here "rough" refers to the underlying continuous mixture; its finite quadrature approximation should be distinguished from that mixture at the level of exact regularity.

We integrate over the output weight analytically and solve the resulting dual population problem by warm-started Newton iterations with Armijo line search, proceeding from larger to smaller λ. The plotted ordinate is the squared L^2 error evaluated on an independent higher-order Gauss–Hermite grid. The center curves are deterministic. The bands are paired 95% bootstrap intervals for the mean of 25 independent evaluation replicates, each using 200,000 draws from the discrete evaluation measure, with evaluation samples shared across targets and λ within each dimension. They therefore quantify evaluation variability for the fixed solutions, not training-sample, optimization, or quadrature uncertainty. The observed plateau is consistent with the compactification barrier, but is not a certified approximation-error lower bound or a finite-time MFLD convergence result.

F.2 OU-GRADIENT CONCENTRATION NEAR THE INDEX SUBSPACE

For Fig. 2, we use $(d, m^\star) = (3, 2)$ and

$$X \sim \mathcal{N}(0, I_3), \qquad Y = h_{\exp}(E_\star^\top X) + \xi, \qquad h_{\exp}(z) = 0.25\,\chi(z_1)\chi(z_2)|z_1|^{11/4},$$

where $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ is independent of X and

$$\chi(t) = \begin{cases} 1, & |t| \le 1, \\ \mathrm{sigmoid}\big((t^2-1)^{-1} - (4-t^2)^{-1}\big), & 1 < |t| < 2, \\ 0, & |t| \ge 2, \end{cases} \qquad \mathrm{sigmoid}(r) = (1+e^{-r})^{-1}.$$

We fix $E_\star$ by QR orthonormalization of a 3×2 standard Gaussian matrix generated with NumPy seed 271828, choosing nonnegative diagonal entries of R. We set $\sigma_\xi = 0.02865722146807541$, which is 0.2 times the target standard deviation estimated from 100,000 independent inputs. Neither $E_\star$ nor the target is supplied to the learner.

We use $M_{\mathrm{p}} = 512$ particles with $f_\theta(x) = \frac{1}{M_{\mathrm{p}}} \sum_{j=1}^{M_{\mathrm{p}}} a_j \tanh(w_j^\top x + b_j)$, initialized from $\mathrm{Unif}([-16, 16]) \otimes \mathcal{N}(0, I_3) \otimes \mathcal{N}(0, 1)$. For $L_N = N^{-1} \sum_{i=1}^N (f_\theta(X_i) - Y_i)^2$, $\lambda_N = N^{-8/11}$, and $\eta = 10^{-3}$, we apply

$$\theta_j^{k+1} = \Pi_a \Big[\theta_j^k - \eta \left(M_{\mathrm{p}} \nabla_{\theta_j} L_N + \lambda_N \nabla V(\theta_j^k)\right) + \sqrt{2\lambda_N \eta}\, \varepsilon_j^k\Big], \qquad V(a, w, b) = \tfrac{1}{2}(\|w\|^2 + b^2),$$

where ε_j^k are independent standard Gaussian vectors and Π_a clips only a to $[-16, 16]$. We use 20,000 updates ($T = 20$). For each $N \in \{100, 1000, 10000, 100000\}$, we run five replicates and display the prespecified replicate 0. Samples are nested across N, but initial particles and Langevin noise are independent, so the panels are not a single training trajectory. The $N = 0$ panel is the initial predictor from the $N = 100$, replicate-0 run. The base seed is 20260916, with separate streams for data, initialization, Langevin noise, and evaluation.

With OU smoothing time fixed to one, we evaluate at 8192 shared Gaussian test inputs

$$v_{N,T}(x) = \nabla T_1^{(3)} f_{N,T}(x), \qquad u_{N,T}(x) = \frac{v_{N,T}(x)}{\|v_{N,T}(x)\|}.$$

The one-dimensional expectation in the particle analogue of (48) is computed by 64-point Gauss–Hermite quadrature. Writing $E_\star = [e_{\star,1}, e_{\star,2}]$ and $n_\star = e_{\star,1} \times e_{\star,2}$, we plot each direction u in coordinates $((E_\star^\top u)_1, (E_\star^\top u)_2, n_\star^\top u)$. Marker area is proportional to gradient energy within each panel, and color shows $|n_\star^\top u|$. All panels use common axes, viewing angle, and color scale, without projection, subsampling, or artificial antipodal copies.

Figure 2 provides qualitative empirical support for the predictor-level subspace alignment predicted by our theory: concentration of OU-smoothed predictor gradients near the unknown index subspace is already visible with 512 particles and finite training time $T = 20$. This observation suggests that the geometric effect identified in our analysis also emerges in a discretized particle implementation.

We also quantify the alignment in Fig. 2 using averages over the five independent replicates. For the OU-smoothed predictor, define the perpendicular energy ratio

$$q_{N,T} := \frac{\mathbb{E}\big[\|(I - P_\star) v_{N,T}(X)\|^2\big]}{\mathbb{E}[\|v_{N,T}(X)\|^2]},$$

which corresponds to the effective directional leakage in (52). We also form the top-two spectral projector $\widehat{P}_{N,T}^{(2)}$ from the second-moment matrix of $v_{N,T}$ and measure its error by $\|\widehat{P}_{N,T}^{(2)} - P_\star\|_{\mathrm{F}}^2$.

The predictor-level alignment becomes sharp much faster than the particle-level directional concentration. The perpendicular OU-gradient energy decreases from 16.7% at $N = 100$ to 0.122% at $N = 100{,}000$, while the squared projector error decreases from 0.929 to 1.48×10^{-4}. Thus, in this finite-particle, finite-time experiment, strong directional alignment of the learned predictor is observed without near-complete concentration of the individual parameter directions.

N	prediction MSE	perp. energy $q_{N,T}$	projector error
100	0.01256	16.711%	0.9293
1,000	0.00532	1.067%	0.02284
10,000	0.00376	0.273%	0.00291
100,000	0.00360	0.122%	0.000148

Table 1: Prediction MSE is evaluated against the noiseless target on 100,000 independent test inputs. The panels in Fig. 2, and any numerical annotations therein, correspond instead to the pre-specified replicate 0.