LEARNING SPHERICAL POLYNOMIALS BY LEARNABLE CHANNEL ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of learning a spherical polynomial of an unknown degree $\ell_0 = \Theta(1) \geq 1$ by training a ReLU network with channel attention by gradient descent (GD) with guarantees for a finite training sample and finite network width. The target functions are general spherical polynomials with bounded L^2 norm and a fixed, unknown degree bound $\ell_0 \geq 1$. The training data are n i.i.d. uniformly distributed inputs on the unit sphere in $\mathbb{R}^d$ with i.i.d. centered noise of finite variance in their responses. The network width m is the number of ReLU channels and the size of the diagonal channel-attention matrix. The training process of the network has two stages. In stage one, one ordinary GD step learns channel attention parameters and strictly enlarges the prediction space on the first half of the training data. In stage two, finitely many output GD steps on the second half of the training data achieve expected regression risk $\mathcal{O}((\log d)^{-2}) \to 0$ as $d \to \infty$. The feature kernel of the ReLU attention network evolves during training and the learnable channel attention provably reduces the regression risk, demonstrating the provable benefit of feature learning. With $n \asymp d^{\ell_0}(\log d)^4$ and network width $m \asymp d^{\ell_0}(\log d)^2$, the sample size and network width m are only logarithmic factors away from their corresponding lower bounds for learning degree-ℓ_0 spherical polynomials. Under sub-Gaussian noise, the same algorithm also satisfies a high-probability risk bound with logarithmic confidence dependence. To the best of our knowledge, this is the first guarantee of vanishing regression risk for a finite ReLU network trained by GD with learnable channel attention, uniformly over bounded-L^2 spherical polynomials of degree at most ℓ_0, under these sample-size and width budgets.

1 INTRODUCTION

In this paper, we study a finite ReLU network with learnable channel attention to learn a target function with noise of finite variance in the response. The target function is a spherical polynomial supported in the unit sphere in $\mathbb{R}^d$, $\mathbb{S}^{d-1}$, with an unknown degree $\ell_0 = \Theta(1) \geq 1$ and bounded L^2 norm under the uniform distribution on $\mathbb{S}^{d-1}$. We observe n independent inputs from this distribution with i.i.d. centered noise of bounded variance, independent of the inputs. Regression risk is the expected mean squared error against the noiseless target. Expected risk also averages over training data and initialization. The width m is the number of biased ReLU channels and the size of the diagonal channel-attention matrix. The ReLU directions remain fixed throughout training. Section 2 specifies the model.

Let r_0 be the dimension of spherical polynomials of degree at most ℓ_0. For fixed $\ell_0 = \Theta(1)$, we have $r_0 \asymp d^{\ell_0}$. Under independent Gaussian noise of variance $\sigma_0^2 > 0$, uniform expected regression risk at most $\varepsilon \in (0, 3/128)$ on the polynomial unit L^2 ball requires $n \geq \Theta(\sigma_0^2 d^{\ell_0}/\varepsilon)$ (Proposition H.1, using minimax arguments (Raskutti et al., 2012; Zhang et al., 2015)). Consequently, uniform consistency requires $n/d^{\ell_0} \to \infty$ for fixed positive noise variance. Even with unlimited data, m square-integrable fixed scalar features with a target-independent joint law require $m \geq \Theta((1-\varepsilon)d^{\ell_0})$ for the same uniform guarantee (Proposition H.2, based on Kamath et al. (2020)). Here the Θ terms give the orders of the necessary lower bounds, with positive constants depending only on the fixed degree.

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.

These results motivate sample n and width m budgets of the order $\widetilde{\Omega}(d^{\ell_0})$, where the tilde suppresses logarithmic factors in terms of $\log d$. We ask the following question.

Question. Under our observation and target assumptions, can a shallow ReLU network learn new prediction functions and achieve vanishing expected regression risk over the full polynomial class, without being given ℓ_0, when both the sample size n and the network width m exceed d^{ℓ_0} only by fixed positive powers of $\log d$?

We answer this question affirmatively with a two-stage training algorithm by ordinary gradient descent (GD) on a carefully designed ReLU neural network with learnable channel attention and the network width is m. For $n \asymp d^{\ell_0}(\log d)^4$ and $m \asymp d^{\ell_0}(\log d)^2$, Theorem 4.1 gives expected risk $O((\log d)^{-2}) \to 0$ as $d \to \infty$. Both resources exceed the polynomial dimension d^{ℓ_0} only by logarithmic factors in terms of $\log d$. At a target risk tolerance $\varepsilon \asymp (\log d)^{-2}$, Proposition H.1 gives a sample lower bound of order $d^{\ell_0}(\log d)^2$ for fixed positive Gaussian noise variance. Our sample budget only exceeds this bound by a factor of order $(\log d)^2$. At the same tolerance, the width lower bound for deterministic or random feature maps with a target-independent law has order d^{ℓ_0} (Proposition H.2), so our channel budget only exceeds this lower bound by a factor of order $(\log d)^2$.

Contributions. To the best of our knowledge, our method establishes the first guarantee that a finite ReLU network trained by GD with the feature learning capability through learnable channel attention achieves vanishing regression risk uniformly over the full bounded-L^2 class of spherical polynomials of degree at most ℓ_0, at these simultaneous training sample size n and network width m budgets. The training algorithm requires finitely many ordinary GD steps, without needing any oracle information such as the unknown degree ℓ_0 and the noise variance.

Feature Learning Capability of Our Method. The training of the ReLU attention network has two stages. Learnable channel attention in stage one preserves the initial features and creates additional features in the same network. In our network design, each biased ReLU feature is split between two input-dependent weighted copies whose weights sum to one. Equal output coefficients recover the original feature, while different coefficients make its product with the learned gate available for fitting the target function.

Proposition D.1 shows that one attention GD step in training stage one on the first half of the training data strictly enlarges the prediction space and changes the kernel used in stage two. Proposition 4.2 in Section 4.2 establishes a statistical benefit of that same update: with probability tending to one in our sample regime, it strictly reduces the regression risk. These results demonstrate the provable benefit of the feature learning capability of learnable attention, where the benefit stems from learning from the response, with evolving feature kernels across the two training stages.

Furthermore, Proposition 4.2 gives a positive quantitative decrease with high probability and a positive expected decrease for sufficiently large dimension under our resource budgets. It covers the full bounded-L^2 polynomial class under finite-variance noise, without imposing an additional condition on the target function's directions or coefficients. The update uses neither the unknown degree nor the noise variance. Stage two then resets and fits the output coefficients on the independent second half of the training data, retaining the original features and using the learned ones after stage one. Theorem 4.1 controls both the expected risk and the regression risk of that final predictor. Sections 3 and 4.1 give the details about the two training stages and their associated kernels.

1.1 Related Work

For spherical polynomial targets, Ghorbani et al. (2021); Mei et al. (2022); Hu et al. (2024) analyze random-feature, linearized-network, and kernel regression in complementary regimes. Random-feature approximation and early-stopped GD also have finite-sample guarantees (Rudi & Rosasco, 2017; Carratino et al., 2018; Raskutti et al., 2014). Our result combines response-dependent enlargement of finite ReLU features with simultaneous logarithmic sample and width overheads and expected-risk control under finite-variance noise. Table 1 compares the closest guarantees under their respective assumptions. Feature-learning results for polynomials of a few unknown projections (Damian et al., 2022; Lee et al., 2024) exploit additional target structure, whereas our guarantee covers the full bounded-L^2 spherical polynomial class. Related attention and projected-GD methods (Yang, 2026; Yang & Li, 2026) use degree selection or constrained training under different

target and noise conditions. Our method learns attention on biased ReLU channels without degree selection. Appendix G.1 gives the detailed literature discussion, including the architectural context and the distinctions in target classes, training procedures, and statistical guarantees.

This paper is organized as follows. Sections 2–4 give the network model, its training algorithm, and main results. The appendices contain the proofs and the empirical comparison with a network without attention in Appendix I.

Notations. Bold symbols denote vectors and matrices. The arrow distinguishes sampled inputs $\vec{\mathbf{x}}_i$ and directions $\vec{\mathbf{q}}_j$ from generic vectors. We use $\|\cdot\|_2$ for the Euclidean or spectral norm and $\mathrm{tr}\,(\cdot)$ for the trace. The symbols $\mathbf{I}_k$, $\mathbf{1}_k$, and $\mathbf{0}$ denote an identity matrix, a vector of ones, and a zero vector or matrix. We write $[k] \coloneqq \{1, \ldots, k\}$ and $\mathbb{I}_{\{E\}}$ for an indicator. For real symmetric matrices, $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semidefinite. Matrix functions act on eigenvalues in an orthonormal eigenbasis. A positive semidefinite kernel has positive semidefinite finite evaluation matrices. For positive sequences, $a_d \sim b_d$ means $a_d/b_d \to 1$, and $a_d \asymp b_d$, or $a_d = \Theta(b_d)$, means that this ratio stays bounded above and away from zero. Tildes suppress logarithmic factors. Limits take $d \to \infty$ with $n = n(d), m = m(d)$. Constants may change between occurrences, with subscripts indicating fixed dependence. In particular, C_L, c_L, d_L depend only on the analysis degree bound L, unless stated otherwise. The relations $X_d = O_{\mathbb{P}}(a_d)$ and $X_d = o_{\mathbb{P}}(a_d)$ mean that X_d/a_d is bounded in probability and converges to zero in probability, respectively. Throughout the paper, we denote the inner product and the norm in L^2 space as $\langle f, h\rangle_{L^2} = \int f h \mathrm{d}\mu$ and $\|f\|_{L^2}^2 = \int f^2 \mathrm{d}\mu$.

1.2 Comparison with Existing Regression Guarantees

Table 1 compares existing relevant guarantees for learning spherical polynomials to our Theorem 4.1. We write $q_n \coloneqq n/d^{\ell_0}$, and let m_{lit} denote a source's neuron or random-feature count. The first three baselines admit the full bounded-L^2 polynomial class, without a restriction to a few directions of the target polynomials or a minimum coefficient size. Our contribution on this class combines finite learned ReLU features, a quantitative population-risk decrease from the attention step, and finite-step GD with explicit logarithmic sample and width overheads. The final predictor has both an unconditional expected-risk bound under finite-variance noise and an additional high-probability bound under sub-Gaussian noise.

The exact-kernel result of Ghorbani et al. (2021, Theorem 4) allows a logarithmic sample overhead, but has infinite feature width. Their finite-width approximation theorem instead assumes population fitting, corresponding to unlimited data. These two results separately control sample size and width. Our theorem controls both finite resources in the same trained network. Mei et al. (2022, Theorem 2) already give a joint finite-sample random-feature guarantee for the same full polynomial class. Their stated regime requires both sample and feature counts to exceed d^{ℓ_0} by a fixed positive power of d, together with a power separation between those counts. Our logarithmic overheads are smaller than these sufficient budgets.

The consistency result from Montanari & Zhong (2022, Theorem 3.3) requires additional spectral and asymptotic conditions beyond their sample and width inequalities. Their theorem compares tangent regression with polynomial ridge regression, whose risk $\mathcal{R}_{\mathrm{poly}}$ must also vanish. Let $\tau_d > 0$ be its effective penalty for the unnormalized sum of squared errors, and let $\rho_{\min,d} > 0$ be its smallest population-kernel eigenvalue over degrees $0, \ldots, \ell_0$. With $r_0 \asymp d^{\ell_0}$ and fixed source remainder exponents $C, C' > 0$, their required sufficient conditions are

$$\frac{r_0 \log(r_0+1)}{n} \to 0, \qquad \frac{\tau_d}{n\rho_{\min,d}} \to 0, \qquad \frac{n[\log(m_{\mathrm{lit}} d)]^{C'}}{m_{\mathrm{lit}} d} \to 0, \qquad \frac{n(\log n)^C}{d^{\ell_0+1}} \to 0. \quad (1)$$

The first two conditions in (1) make $\mathcal{R}_{\mathrm{poly}} \to 0$ in probability by Proposition G.4. The last two make the tangent-model approximation remainders vanish. The source's activation assumptions alone do not ensure positive eigenvalues in every target degree. Table 2 and Appendix G.5 give the full sample, width, activation, and consistency conditions required by Montanari & Zhong (2022). In contrast, our Theorem 4.1, which provides expected-risk and high-probability regression risk bounds, does not require such consistency conditions in Montanari & Zhong (2022).

Yang (2026, Theorems 1–2) also learns channel attention, but on an activation constructed from spherical harmonics, with a supplied degree cap L and a threshold for selecting the target de-

Table 1: Comparison between spherical-polynomial regression guarantees. All limits take $d \to \infty$ with fixed degree and target-norm and noise bounds. The fixed exponent $\delta_0 \in (0, 1/2)$ describes the polynomial separation in the random-feature result, and $c_0, c_1 > 0$ are sufficiently large source constants for tangent features required by Montanari & Zhong (2022). Their linear-degree lower sample bound uses a fixed constant $b_0 > 0$. (Yang, 2026)'s $L \geq \ell_0$ is a supplied upper bound for the true degree ℓ_0. Their result also imposes the stronger target restrictions discussed below. Table 2 in the appendix gives all source assumptions and resource ranges. For our high-probability bound, $t_\delta := \log(8/\delta)$ and $0 < \delta < 1$. Our guarantee combines learned finite features and explicit logarithmic resource (n, m) overheads, and uses a much smaller network width than the harmonic-attention guarantee (Yang, 2026, Theorems 1–2).

Method	Sample ratio q_n	Width ratio	Regression risk
Ghorbani et al. (2021, Theorem 4), exact kernel	$q_n/\log d \to \infty$	$m_{\mathrm{lit}}/d^{\ell_0} = \infty$	In probability: $o_{\mathbb{P}}(1)$
Mei et al. (2022, Theorem 2), fixed random features	$q_n \geq d^{\delta_0}$	$m_{\mathrm{lit}}/d^{\ell_0} \geq d^{\delta_0}$	In probability: $o_{\mathbb{P}}(1)$
Montanari & Zhong (2022, Theorem 3.3), fixed tangent features	$q_n \geq (\log d)^{c_0}$ if $\ell_0 > 1$ $q_n \geq b_0 > 0$ if $\ell_0 = 1$	$m_{\mathrm{lit}}/d^{\ell_0} \geq (q_n/d)[\log(m_{\mathrm{lit}} d)]^{c_1}$	In probability: $o_{\mathbb{P}}(1)$ when (1) and the source conditions hold
Yang (2026, Theorems 1–2), harmonic attention	$q_n \geq c d^{L+\ell_0} \log d$, with full selection conditions	$m_{\mathrm{lit}}/d^{\ell_0} \geq C d^{\ell_0} q_n^4 \log(2n/\delta)$	$O(q_n^{-1})$ with probability $\geq 1 - 2\delta - o(1)$, $0 < \delta < 1/2$
Two-stage training, Theorem 4.1	$q_n \asymp (\log d)^4$	$m/d^{\ell_0} \asymp (\log d)^2$	Expected risk: $O((\log d)^{-2})$. With probability $\geq 1 - \delta - 3n^{-8}$: $O(t_\delta (\log d)^{-2})$.

gree. Their targets satisfy a fixed weighted harmonic norm bound and a positive lower bound on every normalized coefficient. They assume sub-Gaussian noise and obtain the sharper risk bound $O(d^{\ell_0}/n)$, provided both selection and regression conditions hold. The latter require $m_{\mathrm{lit}} \geq C n^4 d^{-2\ell_0} \log(2n/\delta)$. Our sufficient width is instead $m \asymp n/(\log n)^2$ in the stated consistency regime. At $n \asymp d^{\ell_0}(\log d)^4$ and fixed δ, Yang's regression-width condition alone has order $d^{2\ell_0}(\log d)^{17}$, exceeding our width by a factor of order $d^{\ell_0}(\log d)^{15}$. This calculation compares width conditions only. Their selection conditions do not hold at our sample budget with fixed weighted radius: they also imply $n \geq c d^{L+2\ell_0} \log d$, as (163) shows. Their sharper rate therefore cannot be transferred to our budget by dropping those conditions. Our guarantee uses much smaller sufficient sample and width budgets, biased ReLU channels, and no coefficient lower bound or degree-selection threshold. Appendix G.4 retains the full resource calculations.

2 Problem Setup

We consider independent observations $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^n$ with inputs on $\mathcal{X} = \mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$, where $d \geq 3$. Let $P = \mu$ denote the uniform probability law on $\mathcal{X}$. The responses satisfy

$$y_i = f^*(\vec{\mathbf{x}}_i) + w_i, \qquad \vec{\mathbf{x}}_i \overset{i.i.d}{\sim} P, \qquad \mathbb{E}[w_i] = 0, \qquad \mathbb{E}\left[w_i^2\right] \leq \sigma_0^2, \qquad i \in [n]. \tag{2}$$

The noises in (2) are independent and identically distributed and independent of all inputs and initialization. The expected-risk results assume only finite variance, with no fourth-moment requirement. The additional high-probability regression bound assumes (15). The regression risk of the trained predictor $\widehat{f}$ is $\mathcal{R}(\widehat{f}) := \left\|\widehat{f} - f^*\right\|_{L^2}^2$.

For each integer $\ell \geq 0$, we let $Y_{\ell,j}$, $j \in [N(d,\ell)]$, be an orthonormal basis in $L^2(P)$ of degree-ℓ spherical harmonics, where $N(d,\ell)$ is that space's dimension and $Y_{0,1} = 1$. Each basis function is obtained by evaluating on $\mathcal{X}$ a polynomial on $\mathbb{R}^d$ whose monomials all have total degree ℓ and whose Euclidean Laplacian, $\sum_{i=1}^d \partial^2/\partial x_i^2$, is zero. Harmonics of different degrees are orthogonal

in $L^2(P)$. Let $1 \le \ell_0 \le L$ be an unknown integer degree bound, with L fixed in the analysis. For real coefficients $\beta_{\ell,j}$, define the following quantities, with sums over $0 \le \ell \le \ell_0$ and $j \in [N(d,\ell)]$.

$$f^* = \sum_{\ell=0}^{\ell_0} \sum_j \beta_{\ell,j} Y_{\ell,j}, \quad B_*^2 := \sum_{\ell,j} \beta_{\ell,j}^2, \quad \Gamma_*^2 := \sum_{\ell,j} N(d,\ell)\beta_{\ell,j}^2, \quad r_0 := \sum_{\ell=0}^{\ell_0} N(d,\ell). \tag{3}$$

In (3), $B_* = \|f^*\|_{L^2}$, Γ_*^2 is the weighted squared coefficient norm controlling approximation, and $r_0 \asymp d^{\ell_0}$ is the polynomial-space dimension. Since $N(d,\ell) \le r_0$ for $\ell \le \ell_0$, we have $\Gamma_*^2 \le r_0 B_*^2$. The coefficients may vary with d, and consistency is uniform over $B_* \le B$ for fixed B, allowing nonvanishing nonconstant target energy. Training uses neither L, ℓ_0, B_*, Γ_* nor σ_0. Appendices A.2 and G.2 compare target normalizations, and Appendix E.1 proves the dimension formula.

2.1 The ReLU Attention Network

For a deterministic positive integer width m divisible by four, we draw $m/4$ independent sample directions $\vec{\mathbf{q}}_j \sim P$, independently of the data, and set $\vec{\mathbf{q}}_{j+m/4} = -\vec{\mathbf{q}}_j$. We put $\sigma(t) := \max\{t, 0\}$ and $s(t) := (1 + \exp(-t))^{-1}$ as the ReLU function and sigmoid function, then define

$$\begin{aligned} \phi_j(\mathbf{x}) &:= \sigma(\sqrt{d}\, \vec{\mathbf{q}}_j^\top \mathbf{x} + 1/2), & \boldsymbol{\phi}(\mathbf{x}) &:= (\phi_1(\mathbf{x}), \dots, \phi_{m/2}(\mathbf{x}))^\top, \\ \mathbf{h}(\mathbf{x}) &:= [\boldsymbol{\phi}(\mathbf{x})^\top, \boldsymbol{\phi}(\mathbf{x})^\top]^\top, & \mathbf{z}(\mathbf{x}) &:= \mathbf{h}(\mathbf{x})/\sqrt{m}. \end{aligned} \tag{4}$$

The construction (4) first pairs each sampled direction with its negative and then duplicates the resulting feature vector, giving m channels. Opposite directions give a positive channel average, which prevents both candidate attention gradients from vanishing. Duplicated channels let complementary gates retain every initial predictor. Let $\boldsymbol{\tau} \in \mathbb{R}^m$ be the channel-attention parameters and $\boldsymbol{a} \in \mathbb{R}^m$ the output coefficients. The notation diag forms a block diagonal matrix from its arguments. We define the attention gate, attention matrix, and ReLu attention network prediction f by

$$\begin{aligned} g_{\boldsymbol{\tau}}(\mathbf{x}) := s\big(\sigma(1 + \boldsymbol{\tau}^\top \mathbf{z}(\mathbf{x})) - 1\big), \quad \mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x}) &:= \operatorname{diag}\big(g_{\boldsymbol{\tau}}(\mathbf{x})\mathbf{I}_{m/2}, (1 - g_{\boldsymbol{\tau}}(\mathbf{x}))\mathbf{I}_{m/2}\big) \\ f(\boldsymbol{\tau}, \boldsymbol{a}, \mathbf{x}) &:= \sqrt{2/m}\boldsymbol{a}^\top \mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x})\mathbf{h}(\mathbf{x}), \end{aligned} \tag{5}$$

where $\mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x}) \in \mathbb{R}^{m \times m}$ is the diagonal channel attention matrix. The network (5) combines fixed biased ReLU features, a scalar ReLU bottleneck with sigmoid attention, and a linear output. Its diagonal blocks share the two complementary gate values, so no dense matrix is trained. Each trainable vector has m coordinates, with attention depending only on the $m/2$ sums $\tau_j + \tau_{m/2+j}$. Appendices B and F.2 give computation costs and the role of the bias.

3 Two-Stage Gradient Descent

For $n \ge 4$, let $H_n := \log(en)$, $n_1 := \lfloor n/2 \rfloor$, $n_2 := n - n_1$, $I_1 := [n_1]$, and $I_2 := \{n_1 + 1, \dots, n\}$. I_1 and I_2 are the indices of the first training sample and second training sample as the disjoint subsets of the full training sample $\big\{(\vec{\mathbf{x}}_i, y_i)\big\}_{i=1}^n$. The default network width is $m = 4\lceil n/(4H_n^2) \rceil$, although any deterministic width satisfying (12) is permitted in our main result, Theorem 4.1.

Algorithm 1 describes the two-stage training algorithm which trains the attention parameters $\boldsymbol{\tau}$ by one GD step on the first training sample and then trains the output coefficients $\boldsymbol{a}$ by $T = 4(d+1)m$ full-batch GD steps on the second training sample. The sampled directions and biases remain fixed. All training parameters below depend only on the observations, sampled directions, n, and d. The first training sample is used to learn the attention weights. After the output reset, independent observations in the second training sample fit output coefficients on the learned features.

Training Stage One: Learning the Channel Attention Weights.

We define the channel average $b(\mathbf{x}) := (2/m)\sum_{j=1}^{m/2} \phi_j(\mathbf{x})$, which satisfies $b(\mathbf{x}) \ge 1/2$ by the opposite-direction pairing. We initialize $\boldsymbol{\tau} = \mathbf{0}$. For $c \in \{0, 1\}$, form $\boldsymbol{a}_{\text{init},c}$ by concatenating the two blocks $\boldsymbol{a}_{\text{init},c}^{\pm} = (c \pm 1)\sqrt{2/m}\mathbf{1}_{m/2}$. These choices give the initial predictors 0 and b, respectively. Holding these coefficients fixed, define $L_{1,c}(\boldsymbol{\tau}) := (2n_1)^{-1} \sum_{i \in I_1} (f(\boldsymbol{\tau}, \boldsymbol{a}_{\text{init},c}, \vec{\mathbf{x}}_i) -$

Algorithm 1. Two-stage GD with channel attention

Input. The full training sample in (2), with $n \geq 4$ and $d \geq 3$.
Setup. Form the paired channels in (4) using directions sampled independently of the data.

Stage 1. One attention GD step on the first training sample indexed by I_1.

1. Initialize $\boldsymbol{\tau} = \mathbf{0}$. For $c \in \{0, 1\}$, compute the candidate gradient $\mathbf{g}_c$ in (6). Select the larger norm, breaking ties with $c = 0$.
2. Use the selected gradient $\mathbf{g}$, curvature bound Λ_1, and learning rate η_1 in (7) to set $\widehat{\boldsymbol{\tau}} = -\eta_1 \mathbf{g}$.

Stage 2. T output GD steps on the second training sample indexed by I_2.

3. Freeze $\widehat{\boldsymbol{\tau}}$, form $\boldsymbol{\psi}(\mathbf{x}) = \sqrt{2/m}\mathbf{A}_{\widehat{\boldsymbol{\tau}}}(\mathbf{x})\mathbf{h}(\mathbf{x})$, and reset $\boldsymbol{a}(0) = \mathbf{0}$.
4. Set $\eta = [4(d+1)]^{-1}$, $T = 4(d+1)m$. Apply (8) for $t = 0, \ldots, T-1$.

Return. $\widehat{f}(\mathbf{x}) = \boldsymbol{a}(T)^\top \boldsymbol{\psi}(\mathbf{x})$.

$y_i)^2$. Its attention gradient at initialization is

$$\mathbf{g}_c := \nabla L_{1,c}(\mathbf{0}) = \frac{1}{2n_1} \sum_{i \in I_1} (c\, b(\vec{\mathbf{x}}_i) - y_i) b(\vec{\mathbf{x}}_i) \mathbf{z}(\vec{\mathbf{x}}_i), \qquad \widehat{c} \in \arg\max_{c \in \{0,1\}} \|\mathbf{g}_c\|_2. \tag{6}$$

The difference $\mathbf{g}_1 - \mathbf{g}_0 = (2n_1)^{-1} \sum_{i \in I_1} b(\vec{\mathbf{x}}_i)^2 \mathbf{z}(\vec{\mathbf{x}}_i)$ has positive coordinate sum. As a result, the two gradients in (6) cannot both vanish. We select the larger gradient norm, break ties by choosing $\widehat{c} = 0$, and set $\mathbf{g} = \mathbf{g}_{\widehat{c}}$ and $L_1 = L_{1,\widehat{c}}$. These two gradient evaluations select the temporary output initialization and do not update any parameter. We choose the learning rate and update attention by

$$\Lambda_1 := \frac{1}{n_1} \sum_{i \in I_1} \left(\frac{5}{4} b(\vec{\mathbf{x}}_i)^2 + \frac{|y_i|\, b(\vec{\mathbf{x}}_i)}{2} \right) \left\| \mathbf{z}(\vec{\mathbf{x}}_i) \right\|_2^2, \qquad \widehat{\boldsymbol{\tau}} := -\eta_1 \mathbf{g}, \tag{7}$$

where $\eta_1 := \left(1 + 2\sqrt{2(d+1)}\|\mathbf{g}\|_2 + \Lambda_1\right)^{-1}$. The update (7) is one ordinary GD step on L_1. The gradient-norm term keeps the bottleneck ReLU active on the sphere, and Λ_1 bounds loss curvature along the step. Proposition D.1 proves these properties, strict loss descent, and attention weights between $3/8$ and $5/8$.

Training Stage Two: Learning the Output Coefficients.

We freeze $\widehat{\boldsymbol{\tau}}$ and define the learned feature vector $\boldsymbol{\psi}(\mathbf{x}) := \sqrt{2/m}\mathbf{A}_{\widehat{\boldsymbol{\tau}}}(\mathbf{x})\mathbf{h}(\mathbf{x})$. Starting from $\boldsymbol{a}(0) = \mathbf{0}$, we apply full-batch GD to the output coefficients alone, using all observations in the second training sample, indexed by I_2, at every step. We refer to these updates as output GD for all $0 \leq t < T$:

$$\eta := [4(d+1)]^{-1},\ T := 4(d+1)m,\ \boldsymbol{a}(t+1) = \boldsymbol{a}(t) - \frac{\eta}{n_2} \sum_{i \in I_2} \left(\boldsymbol{a}(t)^\top \boldsymbol{\psi}(\vec{\mathbf{x}}_i) - y_i \right) \boldsymbol{\psi}(\vec{\mathbf{x}}_i). \tag{8}$$

We return $\widehat{f}(\mathbf{x}) = \boldsymbol{a}(T)^\top \boldsymbol{\psi}(\mathbf{x})$. The updates (8) do not increase the training loss in stage two. Conditional on the first training sample and directions, the learned map is fixed, while the second training sample retains the observation law (2). Algorithm 1 uses $1 + 4(d+1)m$ GD updates in total and no unknown degree or noise parameter. Appendix B gives computation costs, and Appendix B.1 gives an observed-feature step size with the same expected-risk guarantee and fewer iterations in conditional expectation.

4 Main Results

Our results establish feature learning and regression guarantees for the same algorithm. Theorem 4.1 proves strict feature enlargement and consistency of the final predictor trained by GD. Proposition 4.2 additionally proves that the attention update decreases population regression risk before the output coefficients are reset. The kernel identity below supplies the common representation.

4.1 The Learned Representation and Its Feature Kernel

The kernel identity explains how attention changes the functions available to output training. The learned representation retains the initial ReLU features and adds their products with the response-dependent gate. Retention preserves polynomial approximation, while the added products provide further functions for fitting residual error. Let $\boldsymbol{\psi}_0(\mathbf{x}) := \mathbf{h}(\mathbf{x})/\sqrt{2m}$ be the feature vector at $\boldsymbol{\tau} = \mathbf{0}$, let $a_g(\mathbf{x}) := g_{\widehat{\boldsymbol{\tau}}}(\mathbf{x}) - 1/2$ be the learned gate deviation, and define $K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}') := \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}')$. The initial feature kernel is $K_0(\mathbf{x}, \mathbf{x}') := \boldsymbol{\psi}_0(\mathbf{x})^\top \boldsymbol{\psi}_0(\mathbf{x}') = \frac{1}{2} K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}')$. The learned feature kernel is

$$\widehat{K}(\mathbf{x}, \mathbf{x}') := \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}') = K_0(\mathbf{x}, \mathbf{x}') + 2a_g(\mathbf{x}) a_g(\mathbf{x}') K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}'). \tag{9}$$

Because $\nabla_{\boldsymbol{a}} f(\widehat{\boldsymbol{\tau}}, \boldsymbol{a}, \mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})$, $\widehat{K}$ in (9) is the exact kernel governing output GD in (8). Stage one changes K_0 to $\widehat{K}$, and stage two uses this fixed learned feature kernel. The increment is positive semidefinite as a kernel, although its individual entries may be negative. An orthogonal change of coordinates, taking the sum and difference of the two channel blocks divided by $\sqrt{2}$, gives

$$\boldsymbol{\psi}_+(\mathbf{x}) := \boldsymbol{\phi}(\mathbf{x})/\sqrt{m}, \boldsymbol{\psi}_-(\mathbf{x}) := 2a_g(\mathbf{x})\boldsymbol{\psi}_+(\mathbf{x}), \boldsymbol{\psi}_0 \longmapsto [\boldsymbol{\psi}_+^\top, \mathbf{0}^\top]^\top, \boldsymbol{\psi} \longmapsto [\boldsymbol{\psi}_+^\top, \boldsymbol{\psi}_-^\top]^\top. \tag{10}$$

In (10), an initial predictor $\mathbf{u}^\top \boldsymbol{\psi}_+$ is retained by coefficients $[\mathbf{u}^\top, \mathbf{0}^\top]^\top$ of the same norm, for $\mathbf{u} \in \mathbb{R}^{m/2}$. For initial coefficient blocks $\boldsymbol{a}^\pm \in \mathbb{R}^{m/2}$, take $\mathbf{u} = (\boldsymbol{a}^+ + \boldsymbol{a}^-)/\sqrt{2}$, with $\|\mathbf{u}\|_2^2 \le \|\boldsymbol{a}^+\|_2^2 + \|\boldsymbol{a}^-\|_2^2$. Separate coefficients therefore let output GD retain an initial predictor and fit additional products with the learned gate through $\boldsymbol{\psi}_-$. The coordinate transformation is orthogonal, although the two function blocks need not be orthogonal in $L^2(P)$.

We define $\mathcal{V}_0 := \operatorname{span}\{\phi_j : j \in [m/2]\}$ and $\mathcal{V}_1 := \operatorname{span}\{\psi_j : j \in [m]\}$ for the initial and learned spaces of functions, where ψ_j is the jth coordinate of $\boldsymbol{\psi}$. Proposition D.1 proves $\mathcal{V}_0 \subsetneq \mathcal{V}_1$ by showing $a_g b \in \mathcal{V}_1 \setminus \mathcal{V}_0$. It also proves that $g_{\widehat{\boldsymbol{\tau}}}$ is nonconstant, so $\widehat{K}$ is not a constant multiple of K_0. The squared feature change obeys

$$0 < \int \|\boldsymbol{\psi}(\mathbf{x}) - \boldsymbol{\psi}_0(\mathbf{x})\|_2^2 \mathrm{d}\mu(\mathbf{x}) = \int \big(\widehat{K}(\mathbf{x}, \mathbf{x}) - K_0(\mathbf{x}, \mathbf{x})\big)\mathrm{d}\mu(\mathbf{x}) \le \frac{57}{1024(d+1)}. \tag{11}$$

Regression Guarantee by Theorem 4.1. The theorem combines a deterministic representation guarantee with a statistical prediction guarantee. Strict enlargement holds for every first-stage realization. The expected-risk bound averages over both training samples and initialization, and the additional high-probability regression risk bound assumes sub-Gaussian noise. The admissible width range ensures enough initial directions for approximation and enough observations for output fitting.

Theorem 4.1 (Feature learning and prediction consistency)**.** Fix an integer $L \ge 1$. There are constants $d_L \ge 3$ and $C_L > 0$, depending only on L, such that the following holds. Assume (2)–(3), $d \ge d_L$, and $n \ge 4$. Let m be a deterministic width divisible by four satisfying

$$512(d + H_n) \le m \le n/(512H_n). \tag{12}$$

Suppose the ReLU attention network refeq:summary-network is trained by Algorithm 1, with one attention GD step on the first training sample indexed by I_1 and $T = 4(d+1)m$ output GD steps on the second training sample indexed by I_2, using the learning rates in (7) and (8). For every finite first training sample and every realization of the directions, $\mathcal{V}_0 \subsetneq \mathcal{V}_1$, the training loss in stage one strictly decreases, and (9)–(11) hold. These deterministic conclusions require only $d \ge 3$, independently of (12). With expectation over the samples and initialization, the returned predictor satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le C_L \left(\frac{\Gamma_*^2}{m} + \frac{\sigma_0^2 m}{n} + \frac{B_*^2 + \sigma_0^2}{n^{7/2}}\right). \tag{13}$$

In particular, for fixed $1 \le \ell_0 \le L$, bounded B_*, σ_0, and $d \to \infty$, take $n \asymp d^{\ell_0}(\log d)^4$ and the default width $m = 4\lceil n/(4H_n^2)\rceil$. Then (12) holds eventually and

$$m \asymp d^{\ell_0}(\log d)^2, \qquad \mathbb{E}\left[\mathcal{R}(\widehat{f})\right] = O((\log d)^{-2}) \overset{d\to\infty}{\longrightarrow} 0. \tag{14}$$

The constant in (14) depends only on L, the fixed target and noise bounds, and the sample-budget comparison constants. The expectation bound is uniform over those classes and implies convergence in probability by Markov's inequality. Training uses neither the degree nor unknown target or noise parameters. Proposition B.2 in Appendix B.2 supplies a width satisfying both inequalities whenever their interval contains a multiple of four.

If, in addition, the noise is sub-Gaussian with variance proxy σ_0^2, meaning

$$\mathbb{E}\left[\exp(sw_i)\right] \le \exp(s^2\sigma_0^2/2), \qquad s \in \mathbb{R}, \quad i \in [n], \tag{15}$$

the same predictor has the following high-probability regression risk guarantee. Define the learned population covariance and its effective dimension by

$$\mathbf{\Sigma} := \int \boldsymbol{\psi}\boldsymbol{\psi}^\top \mathrm{d}\mu, \qquad \mathcal{N}_{\widehat{K}}(\lambda) := \mathrm{tr}\left(\mathbf{\Sigma}(\mathbf{\Sigma} + \lambda \mathbf{I}_m)^{-1}\right), \qquad \lambda > 0. \tag{16}$$

For every $\delta \in (0,1)$, put $t_\delta := \log(8/\delta)$. With probability at least $1 - \delta - 3n^{-8}$ over both training samples and initialization,

$$\mathcal{R}(\widehat{f}) \le C_L \left[\frac{\Gamma_*^2 t_\delta}{m} + \frac{\sigma_0^2}{n_2}\left(\mathcal{N}_{\widehat{K}}(m^{-1}) + t_\delta\right)\right] \le 2C_L \left[\frac{\Gamma_*^2 t_\delta}{m} + \frac{\sigma_0^2(m + t_\delta)}{n}\right]. \tag{17}$$

The constants and failure bound are uniform over the stated target and sub-Gaussian noise classes, with probability evaluated separately under each training law. Neither δ nor the noise proxy is supplied to training. At the budgets in (14), (17) is $O((\log d)^{-2})$ for fixed δ. For $\delta = (\log d)^{-a}$ with fixed $a > 0$, it is $O(\log\log d/(\log d)^2)$ with probability tending to one. Assumption (15) is needed only for this additional conclusion. The expected-risk bound (13) retains its finite-variance scope.

Proof roadmap. The proof proceeds through the following five steps. Appendix F.7 gives the complete argument, with the approximation and covariance estimates separated from the properties of the attention update.

Step 1. Feature enlargement and coefficient preservation. Proposition D.1 proves a nonzero descent step and $a_g b \notin \mathcal{V}_0$, with the strict-enlargement argument in Appendix D.2. The preserved block in (10) then allows Step 2 to use the initial approximation without increasing its coefficient norm.

Step 2. Polynomial approximation in the learned representation. For $\lambda > 0$, define the penalized approximation error of the learned features by

$$\mathcal{A}_\lambda(\widehat{K}) := \min_{\boldsymbol{a} \in \mathbb{R}^m} \left(\left\|\boldsymbol{a}^\top \boldsymbol{\psi} - f^*\right\|_{L^2}^2 + \lambda \|\boldsymbol{a}\|_2^2\right).$$

Define $\mathcal{A}_\lambda(K_0)$ using $\boldsymbol{\psi}_0$. The penalty measures coefficient size and makes the minimizers unique, while training remains unpenalized. Lemma F.1 shows that the chosen ReLU bias supplies every harmonic degree through L. Lemma F.2 then constructs an initial predictor with expected squared error at most $C_L\Gamma_*^2/m$ and expected squared coefficient norm at most $C_L\Gamma_*^2$. Step 1 retains both bounds, giving

$$\mathcal{A}_\lambda(\widehat{K}) \le \mathcal{A}_\lambda(K_0), \qquad \mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] \le C_L\Gamma_*^2/m. \tag{18}$$

The stopping rule gives $\eta T = m$, so $\lambda = m^{-1}$ is the optimization scale. On the second-moment event in Step 3, squared optimization error is bounded by a constant times the squared comparison-coefficient norm divided by m. As a result, coefficient preservation connects approximation to finite output GD.

Step 3. Regression bound for independent output training. Condition on the directions and first-stage outcome. Lemma F.5 bounds the feature size, and Lemma F.6 compares empirical and population second moments at scale m^{-1}, with combined failure probability at most $2n^{-8}$. Lemma F.9 controls finite output GD using Step 2's comparison predictor and the conditional noise variance. Appendices F.4 and F.5 give the calculations.

Step 4. Expected risk and the consistency regime. For $\lambda > 0$, define the empirical covariance, meaning the uncentered second-moment matrix, and its effective dimension by

$$\widehat{\mathbf{\Sigma}} := \frac{1}{n_2}\sum_{i \in I_2} \boldsymbol{\psi}(\vec{\mathbf{x}}_i)\boldsymbol{\psi}(\vec{\mathbf{x}}_i)^\top, \qquad \widehat{\mathcal{N}}(\lambda) := \mathrm{tr}\left(\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}} + \lambda \mathbf{I}_m)^{-1}\right). \tag{19}$$

Each eigenvalue $u \ge 0$ contributes $u/(u+\lambda)$ to (19), so $0 \le \widehat{\mathcal{N}}(\lambda) \le m$. At $\lambda = (\eta T)^{-1} = m^{-1}$, this effective dimension measures the noise cost in Step 3. Averaging over the independent inputs in the second training sample turns the comparison predictor's empirical squared error into its population error. Lemma F.8 controls the remaining outcomes with only finite noise variance,

yielding

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le 12\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] + \frac{12\sigma_0^2}{n}\mathbb{E}\left[\widehat{\mathcal{N}}(m^{-1})\right] + \frac{C_L(B_*^2+\sigma_0^2)}{n^{7/2}}. \tag{20}$$

Equation (58) justifies this averaging without conditioning the comparison error on a favorable event. Combining (20) with (18) and $\widehat{\mathcal{N}}(m^{-1}) \le m$ proves (13). Finally, $\Gamma_*^2 \le r_0 B_*^2$ and $r_0/m, m/n = O((\log d)^{-2})$ give (14).

Step 5. High-probability risk under sub-Gaussian noise. The additional argument in Appendix F.7 truncates only the integral coefficients of an analytical comparison predictor, which attention preserves exactly. Concentration controls its population error, empirical error, and coefficient cost by $C_L\Gamma_*^2 t_\delta/m$. Local Rademacher complexity at the actual GD scale m^{-1} and a conditional quadratic-form bound control fitted noise through (16). The exact finite-GD decomposition then proves (17), without changing training or using Markov's inequality on the final risk.

Section 1.2 and Table 1 compare the target classes, resource conditions, and regression guarantees of the baselines.

4.2 Regression Risk Reduction Effect of the Learnable Channel Attention

The attention update in training stage one has a population benefit before output refitting in stage two: under the conditions below, it decreases population loss on fresh inputs with high probability while the selected temporary output coefficients remain fixed. For $c \in \{0,1\}$, we define $R_{1,c}(\boldsymbol{\tau}) := \|f(\boldsymbol{\tau}, \boldsymbol{a}_{\mathrm{init},c}, \cdot) - f^*\|_{L^2}^2/2$ as the population loss with the attention parameter τ, and put $A_* := 1 + B_*^2 + \sigma_0^2$. Appendix D.3 proves the detailed version, Proposition D.2.

Proposition 4.2 (Population attention descent)**.** Under the observation and target assumptions above, let $\widehat{\boldsymbol{\tau}}$ be obtained after the single attention GD step in stage one of Algorithm 1, before the output coefficients are reset. There are constants $C_L \ge 1, c_L > 0$ such that, for every paired direction collection and $d \ge \max\{3, C_L A_*\}$, with probability at least $1 - C_L A_*/n_1$ over the first training sample,

$$R_{1,\widehat{c}}(\mathbf{0}) - R_{1,\widehat{c}}(\widehat{\boldsymbol{\tau}}) \ge \tfrac{1}{2}\eta_1\|\mathbf{g}\|_2^2 \ge \frac{c_L}{\sqrt{A_*(d+1)}}. \tag{21}$$

The constants are not supplied to training. The expected improvement is also positive eventually for bounded A_* and $n_1/\sqrt{d} \to \infty$.

Proof roadmap. First, channel moments and polynomial hypercontractivity control both candidate gradient errors using only the noise variance. Simultaneous concentration permits the selection of $\widehat{c}$. Second, gradient alignment and a population Taylor bound give (21). Finally, a uniform loss bound controls exceptional outcomes. Appendix D.3 proves these steps. Both losses use the same selected temporary outputs, which are subsequently reset.

4.3 Simulation Results

In Appendix I, we provide simulation results comparing the ReLU attention network (5) with the same network without channel attention on a cubic spherical polynomial of unit L^2 norm. Figure 1 reports test mean squared error averaged over 10 random initializations. The ReLU attention network has lower test error at every tested dimension, illustrating a prediction benefit of learnable channel attention in this experiment.

5 Conclusion

We establish a finite-network guarantee that combines response-dependent feature learning with consistent regression over the full bounded-L^2 class of spherical polynomials of unknown degree. One GD step learns channel attention that strictly enlarges the prediction space while preserving the original features. Subsequent output GD achieves vanishing expected regression risk with sample size and network width exceeding the polynomial-space dimension by only logarithmic factors. The guarantee holds under finite-variance noise and strengthens to a high-probability bound under sub-Gaussian noise. These results connect learnable attention, finite-step optimization, and uniform statistical guarantees within a single finite ReLU network.

AI USE STATEMENT

In this work, we used generative AI tools for interpreting results. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, or designing or providing feedback on research methodology or experiments. Generating synthetic data sets, implementing methods, cleaning and reformatting datasets, assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for improving readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4782–4887. PMLR, 2022. URL https://proceedings.mlr.press/v178/abbe22a.html.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946. Curran Associates, Inc., 2022. doi: 10.52202/068431-2749. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f7e7fabd73b3df96c54a320862afcb78-Abstract-Conference.html.

Francis R. 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.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 12873–12884. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/c4ef9c39b300931b69a36fb3dbb8d60e-Abstract.html.

Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10213–10224. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/741a0099c9ac04c7bfc822caf7c7459f-Abstract.html.

Theodore S. Chihara. *An Introduction to Orthogonal Polynomials*. Dover Books on Mathematics. Dover Publications, 2011. ISBN 9780486479293. URL https://store.doverpublications.com/products/9780486479293.

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104630–104693. Curran Associates, Inc., 2024. doi: 10.52202/079017-3323. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bd18189308a4c45c7d71ca83acf3deaa-Abstract-Conference.html.

Costas Efthimiou and Christopher Frye. *Spherical Harmonics in p Dimensions*. World Scientific, 2014. doi: 10.1142/9134. URL https://doi.org/10.1142/9134.

Hengyu Fu, Zihao Wang, Eshaan Nichani, and Jason D. Lee. Learning hierarchical polynomials of multiple nonlinear features. In *International Conference on Learning Representations*, volume 2025, pp. 36166–36220, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/599221d7ebf6b3403190f38a3f282a1c-Abstract-Conference.html.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021. doi: 10.1214/20-aos1990. URL https://doi.org/10.1214/20-aos1990.

Hong Hu, Yue M. Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. arXiv:2403.08160, 2024. URL https://arxiv.org/abs/2403.08160.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 8580–8589. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2236–2262. PMLR, 2020. URL https://proceedings.mlr.press/v125/kamath20b.html.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, volume 37, pp. 58716–58756. Curran Associates, Inc., 2024. doi: 10.52202/079017-1872. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6bd5fca2074dcd9ede9de50f71f7ec28-Abstract-Conference.html.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008. URL https://doi.org/10.1002/cpa.22008.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. doi: 10.1016/j.acha.2021.12.003. URL https://doi.org/10.1016/j.acha.2021.12.003.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36106–36159. PMLR, 2024. URL https://proceedings.mlr.press/v235/moniri24a.html.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022. doi: 10.1214/22-AOS2211. URL https://doi.org/10.1214/22-AOS2211.

Ashley Montanaro. Some applications of hypercontractive inequalities in quantum information theory. *Journal of Mathematical Physics*, 53(12):122206, 2012. doi: 10.1063/1.4769269. URL https://doi.org/10.1063/1.4769269.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13 (13):389–427, 2012. URL https://jmlr.org/papers/v13/raskutti12a.html.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014. URL https://jmlr.org/papers/v15/raskutti14a.html.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3215–3225. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/61b1fb3f59e28c67f3925f3c79be81a1-Abstract.html.

Gábor Szegő. *Orthogonal Polynomials*, volume 23 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, fourth edition, 1975. ISBN 9780821810231. URL https://www.ams.org/books/coll/023/.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://doi.org/10.1007/s10208-011-9099-z.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 11534–11542. Computer Vision Foundation / IEEE, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_ECA-Net_Efficient_Channel_Attention_for_Deep_Convolutional_Neural_Networks_CVPR_2020_paper.html.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue M. Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4558–4570. Curran Associates, Inc., 2022. doi: 10.52202/068431-0329. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1d3591b6746204b332acb464b775d38d-Abstract-Conference.html.

Yingzhen Yang. Shallow neural networks learn low-degree spherical polynomials with feature learning by learnable channel attention. In *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–48. PMLR, 2026. URL https://proceedings.mlr.press/v313/yang26a.html.

Yingzhen Yang and Ping Li. Gradient descent finds over-parameterized neural networks with sharp generalization for nonparametric regression. arXiv:2411.02904, 2024. URL https://arxiv.org/abs/2411.02904v6. Version 6, revised November 6, 2025. First submitted November 5, 2024.

Yingzhen Yang and Ping Li. Gradient descent with projection finds over-parameterized neural networks for learning low-degree polynomials with nearly minimax optimal rate. arXiv:2603.21062, 2026. URL https://arxiv.org/abs/2603.21062.

Yunfei Yang and Ding-Xuan Zhou. Nonparametric regression using over-parameterized shallow ReLU neural networks. *Journal of Machine Learning Research*, 25(165):1–35, 2024. URL https://jmlr.org/papers/v25/23-0918.html.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2. URL https://doi.org/10.1007/s00365-006-0663-2.

Amir Zandieh, Insu Han, and Haim Avron. Near optimal reconstruction of spherical harmonic expansions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23968–23989. Curran Associates, Inc., 2023. doi: 10.52202/075280-1041. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4b719e74623f4fa238ded71b56f0a184-Abstract-Conference.html.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(102):3299–3340, 2015. ISSN 1532-4435. URL https://jmlr.org/papers/v16/zhang15d.html.

Guide to the appendix. The appendix separates the complete statements, their proof ingredients, and the comparisons used to interpret the guarantees. Appendices A–C specify the model and training and collect the full statements. Appendix D proves the effects of the attention step, Appendix E supplies the harmonic and concentration tools, and Appendix F establishes the statistical guarantees. Appendices G and H justify the benchmark comparisons and resource lower bounds. Appendix I reports the numerical experiments. The following proof paths cover each main-text result and its supporting assertions.

1. *Feature learning and prediction consistency, Theorem 4.1.* The complete statement is Theorem C.1 in Appendix C, and Appendix F.7 proves both formulations. Its deterministic part comes from Proposition D.1 in Appendix D.2: the computed attention step decreases the training loss in stage one, strictly enlarges the prediction space, and changes the kernel with the stated feature-energy bound. For the expected-risk bound, Lemmas F.1 and F.2 in Appendix F.1 construct the polynomial comparison predictor. Proposition F.4 in Appendix F.3 retains that predictor and controls its coefficient norm. Proposition F.10 in Appendix F.6 constructs measurable population comparison coefficients and bounds the preserved approximation cost. Lemmas F.5 and F.6 in Appendices F.3 and F.4 control the feature size and compare the empirical and population second moments. Appendix F.5 then separates finite-GD optimization error, approximation residual, and fitted noise, including exceptional outcomes under finite noise variance. Proposition C.2, proved in Appendix C.1, averages these conditional estimates with an explicit exceptional term. Appendix F.7 combines these ingredients, verifies the numerical width conditions, and derives the logarithmic sample and width regime. The same argument proves the kernel-dependent expected-risk inequality used in the main-text roadmap. For the additional high-probability bound (17), Steps 5–9 of the joint proof in Appendix F.7 concentrate a preserved comparison predictor, identify the learned feature kernel's local complexity, and control fitted sub-Gaussian noise through the exact finite-GD filter. Their union bound accounts for both training-sample and initialization randomness.

2. *Population attention descent, Proposition 4.2.* Appendix D.3 proves the stronger Proposition D.2, including an explicit lower bound on the expected improvement. Its proof first bounds channel and target moments, using the spherical-coordinate calculation in Appendix F.1 and polynomial hypercontractivity in Appendix E.2. It next controls both candidate gradient errors simultaneously, so that the data-dependent choice of temporary outputs remains valid. Gradient alignment and a population Taylor estimate give descent on a high-probability event. A uniform loss bound controls its complement and gives the expected improvement stated in the main text. This result concerns the temporary outputs before their reset.

The remaining parts of the appendix support the definitions, training claims, and comparisons accompanying these two results.

- *Model, algorithm, and consistency consequences.* Appendix A defines the observation law, regression risk, network, kernels, and target classes. Appendix B gives the full updates and computation costs. Appendices B.1 and B.2 prove Propositions B.1 and B.2, which justify the observed-feature learning rate and the feasible width rule. Appendix C also proves Corollaries C.3 and C.4, deriving the full bounded-norm and degree-independent consistency conclusions from the risk theorem. Their proofs use that theorem as an input, rather than as a consequence of these corollaries.
- *Representation identities and analytic foundations.* Appendix D.1 derives the output coordinates and kernel decomposition used in the main text. Appendix E.1 proves the spherical polynomial decomposition, harmonic identities, and dimension estimates. Appendix E.2 proves the matrix concentration and polynomial moment bounds used above. Appendix F.2 proves the obstruction for unbiased ReLU features and explains why the biased coefficients from Appendix F.1 remove it.
- *Literature comparisons and resource lower bounds.* Appendix G.1 gives the full literature discussion summarized in Section 1.1. Appendix G.2 proves the target-norm comparisons, and Appendix G.3 records the source assumptions behind the main-text table. The cited works establish their own guarantees. Appendix G.4 derives the resource comparisons and proves Proposition G.3, the supplied-degree benchmark. Appendix G.5 proves Proposition G.4, which gives the stated sufficient condition for the polynomial-ridge comparison risk to vanish. Appendices H.1, H.2,

and H.3 prove, respectively, Proposition H.1, Proposition H.2, and Proposition H.3. These results distinguish the sample lower bound for general estimators from the width restrictions for fixed representations.

- *Numerical comparisons.* Appendix I specifies the target, resource choices, and repetitions for the empirical comparison with a network without attention. These experiments are not used in any theoretical proof.

In the derivations below, a circled number above a relation refers to its nearby justification. The numbering starts at ① in each proof, returns to ① after ⑨, and also starts afresh in each separate calculation outside a proof.

A Detailed Problem Setup

We give the definitions and normalizations used in the approximation and risk proofs, following the setup in Section 2.

Let $\mathcal{X} = \mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ for an integer $d \geq 3$, and let $f^* : \mathcal{X} \to \mathbb{R}$ be the deterministic target. Let $P = \mathrm{Unif}(\mathcal{X})$ be the input law, with probability measure μ. We observe $n \geq 4$ independent pairs $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^n$ satisfying

$$y_i = f^*(\vec{\mathbf{x}}_i) + w_i, \qquad \vec{\mathbf{x}}_i \overset{i.i.d}{\sim} P. \tag{22}$$

The real-valued noise variables w_i are independent and identically distributed and are independent of all inputs. We write $\mathbb{E}[\cdot]$ for expectation and $\Pr[\cdot]$ for probability under the specified random quantities. For an unknown nonnegative number σ_0^2, we assume

$$\mathbb{E}[w_i] = 0, \qquad \mathbb{E}\left[w_i^2\right] \leq \sigma_0^2 < \infty. \tag{23}$$

Write $v_w := \mathbb{E}\left[w_1^2\right]$ for the actual noise variance. The bound σ_0^2 in (23) may equal v_w and is not supplied to training. No exponential or fourth noise moment is required for the expected-risk results. The additional high-probability conclusion uses the sub-Gaussian condition (15), with σ_0^2 then interpreted as a variance proxy.

For a measurable real function f, we set $\|f\|_{L^p} = (\int_{\mathcal{X}} |f|^p \, \mathrm{d}\mu)^{1/p}$ for $1 \leq p < \infty$. The space $L^p(P)$ consists of functions with finite norm, identified when they agree μ-almost everywhere. We also write $L^p(\mu)$ or L^p for this same space and its norm, and set $\langle f, h \rangle_{L^2} = \int_{\mathcal{X}} fh \mathrm{d}\mu$. All network features and polynomial targets are evaluated through their continuous representatives. For any trained predictor $\widehat{f}$, the regression risk is

$$\mathcal{R}(\widehat{f}) := \mathbb{E}_P\left[\left(\widehat{f}(\mathbf{x}) - f^*(\mathbf{x})\right)^2\right] = \left\|\widehat{f} - f^*\right\|_{L^2}^2. \tag{24}$$

The expectation in (24) is over a fresh input conditional on the predictor, so the risk excludes test-response noise. The expected-risk conclusion also averages over the full training sample and initialization. The high-probability conclusion instead bounds this random risk outside an event of explicitly controlled training probability.

A.1 The ReLU Attention Network

Let $\sigma(t) := \max\{t, 0\}$ be ReLU and let $s(t) := (1 + \exp(-t))^{-1}$ be the logistic sigmoid. Let the network width $m \geq 4$ be a deterministic integer divisible by four. There are m ReLU feature channels and an attention matrix of order m. We sample $m/4$ independent input directions and include their negatives to form $m/2$ features:

$$\vec{\mathbf{q}}_1, \ldots, \vec{\mathbf{q}}_{m/4} \overset{i.i.d}{\sim} P, \qquad \vec{\mathbf{q}}_{j+m/4} := -\vec{\mathbf{q}}_j, \quad j \in [m/4]. \tag{25}$$

Independently of the data, draw independent standard Gaussian vectors $\vec{\mathbf{u}}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for $j \in [m/4]$ and set $\vec{\mathbf{q}}_j = \vec{\mathbf{u}}_j / \left\|\vec{\mathbf{u}}_j\right\|_2$, redrawing a zero vector. Rotation invariance gives the uniform law P, so initialization requires only d and internal randomness. Write $\mathbf{Q} = (\vec{\mathbf{q}}_1, \ldots, \vec{\mathbf{q}}_{m/2})$ for the full collection. Each direction has marginal law P, but opposite pairs are dependent. All directions remain fixed during training.

For each fixed unit direction $\mathbf{q}$, rotation invariance gives $\mathbb{E}_P\left[\sqrt{d}\,\mathbf{q}^\top\mathbf{x}\right] = 0$ and $\mathbb{E}_P\left[(\sqrt{d}\,\mathbf{q}^\top\mathbf{x})^2\right] = 1$. The factor $\sqrt{d}$ therefore keeps the linear input to ReLU at unit variance as the dimension changes. With the fixed bias $1/2$, the initial features are

$$\phi_j(\mathbf{x}) \coloneqq \sigma\left(\sqrt{d}\,\vec{\mathbf{q}}_j^{\top}\mathbf{x} + \frac{1}{2}\right), \qquad j \in [m/2], \quad \mathbf{x} \in \mathcal{X}. \tag{26}$$

We define m channel outputs by duplication and their normalization by

$$\mathbf{h}(\mathbf{x}) \coloneqq \left[\phi_1(\mathbf{x}), \ldots, \phi_{m/2}(\mathbf{x}), \phi_1(\mathbf{x}), \ldots, \phi_{m/2}(\mathbf{x})\right]^\top, \qquad \mathbf{z}(\mathbf{x}) \coloneqq m^{-1/2}\mathbf{h}(\mathbf{x}). \tag{27}$$

The $m/4$ independent directions produce $m/2$ opposite-direction features and m duplicated channels. At least one feature in each opposite pair is active at every input. This pairing and duplication are retained throughout training. The bias also permits the polynomial approximation in Lemma F.2, as explained below.

The attention parameter is $\boldsymbol{\tau} \in \mathbb{R}^m$. For a list of square matrices, diag forms their block diagonal matrix. For a vector, it forms the diagonal matrix with that vector on the diagonal. We define

$$g_{\boldsymbol{\tau}}(\mathbf{x}) \coloneqq s\big(\sigma(1 + \boldsymbol{\tau}^\top\mathbf{z}(\mathbf{x})) - 1\big), \qquad \mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x}) \coloneqq \operatorname{diag}\left(g_{\boldsymbol{\tau}}(\mathbf{x})\mathbf{I}_{m/2}, (1 - g_{\boldsymbol{\tau}}(\mathbf{x}))\mathbf{I}_{m/2}\right). \tag{28}$$

As a result, $\mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x})$ is a diagonal matrix of order m, with m diagonal entries but at most two distinct gate values, $g_{\boldsymbol{\tau}}(\mathbf{x})$ and $1 - g_{\boldsymbol{\tau}}(\mathbf{x})$. More explicitly, let $\mathbf{v} = [\mathbf{1}_{m/2}^\top, -\mathbf{1}_{m/2}^\top]^\top \in \mathbb{R}^m$. The excitation block has input matrix $\boldsymbol{\tau}^\top \in \mathbb{R}^{1\times m}$, bias 1, fixed output matrix $\mathbf{v} \in \mathbb{R}^{m\times 1}$, and output bias $-\mathbf{v}$. Its sigmoid output is $s(\mathbf{v}\sigma(\boldsymbol{\tau}^\top\mathbf{z} + 1) - \mathbf{v})$, applied coordinatewise. This is a restricted squeeze-and-excitation construction (Hu et al., 2018), with two shared attention weights that sum to one, as stated in (28).

For output vectors $\boldsymbol{a}^+, \boldsymbol{a}^- \in \mathbb{R}^{m/2}$ with coordinates a_j^+, a_j^- for $j \in [m/2]$, we set $\boldsymbol{a} = [(\boldsymbol{a}^+)^\top, (\boldsymbol{a}^-)^\top]^\top \in \mathbb{R}^m$. The ReLU attention network is defined by

$$f(\boldsymbol{\tau}, \boldsymbol{a}, \mathbf{x}) \coloneqq \sqrt{\frac{2}{m}}\boldsymbol{a}^\top\mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x})\mathbf{h}(\mathbf{x}) = \sqrt{\frac{2}{m}}\sum_{j=1}^{m/2}\phi_j(\mathbf{x})\left(a_j^+ g_{\boldsymbol{\tau}}(\mathbf{x}) + a_j^-(1 - g_{\boldsymbol{\tau}}(\mathbf{x}))\right). \tag{29}$$

The fixed input directions require $O(dm)$ storage, and attention requires $O(m)$ storage and arithmetic per input.

A.2 Kernel and Target Function for Nonparametric Regression

For $\ell \geq 0$, let $\mathcal{H}_\ell$ be the space obtained by restricting real homogeneous harmonic polynomials of degree ℓ to $\mathcal{X}$. Restricting a polynomial means evaluating it only at points of $\mathcal{X}$, with its values there unchanged. Homogeneity means that each monomial has total degree ℓ, and harmonicity means that the polynomial has zero Euclidean Laplacian. Write $N(d, \ell) = \dim\mathcal{H}_\ell$ and choose a real orthonormal basis $Y_{\ell,1}, \ldots, Y_{\ell,N(d,\ell)}$ in $L^2(P)$, with $Y_{0,1} = 1$. Harmonics of distinct degrees are orthogonal in $L^2(P)$. Every polynomial of degree at most ℓ_0, restricted to $\mathcal{X}$, has a unique expansion in these bases for $0 \leq \ell \leq \ell_0$. Appendix E.1 proves the decomposition and dimension formulas.

For an unknown integer $1 \leq \ell_0 \leq L$, where L is a fixed degree bound used only in the analysis, let $\beta_{\ell,j} \in \mathbb{R}$ for $0 \leq \ell \leq \ell_0$ and $j \in [N(d, \ell)]$. We consider the target

$$f^*(\mathbf{x}) = \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)}\beta_{\ell,j}Y_{\ell,j}(\mathbf{x}), \qquad 1 \leq \ell_0 \leq L. \tag{30}$$

The deterministic coefficients may vary with d, while L remains fixed. The integer $\ell_0 \geq 1$ may be the largest nonzero degree or an upper bound, so constant targets are included. Define the nonnegative norms B_*, Γ_* and the dimension r_0 by

$$B_*^2 \coloneqq \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)}\beta_{\ell,j}^2 = \|f^*\|_{L^2}^2, \qquad \Gamma_*^2 \coloneqq \sum_{\ell=0}^{\ell_0}N(d,\ell)\sum_{j=1}^{N(d,\ell)}\beta_{\ell,j}^2, \qquad r_0 \coloneqq \sum_{\ell=0}^{\ell_0}N(d,\ell). \tag{31}$$

The norm B_* measures the target energy, whereas the weighted norm Γ_* controls the coefficients in its ReLU approximation. For $\ell \leq L$ and $d \geq d_L$, Lemma F.1 bounds the inverse squared degree-ℓ ReLU coefficient by $C_L N(d, \ell)$. Lemma F.2 sums these costs over the target coefficients.

For $B_* > 0$, the ratio Γ_*^2/B_*^2 averages $N(d,\ell)$ with weights $\sum_j \beta_{\ell,j}^2/B_*^2$. As a result, a pure degree-ℓ target has $\Gamma_*^2 = N(d,\ell)B_*^2$, including the constant case $N(d,0) = 1$. Both norms vanish for the zero target. The bound (52) retains this degree dependence, whereas (59) uses $\Gamma_*^2 \le r_0 B_*^2$. For nonpolynomial targets, B_* denotes the $L^2(P)$ norm of a fixed Borel measurable real representative.

Benefit of Biased ReLU Features. For every fixed L, Lemma F.1 proves that the bias $1/2$ gives a nonzero scalar coefficient at each degree through L in the spherical-harmonic expansion of a biased ReLU feature when $d \ge d_L$. Appendix F.1 defines these coefficients and shows how their nonvanishing permits recovery of the target's polynomial components. Without a bias, the ReLU span misses all odd degrees above one, so its approximation error cannot vanish when the target retains energy in those degrees. Proposition F.3 in Appendix F.2 proves this obstruction.

A reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions in which point evaluation is continuous. Its kernel represents point evaluation by the Hilbert-space inner product. For a fixed radius $\gamma_0 \ge 0$, the weighted target class considered in Yang (2026) satisfies $\Gamma_*^2 \le \gamma_0^2$. Set $\mu_\ell := N(d,\ell)^{-1}$ for $0 \le \ell \le \ell_0$. The weighted class is the ball of radius γ_0 in the RKHS of the finite-rank kernel

$$K^{(r_0)}(\mathbf{x},\mathbf{x}') := \sum_{\ell=0}^{\ell_0} \mu_\ell \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}'). \tag{32}$$

The integral operator associated with (32) maps $f \in L^2(P)$ to $\mathbf{x} \mapsto \int_{\mathcal{X}} K^{(r_0)}(\mathbf{x},\mathbf{x}') f(\mathbf{x}') \mathrm{d}\mu(\mathbf{x}')$. Orthonormality shows that each $Y_{\ell,j}$ is its eigenfunction with positive eigenvalue μ_ℓ. The squared RKHS norm is therefore $\sum_{\ell=0}^{\ell_0} \sum_{j=1}^{N(d,\ell)} \beta_{\ell,j}^2/\mu_\ell = \Gamma_*^2$. The kernel specifies the target class and is not computed by the algorithm. In particular, it is not identified with the covariance kernel of the ReLU network. We do not impose the minimum-coefficient condition used for harmonic-degree selection in Yang (2026). We also treat the larger class $B_* \le B$ for any fixed $B \ge 0$, using the elementary inequality $\Gamma_*^2 \le r_0 B_*^2$. Proposition G.1 in Appendix G explains how a fixed weighted radius can force nonconstant target energy to vanish with dimension, and contrasts this class with a fixed L^2 radius.

B Detailed Training Algorithm

We give the losses, updates, and computation costs for Algorithm 1. The first training sample determines the attention features through one GD step. The independent second training sample is used in each of the $T = 4(d+1)m$ full-batch GD steps that fit their output coefficients.

With $n_1 = \lfloor n/2 \rfloor$ and $n_2 = n - n_1$, the first training sample and second training sample have index sets $I_1 = [n_1]$ and $I_2 = \{n_1+1,\dots,n\}$, respectively. The empirical distribution P_{n_2} assigns mass $1/n_2$ to each input in the second training sample indexed by I_2. In particular, $\|f\|_{L^2(P_{n_2})}^2 = n_2^{-1} \sum_{i \in I_2} f(\vec{\mathbf{x}}_i)^2$. A complete degree-independent choice of width is

$$H_n := \log(en), \qquad m := 4\lceil \frac{n}{4H_n^2} \rceil. \tag{33}$$

The finite-sample theorem also applies to any deterministic width divisible by four satisfying its stated conditions. Corollary C.4 gives alternative degree-independent width choices and distinguishes their consistency ranges from that of the default rule (33).

Training Stage One: Learning the Channel Attention Weights. We initialize $\boldsymbol{\tau}(0) = \mathbf{0}$ and use the first training sample to choose between two temporary output vectors, ensuring a nonzero attention gradient. For $c \in \{0,1\}$, define

$$\boldsymbol{a}_{\mathrm{init},c}^{+} = (c+1)\sqrt{2/m}\mathbf{1}_{m/2}, \qquad \boldsymbol{a}_{\mathrm{init},c}^{-} = (c-1)\sqrt{2/m}\mathbf{1}_{m/2}. \tag{34}$$

Let $\boldsymbol{a}_{\mathrm{init},c}$ be the concatenation of these two vectors. Recall the ReLU channels ϕ_j and the normalized duplicated vector $\mathbf{z}$ from (26)–(27), and define $b(\mathbf{x}) := (2/m)\sum_{j=1}^{m/2} \phi_j(\mathbf{x})$. Then

$$f(\boldsymbol{\tau}, \boldsymbol{a}_{\mathrm{init},c}, \mathbf{x}) = b(\mathbf{x})\big(c + 2g_{\boldsymbol{\tau}}(\mathbf{x}) - 1\big), \qquad f(\mathbf{0}, \boldsymbol{a}_{\mathrm{init},c}, \mathbf{x}) = c\, b(\mathbf{x}).$$

Define $L_{1,c}(\boldsymbol{\tau}) \coloneqq (2n_1)^{-1} \sum_{i \in I_1} \left(f(\boldsymbol{\tau}, \boldsymbol{a}_{\mathrm{init},c}, \vec{\mathbf{x}}_i) - y_i \right)^2$, the training loss for candidate c, equal to half the empirical mean squared error. Its attention gradient at zero is

$$\mathbf{g}_c \coloneqq \nabla L_{1,c}(\mathbf{0}) \overset{①}{=} \frac{1}{2n_1} \sum_{i \in I_1} (c\, b(\vec{\mathbf{x}}_i) - y_i) b(\vec{\mathbf{x}}_i) \mathbf{z}(\vec{\mathbf{x}}_i). \tag{35}$$

Equality ① uses $\nabla_{\boldsymbol{\tau}} f(\mathbf{0}, \boldsymbol{a}_{\mathrm{init},c}, \mathbf{x}) = b(\mathbf{x})\mathbf{z}(\mathbf{x})/2$, which follows from $s'(0) = 1/4$ and the active bottleneck at its initial argument 1. We select the candidate with larger gradient norm, using $c = 0$ in a tie:

$$\widehat{c} \coloneqq \begin{cases} 0, & \|\mathbf{g}_0\|_2 \geq \|\mathbf{g}_1\|_2, \\ 1, & \|\mathbf{g}_0\|_2 < \|\mathbf{g}_1\|_2, \end{cases} \qquad \mathbf{g} \coloneqq \mathbf{g}_{\widehat{c}}, \qquad \boldsymbol{a}_{\mathrm{init}} \coloneqq \boldsymbol{a}_{\mathrm{init},\widehat{c}}, \qquad L_1 \coloneqq L_{1,\widehat{c}}. \tag{36}$$

Both gradients in (36) use the same m channels. Their comparison selects an output initialization from observed quantities and uses no target degree.

We next choose a scalar learning rate that keeps the bottleneck ReLU in its differentiable region and guarantees a strict decrease of L_1. Define

$$\Lambda_1 \coloneqq \frac{1}{n_1} \sum_{i \in I_1} \left(\frac{5}{4} b(\vec{\mathbf{x}}_i)^2 + \frac{1}{2} |y_i|\, b(\vec{\mathbf{x}}_i) \right) \left\| \mathbf{z}(\vec{\mathbf{x}}_i) \right\|_2^2, \quad \eta_1 \coloneqq \frac{1}{1 + 2\sqrt{2(d+1)}\|\mathbf{g}\|_2 + \Lambda_1}. \tag{37}$$

Holding $\boldsymbol{a}_{\mathrm{init}}$ fixed, we take one ordinary GD step on the original squared training loss:

$$\widehat{\boldsymbol{\tau}} \coloneqq \boldsymbol{\tau}(0) - \eta_1 \nabla L_1(\boldsymbol{\tau}(0)) = -\eta_1 \mathbf{g} = \frac{\eta_1}{2n_1} \sum_{i \in I_1} \big(y_i - \widehat{c}\, b(\vec{\mathbf{x}}_i) \big) b(\vec{\mathbf{x}}_i) \mathbf{z}(\vec{\mathbf{x}}_i). \tag{38}$$

The gradient-norm term in (37) keeps the bottleneck active throughout the step, while $\eta_1 \Lambda_1 < 1$ gives descent. Proposition D.1 proves these assertions, the lower bound $\|\mathbf{g}\|_2 \geq 1/32$, and strict feature enlargement.

The step size is computed once from the first training sample and directions. Duplicated channels give $\widehat{\tau}_j = \widehat{\tau}_{m/2+j}$ for $j \in [m/2]$.

Training Stage Two: Learning the Output Coefficients. We freeze $\widehat{\boldsymbol{\tau}}$ and the initial directions, and define

$$\boldsymbol{\psi}(\mathbf{x}) \coloneqq \sqrt{2/m} \mathbf{A}_{\widehat{\boldsymbol{\tau}}}(\mathbf{x}) \mathbf{h}(\mathbf{x}), \qquad f(\widehat{\boldsymbol{\tau}}, \boldsymbol{a}, \mathbf{x}) = \boldsymbol{a}^\top \boldsymbol{\psi}(\mathbf{x}). \tag{39}$$

We reset $\boldsymbol{a}(0) = \mathbf{0}$ and optimize

$$L_2(\boldsymbol{a}) \coloneqq \frac{1}{2n_2} \sum_{i \in I_2} \left(\boldsymbol{a}^\top \boldsymbol{\psi}(\vec{\mathbf{x}}_i) - y_i \right)^2 \tag{40}$$

by ordinary full-batch GD with respect to $\boldsymbol{a}$ only. We refer to these updates as output GD. Every gradient uses all n_2 observations in the second training sample, indexed by I_2, while $\widehat{\boldsymbol{\tau}}$, $\mathbf{Q}$, and the biases remain fixed. The learning rate, iteration count, and update are

$$\eta \coloneqq \frac{1}{4(d+1)}, \qquad T \coloneqq 4(d+1)m, \qquad \eta T = m, \tag{41}$$

$$\boldsymbol{a}(t+1) = \boldsymbol{a}(t) - \frac{\eta}{n_2} \sum_{i \in I_2} \left(\boldsymbol{a}(t)^\top \boldsymbol{\psi}(\vec{\mathbf{x}}_i) - y_i \right) \boldsymbol{\psi}(\vec{\mathbf{x}}_i), \quad 0 \leq t < T. \tag{42}$$

We return $\widehat{f} = f(\widehat{\boldsymbol{\tau}}, \boldsymbol{a}(T), \cdot)$. Lemma F.7 proves nonincrease of the training loss and the exact finite-GD formula (123).

The reset preserves $\boldsymbol{\psi}$ and starts output fitting from zero. Conditional on the first stage, the independent second training sample retains its original law. Descent concerns L_1 before the reset and L_2 after it.

Algorithm 1 uses only the data, d, and independent Gaussian draws. Its width and positive learning rates are defined for every $n \geq 4$ and finite sample, without the degree, target norm, noise variance, or harmonic coefficients. The relation $n \asymp d^{\ell_0} (\log d)^4$ and the dimension and width conditions of Theorem C.1 specify the guarantee's regime and are not inputs or stopping tests.

The two initial gradients and the sum defining Λ_1 can be accumulated in one pass through the first training sample. The initialization choice and first-stage update therefore use $O(n_1 dm)$ arithmetic and $O(dm)$ working memory. Forming and storing the features for the second training sample costs

$O(n_2 dm)$ arithmetic and $O(n_2 m + dm)$ memory. Each subsequent full-batch step costs $O(n_2 m)$. The total arithmetic cost is therefore $O(ndm + Tn_2 m) = O(ndm^2)$ for the specified $T = 4(d+1)m$. These are real-arithmetic operation counts with constant-cost activation evaluations.

The uniform feature bound gives the conservative horizon in (41). Proposition B.1 replaces it by an observed-feature bound, retaining the expected-risk guarantee with at most $7m/2$ output iterations in conditional expectation. Neither version is claimed to be computationally optimal.

B.1 Output Training with an Observed Feature Bound

The step size in (41) uses a worst-case feature bound over the sphere. The following alternative uses the observed mean squared feature norm to ensure output-GD stability. It remains ordinary full-batch GD with a fixed scalar step size, and its schedule does not use the responses in the second training sample.

Recall that $\boldsymbol{\psi}$ is the frozen feature vector, n_2 is the size of the second training sample, and m is the channel count. Define the observed mean squared feature norm and a positive integer by

$$S_2 := \frac{1}{n_2} \sum_{i \in I_2} \left\| \boldsymbol{\psi}(\vec{\mathbf{x}}_i) \right\|_2^2 = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\right), \qquad J_2 := \max\left\{1, \lceil 2S_2 \rceil\right\}, \qquad \eta := J_2^{-1}, \qquad T := mJ_2. \tag{43}$$

Use (43) in the updates (42), with the same zero initialization, and denote its returned predictor by $\widehat{f}_{\mathrm{obs}}$. This is a variant of Algorithm 1: its single attention GD step is unchanged, while stage two uses $T = mJ_2$ output GD steps instead of $4(d+1)m$. Let $\mathscr{F}_1$ be the sigma-field generated by the directions and observations in the first training sample. Let $\mathscr{F}_2$ be generated by $\mathscr{F}_1$ together with the inputs in the second training sample. As a result, S_2, J_2, η, T are $\mathscr{F}_2$-measurable, and the remaining noises in the second training sample are independent of all four quantities. These sigma-fields are recalled in Appendix C.1.

Let $\mathcal{E}_Q$ be the event that $\frac{2}{m} \sum_{j=1}^{m/2} (\sqrt{d}\, \vec{\mathbf{q}}_j^{\top} \mathbf{x})^2 \le 2$ for every $\mathbf{x} \in \mathcal{X}$. This uniform bound on squared projections is written in matrix form in (109).

Proposition B.1 (Output GD with a step size from observed features)**.** For every finite attention vector and every finite second training sample, the rule (43) satisfies

$$\eta \left\| \widehat{\boldsymbol{\Sigma}} \right\|_2 \le \tfrac{1}{2}, \qquad \eta T = m, \qquad m \le T \le 4(d+1)m. \tag{44}$$

The training loss decreases as in (121). Under the assumptions of Theorem C.1, its expected-risk and consistency bounds in parts (ii)–(iii) hold for $\widehat{f}_{\mathrm{obs}}$ with the same constants. More generally, the explicit-tail inequalities in Proposition C.2 hold under that proposition's assumptions. For every almost surely finite $\mathscr{F}_1$-measurable attention rule,

$$\mathbb{E}\left[T \mid \mathscr{F}_1\right] \le \tfrac{7}{2} m, \qquad T \le 5m \quad \text{on } \mathcal{E}_Q. \tag{45}$$

With stored features for the second training sample, the conditional expected arithmetic cost of the full procedure is $O(ndm + nm^2)$, uniformly over the first-stage outcomes. The procedure uses no unknown degree, target norm, or noise variance.

Proof. **Outline.** Step 1 verifies stability and the unchanged training scale on every observed feature matrix. Step 2 verifies that choosing the step size from the inputs preserves the finite-GD representation and conditional noise calculations. Step 3 carries over the approximation and covariance bounds. Step 4 controls the number of iterations and arithmetic cost.

Step 1. Stability on the observed matrix. The empirical trace controls the largest eigenvalue. Recall $S_2 = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\right)$ and $J_2 = \max\{1, \lceil 2S_2 \rceil\}$. Then $\eta \left\| \widehat{\boldsymbol{\Sigma}} \right\|_2 \overset{①}{\le} S_2/J_2 \overset{②}{\le} 1/2$, where ① uses positive semidefiniteness and ② uses $J_2 \ge 2S_2$, including $S_2 = 0$. Also $\eta T = m$. Proposition F.4 gives $S_2 \le 2(d+1)$, hence $1 \le J_2 \le 4(d+1)$. This proves (44). Step 1 of Lemma F.7 then gives (121) for this step size.

Step 2. A finite polynomial under the same conditioning. Conditional on $\mathscr{F}_2$, the schedule is fixed and the noise in the second training sample remains independent. The recurrence gives (123) with $q_T(u) = \eta \sum_{t=0}^{T-1} (1 - \eta u)^t$ for $u \ge 0$. Step 1 verifies the conditions $0 \le \eta u \le 1/2$ on the empirical

spectrum and $\eta T = m$ used in (125)–(126). These bounds hold unchanged, including $q_T(0) = m$ and $uq_T(u)^2 \le m$.

Given $\mathscr{F}_2$, the noise still has mean zero and covariance $v_w \mathbf{I}_{n_2}$, where $v_w = \mathbb{E}\left[w_1^2\right]$. Hence (130) and (131) apply. The coefficient bound $\|\boldsymbol{a}(T)\|_2^2 \le (m/n_2)\sum_{i\in I_2} y_i^2$ ensures square integrability as in Lemma F.8. The finite recurrence on each measurable event $\{J_2 = j\}$, $1 \le j \le 4(d+1)$, proves measurability despite the random horizon.

Step 3. Transfer of the risk bounds. It remains to check the residual averaging used in the risk proof. The feature map and concentration events are unchanged. The minimizer $\boldsymbol{a}_\lambda$ in (145), at $\lambda = m^{-1}$, is $\mathscr{F}_1$-measurable and does not depend on S_2 or T. For $r = f^* - \boldsymbol{a}_\lambda^\top \boldsymbol{\psi}$, $\mathbb{E}\left[\|r\|_{L^2(P_{n_2})}^2 \mid \mathscr{F}_1\right] \overset{③}{=} \|r\|_{L^2}^2$, where ③ uses the unchanged conditional input law. Steps 2–4 of Lemma F.9 use only the covariance comparison, the scalar bounds verified in Step 2, and the independent centered noise. They therefore give (142) with the same constants.

The coefficient and variance bounds from Step 2 and the unchanged target fourth moment also give Lemma F.8. Averaging as in Proposition C.2 proves its bound for $\widehat{f}_{\mathrm{obs}}$. The same polynomial approximation and width conditions then yield parts (ii)–(iii) of Theorem C.1. None of these steps requires independence between T and the empirical covariance.

Step 4. Iterations and cost. Population feature energy controls the mean iteration count. Since $J_2 \le 1 + 2S_2$, conditioning on $\mathscr{F}_1$ and applying Proposition F.4 give

$$\mathbb{E}\left[S_2 \mid \mathscr{F}_1\right] \overset{④}{=} \int \|\boldsymbol{\psi}\|_2^2 \mathrm{d}\mu = \mathrm{tr}\left(\boldsymbol{\Sigma}\right) \overset{⑤}{\le} \tfrac{5}{4}, \qquad \mathbb{E}\left[J_2 \mid \mathscr{F}_1\right] \overset{⑥}{\le} 1 + 2 \cdot \tfrac{5}{4} = \tfrac{7}{2}.$$

Equality ④ averages over the conditional input law, ⑤ is the population feature bound, and ⑥ uses $J_2 \le 1 + 2S_2$. Since $T = mJ_2$, this proves the first bound in (45). On $\mathcal{E}_Q$, Lemma F.5 gives $S_2 \le 9/4$ and hence $J_2 \le 5$, proving the second. Forming S_2 with the stored features and performing each GD step cost $O(n_2 m)$. The conditional expected output cost is therefore $O(n_2 m^2)$. Including feature construction and stage one gives $O(ndm + nm^2)$ operations and $O(n_2 m + dm)$ memory, under Algorithm 1's real-arithmetic conventions. □

B.2 Selecting a Width Inside the Finite-Sample Interval

The default rule (33) gives $m \ge n/H_n^2$, so the upper width condition $m \le n/(512H_n)$ requires $H_n \ge 512$. This restriction comes from the sufficient constants and does not constitute a statistical lower bound. We give an alternative that satisfies both width inequalities whenever their interval contains a multiple of four, while retaining the default rule eventually in the logarithmic regime.

For $n \ge 4$ and $d \ge 3$, recall $H_n = \log(en)$ and put

$$m_{\mathrm{lo}} := 4\lceil 512(d + H_n)/4 \rceil, \qquad m_{\mathrm{hi}} := 4\lfloor n/(2048H_n) \rfloor, \qquad m_{\mathrm{def}} := 4\lceil n/(4H_n^2) \rceil. \tag{46}$$

In (46), m_{lo} and m_{hi} round the lower and upper width thresholds inward to multiples of four. Define

$$m_{\mathrm{adm}} := \begin{cases} \min\left\{m_{\mathrm{hi}}, \max\left\{m_{\mathrm{lo}}, m_{\mathrm{def}}\right\}\right\}, & m_{\mathrm{lo}} \le m_{\mathrm{hi}}, \\ m_{\mathrm{def}}, & m_{\mathrm{lo}} > m_{\mathrm{hi}}. \end{cases} \tag{47}$$

When the admissible interval is empty, (47) returns the default width without asserting the theorem's width conditions. The rule uses only n, d and the numerical constant 512.

Proposition B.2 (A width rule with an exact feasibility check)**.** There is a positive multiple of four satisfying (12) if and only if $m_{\mathrm{lo}} \le m_{\mathrm{hi}}$. Whenever this condition holds, $m = m_{\mathrm{adm}}$ satisfies both inequalities. Under the remaining assumptions of Theorem C.1, its expected-risk bound therefore applies, with either the original output parameters or (43). For each fixed $1 \le \ell_0 \le L$ and $n \asymp d^{\ell_0}(\log d)^4$ as $d \to \infty$, one has $m_{\mathrm{adm}} = m_{\mathrm{def}}$ eventually. As a result, this rule retains the same channel count and consistency rate in (14).

Proof. **Outline.** We first identify the smallest and largest permitted multiples of four. We then check the selection rule and verify that its adjustments disappear in the stated asymptotic regime.

Step 1. The integer interval. Rounding inward gives the least and greatest admissible multiples of four, m_{lo} and m_{hi}. An admissible width exists exactly when $m_{\mathrm{lo}} \le m_{\mathrm{hi}}$. In that case, (47) clips

the default multiple of four to this interval and depends only on n, d. This proves the finite-sample assertion and retains the deterministic-width hypothesis. Proposition B.1 justifies the alternative output schedule.

Step 2. The logarithmic regime. To show that the adjustment eventually disappears, we compare the default width with both endpoints. At $n \asymp d^{\ell_0}(\log d)^4$, one has $H_n \asymp \log d$ and $m_{\mathrm{def}} \sim n/H_n^2$, so

$$\frac{m_{\mathrm{def}}}{512(d+H_n)} \xrightarrow{①} \infty, \qquad \frac{m_{\mathrm{def}}}{n/(512H_n)} \xrightarrow{②} 0.$$

For ①, the ratio has order $d^{\ell_0-1}(\log d)^2$, which diverges also when $\ell_0 = 1$. For ②, the ratio is asymptotic to $512/H_n$, which tends to zero. Both endpoints diverge, and rounding changes either by less than four. As a result, $m_{\mathrm{lo}} \le m_{\mathrm{def}} \le m_{\mathrm{hi}}$ eventually. The first branch of (47) then returns m_{def} exactly. The consistency conclusion follows by applying the calculation for the original width rule. □

The feasibility test certifies the width inequalities only. It does not estimate the target degree or verify the degree-dependent threshold d_L. If the test fails, Proposition C.2 still gives its bound with an explicit exceptional term for the returned width, but it need not imply a small risk. No wider network is trained and then reduced under either rule.

C Full Statements and Additional Consequences

This appendix states the complete risk theorem and its consistency consequences. Appendix D develops the representation identities and proves feature enlargement. Appendix F supplies the approximation and statistical estimates.

C.1 Approximation, Covariance, and the GD Scale

Let $\mathscr{F}_1$ be the sigma-field generated by the directions and the observations in the first training sample indexed by I_1, and let $\mathscr{F}_2$ also include the inputs in the second training sample indexed by I_2. Conditioning on $\mathscr{F}_1$ fixes the learned map $\boldsymbol{\psi}$ and leaves the second training sample independent with its original law. Conditioning further on $\mathscr{F}_2$ fixes its inputs while preserving the centered noise law with variance v_w. The responses in the first training sample remain part of both sigma-fields. All conditional assertions are understood almost surely.

Write $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ for the population uncentered second-moment matrix of $\boldsymbol{\psi}$ and its empirical counterpart on the second training sample. For $\lambda > 0$, the empirical effective dimension $\widehat{\mathcal{N}}(\lambda)$ sums $u/(u+\lambda)$ over the eigenvalues u of $\widehat{\boldsymbol{\Sigma}}$, with multiplicity, and is at most m. Appendix F.4 gives their formulas. We use the two conditioning steps above for covariance concentration and the exact noise calculation, respectively.

Regression risk with a coefficient penalty. The finite-GD bound depends on both the risk of a comparison predictor and the size of its coefficients. We therefore use the following penalized approximation quantity, whose exact improvement from the added features is proved in Proposition F.10.

For each fixed first-stage outcome, set

$$\mathcal{A}_\lambda(\widehat{K}) = \min_{\boldsymbol{a}\in\mathbb{R}^m} \left(\left\| \boldsymbol{a}^\top \boldsymbol{\psi} - f^* \right\|_{L^2}^2 + \lambda \|\boldsymbol{a}\|_2^2 \right), \qquad \lambda > 0. \tag{48}$$

Replacing $\boldsymbol{\psi}$ by $\boldsymbol{\psi}_0$ gives the corresponding quantity for the initial feature kernel $\mathcal{A}_\lambda(K_0)$, with $K_0 = K_{\mathrm{RF}}/2$ as in (9). The minimizers in (48) exist and are unique because $\lambda > 0$, even for dependent features. This penalty is used in the analysis and is absent from the training loss. Since $\eta T = m$, finite GD uses the scale $\lambda = m^{-1}$.

The preservation identity (78) gives $\mathcal{A}_\lambda(\widehat{K}) \le \mathcal{A}_\lambda(K_0)$. Proposition F.10 proves this preservation inequality and constructs the measurable comparison coefficients used in the risk analysis. Appendix F.7 bounds the preserved approximation cost, and Appendices F.4 and F.5 control fitting on the second training sample.

The finite-sample guarantee. The next theorem uses a deterministic width m. Its lower width condition controls the sampled directions, and its upper condition permits covariance comparison at scale m^{-1}. Lemmas F.5 and F.6 give these sufficient conditions. We write $O_L(a)$ for a bound by $C_L a$, where C_L depends only on L. Combining these estimates with (147) yields the approximation and optimization term $O_L(\Gamma_*^2/m)$ and noise term $O(\sigma_0^2 m/n)$. Part (iii) states the resulting consistency regime.

Theorem C.1 (Feature learning and regression risk)**.** Fix an integer $L \geq 1$. There exist a numerical constant $C_0 > 0$, an integer $d_L \geq 3$, and a constant $C_L > 0$ depending only on L, with the following property. Suppose $d \geq d_L$, $n \geq 4$, the model (22)–(23) holds, and f^* is given by (30) with $1 \leq \ell_0 \leq L$. Let m be a deterministic positive integer divisible by four satisfying

$$C_0\big(d + \log(en)\big) \leq m \leq \frac{n}{C_0 \log(en)}. \tag{49}$$

Run Algorithm 1: perform one attention GD step with the initialization and learning rate (36)–(37), then $T = 4(d+1)m$ output GD steps from zero with learning rate $\eta = [4(d+1)]^{-1}$ as in (41). Section B gives the detailed updates.

(i) Deterministic representation change. For every finite first training sample and every realization of the directions, the attention gradient satisfies $\|\mathbf{g}\|_2 \geq 1/32$, the loss strictly decreases as in (80), and $g_{\widehat{\tau}}$ is nonconstant. The learned prediction space strictly contains the initial ReLU span, with the explicit witness (77). The learned feature kernel $\widehat{K}$ differs from every constant multiple of K_{RF}, as in (81). Its population squared feature change and kernel-increment trace satisfy (75), with $c_d = 57/[1024(d+1)]$. These deterministic assertions require only $d \geq 3$ and the stated architecture, not the finite-sample inequalities.

(ii) Expected regression risk. The second stage uses the learned feature kernel, held fixed. Its returned predictor satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \leq 12\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] + \frac{12\sigma_0^2}{n}\mathbb{E}\left[\widehat{\mathcal{N}}(m^{-1})\right] + \frac{C_L(B_*^2+\sigma_0^2)}{n^{7/2}}. \tag{50}$$

The numerical coefficients in (50) are independent of the degree and the attention update. The bound is uniform over the centered noise laws in (23), including laws with infinite fourth moment. The constant C_L in its last term bounds exceptional outcomes and depends only on L. In particular,

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \leq C_L\left(\frac{\Gamma_*^2}{m} + \frac{\sigma_0^2}{n}\mathbb{E}\left[\widehat{\mathcal{N}}(m^{-1})\right] + \frac{B_*^2+\sigma_0^2}{n^{7/2}}\right). \tag{51}$$

Consequently,

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \leq C_L\left(\frac{\Gamma_*^2}{m} + \frac{\sigma_0^2 m}{n} + \frac{B_*^2+\sigma_0^2}{n^{7/2}}\right). \tag{52}$$

In particular, uniformly over the target class $\Gamma_* \leq \gamma_0$,

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \leq C_L\left(\frac{\gamma_0^2}{m} + \frac{\sigma_0^2 m}{n} + \frac{\gamma_0^2+\sigma_0^2}{n^{7/2}}\right). \tag{53}$$

(iii) Consistency. Consider joint sequences with $d \to \infty$, fixed $1 \leq \ell_0 \leq L$, and deterministic widths satisfying (49). These inequalities imply $m, n \to \infty$ and $m/n \to 0$. For the weighted class with fixed γ_0, σ_0, equation (53) therefore gives $\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \to 0$ and $\mathcal{R}(\widehat{f}) \to 0$ in probability without an additional degree-scale condition. For the full class with bounded B_*, σ_0, the additional condition $r_0/m \to 0$ gives the same conclusions. Corollary C.4 specifies the corresponding sample-dimension regimes. In particular, take the width rule (33) and $n \asymp d^{\ell_0}(\log d)^4$. Then, for all sufficiently large d, the finite-sample assumptions hold and

$$\frac{n}{d^{\ell_0}} \asymp (\log d)^4, \qquad \frac{m}{d^{\ell_0}} \asymp (\log d)^2, \qquad \mathbb{E}\left[\mathcal{R}(\widehat{f})\right] = O_L\left((B_*^2+\sigma_0^2)(\log d)^{-2}\right) \longrightarrow 0. \tag{54}$$

The constant in (54) may also depend on the fixed comparison constants in $n \asymp d^{\ell_0}(\log d)^4$. Both displayed ratios diverge and are $o(d)$. The number of ReLU channels and the order of the attention matrix remain equal to this same m throughout training. The numerical choice $C_0 = 512$ is sufficient. The constants C_0, d_L, C_L specify sufficient conditions and bounds, and need not be supplied to the training algorithm. The width m is fixed before observing the data. In particular, the theorem does not cover choosing m from the responses in the first training sample without a separate argument.

(iv) High-probability regression risk. If the noises additionally satisfy (15), the same predictor satisfies (17) for every $\delta \in (0,1)$, with probability at least $1-\delta-3n^{-8}$ and $t_\delta = \log(8/\delta)$. The population effective dimension is defined in (16). At the budgets in (54), the bound is $O((\log d)^{-2})$ for fixed δ, and $O(\log\log d/(\log d)^2)$ with probability tending to one when $\delta = (\log d)^{-a}$ for fixed $a > 0$. Parts (ii)–(iii) retain their finite-variance scope, and the training parameters are unchanged.

Proof roadmap for Theorem C.1. The proof in Appendix F.7 follows the five-step roadmap in Section 4. Its first four steps establish parts (i)–(iii). Proposition D.1 proves representation change. Lemmas F.1 and F.2, together with coefficient preservation, then give (147). Lemmas F.5 and F.6 transfer the finite-GD estimate of Lemma F.9 to population risk outside an event of probability at most $2n^{-8}$. Proposition C.2 averages this estimate and controls the complement using Lemma F.8. Finally, $\widehat{\mathcal{N}}(m^{-1}) \le m$ and $\Gamma_*^2 \le r_0 B_*^2$ give the stated rates, and Markov's inequality gives convergence in probability. The fifth roadmap step proves part (iv) through concentration of the preserved comparison predictor and the learned feature kernel's localized noise term, using Steps 5–9 of the joint appendix proof.

A finite-sample bound with explicit tails. The next result retains the exceptional contribution for every deterministic width. In (55), the first exponential term controls the directions and the second controls the empirical covariance. It applies to a fixed target in $L^4(P)$ under finite noise variance and is the estimate used to prove the expected-risk part of Theorem C.1.

Proposition C.2 (Risk with an explicit exceptional term)**.** For integers $d \ge 3$, $n \ge 4$, and a deterministic integer $m \ge 4$ divisible by four, consider the following finite-sample assertion. Use the observation model (22)–(23) with a fixed measurable real target $f^* \in L^4(P)$. Draw the directions by (25), independently of the observations. Let the frozen attention be any measurable, almost surely finite function of the sampled directions and observations in the first training sample. Perform stage two of Algorithm 1 from zero, namely $T = 4(d+1)m$ output GD steps with $\eta = [4(d+1)]^{-1}$ as in (41)–(42). This assertion does not require the attention rule from stage one. Recall $B_* = \|f^*\|_{L^2}$, $v_w = \mathbb{E}\left[w_1^2\right]$, and $n_2 = n - \lfloor n/2 \rfloor$. The quantity $\|f^*\|_{L^4}^4 - B_*^4$ is the variance of $f^*(\mathbf{x})^2$ for $\mathbf{x} \sim P$, so it is nonnegative. Define

$$\delta_{d,n,m} := \min\left\{1, \exp(d\log 21 - m/64) + 2m\exp\left(-\frac{n_2}{21m}\right)\right\}. \tag{55}$$

Then

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le 12\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] + \frac{12v_w}{n}\mathbb{E}\left[\widehat{\mathcal{N}}(m^{-1})\right] + \frac{5m}{2}\sqrt{\frac{(\|f^*\|_{L^4}^4 - B_*^4)\delta_{d,n,m}}{n_2}} + 3mB_*^2\delta_{d,n,m} + \frac{5mv_w}{4n_2}\delta_{d,n,m}. \tag{56}$$

If f^* has the polynomial expansion (30) for a fixed $L \ge 1$, then

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le 12\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] + \frac{12\sigma_0^2}{n}\mathbb{E}\left[\widehat{\mathcal{N}}(m^{-1})\right] + C_L mB_*^2\left(\delta_{d,n,m} + \sqrt{\frac{\delta_{d,n,m}}{n_2}}\right) + \frac{5m\sigma_0^2}{4n_2}\delta_{d,n,m}, \tag{57}$$

where C_L depends only on L. Neither bound requires (49) or a lower dimension threshold d_L. The constants are uniform over the first-stage rules and noise laws with the stated variance bound. Under (49) with $C_0 = 512$, the exceptional contributions in (56) sum to at most $11\|f^*\|_{L^4}^2 n^{-7/2} + 5v_w n^{-8}$. In (57), their sum is at most $C_L B_*^2 n^{-7/2} + 5\sigma_0^2 n^{-8}$ after enlarging C_L.

Proof. **Outline.** Step 1 bounds the probability of covariance failure. Step 2 averages the conditional GD estimate on the covariance event. Step 3 uses the variance of the empirical target square and the fitted-noise bound on the complement, then specializes to polynomial targets.

Step 1. The covariance event. Recall $\mathcal{E} = \mathcal{E}_Q \cap \mathcal{E}_S$ from Lemma F.6. Equations (110) and (118) give $\Pr\left[\mathcal{E}_Q^c\right] \le \exp(d\log 21 - m/64)$ and, on $\mathcal{E}_Q$, $\Pr\left[\mathcal{E}_S^c \mid \mathscr{F}_1\right] \le 2m\exp(-n_2/(21m))$. Averaging the second bound on $\mathcal{E}_Q$ and adding the first yields $\Pr\left[\mathcal{E}^c\right] \overset{①}{\le} \delta_{d,n,m}$ by the union bound and tower property. These estimates hold for every finite $\mathscr{F}_1$-measurable attention rule. Also $\mathcal{E} \in \mathscr{F}_2$, which leaves the noise in the second training sample independent.

Step 2. Averaging on the event. We average the residual before imposing any event determined by the second training sample, so that its empirical square still has the correct population mean. Set $\lambda = m^{-1}$ and let $\boldsymbol{a}_\lambda$ be the population minimizer in (145). That formula applies to every target in $L^2(P)$, makes the minimizer $\mathscr{F}_1$-measurable, and gives $\|\boldsymbol{a}_\lambda\|_2^2 \le mB_*^2$. Its residual $r = f^* - \boldsymbol{a}_\lambda^\top \boldsymbol{\psi}$ has $\|r\|_{L^2}^2 \le B_*^2$.

Lemma F.9 applies with the variance bound $\sigma_0^2 = v_w$ in that lemma. We multiply (142) by $\mathbb{1}_{\{\mathcal{E}\}}$ and take expectations. Since $\mathcal{E}$ is $\mathscr{F}_2$-measurable, the tower property identifies the left side as $\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{\mathcal{E}\}}\right]$. Removing the event indicator from the nonnegative terms on the right gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{\mathcal{E}\}}\right] \overset{②}{\le} \mathbb{E}\left[3\|r\|_{L^2}^2 + \frac{3(1+e^{-1})}{m}\|\boldsymbol{a}_\lambda\|_2^2 + 9\|r\|_{L^2(P_{n_2})}^2 + \frac{6v_w}{n_2}\widehat{\mathcal{N}}(m^{-1})\right]. \tag{58}$$

In ②, the tower property and nonnegativity remove the event indicator. This is necessary before averaging the empirical residual, since conditioning on $\mathcal{E}$ need not preserve the input law. Given $\mathscr{F}_1$, the jointly measurable residual is fixed and the inputs in the second training sample have law P. Its squared population norm is at most B_*^2, so conditional integration gives

$$\mathbb{E}\left[\|r\|_{L^2(P_{n_2})}^2 \mid \mathscr{F}_1\right] \overset{③}{=} \frac{1}{n_2}\sum_{i\in I_2}\int_{\mathcal{X}} r(\mathbf{x})^2 \mathrm{d}\mu(\mathbf{x}) = \|r\|_{L^2}^2.$$

Equality ③ uses the conditional input law and Tonelli's theorem. Averaging (58) and evaluating the population objective gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{\mathcal{E}\}}\right] \overset{④}{\le} 12\mathbb{E}\left[\|r\|_{L^2}^2 + m^{-1}\|\boldsymbol{a}_\lambda\|_2^2 + \tfrac{v_w}{n}\widehat{\mathcal{N}}(m^{-1})\right] \overset{⑤}{=} 12\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K}) + \tfrac{v_w}{n}\widehat{\mathcal{N}}(m^{-1})\right].$$

In ④, we use ③, the tower property, $3(1+\exp(-1)) < 12$, and $n_2 \ge n/2$. Equality ⑤ evaluates the penalized population objective at its minimizer.

Step 3. Averaging on the complement. On the exceptional event, we replace covariance control by a deterministic GD bound and a fourth-moment bound on the target. The event $\mathcal{E}^c$ is $\mathscr{F}_2$-measurable. Lemma F.8, specifically (132), therefore gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{\mathcal{E}^c\}}\right] \overset{⑥}{\le} 3mB_*^2\Pr\left[\mathcal{E}^c\right] + \frac{5m}{2}\sqrt{\frac{(\|f^*\|_{L^4}^4 - B_*^4)\Pr\left[\mathcal{E}^c\right]}{n_2}} + \frac{5mv_w}{4n_2}\Pr\left[\mathcal{E}^c\right].$$

Inequality ⑥ applies (132) and bounds $1 - \Pr\left[\mathcal{E}^c\right]$ by one. Its right side is increasing in $\Pr\left[\mathcal{E}^c\right]$, so Steps 1 and 2 give (56). For polynomial targets, substitution of $\|f^*\|_{L^4}^2 \le 3^L B_*^2$ from Lemma E.6 and $v_w \le \sigma_0^2$ gives (57).

Under (49), Lemmas F.5 and F.6 give $\delta_{d,n,m} \le 2n^{-8}$. Using $m \le n$ and $n_2 \ge n/2$, the target terms are at most $6B_*^2 n^{-7} + 5\sqrt{\|f^*\|_{L^4}^4 - B_*^4}\, n^{-7/2} \le 11\|f^*\|_{L^4}^2 n^{-7/2}$, and the noise term is at most $5v_w n^{-8}$. The same fourth-moment bound gives the polynomial remainder. No fourth response moment or second moment of the risk is needed.

□

The next corollary uses $\Gamma_*^2 \le r_0 B^2$ to obtain a bound uniform over the full degree-bounded class with $B_* \le B$. Proposition G.1 explains why this class permits nonvanishing energy in every fixed degree.

Corollary C.3. Under the assumptions of Theorem C.1, for every fixed $B \ge 0$ and every target of degree at most ℓ_0 with $\|f^*\|_{L^2} \le B$,

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le C_L\left(\frac{B^2 r_0}{m} + \frac{\sigma_0^2 m}{n} + \frac{B^2 + \sigma_0^2}{n^{7/2}}\right). \tag{59}$$

Proof. **Outline.** We bound the weighted norm by the polynomial-space dimension times the squared L^2 norm, then substitute into the theorem. By (31), $\Gamma_*^2 \overset{①}{\le} r_0 B_*^2 \overset{②}{\le} r_0 B^2$, where ① uses $N(d,\ell) \le r_0$ and ② uses $B_* \le B$. Substitution into (52) proves (59). □

Under the theorem's conditions, $1 \le m \le n$ and $r_0 \ge 1$, so $n^{-7/2} \le r_0/m$ and $n^{-7/2} \le m/n$. The remainder in (59) is therefore absorbed into its first two terms. Since $r_0 \asymp d^{\ell_0}$ by Lemma E.4, this gives the introduction's order $d^{\ell_0}/m + m/n$ for fixed target and noise bounds.

For the default rule (33), the same substitution gives, under the theorem's assumptions and for sufficiently large n,

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le C_L \left(\frac{B_*^2 d^{\ell_0} H_n^2}{n} + \frac{\sigma_0^2}{H_n^2} \right). \tag{60}$$

Indeed, $n/H_n^2 \le m < n/H_n^2 + 4$ and $\Gamma_*^2 \le C_L d^{\ell_0} B_*^2$. These inequalities bound the first two terms of (52), while $n^{-7/2} \le H_n^2/n$ and $n^{-7/2} \le H_n^{-2}$ absorb its remainder. Equation (60) retains the theorem's width conditions and expectation over all training randomness.

The next corollary gives the consistency ranges for weighted-norm targets, bounded-L^2 targets, and the default width rule. Its constants are independent of n, d.

Corollary C.4 (Joint consistency regimes)**.** Let $d \to \infty$, let $1 \le \ell_0 \le L$ be fixed, and retain the model, target expansion, initialization, and training assumptions of Theorem C.1. The targets and noise laws may vary with d. Their stated norm and variance bounds are fixed uniformly along the sequence. Put $H_n = \log(en)$ and take $C_0 = 512$.

(i) Fixed weighted-norm targets. Suppose $\Gamma_* \le \gamma_0$ and the noise variance is at most σ_0^2, where γ_0, σ_0 are fixed. Every deterministic width satisfying (49) eventually gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \longrightarrow 0, \qquad \mathcal{R}(\widehat{f}) \longrightarrow 0 \quad \text{in probability as } d \to \infty. \tag{61}$$

Such a width exists if and only if

$$4\left\lceil \frac{C_0(d + H_n)}{4} \right\rceil \le \frac{n}{C_0 H_n}. \tag{62}$$

As a result, (62), eventually along the sequence, is the complete condition on n, d needed to apply this weighted-class consequence of the theorem. In particular, for any fixed $A > C_0^2$, the choices

$$n = \lceil Ad \log d \rceil, \qquad m = 4\left\lceil \frac{C_0(d + H_n)}{4} \right\rceil \asymp d \tag{63}$$

are admissible for all sufficiently large d and satisfy

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le C_{L,A}\left(\frac{\gamma_0^2}{d} + \frac{\sigma_0^2}{\log d} \right) \longrightarrow 0. \tag{64}$$

Here $C_{L,A}$ depends only on the fixed L, A. Training uses n, d in this width choice and receives no target degree.

(ii) Full bounded-L^2 targets. Suppose $B_* \le B$ and the noise variance is at most σ_0^2, with fixed $B > 0$ and $\sigma_0 \ge 0$. Assume

$$\frac{n}{d^{\ell_0} \log d} \longrightarrow \infty, \qquad \text{equivalently} \qquad \frac{n}{d^{\ell_0} H_n} \longrightarrow \infty. \tag{65}$$

An alternative degree-independent width rule, permitted by Theorem C.1, is

$$m = 4\left\lfloor \frac{n}{4C_0 H_n} \right\rfloor \sim \frac{n}{C_0 H_n}. \tag{66}$$

This rule is used for all sufficiently large d, when its value is positive. All other training parameters remain unchanged. It satisfies (49) eventually and gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \le C_L\left(\frac{B^2 d^{\ell_0} H_n}{n} + \frac{\sigma_0^2}{H_n} + \frac{B^2 + \sigma_0^2}{n^{7/2}} \right) \longrightarrow 0. \tag{67}$$

The probability conclusion in (61) follows as well. Condition (65) is equivalent to the existence of an admissible deterministic width with $r_0/m \to 0$. It is therefore the weakest sample-dimension condition for making the coarse uniform bounded-L^2 risk bound vanish by choosing a width inside (49). This is not a necessary sample-size condition for learning.

(iii) The original width rule. The rule (33) remains valid. Under that rule, part (ii)'s consistency conclusions hold if

$$\frac{n}{d^{\ell_0} H_n^2} \longrightarrow \infty. \tag{68}$$

For the weighted class in part (i), the weaker sufficient condition $\liminf_{d\to\infty} n/(dH_n^2) > C_0$ already makes that original rule admissible and consistent. Neither class requires $n/d^{\ell_0+1} \to 0$ for consistency. For the bounded-L^2 class, both width choices give the first limit below. If $n/d^{\ell_0+1} \to 0$ is additionally imposed, they also give the second:

$$\frac{m}{d^{\ell_0}} \longrightarrow \infty, \qquad \frac{m}{d^{\ell_0+1}} \longrightarrow 0. \tag{69}$$

The second limit is a width ceiling, not a requirement for vanishing risk.

Proof. **Outline.** Step 1 applies the weighted-class bound to any admissible width and identifies exactly when such a width exists. Step 2 verifies the alternative width rule (66) and the full bounded-L^2 rate. Step 3 characterizes when an admissible width can make the coarse approximation term vanish. Step 4 treats the original width rule, the optional width ceiling, and probability convergence.

Step 1. The weighted class. For the weighted class, feasibility alone makes the approximation and noise terms vanish. Recall that $d \to \infty$, $H_n = \log(en)$, and admissibility means (49). Its lower bound gives $m \geq C_0 d \to \infty$, and its upper bound gives $m/n \leq 1/(C_0 H_n)$. These inequalities also imply $n \to \infty$, so $H_n \to \infty$. With fixed γ_0, σ_0, all three terms on the right of (53) therefore vanish. The least multiple of four above the lower width threshold is $4\lceil C_0(d + H_n)/4\rceil$. It lies in the width interval exactly when (62) holds.

For (63), one has $H_n/\log d \to 1$, $m/d \to C_0$, and $n/(C_0 H_n d) \to A/C_0 > C_0$. These limits verify both width inequalities eventually, including integer rounding. They give $\gamma_0^2/m = O(\gamma_0^2/d)$ and $\sigma_0^2 m/n = O_A(\sigma_0^2/\log d)$. The remainder in (53) is absorbed into their sum with a constant depending only on L, A, which proves (64). The dimension condition also holds eventually, since d_L is fixed.

Step 2. The alternative width for bounded-energy targets. We verify the logarithmic equivalence in (65) before choosing a width near the permitted upper bound. Either condition eventually implies $n \geq d^{\ell_0}$, where $H_n \geq \ell_0 \log d$, so the condition with H_n implies the one with $\log d$. Conversely, suppose $n/(d^{\ell_0} \log d) \to \infty$. On indices with $n \leq d^{\ell_0+1}$, the bound $H_n \leq 1 + (\ell_0 + 1)\log d$ gives the condition with H_n. On the remaining indices, monotonicity of $x/\log(ex)$ for $x > 1$ gives $\frac{n}{d^{\ell_0} H_n} \overset{①}{\geq} \frac{d}{1+(\ell_0+1)\log d} \longrightarrow \infty$. Inequality ① evaluates that increasing function at $x = d^{\ell_0+1}$. Together the two index sets prove the equivalence.

We now take the width in (66). The rounding operation gives $n/(C_0 H_n) - 4 < m \leq n/(C_0 H_n)$. Condition (65) implies $m \sim n/(C_0 H_n)$, $m/d^{\ell_0} \to \infty$, and $m/d \to \infty$. Also $m/H_n \sim n/(C_0 H_n^2) \to \infty$, since every fixed power of $\log(en)$ is $o(n)$. As a result, $m/(d + H_n) \to \infty$, proving the lower width condition, while the upper one holds by construction.

Lemma E.4 gives $r_0 \leq C_L d^{\ell_0}$. Substituting this and the bounds on m into (59) proves (67), with constants independent of n, d. Every term vanishes under (65).

Step 3. The scope of the condition. The upper width constraint gives the converse. If an admissible width satisfies $r_0/m \to 0$, then $\frac{r_0 H_n}{n} \overset{②}{\leq} \frac{r_0}{C_0 m} \to 0$, where ② uses $C_0 m H_n \leq n$. Since $r_0 \asymp d^{\ell_0}$, condition (65) follows. Step 2 constructs a width under this condition, proving the claimed equivalence for the theorem's coarse approximation bound.

Step 4. The original rule and the remaining limits. We now check the original width separately because it spends one additional logarithmic factor on noise control. Under (33), rounding gives $n/H_n^2 \leq m < n/H_n^2 + 4$. Condition (68) implies $m \sim n/H_n^2$, $m/d \to \infty$, $m/H_n \to \infty$, and $mH_n/n \to 0$. Therefore both finite-sample width inequalities hold eventually. Moreover, $r_0/m = O_L(d^{\ell_0} H_n^2/n) \to 0$ and $m/n \to 0$, so (59) proves consistency.

For the weighted class, $\liminf n/(dH_n^2) > C_0$ gives $m/d \geq C_0 + \varepsilon$ eventually for some fixed $\varepsilon > 0$. Also $H_n/m \leq H_n^3/n \to 0$. Consequently $C_0(d + H_n)/m \leq C_0/(C_0 + \varepsilon) + o(1) < 1$ eventually, which proves the lower width condition without restricting how quickly n grows relative to d. Its upper width condition follows from $mH_n/n \to 0$, and Step 1 proves weighted consistency.

For both bounded-L^2 width choices, the already established $m/d^{\ell_0} \to \infty$ gives the first limit in (69). If $n/d^{\ell_0+1} \to 0$, the fact that $m \leq n$ eventually gives the second limit. For every case above

and every $\varepsilon > 0$, Markov's inequality gives $\Pr\left[\mathcal{R}(\widehat{f}) > \varepsilon\right] \overset{③}{\leq} \mathbb{E}\left[\mathcal{R}(\widehat{f})\right]/\varepsilon \to 0$, where ③ uses nonnegativity of the risk, proving (61). The architecture retains its chosen width throughout training. □

For $n \asymp d^{\ell_0}(\log d)^p$, the alternative rule (66) is consistent for every fixed $p > 1$. For the default rule (33), retain the observation and target assumptions of Corollary C.4, with $B_* \leq B$ for fixed $B \geq 0$ and bounded σ_0, and take any fixed $p > 2$. Then $H_n \asymp \log d$, $m \asymp d^{\ell_0}(\log d)^{p-2}$, and $n/(d^{\ell_0} H_n^2) \to \infty$. Part (iii) of that corollary applies, and substitution into (59) gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] = O_{L,p}\left(B^2(\log d)^{2-p} + \sigma_0^2(\log d)^{-2}\right). \tag{70}$$

The remainder is absorbed into these two terms. The constant in (70) may also depend on the fixed comparison constants in the choice of n. The algorithm receives neither ℓ_0 nor p.

We next distinguish prediction consistency from stronger assertions about rates and learned eigenspaces.

Remark C.5. The theorem gives an expected-risk upper bound and consistency in probability. It does not assert a matching rate, almost-sure convergence, or recovery of a leading harmonic eigenspace of the learned feature kernel. The risk proof uses $\widehat{\mathcal{N}}(m^{-1}) \leq m$, with no bound of order r_0. Section G.4 compares target assumptions and resource budgets. All results concern the specified staged GD with fixed input directions, rather than simultaneous training of every layer.

D Feature Learning and Kernel Evolution

D.1 The Learned Representation and Its Feature Kernel

We recall the biased ReLU features ϕ_j, their duplicated vector $\mathbf{h}$, and the normalization $\mathbf{z} = \mathbf{h}/\sqrt{m}$. At zero attention, $\boldsymbol{\psi}_0 = \mathbf{h}/\sqrt{2m}$, so the initial feature kernel is $K_0 = K_{\mathrm{RF}}/2$, where $K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}') := \frac{2}{m}\sum_{j=1}^{m/2} \phi_j(\mathbf{x})\phi_j(\mathbf{x}') = \mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}')$. The temporary output initialization does not affect this kernel. Here the feature span is the space of all real linear combinations of the coordinate functions.

The score learned from the first training sample. We recall the channel average $b = (2/m)\sum_{j=1}^{m/2} \phi_j$, the selected initialization $\widehat{c} \in \{0, 1\}$, and the corresponding gradient $\mathbf{g} = \nabla L_1(\mathbf{0})$. Stage one computes $\widehat{\boldsymbol{\tau}} = -\eta_1 \mathbf{g}$, with the positive learning rate in (37). Taking its inner product with $\mathbf{z}(\mathbf{x})$ gives

$$\widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\mathbf{x}) \overset{①}{=} \frac{\eta_1}{2n_1} \sum_{i \in I_1} \big(y_i - \widehat{c}\, b(\vec{\mathbf{x}}_i)\big) b(\vec{\mathbf{x}}_i) K_{\mathrm{RF}}(\vec{\mathbf{x}}_i, \mathbf{x}). \tag{71}$$

Equality ① substitutes the first-stage update and uses $K_{\mathrm{RF}}(\vec{\mathbf{x}}_i, \mathbf{x}) = \mathbf{z}(\vec{\mathbf{x}}_i)^\top \mathbf{z}(\mathbf{x})$. It expresses the score as a response-weighted combination of initial feature kernel sections. The score in (71) lies in the initial ReLU span. Its sigmoid, multiplied by the channels, supplies the added features. The update is finite and measurable, with $\widehat{\tau}_j = \widehat{\tau}_{m/2+j}$ for $j \in [m/2]$.

Proposition D.1 proves $\|\mathbf{g}\|_2 \geq 1/32$, $0 < \|\widehat{\boldsymbol{\tau}}\|_2 < 1/[2\sqrt{2(d+1)}]$, and $\sup_{\mathbf{x}} \left|\widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\mathbf{x})\right| < 1/2$. As a result, the bottleneck remains active and $g_{\widehat{\boldsymbol{\tau}}} = s(\widehat{\boldsymbol{\tau}}^\top \mathbf{z})$, where s is the sigmoid from Section A.1. That proposition also proves strict descent and nonconstancy of the score, which do not follow from its size bound alone.

Kernel change across the two stages of training. Define the learned gate deviation $a_g = g_{\widehat{\boldsymbol{\tau}}} - 1/2$ and the learned feature kernel $\widehat{K}$ through the frozen feature vector $\boldsymbol{\psi}$ in (39). Taking the inner product of the two attention-weighted feature vectors gives

$$\widehat{K}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}') \overset{①}{=} \frac{1}{2} K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}') + 2a_g(\mathbf{x}) a_g(\mathbf{x}') K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}'). \tag{72}$$

Equality ① uses $g_{\widehat{\boldsymbol{\tau}}} = 1/2 + a_g$ and $1 - g_{\widehat{\boldsymbol{\tau}}} = 1/2 - a_g$. The two channel products sum to $1/2 + 2a_g(\mathbf{x})a_g(\mathbf{x}')$. The increment is the kernel of the $m/2$ functions $\frac{2}{\sqrt{m}} a_g \phi_j$, so it is positive semidefinite. For a kernel K here, its integral operator sends $f \in L^2(P)$ to $\mathbf{x} \mapsto$

$\int_{\mathcal{X}} K(\mathbf{x},\mathbf{x}')f(\mathbf{x}')\mathrm{d}\mu(\mathbf{x}')$. These operators are bounded because the feature functions are bounded. Consequently, (72) gives an ordering of both the finite Gram matrices and the integral operators through their quadratic forms. It does not require the increment to be nonnegative at every pair of distinct inputs.

The learned operator has range in the span of the m coordinates of $\boldsymbol{\psi}$, hence rank at most m. Its eigenfunctions need not be spherical harmonics.

The score bound and $|s'| \le 1/4$ give

$$\sup_{\mathbf{x}\in\mathcal{X}} |a_g(\mathbf{x})| < \frac{1}{8}, \qquad \frac{3}{8} < g_{\widehat{\boldsymbol{\tau}}}(\mathbf{x}) < \frac{5}{8} \quad (\mathbf{x}\in\mathcal{X}). \tag{73}$$

Since $K_{\mathrm{RF}}(\mathbf{x},\mathbf{x}) \ge b(\mathbf{x})^2 \ge 1/4$, equation (72) implies

$$1 \le \frac{\widehat{K}(\mathbf{x},\mathbf{x})}{K_{\mathrm{RF}}(\mathbf{x},\mathbf{x})/2} \overset{①}{=} 1 + 4a_g(\mathbf{x})^2 \overset{②}{<} \frac{17}{16}. \tag{74}$$

Equality ① divides the diagonal kernel decomposition by the positive initial diagonal, and inequality ② uses (73). Proposition D.1 proves that the ratio in (74) is nonconstant, excluding every constant rescaling of K_{RF}.

For the population size of the change, set $c_d = 57/[1024(d+1)]$. The same proposition establishes

$$0 < \int_{\mathcal{X}} \|\boldsymbol{\psi} - \boldsymbol{\psi}_0\|_2^2 \mathrm{d}\mu = \int_{\mathcal{X}} \left(\widehat{K}(\mathbf{x},\mathbf{x}) - \frac{1}{2}K_{\mathrm{RF}}(\mathbf{x},\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x}) \le c_d. \tag{75}$$

Equation (75) quantifies the change in the learned representation while retaining a uniform feature-energy bound.

Strict enlargement and output coordinates. The output coordinates described below connect feature enlargement to the risk proof. They retain the polynomial comparison predictor and its coefficient norm, while allowing output GD to use the added products.

Write ψ_j for the jth coordinate of $\boldsymbol{\psi}$. For $j \in [m/2]$, the identities $\psi_j + \psi_{m/2+j} = \phi_j/\sqrt{m/2}$ and $\psi_j - \psi_{m/2+j} = 2a_g\phi_j/\sqrt{m/2}$ give

$$f(\widehat{\boldsymbol{\tau}}, \boldsymbol{a}, \mathbf{x}) = \sqrt{\frac{2}{m}} \sum_{j=1}^{m/2} \phi_j(\mathbf{x}) \left[\frac{a_j^+ + a_j^-}{2} + a_g(\mathbf{x})(a_j^+ - a_j^-)\right]. \tag{76}$$

Consequently, the learned span is $\operatorname{span}\{\phi_j, a_g\phi_j : j \in [m/2]\}$. Proposition D.1 proves the strict inclusion with the witness

$$a_g b \in \operatorname{span}\{\psi_j : j \in [m]\} \setminus \operatorname{span}\{\phi_j : j \in [m/2]\}. \tag{77}$$

The coefficients $a_j^+ = 1/\sqrt{2m}$ and $a_j^- = -1/\sqrt{2m}$ represent the witness in (76). Step 5 of the proof of Proposition D.1 excludes a representation by the initial features by restricting to a region where every ReLU has a fixed activation status. Continuity gives (77) both pointwise and in $L^2(P)$. The dimension increases by at least one and at most the initial dimension, since multiplication by a_g is linear on the initial span. The product coordinates need not be independent.

Define $\boldsymbol{\psi}_+ = [\phi_j]_{j=1}^{m/2}/\sqrt{m}$ and $\boldsymbol{\psi}_- = 2a_g\boldsymbol{\psi}_+$. The orthogonal matrix

$$\mathbf{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{I}_{m/2} & \mathbf{I}_{m/2} \\ \mathbf{I}_{m/2} & -\mathbf{I}_{m/2} \end{pmatrix} \quad \text{satisfies} \quad \mathbf{U}\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\psi}_+ \\ \boldsymbol{\psi}_- \end{pmatrix}, \qquad \mathbf{U}\boldsymbol{\psi}_0 = \begin{pmatrix} \boldsymbol{\psi}_+ \\ \mathbf{0} \end{pmatrix}. \tag{78}$$

In (78), the first block retains the initial features and the second supplies their learned products. Orthogonality preserves predictions and coefficient norms: $(\mathbf{U}\boldsymbol{a})^\top \mathbf{U}\boldsymbol{\psi} = \boldsymbol{a}^\top\boldsymbol{\psi}$ and $\|\mathbf{U}\boldsymbol{a}\|_2 = \|\boldsymbol{a}\|_2$. For initial output blocks $\boldsymbol{a}^+, \boldsymbol{a}^- \in \mathbb{R}^{m/2}$, replacing both by $(\boldsymbol{a}^+ + \boldsymbol{a}^-)/2$ preserves the initial predictor under every learned gate and gives

$$\frac{1}{2}\left\|\boldsymbol{a}^+ + \boldsymbol{a}^-\right\|_2^2 \overset{①}{=} \|\boldsymbol{a}\|_2^2 - \frac{1}{2}\left\|\boldsymbol{a}^+ - \boldsymbol{a}^-\right\|_2^2 \le \|\boldsymbol{a}\|_2^2.$$

Equality ① is the parallelogram identity. Equality in the norm bound holds exactly when the blocks agree. Every minimum-norm initial representation has this form because its difference block represents zero, so those coefficients retain the predictor under every learned gate.

The Euclidean orthogonality in (78) does not assert orthogonality of the two groups of functions in $L^2(P)$. The algorithm itself continues to update the original coordinates.

D.2 Strict Enlargement for the Performed Attention Update

The next proposition proves that the specified update creates a nonconstant score and a gated channel average outside the initial span. It also quantifies the resulting kernel change. These deterministic conclusions supply part (i) of Theorem C.1.

Proposition D.1. Let $d \geq 3$, let $m \geq 4$ be divisible by four, and use the opposite directions in (25), learnable channel attention, and the initialization and single attention GD step from stage one of Algorithm 1, detailed in Section B. For every realization of the directions and every first training sample of size $n_1 \geq 1$ with inputs on $\mathcal{X}$ and finite real responses, the initial attention gradient and updated parameter satisfy

$$\|\mathbf{g}\|_2 \geq \frac{1}{32}, \qquad \|\widehat{\boldsymbol{\tau}}\|_2 \geq \frac{\eta_1}{32} > 0, \qquad \sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\mathbf{x})\right| < \frac{1}{2}. \tag{79}$$

The training loss in stage one strictly decreases:

$$L_1(\widehat{\boldsymbol{\tau}}) \leq L_1(\mathbf{0}) - \frac{\eta_1}{2}\|\mathbf{g}\|_2^2 \leq L_1(\mathbf{0}) - \frac{\eta_1}{2048}. \tag{80}$$

The learned gate $g_{\widehat{\boldsymbol{\tau}}}$ is nonconstant on the sphere. The function $a_g b$ satisfies (77), so the learned prediction space strictly contains the initial ReLU span, also as subspaces of $L^2(P)$. Here $a_g = g_{\widehat{\boldsymbol{\tau}}} - 1/2$ and $b = (2/m)\sum_{j=1}^{m/2} \phi_j$. With K_{RF} and a_g defined in Section D.1, the kernel identity (72) holds, and $\widehat{K} - K_{\mathrm{RF}}/2$ is a nonzero positive semidefinite kernel. More strongly,

$$\widehat{K} \neq cK_{\mathrm{RF}} \quad \text{for every constant } c \in \mathbb{R}. \tag{81}$$

Here inequality means that the two kernels differ at some pair of inputs. The same assertion holds for their integral operators on $L^2(P)$. The attention score satisfies (71). These assertions require no condition on the target coefficients, noise variance, or sample-to-dimension ratio. They hold for every finite first training sample, without discarding a probability-zero event. They also give the uniform gate bound (73) and diagonal comparison (74). The population feature change satisfies (75), and the operator norm of the kernel increment is at most c_d. In particular, these population measures tend to zero uniformly over the first-stage outcomes. No dimension-independent positive lower bound in these measures is possible for the specified learning rate.

Proof. **Outline.** Step 1 uses opposite ReLU directions and the two candidate output initializations to bound the chosen gradient away from zero. Step 2 checks the learning rate, proves strict loss decrease, and shows that the bottleneck ReLU remains active throughout the step. Step 3 proves that the attention score, and hence the learned gate, cannot be constant on the sphere. Step 4 uses the exact kernel identity to prove a nonzero change that cannot be a constant rescaling, bounds its diagonal ratio above, and verifies the corresponding statement for population integral operators. Step 5 proves that the learned prediction space is strictly larger by excluding an affine representation of the gated channel average on an open part of the sphere, then bounds the possible dimension increase. Step 6 combines the attention-parameter norm bound with the fourth moment of the paired channel vector to bound the population size of the change, then bounds its integral-operator norm.

Step 1. A nonzero attention gradient for every first training sample. Opposite directions provide a positive channel average, which prevents both candidate gradients from vanishing through response cancellation. We recall $b = (2/m)\sum_{j=1}^{m/2} \phi_j$ and $\mathbf{z} = \mathbf{h}/\sqrt{m}$, where $\mathbf{h}$ duplicates the $m/2$ biased ReLU features as in (27). For $j \leq m/4$, the opposite directions give

$$\phi_j(\mathbf{x}) + \phi_{j+m/4}(\mathbf{x}) = \sigma\big(\sqrt{d}\,\vec{\mathbf{q}}_j^{\,\top}\mathbf{x} + 1/2\big) + \sigma\big(-\sqrt{d}\,\vec{\mathbf{q}}_j^{\,\top}\mathbf{x} + 1/2\big) \overset{①}{\geq} 1.$$

Inequality ① follows from $\sigma(t) \geq t$ applied to the two arguments, whose sum is one. As a result, $b(\mathbf{x}) \geq 1/2$. Define the unit vector $\mathbf{u} := \mathbf{1}_m/\sqrt{m}$. Then $\mathbf{u}^\top \mathbf{z}(\mathbf{x}) = b(\mathbf{x})$. Subtracting the two gradients in (35) gives

$$\mathbf{u}^\top(\mathbf{g}_1 - \mathbf{g}_0) \overset{②}{=} \frac{1}{2n_1}\sum_{i\in I_1} b(\vec{\mathbf{x}}_i)^3 \overset{③}{\geq} \frac{1}{16}, \qquad \max\{\|\mathbf{g}_0\|_2, \|\mathbf{g}_1\|_2\} \overset{④}{\geq} \frac{1}{2}\|\mathbf{g}_1 - \mathbf{g}_0\|_2 \overset{⑤}{\geq} \frac{1}{32}.$$

Equality ② substitutes the gradient difference and $\mathbf{u}^\top \mathbf{z} = b$. Inequality ③ uses $b \geq 1/2$, ④ is the triangle inequality, and ⑤ uses $\|\mathbf{u}\|_2 = 1$ and Cauchy–Schwarz. The initialization rule (36) selects this maximum. The finite positive denominator in (37) gives $\eta_1 > 0$, so (38) proves the first two bounds in (79).

Step 2. A smooth step and strict empirical descent. We keep the bottleneck away from its kink so that a Hessian bound can certify a genuine loss decrease along the entire update. Put $R_d \coloneqq \sqrt{2(d+1)}$ in this proof. The channel definition gives $\|\mathbf{z}(\mathbf{x})\|_2^2 = (2/m)\sum_j \phi_j(\mathbf{x})^2 \le R_d^2$ for every input. For $0 \le t \le 1$, the line segment from zero to the GD update therefore satisfies $\sup_{\mathbf{x}\in\mathcal{X}} \left|(t\widehat{\boldsymbol{\tau}})^\top \mathbf{z}(\mathbf{x})\right| \overset{⑥}{\le} \eta_1 R_d \|\mathbf{g}\|_2 \overset{⑦}{<} \frac{1}{2}$. Inequality ⑥ is Cauchy–Schwarz and $t \le 1$. The strict inequality ⑦ follows from the term $2R_d\|\mathbf{g}\|_2$ and the positive constant 1 in the denominator of η_1. This proves the last assertion in (79). The bottleneck argument $1 + (t\widehat{\boldsymbol{\tau}})^\top \mathbf{z}(\mathbf{x})$ exceeds $1/2$, so its ReLU is affine along this entire segment. Consequently $g_{t\widehat{\boldsymbol{\tau}}}(\mathbf{x}) = s((t\widehat{\boldsymbol{\tau}})^\top \mathbf{z}(\mathbf{x}))$, and the loss is twice continuously differentiable in a neighborhood of the segment.

For a fixed input in the first training sample, we use the abbreviations $b = b(\vec{\mathbf{x}}_i)$, $\mathbf{z} = \mathbf{z}(\vec{\mathbf{x}}_i)$, $y = y_i$, and $f = f(\boldsymbol{\tau}, \boldsymbol{a}_{\text{init}}, \vec{\mathbf{x}}_i)$. All gradients and Hessians in this step are with respect to $\boldsymbol{\tau}$. On the segment, $f = b(\widehat{c} + 2s(\boldsymbol{\tau}^\top \mathbf{z}) - 1)$. The sigmoid bounds $0 < s < 1$, $0 < s' \le 1/4$, and $|s''| \le 1/4$ imply

$$|f| \overset{⑧}{\le} 2b, \qquad \|\nabla f\|_2 \overset{⑨}{\le} \frac{b}{2}\|\mathbf{z}\|_2, \qquad \left\|\nabla^2 f\right\|_2 \overset{①}{\le} \frac{b}{2}\|\mathbf{z}\|_2^2.$$

In ⑧, we use $\widehat{c} \in \{0, 1\}$ and $0 < s < 1$. The chain rule and $s' \le 1/4$ give ⑨, and the rank-one Hessian with $|s''| \le 1/4$ gives ①. The Hessian of $(f-y)^2/2$ equals $\nabla f(\nabla f)^\top + (f-y)\nabla^2 f$. Its operator norm is therefore at most $(5b^2/4 + |y|\, b/2)\|\mathbf{z}\|_2^2$. Averaging shows that the Hessian norm of L_1 is at most Λ_1 along the segment. Taylor's formula with integral remainder gives

$$L_1(-\eta_1 \mathbf{g}) \overset{②}{\le} L_1(\mathbf{0}) - \eta_1\|\mathbf{g}\|_2^2 + \frac{\Lambda_1\eta_1^2}{2}\|\mathbf{g}\|_2^2 \overset{③}{\le} L_1(\mathbf{0}) - \frac{\eta_1}{2}\|\mathbf{g}\|_2^2.$$

Inequality ② uses the Hessian bound in the integral remainder. Inequality ③ uses $\eta_1\Lambda_1 < 1$, which follows from (37). Since Step 1 shows that the gradient is nonzero, this proves the strict decrease in (80).

The bound $|s'| \le 1/4$, $s(0) = 1/2$, and the uniform score bound give (73).

Step 3. A nonconstant attention score and gate. A nonzero parameter update need not change a function, so we must also rule out a constant or identically zero attention score. We first prove an elementary property of the biased ReLU functions on the sphere. A finite real linear combination of functions $\mathbf{x} \mapsto \sigma(\sqrt{d}\,\mathbf{q}^\top\mathbf{x} + 1/2)$, with unit directions $\mathbf{q}$, can be constant on $\mathcal{X}$ only if it is identically zero. We group identical directions and let $A_{\mathbf{q}}$ denote the sum of the coefficients for one distinct direction. Its ReLU changes derivative on $S_{\mathbf{q}} \coloneqq \left\{\mathbf{x} \in \mathcal{X} : \mathbf{q}^\top\mathbf{x} = -1/(2\sqrt{d})\right\}$. This is a sphere of dimension $d-2 \ge 1$ with positive radius, whose affine hull is the displayed hyperplane. For any other distinct unit direction $\mathbf{q}'$, the intersection $S_{\mathbf{q}} \cap S_{\mathbf{q}'}$ has empty relative interior in $S_{\mathbf{q}}$.

To justify this assertion, write a point of $S_{\mathbf{q}}$ as $-\mathbf{q}/(2\sqrt{d}) + r\mathbf{v}$, where $r = \sqrt{1 - 1/(4d)} > 0$ and $\mathbf{v}$ is a unit vector in the $(d-1)$-dimensional space perpendicular to $\mathbf{q}$. A nonconstant affine function of $\mathbf{v}$ cannot vanish on an open subset of that unit sphere. Indeed, its tangential derivatives would force its linear coefficient to be parallel to every $\mathbf{v}$ in that open subset, which contains two nonparallel vectors because $d-2 \ge 1$. That coefficient must then be zero, and the affine function is constant. Consequently, an affine linear equation holding on an open subset of $S_{\mathbf{q}}$ holds on its entire affine hull. Coincidence of the two hyperplanes would require $\mathbf{q}' = \mathbf{q}$, because their unit normals have the same nonzero offset. More explicitly, restricting the equation for $S_{\mathbf{q}'}$ to the affine hull of $S_{\mathbf{q}}$ gives either a nonzero constant or a nonconstant affine function on that sphere. Its zero set has empty interior.

Each such intersection is closed. A finite union of these closed sets cannot cover $S_{\mathbf{q}}$, since the complements are open dense subsets whose finite intersection is dense and nonempty.

Choose $\mathbf{x}^\circ \in S_{\mathbf{q}}$ outside all the other sets. Let $\mathbf{v} \coloneqq \mathbf{q} - (\mathbf{q}^\top\mathbf{x}^\circ)\mathbf{x}^\circ$. Then $\mathbf{v}$ is a nonzero tangent vector to $\mathcal{X}$ at $\mathbf{x}^\circ$ and $\mathbf{q}^\top\mathbf{v} = \|\mathbf{v}\|_2^2 > 0$. The curve $\mathbf{x}(t) \coloneqq (\mathbf{x}^\circ + t\mathbf{v})/\|\mathbf{x}^\circ + t\mathbf{v}\|_2$ lies on the sphere and crosses $S_{\mathbf{q}}$ transversely at zero. Every ReLU with a different direction is differentiable along this curve near zero. The derivative of $\sqrt{d}\,\mathbf{q}^\top\mathbf{x}(t) + 1/2$ at zero is $\sqrt{d}\,\mathbf{q}^\top\mathbf{v} > 0$. Hence this ReLU is inactive immediately to the left of zero and active immediately to the right. Its right derivative minus its left derivative, after multiplication by the grouped coefficient, is $A_{\mathbf{q}}\sqrt{d}\,\mathbf{q}^\top\mathbf{v}$. If

the linear combination is constant, its derivative has no jump. Therefore $A_{\mathbf{q}} = 0$. Repeating this argument for each distinct direction proves the claimed property.

Now consider the actual attention score $\mathbf{x} \mapsto \widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\mathbf{x})$. It is a finite linear combination of these biased ReLU functions. It is not identically zero. Indeed, taking the inner product of the update (38) with $\widehat{\boldsymbol{\tau}}$ gives

$$0 < \|\widehat{\boldsymbol{\tau}}\|_2^2 \overset{④}{=} \frac{\eta_1}{2n_1} \sum_{i \in I_1} \big(y_i - \widehat{c}\, b(\overrightarrow{\mathbf{x}}_i)\big) b(\overrightarrow{\mathbf{x}}_i) \widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\overrightarrow{\mathbf{x}}_i),$$

where ④ substitutes the exact attention update. Hence $\widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\overrightarrow{\mathbf{x}}_i) \neq 0$ for at least one $i \in I_1$. The property of finite ReLU combinations proved above now implies that this nonzero score is nonconstant. Step 2 gives $g_{\widehat{\boldsymbol{\tau}}}(\mathbf{x}) = s(\widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\mathbf{x}))$ throughout the sphere. Strict monotonicity of the sigmoid implies that the learned gate is nonconstant as well.

Step 4. Kernel evolution. The kernel identity separates preservation of the initial functions from the additional functions created by the learned gate. We recall $a_g = g_{\widehat{\boldsymbol{\tau}}} - 1/2$ and the initial random-feature kernel K_{RF} from Section D.1. For fixed inputs, write $g = g_{\widehat{\boldsymbol{\tau}}}(\mathbf{x})$ and $g' = g_{\widehat{\boldsymbol{\tau}}}(\mathbf{x}')$ in the following calculation. Multiplication of the two attention-weighted feature blocks gives

$$\widehat{K}(\mathbf{x}, \mathbf{x}') \overset{⑤}{=} \big[gg' + (1-g)(1-g')\big] K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}') = \big[1/2 + 2a_g(\mathbf{x}) a_g(\mathbf{x}')\big] K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}').$$

Equality ⑤ sums the inner products of the two feature blocks. Substituting $g = 1/2 + a_g(\mathbf{x})$ and $g' = 1/2 + a_g(\mathbf{x}')$ gives (72). Its additional term is a positive semidefinite kernel because it is the sum of the products of the functions $\frac{2}{\sqrt{m}}\, a_g \phi_j$ with themselves. The identity $\mathbf{z}(\mathbf{x})^\top \mathbf{z}(\mathbf{x}') = K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}')$, substituted in the attention update, also proves (71).

By Cauchy–Schwarz and Step 1, $K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}) = (2/m) \sum_j \phi_j(\mathbf{x})^2 \geq b(\mathbf{x})^2 \geq 1/4$ for every input.

To prove (81), suppose $\widehat{K} = cK_{\mathrm{RF}}$. Since $K_{\mathrm{RF}}(\mathbf{x}, \mathbf{x}) > 0$, the diagonal identity forces $a_g(\mathbf{x})^2 = (c - 1/2)/2$ everywhere. Continuity on the connected sphere then makes a_g constant, contrary to Step 3. The same diagonal identity and $|a_g| < 1/8$ prove (74), with strict increase wherever the attention score is nonzero.

To obtain the corresponding assertion for integral operators, we use the fact that continuous kernels on $\mathcal{X} \times \mathcal{X}$ define the same operator only when the kernels agree. To see this directly, let D be the difference of two such kernels and suppose $D(\mathbf{x}^\circ, \mathbf{x}'^\circ) \neq 0$. Continuity supplies relatively open neighborhoods U, V of these two inputs on which D has one sign and its absolute value is bounded below by a positive constant. Both neighborhoods have positive uniform measure. If the associated integral operator were zero on $L^2(P)$, its inner product between the indicator functions of V and U would be zero. This inner product equals $\int_U \int_V D(\mathbf{x}, \mathbf{x}') \mathrm{d}\mu(\mathbf{x}') \mathrm{d}\mu(\mathbf{x})$, which is nonzero by construction. This contradiction proves the operator assertion, both for the initial feature kernel and for every constant rescaling. All kernels considered here are continuous by their finite feature representations.

Step 5. Strict enlargement of the prediction space. To show that the extra kernel term adds a new function, we compare the nonlinear gated product with the locally affine form of every initial ReLU combination. We recall that $a_g = s(\widehat{\boldsymbol{\tau}}^\top \mathbf{z}) - 1/2$ and $b = (2/m) \sum_j \phi_j \geq 1/2$. The coordinate identities $\psi_j + \psi_{m/2+j} = \phi_j / \sqrt{m/2}$ and $\psi_j - \psi_{m/2+j} = 2a_g \phi_j / \sqrt{m/2}$ show that the learned span contains the initial span and the function $a_g b$. It remains to prove that $a_g b$ is outside the initial span. We prove this exclusion in three parts, first choosing an activation region, then extending a proposed identity, and finally obtaining a contradiction.

First, we show that the attention score is affine with a nonzero linear part on a nonempty relatively open subset U of the sphere where every ReLU has a fixed active or inactive status. The complement of the finitely many sets $S_{\mathbf{q}}$ from Step 3 is dense and open, and each nonempty region defined by strict choices of these statuses is open. On each such region every ReLU is either its affine argument or zero. As a result, the score is the restriction of an affine function there.

If this restriction were constant on every such region, its affine linear part on each nonempty region would be zero. This follows from the affine-function argument in Step 3, since a nonzero linear function cannot be constant on an open subset of a sphere. There are only finitely many active-status choices. The score would therefore take values in a finite set on the dense complement of the kink

sets, and by continuity on the whole sphere. A continuous function on a connected sphere with finite image is constant, contrary to Step 3. This proves the assertion about U.

Second, suppose, for a contradiction, that $a_g b$ belongs to the initial span. On U, let H, B, A be affine functions on $\mathbb{R}^d$ whose restrictions equal the score, the channel average, and this proposed ReLU representation, respectively. Here these three letters denote local affine extensions, and $\nabla H, \nabla B, \nabla A \in \mathbb{R}^d$ are their constant Euclidean gradients. The choice of U gives $\nabla H \neq \mathbf{0}$, and $B \geq 1/2$ on U. The proposed equality says $B(\mathbf{x})\big(s(H(\mathbf{x})) - 1/2\big) = A(\mathbf{x})$ for $\mathbf{x} \in U$. This identity extends to every point of the sphere for these same affine functions. Indeed, fix a point in U and any other point on the sphere. Choose a great circle through them, that is, the intersection of the sphere with a two-dimensional linear subspace. The difference of the two sides along this circle is real analytic in the circle parameter, because affine functions, sine, cosine, and the sigmoid are real analytic on their real domains. It vanishes on an open arc around the fixed point. The one-variable identity theorem for real analytic functions implies that it vanishes on the entire circle. Since the other point was arbitrary, the identity holds on all of $\mathcal{X}$. The analytic-continuation argument applies only to the affine extensions just defined. It does not assume that the actual ReLU functions are analytic across their kinks.

Third, we show that this affine identity is impossible. Define $\rho = \|\nabla H\|_2 > 0$, $\mathbf{e} = \nabla H/\rho$, and $h_0 = H(\mathbf{0})$. For $-1 < t < 1$ and any unit vector $\mathbf{v}$ perpendicular to $\mathbf{e}$, the two points $t\mathbf{e} \pm \sqrt{1-t^2}\mathbf{v}$ lie on the sphere and have the same score $h_0 + \rho t$. Subtracting their affine identities and dividing by $2\sqrt{1-t^2}$ gives

$$\big(s(h_0 + \rho t) - 1/2\big)(\nabla B)^\top \mathbf{v} \overset{⑥}{=} (\nabla A)^\top \mathbf{v}.$$

Equality ⑥ cancels the affine constants by comparing the two symmetric points. The scalar $s(h_0 + \rho t) - 1/2$ is strictly increasing in t. Evaluating the last equality at two distinct values of t forces $(\nabla B)^\top \mathbf{v} = (\nabla A)^\top \mathbf{v} = 0$ for every such $\mathbf{v}$. Hence both gradients are parallel to $\mathbf{e}$. Put $b_0 = B(\mathbf{0})$, $b_1 = (\nabla B)^\top \mathbf{e}$, $a_0 = A(\mathbf{0})$, and $a_1 = (\nabla A)^\top \mathbf{e}$. On the sphere, the identity now reduces to

$$(b_0 + b_1 t)\big(s(h_0 + \rho t) - 1/2\big) \overset{⑦}{=} a_0 + a_1 t, \qquad -1 < t < 1.$$

Equality ⑦ substitutes the affine functions whose gradients are parallel to $\mathbf{e}$, with $t = \mathbf{e}^\top \mathbf{x}$. Using $s(u) - 1/2 = (\exp(u) - 1)/(2(\exp(u) + 1))$, define the affine functions of t by $p(t) = b_0 + b_1 t - 2(a_0 + a_1 t)$ and $q(t) = b_0 + b_1 t + 2(a_0 + a_1 t)$. The preceding identity becomes $p(t)\exp(h_0 + \rho t) = q(t)$. Differentiating twice and using $p'' = q'' = 0$ yields

$$\exp(h_0 + \rho t)\big(\rho^2 p(t) + 2\rho p'(t)\big) \overset{⑧}{=} 0, \qquad -1 < t < 1.$$

Equality ⑧ differentiates $p\exp(h_0 + \rho t) = q$ twice, using $p'' = q'' = 0$. After division by the positive exponential, comparison of the linear and constant coefficients in $\rho^2 p + 2\rho p' = 0$ gives $p = 0$, since $\rho > 0$. The original identity then gives $q = 0$. Hence $b_0 + b_1 t = (p + q)/2 = 0$, contradicting $B \geq 1/2$ on U. This proves the pointwise assertion (77).

All functions in this argument are continuous on $\mathcal{X}$. Equality in $L^2(P)$ between two such functions would imply pointwise equality, because a nonzero value of their continuous difference has a neighborhood of positive uniform measure. As a result, (77) holds in $L^2(P)$ as well. The initial span is finite dimensional and therefore closed in $L^2(P)$. Consequently the distance of $a_g b$ from that span is strictly positive for each fixed training outcome. Neither this argument nor the theorem bounds that distance below uniformly over outcomes or dimensions.

Since the initial span is contained in the learned span and both are finite dimensional, their dimensions are strictly ordered. Multiplication by the fixed function a_g defines a linear map from the initial span into $L^2(P)$. It is well defined and bounded because $|a_g| \leq 1/2$. Its image is precisely the span of the products $a_g \phi_j$, whose dimension is therefore at most the initial dimension. The learned span is the sum of these two spaces, so its dimension is at most twice the initial dimension. Together with strict inclusion, this proves both dimension assertions in Section D.1.

Step 6. The population size of the change. The pointwise score bound alone is too coarse for the population scale. We instead combine the parameter size with a dimension-independent fourth moment of the feature vector. We recall $\mathbf{z} = \mathbf{h}/\sqrt{m}$, $\boldsymbol{\psi}_0 = \mathbf{h}/\sqrt{2m}$, $a_g = g_{\widehat{\boldsymbol{\tau}}} - 1/2$, and $c_d = 57/[1024(d+1)]$. The learning rate gives the deterministic coefficient bound $\|\widehat{\boldsymbol{\tau}}\|_2 = \eta_1 \|\mathbf{g}\|_2 < \frac{1}{2\sqrt{2(d+1)}}$. Put $k = m/4$ and $U_j(\mathbf{x}) = \sqrt{d}\,\vec{\mathbf{q}}_j^{\,\top}\mathbf{x}$ for $j \in [k]$. The pair identity (112), applied to each

direction and its negative, gives

$$\|\mathbf{z}(\mathbf{x})\|_2^2 = \frac{1}{k}\sum_{j=1}^{k}\frac{\sigma(U_j(\mathbf{x})+1/2)^2+\sigma(-U_j(\mathbf{x})+1/2)^2}{2} \overset{⑨}{\leq} \frac{1}{k}\sum_{j=1}^{k}\left(U_j(\mathbf{x})^2+\frac{1}{4}\right).$$

Inequality ⑨ is (112). To control the integrated squared change, we next bound the fourth moment:

$$\int \|\mathbf{z}(\mathbf{x})\|_2^4 \mathrm{d}\mu(\mathbf{x}) \overset{①}{\leq} \frac{1}{k}\sum_{j=1}^{k}\int\left(U_j(\mathbf{x})^2+\frac{1}{4}\right)^2 \mathrm{d}\mu(\mathbf{x}) \overset{②}{=} \frac{3d}{d+2}+\frac{1}{2}+\frac{1}{16} \leq \frac{57}{16}.$$

Inequality ① uses Jensen after ⑨. For equality ②, rotation invariance and (100) give $\mathbb{E}_P\left[U_j^2\right]=1$ and $\mathbb{E}_P\left[U_j^4\right]=3d/(d+2)$ for each fixed unit direction. These moments are taken over a fresh input, so the argument applies to every fixed paired direction collection and requires no independence between its directions. The sigmoid bound from Step 2 gives $|a_g(\mathbf{x})| \leq \left|\widehat{\boldsymbol{\tau}}^\top \mathbf{z}(\mathbf{x})\right|/4$. Also, subtracting the two blocks of $\boldsymbol{\psi}_0$ from $\boldsymbol{\psi}$ gives

$$\|\boldsymbol{\psi}(\mathbf{x})-\boldsymbol{\psi}_0(\mathbf{x})\|_2^2 \overset{③}{=} 2a_g(\mathbf{x})^2\|\mathbf{z}(\mathbf{x})\|_2^2 = \widehat{K}(\mathbf{x},\mathbf{x})-\frac{1}{2}K_{\mathrm{RF}}(\mathbf{x},\mathbf{x}).$$

Equality ③ sums the squared changes of the two duplicated blocks. After integrating this identity, Cauchy–Schwarz for the scalar score and the preceding bounds yield

$$\int \|\boldsymbol{\psi}-\boldsymbol{\psi}_0\|_2^2 \mathrm{d}\mu \overset{④}{\leq} \frac{\|\widehat{\boldsymbol{\tau}}\|_2^2}{8}\int \|\mathbf{z}\|_2^4 \mathrm{d}\mu \overset{⑤}{\leq} \frac{57}{1024(d+1)} = c_d.$$

Inequality ④ uses $|a_g| \leq |\widehat{\boldsymbol{\tau}}^\top \mathbf{z}|/4$ and Cauchy–Schwarz. Inequality ⑤ substitutes the parameter-norm and fourth-moment bounds. The integral is strictly positive. Step 3 makes a_g a nonzero continuous function, and $\|\mathbf{z}(\mathbf{x})\|_2^2 = K_{\mathrm{RF}}(\mathbf{x},\mathbf{x}) \geq 1/4$. As a result, its nonnegative integrand is positive on a nonempty open set, which has positive uniform measure. This proves (75).

To bound the increment operator, put $\boldsymbol{\delta} := \boldsymbol{\psi}-\boldsymbol{\psi}_0$. The two attention-weighted blocks satisfy $\boldsymbol{\psi}_0(\mathbf{x})^\top \boldsymbol{\delta}(\mathbf{x}') = 0$ at every pair of inputs. Expanding $\boldsymbol{\psi} = \boldsymbol{\psi}_0 + \boldsymbol{\delta}$ therefore represents the kernel increment as the inner-product kernel of $\boldsymbol{\delta}$. For any $f \in L^2(P)$, two applications of Cauchy–Schwarz give

$$\left\|\int_{\mathcal{X}} \boldsymbol{\delta}(\cdot)^\top \boldsymbol{\delta}(\mathbf{x}') f(\mathbf{x}')\mathrm{d}\mu(\mathbf{x}')\right\|_{L^2} \overset{⑥}{\leq} \left(\int_{\mathcal{X}} \|\boldsymbol{\delta}(\mathbf{x})\|_2^2 \mathrm{d}\mu(\mathbf{x})\right)\|f\|_{L^2} \overset{⑦}{\leq} c_d\|f\|_{L^2}.$$

Inequality ⑦ uses the population feature-change bound just proved. To justify ⑥, we set $\mathbf{v} = \int \boldsymbol{\delta} f \mathrm{d}\mu$, use the bound $\|\mathbf{v}\|_2 \leq (\int \|\boldsymbol{\delta}\|_2^2 \mathrm{d}\mu)^{1/2}\|f\|_{L^2}$, and then bound $\left\|\boldsymbol{\delta}^\top \mathbf{v}\right\|_{L^2}$ by the same square-root factor times $\|\mathbf{v}\|_2$. This proves the asserted operator-norm bound. It is uniform in the directions and responses in the first training sample. It does not assert a uniform pointwise bound of order $1/d$ on every kernel evaluation.

We finally verify the trace interpretation used in Section D.1. Let $V : \mathbb{R}^m \to L^2(P)$ send $\mathbf{c}$ to $\boldsymbol{\delta}^\top \mathbf{c}$. Denote its Hilbert-space adjoint by V^*, characterized by $\langle V\mathbf{c}, f\rangle_{L^2} = \mathbf{c}^\top V^* f$, where $\langle h, f\rangle_{L^2} = \int hf \mathrm{d}\mu$. This identity gives $V^* f = \int \boldsymbol{\delta} f \mathrm{d}\mu$. As a result, the increment operator is VV^*, while $V^*V = \int \boldsymbol{\delta}\boldsymbol{\delta}^\top \mathrm{d}\mu$. The two operators have the same nonzero eigenvalues, with multiplicities. Indeed, for a positive eigenvalue, V and V^* map the corresponding eigenspaces into each other, and their compositions on these eigenspaces equal multiplication by that eigenvalue. The equality of the nonzero eigenvalues gives equality of the traces. The matrix trace is $\int \|\boldsymbol{\delta}\|_2^2 \mathrm{d}\mu$, which is the second integral in (75). All these operators have finite rank, so the trace identity requires no limiting argument.

□

Proposition F.10 bounds the preserved approximation cost. Appendix F.5 separately controls fitted noise for the learned representation.

D.3 Population Loss after the Attention Step

We now prove population descent for the performed attention step with its temporary output coefficients fixed. The proof combines polynomial target moments with finite noise variance.

For $c \in \{0,1\}$, recall the temporary output vector $\boldsymbol{a}_{\mathrm{init},c}$ from (34). Define the population loss as half the regression risk and define a combined target and noise size by

$$R_{1,c}(\boldsymbol{\tau}) := \frac{1}{2}\|f(\boldsymbol{\tau}, \boldsymbol{a}_{\mathrm{init},c}, \cdot) - f^*\|_{L^2}^2, \qquad A_* := 1 + B_*^2 + \sigma_0^2. \tag{82}$$

The loss in (82) integrates over a fresh input with law P and does not include fresh response noise. The vectors $\mathbf{g}_c$, the selected index $\widehat{c}$, the step size η_1, and the updated attention $\widehat{\boldsymbol{\tau}}$ are exactly those used by training. The constants below are used only to state the guarantee.

Proposition D.2 (Population descent of the performed attention update)**.** Under the observation and polynomial assumptions of Theorem C.1, perform the single attention GD step from stage one of Algorithm 1. There are constants $C_L \geq 1$ and $c_L > 0$, depending only on L, with the following property. For every deterministic paired collection of unit directions, every positive width divisible by four, and $d \geq \max\{3, C_L A_*\}$, an event determined by the first training sample has probability at least $1 - C_L A_*/n_1$, and on that event

$$R_{1,\widehat{c}}(\mathbf{0}) - R_{1,\widehat{c}}(\widehat{\boldsymbol{\tau}}) \geq \frac{\eta_1}{2}\|\mathbf{g}\|_2^2 \geq \frac{c_L}{\sqrt{A_*(d+1)}}. \tag{83}$$

The probability is with respect to the first training sample for the fixed direction collection. With the same conditioning, the expected improvement satisfies

$$\mathbb{E}\left[R_{1,\widehat{c}}(\mathbf{0}) - R_{1,\widehat{c}}(\widehat{\boldsymbol{\tau}}) \mid \mathbf{Q}\right] \geq \frac{c_L}{\sqrt{A_*(d+1)}} - \frac{C_L A_*^2}{n_1}. \tag{84}$$

Here $\mathbf{Q}$ denotes the sampled directions when they are random. Both assertions also hold after averaging over those directions. For bounded B_*, σ_0 and $n_1/\sqrt{d} \to \infty$, the right side of (84) is positive for all sufficiently large d. In particular, this conclusion applies to the logarithmic sample regime of Theorem 4.1. No lower bound on any target coefficient and no fourth moment of the noise are required.

Proof roadmap. Proposition D.2 proves (21) in four steps. Step 1 bounds moments of the paired channels and polynomial target uniformly in the directions and width. Step 2 uses these moments to control both candidate gradient errors and the empirical curvature bound. Simultaneous concentration remains valid for the selected index $\widehat{c}$ and aligns its empirical step with the corresponding population gradient. Step 3 combines this alignment with a population Taylor bound to obtain strict descent. Step 4 controls the loss on exceptional outcomes and averages the improvement. The output reset means that this result does not order final refitted risks.

Proof. Step 1. Uniform channel moments. Uniform moments let us concentrate gradients without bounding every observed response or depending on the width. We fix the directions throughout the proof. Recall $\mathbf{z} = \mathbf{h}/\sqrt{m}$, $b = (2/m)\sum_j \phi_j$, and $\mathbf{u} = \mathbf{1}_m/\sqrt{m}$, so $b = \mathbf{u}^\top \mathbf{z}$ and $1/2 \leq b \leq \|\mathbf{z}\|_2$. Put $k = m/4$ and $U_j(\mathbf{x}) = \sqrt{d}\,\overrightarrow{\mathbf{q}}_j^\top \mathbf{x}$ for $j \in [k]$. The opposite-direction identity (112) gives $\|\mathbf{z}(\mathbf{x})\|_2^2 \leq \frac{1}{k}\sum_{j=1}^k \left(U_j(\mathbf{x})^2 + \frac{1}{4}\right)$. For every integer $p \geq 1$, the spherical coordinate moment formula is

$$\int U_j^{2p} \mathrm{d}\mu \overset{①}{=} \frac{d^p(1\cdot 3\cdots(2p-1))}{d(d+2)\cdots(d+2p-2)} \leq 1\cdot 3\cdots(2p-1).$$

Equality ① is the spherical coordinate formula (100), proved by a beta integral in Appendix F.1. Write $(2r-1)!! = 1\cdot 3\cdots(2r-1)$ for $r \geq 1$ and $(-1)!! = 1$. Raising the preceding channel inequality to the power p, using convexity for the finite average, and expanding $(U_j^2 + 1/4)^p$ therefore proves

$$\int \|\mathbf{z}\|_2^{2p} \mathrm{d}\mu \overset{②}{\leq} D_p \quad (1 \leq p \leq 6), \qquad \int b^2 \mathrm{d}\mu \leq \frac{5}{4}, \tag{85}$$

Inequality ② uses convexity and the coordinate moments in ①. Expanding the power gives the width-independent choice $D_p = \sum_{r=0}^p \binom{p}{r} 4^{-(p-r)}(2r-1)!!$, without independence or distinctness of the fixed directions. Lemma E.6 also gives $\|f^*\|_{L^4}^2 \leq 3^L B_*^2$.

Step 2. Both candidate gradients concentrate. Both candidate gradients must concentrate on the same event because the algorithm selects one using the data. Let $(\mathbf{x}, y)$ be one observation, independent of the fixed directions. Define

$$\mathbf{W}_c(\mathbf{x}, y) := \tfrac{1}{2}(cb(\mathbf{x}) - y)b(\mathbf{x})\mathbf{z}(\mathbf{x}), \qquad U(\mathbf{x}, y) := \left(\tfrac{5}{4}b(\mathbf{x})^2 + \tfrac{1}{2}|y|\, b(\mathbf{x})\right)\|\mathbf{z}(\mathbf{x})\|_2^2.$$

Here U is a scalar curvature bound, distinct from the coordinate variables U_j in Step 1. Write $\overline{\mathbf{g}}_c = \mathbb{E}[\mathbf{W}_c]$ and $\overline{\Lambda} = \mathbb{E}[U]$. Then $\mathbf{g}_c$ and Λ_1 are the averages of these variables over the first training sample. Since the noise is centered and independent of the input, $\overline{\mathbf{g}}_c = \nabla R_{1,c}(\mathbf{0})$. Differentiation under the integral is justified by the integrable bounds below and the uniform smoothness of the bottleneck in a neighborhood of zero.

The inequalities $b \le \|\mathbf{z}\|_2$ and $(a+b)^2 \le 2a^2 + 2b^2$ give

$$\|\mathbf{W}_c\|_2^2 \overset{③}{\le} \tfrac{1}{2}\left(\|\mathbf{z}\|_2^6 + y^2\|\mathbf{z}\|_2^4\right), \qquad U^2 \overset{④}{\le} \tfrac{25}{8}\|\mathbf{z}\|_2^8 + \tfrac{1}{2}y^2\|\mathbf{z}\|_2^6.$$

Inequalities ③ and ④ square the two summands separately and use $b \le \|\mathbf{z}\|_2$. For $r = 4, 6$, conditioning on $\mathbf{x}$ and using Cauchy–Schwarz yield

$$\mathbb{E}\left[y^2\|\mathbf{z}\|_2^r\right] \overset{⑤}{\le} \|f^*\|_{L^4}^2 \left(\int \|\mathbf{z}\|_2^{2r}\,\mathrm{d}\mu\right)^{1/2} + \sigma_0^2 \int \|\mathbf{z}\|_2^r \mathrm{d}\mu.$$

In ⑤, $\mathbb{E}\left[y^2 \mid \mathbf{x}\right] = (f^*(\mathbf{x}))^2 + \mathbb{E}\left[w^2\right]$ removes the mixed noise term, and Cauchy–Schwarz bounds the target product. As a result, (85) supplies a number $V = C_L A_* \ge 1$, after choosing C_L sufficiently large, such that $\mathbb{E}\left[\|\mathbf{W}_c\|_2^2\right] \le V$ for both c, $\mathbb{E}\left[U^2\right] \le V$, $\|\overline{\mathbf{g}}_c\|_2 \le \sqrt{V}$, and $\overline{\Lambda} \le \sqrt{V}$. Independence of the observations gives

$$\mathbb{E}\left[\|\mathbf{g}_c - \overline{\mathbf{g}}_c\|_2^2\right] \overset{⑥}{\le} V/n_1, \qquad \mathbb{E}\left[(\Lambda_1 - \overline{\Lambda})^2\right] \overset{⑦}{\le} V/n_1.$$

Both ⑥ and ⑦ use cancellation of cross terms between independent centered observations. Markov's inequality for the squared gradient errors and Chebyshev's inequality for Λ_1 show that the simultaneous event $\mathcal{E}_1 := \left\{\max_{c\in\{0,1\}} \|\mathbf{g}_c - \overline{\mathbf{g}}_c\|_2 \le 1/128, \quad \Lambda_1 \le 2\sqrt{V}\right\}$ has complement of probability at most $(2\cdot 128^2 + 1)V/n_1$. The same event controls both candidates before the data-dependent selection of $\widehat{c}$. On $\mathcal{E}_1$, $\|\mathbf{g}\|_2 \le \sqrt{V} + 1/128$, while the deterministic gradient bound gives $\|\mathbf{g}\|_2 \ge 1/32$ on every outcome.

Step 3. Population descent for the selected candidate. We compare the empirical direction with its population gradient, then turn their positive inner product into descent using smoothness. Put $R_d = \sqrt{2(d+1)}$. Proposition D.1 keeps the entire segment from zero to $-\eta_1\mathbf{g}$ in the region where the bottleneck ReLU is active for every input. For $c \in \{0,1\}$, write $F_c(\boldsymbol{\tau}, \mathbf{x}) = b(\mathbf{x})(c + 2s(\boldsymbol{\tau}^\top \mathbf{z}(\mathbf{x})) - 1)$ for the temporary predictor on this region. All derivatives below are with respect to $\boldsymbol{\tau}$. Differentiating the population loss gives

$$\nabla^2 R_{1,c}(\boldsymbol{\tau}) \overset{⑧}{=} \int \left[\nabla F_c (\nabla F_c)^\top + (F_c - f^*)\nabla^2 F_c\right] \mathrm{d}\mu.$$

Equality ⑧ differentiates the squared population loss twice. The derivative bounds from Step 2 of Proposition D.1's proof therefore imply

$$\left\|\nabla^2 R_{1,c}(\boldsymbol{\tau})\right\|_2 \overset{⑨}{\le} \int \left(\tfrac{5}{4}b^2 + \tfrac{1}{2}|f^*|\, b\right)\|\mathbf{z}\|_2^2 \mathrm{d}\mu \overset{①}{\le} \mathbb{E}[U] = \overline{\Lambda} \le \sqrt{V}.$$

Inequality ⑨ applies the derivative bounds inside the integral. For ①, centering and independence of the noise give $f^*(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$ and hence $|f^*(\mathbf{x})| \le \mathbb{E}[|y| \mid \mathbf{x}]$. The envelope in this display is integrable by Steps 1 and 2 and is uniform on a neighborhood of the segment. It justifies differentiation under the integral. This Hessian bound holds for both fixed candidates throughout the smooth region, so the sample-dependent choice of $\widehat{c}$ and of the segment preserves it.

Because $\eta_1 < (2R_d\|\mathbf{g}\|_2)^{-1}$ and $\|\mathbf{g}\|_2 \ge 1/32$, one has $\eta_1\sqrt{V} \le 1/2$ whenever $R_d \ge 32\sqrt{V}$. The assumption $d \ge C_L A_*$ ensures this after enlarging C_L. On $\mathcal{E}_1$, the gradient error is at most $\|\mathbf{g}\|_2/4$, so

$$\overline{\mathbf{g}}_{\widehat{c}}^\top \mathbf{g} \overset{②}{\ge} \|\mathbf{g}\|_2^2 - \|\overline{\mathbf{g}}_{\widehat{c}} - \mathbf{g}\|_2\|\mathbf{g}\|_2 \overset{③}{\ge} \tfrac{3}{4}\|\mathbf{g}\|_2^2.$$

Inequality ② is Cauchy–Schwarz, and ③ uses the relative gradient error on $\mathcal{E}_1$. Taylor's formula with integral remainder consequently yields

$$R_{1,\widehat{c}}(-\eta_1\mathbf{g}) \overset{④}{\le} R_{1,\widehat{c}}(\mathbf{0}) - \tfrac{3}{4}\eta_1\|\mathbf{g}\|_2^2 + \tfrac{1}{2}\eta_1^2\sqrt{V}\|\mathbf{g}\|_2^2 \overset{⑤}{\le} R_{1,\widehat{c}}(\mathbf{0}) - \tfrac{1}{2}\eta_1\|\mathbf{g}\|_2^2.$$

Inequality ④ combines the Hessian bound with ③, and ⑤ uses $\eta_1\sqrt{V} \le 1/2$. On the same event, the denominator of the step size is at most $1 + 2R_d(\sqrt{V} + 1/128) + 2\sqrt{V} \le 6R_d\sqrt{V}$. Hence

$\eta_1 \geq (6R_d\sqrt{V})^{-1}$, and the improvement is at least $(12288R_d\sqrt{V})^{-1}$. Since $V = C_L A_*$, this proves (83), including its stated probability, with constants depending only on L.

Step 4. Expected improvement. A uniform loss bound controls the exceptional event when converting high-probability descent to expected descent. On every outcome and for both c, the temporary predictor satisfies $|f(\boldsymbol{\tau}, \boldsymbol{a}_{\mathrm{init},c}, \mathbf{x})| \leq 2b(\mathbf{x})$ for every finite $\boldsymbol{\tau}$. This follows directly from the complementary weights in $(0, 1)$. Equation (85) gives

$$0 \leq R_{1,c}(\boldsymbol{\tau}) \overset{⑥}{\leq} 4 \int b^2 \mathrm{d}\mu + B_*^2 \overset{⑦}{\leq} 5 + B_*^2 \leq 5A_*.$$

Inequality ⑥ uses $(F_c - f^*)^2/2 \leq F_c^2 + (f^*)^2$ and $|F_c| \leq 2b$. Inequality ⑦ uses (85). The loss difference is therefore at least $-5A_*$ even on $\mathcal{E}_1^c$. Set $\delta = c_L/\sqrt{A_*(d+1)}$, with $c_L \leq 1$. Combining this deterministic lower bound with Step 3 gives expected improvement at least $\delta - (5A_* + \delta)\Pr[\mathcal{E}_1^c]$. Since $A_* \geq 1$, enlarging C_L proves (84). For bounded A_*, the negative term is $o(d^{-1/2})$ when $n_1/\sqrt{d} \to \infty$. This proves the final assertion. Every estimate is uniform in the fixed directions. Taking expectations over their distribution therefore preserves the bounds. □

Proposition 4.2 follows by identifying (83) with (21) and applying (84) to its expectation claim. The comparison uses the same selected temporary outputs on both sides of (83).

E Harmonic Analysis and Concentration Tools

E.1 Harmonic Identities and Polynomial Dimension

In this section, we provide the spherical harmonic analysis needed for the target class and the ReLU approximation. These identities quantify the polynomial comparison predictor and its coefficient bound, both of which attention preserves while adding nonlinear products. Appendix F.1 converts these identities into the finite-width comparison used in the risk bound for the learned feature kernel. Background on spherical harmonics and orthogonal polynomials is available in Efthimiou & Frye (2014), Chihara (2011), and Szegő (1975). Expansions in spherical harmonics have also been used in statistical analyses of neural-network kernels (Bach, 2017; Bietti & Mairal, 2019; Ghorbani et al., 2021). All spherical integrals below use the probability measure μ from Section A.

We use the polynomial restrictions defined in Appendix A.2, always through continuous representatives. Restriction need not be injective on all polynomials, as $\|\mathbf{x}\|_2^2$ and 1 agree on $\mathcal{X}$.

We first fix the algebraic notation. A multi-index is a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ of nonnegative integers. We write $|\boldsymbol{\alpha}| = \sum_{i=1}^d \alpha_i$, $\boldsymbol{\alpha}! = \prod_{i=1}^d \alpha_i!$, and $\mathbf{x}^{\boldsymbol{\alpha}} = \prod_{i=1}^d x_i^{\alpha_i}$. The ambient Laplacian is $\Delta = \sum_{i=1}^d \partial^2/\partial x_i^2$. A homogeneous polynomial of degree ℓ is a linear combination of monomials with $|\boldsymbol{\alpha}| = \ell$. Its restriction is a spherical harmonic when the ambient polynomial has zero Laplacian. With real coefficients $c_{\boldsymbol{\alpha}}$, this gives $\mathcal{H}_\ell = \left\{ p \restriction_{\mathcal{X}} : p(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|=\ell} c_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}},\ \Delta p = 0 \right\}$. Proposition E.1 proves that different harmonic degrees are orthogonal in $L^2(P)$ and gives their dimensions. For $d \geq 3$,

$$N(d,0) = 1, \qquad N(d,\ell) = \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2} = \frac{(2\ell+d-2)(\ell+d-3)!}{\ell!(d-2)!}, \quad \ell \geq 1, \tag{86}$$

where a binomial coefficient with a negative lower argument is zero. Theorem E.2 also proves that the union of these orthonormal bases over all degrees is complete in $L^2(P)$, meaning that its finite linear combinations are dense in that space.

For later target-norm comparisons, the dimensions in (86) also give $N(d,1) = d$ and

$$\frac{N(d,\ell+1)}{N(d,\ell)} \overset{①}{=} \frac{2\ell+d}{2\ell+d-2} \frac{\ell+d-2}{\ell+1} \geq 1, \qquad d \geq 3, \quad \ell \geq 1. \tag{87}$$

Equality ① divides the dimension formulas at consecutive degrees. Both factors are at least one, so (87) implies $N(d,\ell) \geq d$ for every $\ell \geq 1$.

For each nonnegative integer ℓ, let $P_\ell^{(d)}$ be the degree-ℓ Gegenbauer polynomial with parameter $(d-2)/2$, normalized by $P_\ell^{(d)}(1) = 1$. Equivalently, it is the degree-ℓ polynomial orthogonal to every lower-degree polynomial for the weight $(1-t^2)^{(d-3)/2}$ on $[-1,1]$, with value one at $t = 1$. The addition formula is

$$\sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}') = N(d,\ell) P_\ell^{(d)}(\langle \mathbf{x}, \mathbf{x}' \rangle). \tag{88}$$

In particular, $\sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x})^2 = N(d,\ell)$. Let ω_k be the surface area of $\mathbb{S}^k$, and let Γ denote the gamma function. We define the probability measure on $[-1,1]$ by

$$\mathrm{d}\nu_d(t) := \frac{\omega_{d-2}}{\omega_{d-1}} (1-t^2)^{(d-3)/2} \mathrm{d}t, \qquad \omega_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}. \tag{89}$$

The measure ν_d is the law of a fixed coordinate of a uniform unit vector. Its total mass is one because the integral of $(1-t^2)^{(d-3)/2}$ over $[-1,1]$ equals $\Gamma(1/2)\Gamma((d-1)/2)/\Gamma(d/2)$. As a result, the factors in (88) and (92) correspond to probability measure, rather than unnormalized surface area. For nonnegative integers j, k, let δ_{jk} be one when $j = k$ and zero otherwise. The normalized Gegenbauer polynomials satisfy

$$\int_{-1}^{1} P_j^{(d)}(t) P_k^{(d)}(t) \mathrm{d}\nu_d(t) = \frac{\delta_{jk}}{N(d,k)}. \tag{90}$$

The recurrence relation is

$$t P_k^{(d)}(t) = \frac{k}{2k+d-2} P_{k-1}^{(d)}(t) + \frac{k+d-2}{2k+d-2} P_{k+1}^{(d)}(t), \quad k \geq 1, \tag{91}$$

with $P_0^{(d)} = 1$ and $P_1^{(d)}(t) = t$. The Funk–Hecke formula computes the harmonic coefficient of a feature that depends on an inner product. For $\mathbf{x} \in \mathcal{X}$, an integer $\ell \geq 0$, $j \in [N(d,\ell)]$, and a measurable function a with $\int_{-1}^{1} |a(t)| \, \mathrm{d}\nu_d(t) < \infty$,

$$\int_{\mathcal{X}} a(\langle \mathbf{x}, \mathbf{q} \rangle) Y_{\ell,j}(\mathbf{q}) \mathrm{d}\mu(\mathbf{q}) = Y_{\ell,j}(\mathbf{x}) \int_{-1}^{1} a(t) P_\ell^{(d)}(t) \mathrm{d}\nu_d(t). \tag{92}$$

The coordinate law (89) and boundedness of each harmonic and each $P_\ell^{(d)}$ ensure absolute convergence in (92). Proposition E.3 below derives (88), (90), (91), and (92) from the harmonic decomposition and polynomial orthogonality. This derivation verifies the probability normalization independently of the ReLU approximation argument.

Proposition E.1 (Homogeneous harmonic decomposition (Efthimiou & Frye, 2014))**.** For every nonnegative integer ℓ, each homogeneous polynomial p of degree ℓ on $\mathbb{R}^d$ has a unique decomposition

$$p(\mathbf{x}) = \sum_{j=0}^{\lfloor \ell/2 \rfloor} \|\mathbf{x}\|_2^{2j} h_{\ell-2j}(\mathbf{x}), \tag{93}$$

where each $h_{\ell-2j}$ is homogeneous and harmonic of degree $\ell - 2j$. On the unit sphere the factors $\|\mathbf{x}\|_2^{2j}$ equal one, and the terms of different harmonic degrees are orthogonal in $L^2(P)$.

Proof. **Outline.** We first show that multiplication by $\|\mathbf{x}\|_2^2$ is adjoint to the Laplacian for a coefficient inner product. We iterate the resulting decomposition and count dimensions. Finally, the divergence theorem proves orthogonality of different degrees for the spherical probability measure.

Step 1. A finite-dimensional orthogonal decomposition. An inner product on coefficients turns the desired decomposition into an orthogonal-complement calculation. For real polynomials $p = \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$ and $q = \sum_{\boldsymbol{\alpha}} q_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$, we define $\langle p, q \rangle_F = \sum_{\boldsymbol{\alpha}} \boldsymbol{\alpha}! p_{\boldsymbol{\alpha}} q_{\boldsymbol{\alpha}}$, where missing coefficients are zero. This is a positive definite inner product on each finite-degree polynomial space. Calculating the coefficients gives $\left\langle \|\mathbf{x}\|_2^2 q, p \right\rangle_F \overset{①}{=} \langle q, \Delta p \rangle_F$. To verify ①, let $\mathbf{e}_i$ be the ith standard coordinate vector. The coefficient identity for that coordinate is $\boldsymbol{\alpha}!(\alpha_i+2)(\alpha_i+1) = (\boldsymbol{\alpha}+2\mathbf{e}_i)!$. Consequently, within the homogeneous degree-ℓ space, the orthogonal complement of $\|\mathbf{x}\|_2^2$ times the degree-$(\ell-2)$ space is exactly the harmonic subspace. For $\ell = 0, 1$, every homogeneous polynomial is already harmonic.

Step 2. Iteration and uniqueness. For $\ell \geq 2$, Step 1 gives $p = h_\ell + \|\mathbf{x}\|_2^2 q_{\ell-2}$ with h_ℓ harmonic. Orthogonal decomposition and injectivity of multiplication by $\|\mathbf{x}\|_2^2$ determine both terms uniquely. Iterating on $q_{\ell-2}$ proves (93). Restriction to $\mathcal{X}$ sets every radial factor to one.

Counting monomials gives dimension $\binom{d+\ell-1}{\ell}$ for homogeneous degree-ℓ polynomials. Step 1 therefore gives harmonic dimension $\binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}$ for $\ell \geq 2$. Restriction to the sphere is injective on homogeneous polynomials, since homogeneity recovers their values away from zero and continuity recovers the value at zero. The ambient and spherical dimensions agree. Together with $\ell = 0, 1$, this proves (86).

Step 3. Orthogonality on the sphere. The boundary normal derivative of a homogeneous polynomial records its degree, which separates different degrees in the divergence identity. Let h_ℓ, h_k be homogeneous harmonic polynomials of distinct nonnegative degrees ℓ, k. Here ∇ is the Euclidean gradient. The polynomial vector field $h_\ell \nabla h_k - h_k \nabla h_\ell$ has divergence $h_\ell \Delta h_k - h_k \Delta h_\ell = 0$. Differentiating $h_j(t\mathbf{x}) = t^j h_j(\mathbf{x})$ at $t = 1$ gives $\mathbf{x}^\top \nabla h_j(\mathbf{x}) = j h_j(\mathbf{x})$ for $j = \ell, k$. The outward unit normal at $\mathbf{x} \in \mathcal{X}$ is $\mathbf{x}$. Applying the divergence theorem on the unit ball and writing surface area measure as $\omega_{d-1}\mathrm{d}\mu$, we obtain

$$0 \overset{②}{=} \omega_{d-1} \int_{\mathcal{X}} (h_\ell \nabla h_k - h_k \nabla h_\ell)^\top \mathbf{x} \mathrm{d}\mu(\mathbf{x}) \overset{③}{=} (k - \ell)\omega_{d-1} \int_{\mathcal{X}} h_\ell(\mathbf{x}) h_k(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}).$$

Equality ② is the divergence theorem, valid for these polynomial fields on the closed ball, and ③ uses the homogeneous normal derivatives. Since $(k - \ell)\omega_{d-1} \neq 0$, the integral vanishes, proving orthogonality. □

Theorem E.2. For every nonnegative integer k, the restriction to $\mathbb{S}^{d-1}$ of every ambient polynomial of degree at most k belongs to $\bigoplus_{\ell=0}^{k} \mathcal{H}_\ell$. Moreover, every function in $L^2(P)$ can be approximated arbitrarily closely in L^2 norm by finite linear combinations of spherical harmonics.

Proof. **Outline.** We first restrict the homogeneous decomposition to the sphere. We then use polynomial approximation of continuous functions to prove completeness in $L^2(P)$.

Step 1. Finite polynomial expansion. Apply Proposition E.1 to each homogeneous part of the polynomial. Restriction to $\mathcal{X}$ sets every radial factor to one, giving an element of $\bigoplus_{\ell=0}^{k} \mathcal{H}_\ell$. Orthogonality makes the resulting harmonic expansion unique.

Step 2. Completeness. Polynomial restrictions form an algebra containing constants and separating points on the compact sphere. Stone–Weierstrass therefore gives uniform density in the continuous functions. For continuous f, g, the normalization $\mu(\mathcal{X}) = 1$ gives $\|f - g\|_{L^2} \overset{①}{\leq} \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - g(\mathbf{x})|$, where ① integrates the pointwise squared bound. Continuous functions are dense in $L^2(P)$ for this finite Borel measure. Step 1 therefore proves density of finite harmonic sums, and the orthonormal bases of the mutually orthogonal spaces $\mathcal{H}_\ell$ form a complete system. □

Proposition E.3 (Harmonic identities with probability normalization)**.** For $d \geq 3$ and the harmonic bases orthonormal in $L^2(P)$, the addition formula (88), the norm identity (90), the recurrence (91), and the Funk–Hecke formula (92) hold with the normalizations stated above. The last formula requires only $\int_{-1}^{1} |a(t)| \, \mathrm{d}\nu_d(t) < \infty$.

Proof. **Outline.** We derive the spherical-coordinate density and characterize harmonic functions invariant under orthogonal transformations fixing one input. We then identify the basis sum in the addition formula and compute its norm. Next, polynomial orthogonality determines the recurrence coefficients. Finally, we use the basis sum to evaluate the harmonic coefficients of any integrable function of an inner product. These are standard identities (Efthimiou & Frye, 2014), whose normalization we derive here.

Step 1. The coordinate law and invariant harmonics. The coordinate law fixes all normalization constants. Invariance then reduces a spherical harmonic to a one-variable orthogonal polynomial. We first verify the probability law (89) used in the integrals. Apart from its two poles, the sphere is parametrized by $\mathbf{q} = \left(t, \sqrt{1 - t^2}\, \mathbf{v}\right)$, where $-1 < t < 1$ and $\mathbf{v} \in \mathbb{S}^{d-2} \subset \mathbb{R}^{d-1}$. The derivative in t has length $(1 - t^2)^{-1/2}$ and is perpendicular to the $d - 2$ tangent directions in $\mathbf{v}$, each scaled

by $\sqrt{1-t^2}$. The surface element consequently has factor $(1-t^2)^{(d-3)/2}$. Integrating over $\mathbf{v}$ and dividing by the total area gives the density factor $\omega_{d-2}/\omega_{d-1}$ in (89). The poles have zero measure. Rotation invariance gives the same law for $\mathbf{x}^\top\mathbf{q}$ at every fixed $\mathbf{x} \in \mathcal{X}$.

We now fix $\mathbf{x} \in \mathcal{X}$ and a nonnegative integer ℓ. An orthogonal transformation is a matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ with $\mathbf{R}^\top\mathbf{R} = \mathbf{I}_d$. The chain rule shows that such transformations preserve homogeneous harmonic polynomials, and they also preserve μ. Let $h \in \mathcal{H}_\ell$ satisfy $h(\mathbf{Rq}) = h(\mathbf{q})$ whenever $\mathbf{Rx} = \mathbf{x}$. Two unit vectors with the same inner product with $\mathbf{x}$ have perpendicular components of equal length. An orthogonal transformation fixing $\mathbf{x}$ can send one component to the other. As a result, $h(\mathbf{q})$ depends only on $t = \mathbf{x}^\top\mathbf{q}$.

We next show that this dependence is polynomial of degree at most ℓ. Choose orthonormal coordinates in which $\mathbf{x} = (1, 0, \dots, 0)$. The homogeneous polynomial representing h has the same invariance off the sphere by homogeneity. Its restriction to $(t, s, 0, \dots, 0)$ is even in s, because reflection of the second coordinate fixes $\mathbf{x}$. It is therefore a sum of monomials $t^a s^{2b}$ with nonnegative integers a, b satisfying $a + 2b = \ell$. On the sphere we substitute $s^2 = 1 - t^2$, obtaining a polynomial $p(t)$ of degree at most ℓ. The invariance just proved gives $h(\mathbf{q}) = p(\mathbf{x}^\top\mathbf{q})$ for every $\mathbf{q} \in \mathcal{X}$, including the endpoints by continuity. For $\ell = 0$, this is already a constant multiple of $P_0^{(d)} = 1$. For $\ell \geq 1$ and every one-variable polynomial v of degree less than ℓ, Theorem E.2 puts $v(\mathbf{x}^\top\mathbf{q})$ in the lower-degree harmonic spaces. Their orthogonality to h, together with the coordinate law ν_d, gives

$$\int_{-1}^{1} p(t)v(t)\mathrm{d}\nu_d(t) \overset{①}{=} \int_{\mathcal{X}} h(\mathbf{q})v(\mathbf{x}^\top\mathbf{q})\mathrm{d}\mu(\mathbf{q}) \overset{②}{=} 0.$$

Equality ① uses the coordinate law, and ② uses harmonic orthogonality. The positive density of ν_d on $(-1, 1)$ makes the polynomial inner product positive definite. Within the $(\ell+1)$-dimensional space of polynomials of degree at most ℓ, the lower-degree space has dimension ℓ and hence a one-dimensional orthogonal complement.

We next verify that a nonzero polynomial in this orthogonal complement can be normalized at 1. Let Q be such a member. It has degree exactly ℓ, because a lower-degree member would be orthogonal to itself. If $Q(1) = 0$, then $Q(t) = (t-1)R(t)$ for a nonzero polynomial R of degree $\ell - 1$. Orthogonality to R would give

$$0 \overset{③}{=} \int_{-1}^{1} Q(t)R(t)\mathrm{d}\nu_d(t) = \int_{-1}^{1} (t-1)R(t)^2\mathrm{d}\nu_d(t) \overset{④}{<} 0.$$

Equality ③ is orthogonality to R. Inequality ④ uses the positive density and $(t-1)R(t)^2 < 0$ away from the finitely many roots of R in $(-1, 1)$. Hence $Q(1) \neq 0$ and $Q/Q(1) = P_\ell^{(d)}$. Together with the case $\ell = 0$, this identifies p as a scalar multiple of $P_\ell^{(d)}$, including $p = 0$, without using the addition formula.

Step 2. Addition formula and normalization. The diagonal of the basis sum fixes its normalization. Set $Z_\ell(\mathbf{x}, \mathbf{q}) := \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{q})$. This sum is unchanged by orthogonal changes of basis. Since rotations preserve the harmonic space and its inner product, $Z_\ell(\mathbf{Rx}, \mathbf{Rq}) = Z_\ell(\mathbf{x}, \mathbf{q})$. Its diagonal is therefore constant, with integral $\sum_j \|Y_{\ell,j}\|_{L^2}^2 = N(d, \ell)$. Since $\mu(\mathcal{X}) = 1$, the constant is $N(d, \ell)$. Step 1 and $P_\ell^{(d)}(1) = 1$ now give $Z_\ell(\mathbf{x}, \mathbf{q}) = N(d, \ell)P_\ell^{(d)}(\mathbf{x}^\top\mathbf{q})$, proving (88). Orthonormality also gives

$$\int_{\mathcal{X}} Z_\ell(\mathbf{x}, \mathbf{q})^2\mathrm{d}\mu(\mathbf{q}) \overset{⑤}{=} \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x})^2 \overset{⑥}{=} N(d, \ell).$$

Equality ⑤ is orthonormality, and ⑥ uses the computed diagonal. Substituting the addition formula and the coordinate law gives $\int (P_\ell^{(d)})^2\mathrm{d}\nu_d = 1/N(d, \ell)$. Together with orthogonality of distinct degrees, this proves (90).

Step 3. Recurrence coefficients. Orthogonality and parity leave only two degrees in multiplication by t, and the norms determine their coefficients. We recall that $P_k^{(d)}(1) = 1$ and that its squared norm under ν_d is $1/N(d, k)$. Symmetry of ν_d gives $P_k^{(d)}(-t) = (-1)^k P_k^{(d)}(t)$. Indeed, $(-1)^k P_k^{(d)}(-t)$ has the same leading coefficient and lower-degree orthogonality as $P_k^{(d)}$, so their difference is a lower-degree polynomial orthogonal to itself. For $k \geq 1$ and $0 \leq j \leq k-2$, the polynomial $tP_j^{(d)}(t)$

has degree less than k. Hence $tP_k^{(d)}$ is orthogonal to $P_j^{(d)}$. Parity also makes it orthogonal to $P_k^{(d)}$. Its expansion in the orthogonal polynomial basis therefore has the form $tP_k^{(d)} = a_k P_{k+1}^{(d)} + b_k P_{k-1}^{(d)}$, defining real coefficients a_k, b_k. We put $a_0 = 1$, since $tP_0^{(d)} = P_1^{(d)}$. Evaluation at $t = 1$ and integration against $P_{k-1}^{(d)}$ give, respectively,

$$a_k + b_k = 1, \qquad \frac{b_k}{N(d,k-1)} \overset{⑦}{=} \int_{-1}^{1} tP_k^{(d)}(t)P_{k-1}^{(d)}(t)\mathrm{d}\nu_d(t) \overset{⑧}{=} \frac{a_{k-1}}{N(d,k)}.$$

Equality ⑦ isolates the degree-$(k-1)$ coefficient using its squared norm. Equality ⑧ follows by applying multiplication by t to $P_{k-1}^{(d)}$. Orthogonality leaves only its degree-k component. Since $N(d,0) = 1$ and $N(d,1) = d$, we obtain $b_1 = 1/d$ and $a_1 = (d-1)/d$. For $k \geq 2$, formula (86) gives $N(d,k-1)/N(d,k) = k(2k+d-4)/[(k+d-3)(2k+d-2)]$. Induction in the preceding identities consequently gives $b_k = k/(2k+d-2)$ and $a_k = (k+d-2)/(2k+d-2)$. Substitution into the polynomial expansion proves (91). This derivation uses the norm identity already established and does not depend on the subsequent ReLU coefficient calculation.

Step 4. Integrable functions of inner products. A finite harmonic projection avoids imposing a square-integrability assumption on the function of the inner product. We fix a measurable a with $\int |a|\,\mathrm{d}\nu_d < \infty$ and define, for $\mathbf{q} \in \mathcal{X}$,

$$F_{\mathbf{x}}(\mathbf{q}) := \int_{\mathcal{X}} a(\mathbf{x}^\top \mathbf{v}) Z_\ell(\mathbf{q},\mathbf{v})\mathrm{d}\mu(\mathbf{v}) \overset{⑨}{=} \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{q}) \int_{\mathcal{X}} a(\mathbf{x}^\top \mathbf{v}) Y_{\ell,j}(\mathbf{v})\mathrm{d}\mu(\mathbf{v}).$$

Equality ⑨ expands Z_ℓ in its finite orthonormal basis sum and interchanges that sum with the integral. The vector $\mathbf{v}$ is the integration variable. Boundedness of the basis functions and the fact that $\mathbf{x}^\top \mathbf{v}$ has law ν_d ensure absolute convergence of every integral. The finite sum requires only integrability, rather than square integrability, of a. The finite expansion shows that $F_{\mathbf{x}}$ belongs to $\mathcal{H}_\ell$. For $\mathbf{Rx} = \mathbf{x}$, the change of variables $\mathbf{v} = \mathbf{Ru}$ and the simultaneous invariance of Z_ℓ give $F_{\mathbf{x}}(\mathbf{Rq}) = F_{\mathbf{x}}(\mathbf{q})$. Step 1 therefore shows that $F_{\mathbf{x}}$ is a scalar multiple of $P_\ell^{(d)}(\mathbf{x}^\top \cdot)$. Its value at $\mathbf{x}$ is $N(d,\ell)\int_{-1}^{1} a(t)P_\ell^{(d)}(t)\mathrm{d}\nu_d(t)$. Using $P_\ell^{(d)}(1) = 1$ and the addition formula, we obtain $F_{\mathbf{x}}(\mathbf{q}) \overset{①}{=} Z_\ell(\mathbf{x},\mathbf{q}) \int a(t)P_\ell^{(d)}(t)\mathrm{d}\nu_d(t)$. Equality ① identifies the invariant harmonic by its value at $\mathbf{x}$. Comparing the coefficients of $Y_{\ell,j}(\mathbf{q})$ in these two finite expansions proves (92) for every $\mathbf{x}$ and j. □

Lemma E.4. For every fixed integer $\ell_0 \geq 1$, as $d \to \infty$,

$$N(d,\ell) = \frac{d^\ell}{\ell!}(1 + O_\ell(d^{-1})), \quad 1 \leq \ell \leq \ell_0, \qquad r_0 = \frac{d^{\ell_0}}{\ell_0!}(1 + O_{\ell_0}(d^{-1})). \tag{94}$$

Proof. **Outline.** We identify the leading power of d in each harmonic dimension and then sum the finitely many degrees. Recall $r_0 = \sum_{\ell=0}^{\ell_0} N(d,\ell)$. For fixed $\ell \geq 1$, the first binomial coefficient in (86) is $d(d+1)\cdots(d+\ell-1)/\ell! \overset{①}{=} d^\ell(1+O_\ell(d^{-1}))/\ell!$, where ① factors d from each term. The subtracted coefficient is $O_\ell(d^{\ell-2})$ for $\ell \geq 2$ and zero for $\ell = 1$, proving the first assertion of (94). The sum of lower-degree dimensions, including $N(d,0) = 1$, is $O_{\ell_0}(d^{\ell_0-1})$, giving the formula for r_0. □

E.2 Concentration and Polynomial Moments

Matrix Bernstein controls the empirical covariance conditional on the learned representation, while polynomial hypercontractivity controls exceptional target evaluations. In the matrix statement, k, p are positive integers and sums range over $i = 1, \ldots, k$. Lemma F.5 gives the separate direction bound using spherical moments and a finite covering.

Theorem E.5 (Matrix Bernstein inequality (Tropp, 2012))**.** Let $\mathbf{X}_1, \ldots, \mathbf{X}_k$ be independent centered real symmetric matrices of order p, with $\|\mathbf{X}_i\|_2 \leq R$ almost surely for $R > 0$. If $v \geq \left\|\sum_i \mathbb{E}\left[\mathbf{X}_i^2\right]\right\|_2$, then for $t > 0$,

$$\Pr\left[\left\|\sum_i \mathbf{X}_i\right\|_2 \geq t\right] \leq 2p \exp\left(-\frac{t^2}{2v + (2/3)Rt}\right). \tag{95}$$

Proof. **Outline.** Apply the one-sided matrix Bernstein bound to the sum and its negative, then take a union bound. Let $\mathbf{S} = \sum_i \mathbf{X}_i$ and write $\lambda_{\max}$ for the largest eigenvalue. Tropp (2012, Theorem 6.1) applies to both signs because the centering, norm, and variance hypotheses are unchanged. Consequently,

$$\Pr\left[\|\mathbf{S}\|_2 \geq t\right] \overset{①}{\leq} \Pr\left[\lambda_{\max}(\mathbf{S}) \geq t\right] + \Pr\left[\lambda_{\max}(-\mathbf{S}) \geq t\right] \overset{②}{\leq} 2p \exp\left(-\frac{t^2}{2v + (2/3)Rt}\right).$$

Inequality ① uses the spectral characterization of the norm and the union bound. Inequality ② applies the cited theorem, proving (95). If $v = 0$, every centered summand vanishes almost surely and the conclusion still holds. □

Lemma E.6 (Polynomial moments on the sphere)**.** If p is the restriction to $\mathbb{S}^{d-1}$ of a polynomial of degree at most L, then

$$\|p\|_{L^q} \leq (q-1)^{L/2} \|p\|_{L^2}, \qquad 2 \leq q < \infty. \tag{96}$$

In particular, the polynomial target in (30) satisfies $\|f^*\|_{L^4}^2 \leq 3^L B_*^2$. This bound concerns the target values, independently of the noise law.

Proof. **Outline.** Step 1 obtains a dimension-independent polynomial moment bound from spherical hypercontractivity. Step 2 specializes it to the fourth moment of the target.

Step 1. Polynomial moments. We reverse Poisson smoothing on the finite harmonic expansion before applying hypercontractivity. By Theorem E.2, write $p = \sum_{\ell=0}^{L} p_\ell$ with $p_\ell \in \mathcal{H}_\ell$, orthogonal under the uniform probability measure μ. On finite harmonic sums, define the spherical Poisson operator by $\mathcal{P}_t(\sum_\ell h_\ell) = \sum_\ell t^\ell h_\ell$ for $0 \leq t \leq 1$, where $h_\ell \in \mathcal{H}_\ell$ and the degree-zero factor is always one.

The hypercontractive inequality for this operator is $\|\mathcal{P}_t h\|_{L^q} \leq \|h\|_{L^2}$ for $2 \leq q < \infty$ and $0 \leq t \leq (q-1)^{-1/2}$, as stated in Montanaro (2012, Theorem 7). We apply it with $h = \sum_{\ell=0}^{L} (q-1)^{\ell/2} p_\ell$ and $t = (q-1)^{-1/2}$. Then $\mathcal{P}_t h = p$, and orthogonality gives

$$\|p\|_{L^q}^2 \overset{①}{\leq} \|h\|_{L^2}^2 \overset{②}{=} \sum_{\ell=0}^{L} (q-1)^\ell \|p_\ell\|_{L^2}^2 \overset{③}{\leq} (q-1)^L \|p\|_{L^2}^2.$$

Inequality ① is hypercontractivity, equality ② is orthogonality, and ③ bounds every weight by $(q-1)^L$ and uses orthogonality again. This proves (96).

Step 2. The target fourth moment. Setting $p = f^*$ and $q = 4$ in (96) gives $\|f^*\|_{L^4}^2 \leq 3^L B_*^2$. Lemma F.8 uses this bound after averaging over the noise, so no fourth noise moment is needed. □

F Detailed Proofs

The proof order follows the approximation and noise costs identified in the introduction. Appendices F.1–F.3 construct and retain the polynomial comparison predictor. Appendices F.4 and F.5 then control empirical second moments and finite output GD. Appendix F.6 proves the population approximation comparisons. With these ingredients in place, Appendix F.7 proves Theorems C.1 and 4.1.

Throughout, $\mathscr{F}_1$ fixes the directions and observations in the first training sample, and $\mathscr{F}_2$ additionally fixes the inputs in the second training sample. Conditioning on $\mathscr{F}_2$ leaves only the noises in the second training sample to average.

F.1 Biased ReLU Coefficients and Approximation

We first estimate the harmonic coefficients needed to construct the polynomial comparison predictor and bound the coefficient norm retained by learnable channel attention. Appendix F.6 uses this comparison to bound the approximation cost for the learned representation.

Throughout this subsection, ℓ is a nonnegative integer and $j \in [N(d,\ell)]$ whenever a harmonic coordinate is indexed. Recall that $\sigma(t) = \max\{t, 0\}$, $P_\ell^{(d)}$ is the normalized degree-ℓ Gegenbauer

polynomial, and ν_d is the probability law of one coordinate of a uniform unit vector. For $\mathbf{q} \in \mathcal{X}$, put $\phi_{\mathbf{q}}(\mathbf{x}) = \sigma(\sqrt{d}\,\mathbf{q}^\top \mathbf{x} + 1/2)$. For every $\ell \geq 0$, define its harmonic coefficient by

$$\xi_{\ell,d} := \int_{-1}^{1} \sigma(\sqrt{d}\,t + 1/2) P_\ell^{(d)}(t) \mathrm{d}\nu_d(t). \tag{97}$$

The Funk–Hecke formula gives

$$\int_{\mathcal{X}} \phi_{\mathbf{q}}(\mathbf{x}) Y_{\ell,j}(\mathbf{q}) \mathrm{d}\mu(\mathbf{q}) = \xi_{\ell,d} Y_{\ell,j}(\mathbf{x}). \tag{98}$$

Lemma F.1. For every fixed integer $L \geq 1$, there are a constant $c_L > 0$ and an integer $d_L \geq 3$ such that

$$\xi_{\ell,d}^2 \geq \frac{c_L}{N(d,\ell)}, \qquad 0 \leq \ell \leq L, \quad d \geq d_L. \tag{99}$$

For every $d \geq 3$, the second moment $\kappa_d := \|\phi_{\mathbf{q}}\|_{L^2}^2$ is independent of $\mathbf{q} \in \mathcal{X}$ and satisfies $\kappa_d \leq 5/4$. This moment bound does not require $d \geq d_L$.

Proof. **Outline.** We first show that the coefficients of normalized Gegenbauer polynomials converge to those of normalized Hermite polynomials. We then justify convergence of their integrals against biased ReLU. We compute the limiting coefficient and prove that it is nonzero. Finally, we bound the feature second moment.

Step 1. Convergence of the orthogonal polynomials. A Gaussian limit turns the high-dimensional spherical coefficient into a fixed one-dimensional integral whose sign can be checked. We recall that $P_\ell^{(d)}(1) = 1$ and $N(d,\ell)$ is the harmonic dimension. We define $p_{\ell,d}(u) = \sqrt{N(d,\ell)} P_\ell^{(d)}(u/\sqrt{d})$. Formula (90) normalizes its squared integral to one under the law of a scaled spherical coordinate. We define the probabilists' Hermite polynomials by $\mathrm{He}_0 = 1$, $\mathrm{He}_1(u) = u$, and $\mathrm{He}_{k+1}(u) = u\,\mathrm{He}_k(u) - k\,\mathrm{He}_{k-1}(u)$. The recurrence (91) gives, for $k \geq 1$,

$$p_{k+1,d}(u) \overset{①}{=} \frac{2k+d-2}{\sqrt{d}(k+d-2)} \sqrt{\frac{N(d,k+1)}{N(d,k)}}\, u p_{k,d}(u) - \frac{k}{k+d-2} \sqrt{\frac{N(d,k+1)}{N(d,k-1)}}\, p_{k-1,d}(u).$$

Equality ① rescales the Gegenbauer recurrence by $u/\sqrt{d}$ and the harmonic normalization. For each fixed $k \geq 1$, Lemma E.4 and $N(d,0) = 1$ give $\frac{N(d,k+1)}{N(d,k)} \sim \frac{d}{k+1}$ and $\frac{N(d,k+1)}{N(d,k-1)} \sim \frac{d^2}{k(k+1)}$. Substitution into the preceding recurrence shows that its two coefficients tend to $1/\sqrt{k+1}$ and $\sqrt{k/(k+1)}$, respectively. The initial polynomials are exactly $p_{0,d} = 1$ and $p_{1,d}(u) = u$. The limiting recurrence coefficients are those of the normalized Hermite polynomials. Induction on k therefore shows that the coefficient of each power of u in $p_{k,d}$ converges to the corresponding coefficient in $\mathrm{He}_k/\sqrt{k!}$. This is the coefficientwise convergence used below. In particular, all coefficients are uniformly bounded for each fixed degree when d is sufficiently large.

Step 2. Convergence of the ReLU integrals. Convergence in distribution alone does not control the unbounded ReLU integrand, so we also need a uniform tail estimate. We let $\mathbf{q}$ be a single uniform unit vector and set $U_d = \sqrt{d}\,q_1$. This $\mathbf{q}$ is a generic integration variable, distinct from the ordered direction collection $\mathbf{Q}$. Let $G_1, G_2, \ldots$ be independent standard normal variables on one auxiliary probability space, with $G = G_1$. The standard Gaussian vector $(G_1, \ldots, G_d)$ has a radial density, so its direction is uniform on the sphere. Consequently, $G_1/(d^{-1}\sum_{i=1}^d G_i^2)^{1/2}$ has the law of U_d. Its denominator converges almost surely to one by the law of large numbers, so this representation proves $U_d \to G$ in distribution.

Symmetry gives zero odd moments. For an integer $k \geq 1$, the change of variable $s = t^2$ in the integral over $t \in [-1,1]$ gives $\int_{-1}^1 t^{2k}(1-t^2)^{(d-3)/2}\mathrm{d}t = \Gamma(k+1/2)\Gamma((d-1)/2)/\Gamma(k+d/2)$. Write $(2k-1)!! = 1 \cdot 3 \cdots (2k-1)$. Multiplication by the density constant in (89) and by d^k therefore gives

$$\mathbb{E}\left[U_d^{2k}\right] \overset{②}{=} \frac{d^k \Gamma(k+1/2)\Gamma(d/2)}{\Gamma(1/2)\Gamma(d/2+k)} \overset{③}{=} \frac{d^k (2k-1)!!}{d(d+2)\cdots(d+2k-2)} \leq (2k-1)!!, \qquad k \geq 1. \tag{100}$$

Equality ② uses the beta integral and the coordinate density. Equality ③ uses the Gamma recurrence, after which each denominator factor is at least d. Step 1 and the linear growth of ReLU give $|\sigma(u+1/2)p_{\ell,d}(u)| \leq C_\ell(1+|u|^{\ell+1})$ for all sufficiently large d. Formula (100) bounds the second

moments of these integrands uniformly in d. The integrands are therefore uniformly integrable, and they converge uniformly on each compact interval. More explicitly, for $R \geq 1$, Cauchy–Schwarz and $\mathbb{E}\left[U_d^2\right] = 1$ give

$$\mathbb{E}\left[|\sigma(U_d + 1/2)p_{\ell,d}(U_d)|\, \mathbb{1}_{\{|U_d|>R\}}\right] \overset{④}{\leq} C_\ell \Pr\left[|U_d| > R\right]^{1/2} \overset{⑤}{\leq} C_\ell / R.$$

Inequality ④ is Cauchy–Schwarz, and ⑤ uses $\mathbb{E}\left[U_d^2\right] = 1$ and Markov's inequality. The Gaussian integrand has the same tail bound because its second moment is finite. On compact intervals, coefficientwise convergence and convergence in distribution apply after inserting a continuous cutoff. Removing the cutoff then gives

$$\sqrt{N(d,\ell)}\xi_{\ell,d} \xrightarrow{⑥} \zeta_\ell := \frac{\mathbb{E}\left[\sigma(G + 1/2)\,\mathrm{He}_\ell(G)\right]}{\sqrt{\ell!}}.$$

Limit ⑥ first passes to the limit with a continuous cutoff, then removes it using ④–⑤ and the Gaussian tail bound.

Step 3. Nonvanishing of the limit. The bias places the boundary evaluation at a noninteger rational point, where a monic integer-coefficient Hermite polynomial cannot vanish. We let $\varphi(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ and $\Phi(u) = \int_{-\infty}^{u} \varphi(t)\mathrm{d}t$. Direct integration gives $\zeta_0 = \varphi(1/2) + \Phi(1/2)/2 > 0$ and $\zeta_1 = \Phi(1/2) > 0$. Differentiating the Hermite recurrence gives $\mathrm{He}_j' = j\,\mathrm{He}_{j-1}$ for $j \geq 1$ by induction. Together with $\varphi'(u) = -u\varphi(u)$, this proves $(\varphi\,\mathrm{He}_{j-1})' = -\varphi\,\mathrm{He}_j$. For $\ell \geq 2$, we apply this identity twice on $[-1/2, \infty)$. In the first integration by parts, the lower boundary term vanishes because $u + 1/2 = 0$ there. All upper boundary terms vanish by Gaussian decay. As a result,

$$\mathbb{E}\left[\sigma(G + 1/2)\,\mathrm{He}_\ell(G)\right] \overset{⑦}{=} \int_{-1/2}^{\infty} \varphi(u)\,\mathrm{He}_{\ell-1}(u)\mathrm{d}u \overset{⑧}{=} \varphi(1/2)\,\mathrm{He}_{\ell-2}(-1/2).$$

Equality ⑦ is the first integration by parts, and ⑧ is the second, leaving only the lower boundary term. The Hermite recurrence gives monic polynomials with integer coefficients, so the rational-root theorem excludes $-1/2$ as a root, also for $\mathrm{He}_0 = 1$. Hence every ζ_ℓ is nonzero. For the finitely many degrees $\ell \leq L$, choose a common $d_L \geq 3$ such that $|\sqrt{N(d,\ell)}\xi_{\ell,d}| \geq |\zeta_\ell|/2$ when $d \geq d_L$. Taking $c_L = \min_{0\leq\ell\leq L} \zeta_\ell^2/4$ proves (99). Here L is fixed. Appendix F.2 contrasts this result with the exact odd-degree obstruction without a bias.

Step 4. Feature second moment. The second moment controls the cost of the later Monte Carlo construction. Rotation invariance makes $\kappa_d = \|\phi_{\mathbf{q}}\|_{L^2}^2$ independent of the fixed direction, and $\kappa_d \overset{⑨}{\leq} \mathbb{E}_P\left[(\sqrt{d}\,\mathbf{q}^\top\mathbf{x} + 1/2)^2\right] \overset{①}{=} 5/4$. Here ⑨ uses $\sigma(t)^2 \leq t^2$, and ① uses the centered, unit-variance spherical coordinate. □

Lemma F.2. Under the target and dimension assumptions of the main theorem, there are comparison coefficients $\boldsymbol{a}^\circ \in \mathbb{R}^{m/2}$, depending only on the target and the sampled directions, such that

$$f^\circ(\mathbf{x}) := \sqrt{\frac{2}{m}} \sum_{j=1}^{m/2} a_j^\circ \phi_j(\mathbf{x}), \tag{101}$$

$$\mathbb{E}_{\mathbf{Q}}\left[\|f^\circ - f^*\|_{L^2}^2\right] \leq \frac{2C_L\Gamma_*^2}{m}, \qquad \mathbb{E}_{\mathbf{Q}}\left[\|\boldsymbol{a}^\circ\|_2^2\right] \leq C_L\Gamma_*^2. \tag{102}$$

Proof. **Outline.** We recall that $\mathbb{E}_{\mathbf{Q}}[\cdot]$ averages over the sampled directions $\mathbf{Q} = (\vec{\mathbf{q}}_1, \ldots, \vec{\mathbf{q}}_{m/2})$ only. We represent the target as an integral of ReLU features, approximate that integral by independent opposite-direction pairs, and bound its mean squared error. The two features within each pair are dependent and are treated together.

Step 1. Integral representation. We divide each harmonic coefficient by the ReLU coefficient so that integrating the features reproduces the target exactly. We recall that $\xi_{\ell,d}$ is defined in (97) and is nonzero for $\ell \leq \ell_0$ under Lemma F.1. We define $v(\mathbf{q}) := \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)} \frac{\beta_{\ell,j}}{\xi_{\ell,d}} Y_{\ell,j}(\mathbf{q})$. The Funk–Hecke identity (98), applied to this finite sum, gives

$$f^*(\mathbf{x}) \overset{①}{=} \int_{\mathcal{X}} v(\mathbf{q})\phi_{\mathbf{q}}(\mathbf{x})\mathrm{d}\mu(\mathbf{q}), \qquad \|v\|_{L^2}^2 \overset{②}{=} \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)} \frac{\beta_{\ell,j}^2}{\xi_{\ell,d}^2} \overset{③}{\leq} C_L\Gamma_*^2. \tag{103}$$

Equality ① is Funk–Hecke, ② is orthonormality, and ③ is Lemma F.1. The ReLU function is 1-Lipschitz and $\int \mathbf{x}\mathbf{x}^\top \mathrm{d}\mu = \mathbf{I}_d/d$, so $\|\phi_{\mathbf{q}} - \phi_{\mathbf{q}'}\|_{L^2}^2 \le d\int[(\mathbf{q}-\mathbf{q}')^\top \mathbf{x}]^2 \mathrm{d}\mu = \|\mathbf{q}-\mathbf{q}'\|_2^2$. Together with continuity of v, this makes $\mathbf{q} \mapsto v(\mathbf{q})\phi_{\mathbf{q}}$ continuous into the separable space $L^2(P)$. Its second moment is $\kappa_d \|v\|_{L^2}^2 < \infty$, which also implies integrability and justifies the Hilbert-space expectations below.

The reciprocal coefficients $1/\xi_{\ell,d}$ compensate for attenuation of each degree. Equation (103) therefore retains the weighted cost Γ_*^2 from Section A.2, rather than replacing it by $r_0 B_*^2$.

Step 2. Approximation by independent pairs. Averaging independent pairs preserves the architecture while allowing the mean-square error to decrease inversely with the number of pairs. Set $a_j^\circ = v(\vec{\mathbf{q}}_j)/\sqrt{m/2}$ for all $j \in [m/2]$. For a uniform direction $\mathbf{q}$, define the $L^2(P)$-valued random function $V(\mathbf{q}) := \frac{1}{2}\big[v(\mathbf{q})\phi_{\mathbf{q}} + v(-\mathbf{q})\phi_{-\mathbf{q}}\big]$. Then $f^\circ = (4/m)\sum_{j=1}^{m/4} V(\vec{\mathbf{q}}_j)$ averages independent $L^2(P)$-valued functions with mean f^*, since $\mathbf{q}$ and $-\mathbf{q}$ are both uniform. Convexity gives $\mathbb{E}\left[\|V(\mathbf{q})\|_{L^2}^2\right] \le \kappa_d \|v\|_{L^2}^2 < \infty$. This bound justifies successive integration of the cross terms in the squared error:

$$\mathbb{E}_{\mathbf{Q}}\left[\|f^\circ - f^*\|_{L^2}^2\right] \overset{④}{=} \frac{4}{m}\Big(\mathbb{E}\left[\|V(\mathbf{q})\|_{L^2}^2\right] - B_*^2\Big) \overset{⑤}{\le} \frac{4\kappa_d\|v\|_{L^2}^2}{m} \overset{⑥}{\le} \frac{5\|v\|_{L^2}^2}{m}. \tag{104}$$

Equality ④ cancels cross terms between independent centered pairs. Inequality ⑤ uses convexity and drops $B_*^2 \ge 0$, and ⑥ uses $\kappa_d \le 5/4$. Also, every $\vec{\mathbf{q}}_j$ has uniform marginal law, so $\mathbb{E}_{\mathbf{Q}}\left[\|\boldsymbol{a}^\circ\|_2^2\right] = (2/m)\sum_{j=1}^{m/2}\mathbb{E}\left[v(\vec{\mathbf{q}}_j)^2\right] = \|v\|_{L^2}^2$. This identity needs no independence within an opposite pair. Combining it with (104) and (103) proves (102), with the numerical factor absorbed in C_L. The comparison function depends only on the target and directions and is not computed by the training algorithm. □

F.2 Why the ReLU Bias Is Needed

The approximation in Lemma F.2 uses linear combinations of the initial ReLU features. We first show precisely which target components this approach would miss without the bias. We then explain why the chosen bias $1/2$ supplies those components, with the quantitative coefficient bounds needed for finite width. Throughout this subsection, $\mathcal{X} = \mathbb{S}^{d-1}$, μ is the uniform probability measure, and $\mathcal{H}_\ell$ is the degree-ℓ harmonic space from Section A.2.

Let $\mathcal{V}_{\mathrm{ub}}$ be the space of all finite real linear combinations of the constant function 1 and the unbiased ReLU features $\mathbf{x} \mapsto \sigma(\sqrt{d}\,\mathbf{q}^\top \mathbf{x})$ with $\mathbf{q} \in \mathcal{X}$. Allowing a constant enlarges the prediction space, so an approximation lower bound for $\mathcal{V}_{\mathrm{ub}}$ also applies when no output constant is allowed.

Proposition F.3 (The odd-degree obstruction without a ReLU bias)**.** For every $f \in \mathcal{V}_{\mathrm{ub}}$, the odd part $f_{\mathrm{odd}}(\mathbf{x}) := (f(\mathbf{x}) - f(-\mathbf{x}))/2$ belongs to $\mathcal{H}_1$. For every odd $\ell \ge 3$, the orthogonal projection of f onto $\mathcal{H}_\ell$ is zero. Equivalently, $\langle f, Y_{\ell,j}\rangle_{L^2} = 0$ for all $j \in [N(d,\ell)]$. For the polynomial target in (30),

$$\inf_{f \in \mathcal{V}_{\mathrm{ub}}} \|f - f^*\|_{L^2}^2 \ge \sum_{\substack{3 \le \ell \le \ell_0 \\ \ell \text{ odd}}} \sum_{j=1}^{N(d,\ell)} \beta_{\ell,j}^2. \tag{105}$$

An empty sum is zero. The same bound holds for the $L^2(P)$ closure of $\mathcal{V}_{\mathrm{ub}}$.

Proof. **Outline.** Step 1 separates a finite ReLU combination into its linear odd part and its even part. Step 2 uses the symmetry of the uniform measure and orthogonality of distinct harmonic spaces to show that every higher odd-degree projection is zero. Step 3 converts these missing projections into the approximation lower bound and extends it to $L^2(P)$ limits.

Step 1. The odd and even parts. Parity determines the obstruction. Fix $f \in \mathcal{V}_{\mathrm{ub}}$ and write $f(\mathbf{x}) = c + \sum_{j=1}^r a_j \sigma(\sqrt{d}\,\vec{\mathbf{q}}_j^\top \mathbf{x})$ with $r \ge 1$, real coefficients, and directions $\vec{\mathbf{q}}_j \in \mathcal{X}$. These quantities are local to this representation. Continuity on the compact sphere gives $f \in L^2(P)$. Subtracting its

values at opposite inputs gives

$$f_{\mathrm{odd}}(\mathbf{x}) \overset{①}{=} \frac{\sqrt{d}}{2} \sum_{j=1}^{r} a_j \overrightarrow{\mathbf{q}}_j^{\top} \mathbf{x}. \tag{106}$$

Equality ① uses $\sigma(t) - \sigma(-t) = t$. The linear form in (106) belongs to $\mathcal{H}_1$. Define $f_{\mathrm{even}}(\mathbf{x}) := (f(\mathbf{x}) + f(-\mathbf{x}))/2$. Adding the two values gives

$$f_{\mathrm{even}}(\mathbf{x}) \overset{②}{=} c + \frac{\sqrt{d}}{2} \sum_{j=1}^{r} a_j \left| \overrightarrow{\mathbf{q}}_j^{\top} \mathbf{x} \right|, \qquad f = f_{\mathrm{odd}} + f_{\mathrm{even}}.$$

Equality ② uses $\sigma(t)+\sigma(-t) = |t|$. As a result, $f_{\mathrm{even}}(-\mathbf{x}) = f_{\mathrm{even}}(\mathbf{x})$ and $f_{\mathrm{odd}}(-\mathbf{x}) = -f_{\mathrm{odd}}(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$.

Step 2. Higher odd-degree projections vanish. For odd $\ell \geq 3$ and $j \in [N(d,\ell)]$, homogeneity gives $Y_{\ell,j}(-\mathbf{x}) = -Y_{\ell,j}(\mathbf{x})$. Invariance of μ under $\mathbf{x} \mapsto -\mathbf{x}$ therefore eliminates the even part:

$$\langle f_{\mathrm{even}}, Y_{\ell,j} \rangle_{L^2} \overset{③}{=} \int_{\mathcal{X}} f_{\mathrm{even}}(-\mathbf{x}) Y_{\ell,j}(-\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \overset{④}{=} - \langle f_{\mathrm{even}}, Y_{\ell,j} \rangle_{L^2}.$$

Equality ③ changes variables by $\mathbf{x} \mapsto -\mathbf{x}$, and ④ uses opposite parities. Cauchy–Schwarz ensures integrability, so $\langle f_{\mathrm{even}}, Y_{\ell,j} \rangle_{L^2} = 0$. Also $\langle f_{\mathrm{odd}}, Y_{\ell,j} \rangle_{L^2} = 0$ because $f_{\mathrm{odd}} \in \mathcal{H}_1$ and distinct harmonic degrees are orthogonal by Proposition E.1. Adding gives $\langle f, Y_{\ell,j} \rangle_{L^2} = 0$. The projection $\sum_j \langle f, Y_{\ell,j} \rangle_{L^2} Y_{\ell,j}$ onto $\mathcal{H}_\ell$ is therefore zero. Degree one is excluded because it contains the possibly nonzero odd part in (106).

Step 3. The lower bound on regression risk. Every missing harmonic coefficient contributes unavoidable squared error, and continuity of the coefficient functionals extends this obstruction to limits. For each odd ℓ with $3 \leq \ell \leq \ell_0$, Step 2 and (30) imply $\langle f - f^*, Y_{\ell,j} \rangle_{L^2} = -\beta_{\ell,j}$. Bessel's inequality, applied to this finite orthonormal family, gives

$$\|f - f^*\|_{L^2}^2 \overset{⑤}{\geq} \sum_{\substack{3 \leq \ell \leq \ell_0 \\ \ell \text{ odd}}} \sum_{j=1}^{N(d,\ell)} \left| \langle f - f^*, Y_{\ell,j} \rangle_{L^2} \right|^2 \overset{⑥}{=} \sum_{\substack{3 \leq \ell \leq \ell_0 \\ \ell \text{ odd}}} \sum_{j=1}^{N(d,\ell)} \beta_{\ell,j}^2.$$

Inequality ⑤ is Bessel's inequality, and ⑥ uses the coefficients from Step 2. Taking the infimum proves (105), including when the sum is empty.

Finally, let $f_k \in \mathcal{V}_{\mathrm{ub}}$ converge to f in $L^2(P)$. For any fixed odd $\ell \geq 3$ and $j \in [N(d,\ell)]$, Step 2 and the unit norm of $Y_{\ell,j}$ give $\left| \langle f, Y_{\ell,j} \rangle_{L^2} \right| = \left| \langle f - f_k, Y_{\ell,j} \rangle_{L^2} \right| \overset{⑦}{\leq} \|f - f_k\|_{L^2} \longrightarrow 0$, where ⑦ is Cauchy–Schwarz and $\|Y_{\ell,j}\|_{L^2} = 1$. As a result, the same coefficients vanish for every f in the closure of $\mathcal{V}_{\mathrm{ub}}$, and Bessel's inequality gives the same approximation lower bound. □

For a nonzero target lying entirely in odd harmonic degrees three and above, the infimum in (105) equals $\|f^*\|_{L^2}^2$, since the zero predictor attains the lower bound. More generally, if the energy on the right-hand side stays bounded away from zero as d grows, the regression risk cannot vanish. This holds regardless of the number of unbiased ReLU features or how their directions and output coefficients are chosen.

The bias $1/2$ removes this obstruction for the initial approximation: Lemma F.1 proves nonzero coefficients at every fixed degree for sufficiently large d, and Lemma F.2 converts them into a comparison predictor with controlled coefficient norm. Proposition F.4 retains that predictor under attention. The obstruction in Proposition F.3 concerns linear combinations of unbiased ReLU features, rather than networks applying additional nonlinear attention to them.

F.3 Preservation by Attention and Feature Bounds

The benefit of the paired attention construction is that learning new products does not remove an available predictor or increase the coefficient norm needed to retain it. We prove the preservation identity first and then bound the feature energy needed for the statistical analysis.

Proposition F.4. For every $\boldsymbol{\tau} \in \mathbb{R}^m$ and every $\boldsymbol{a}^\circ \in \mathbb{R}^{m/2}$, define $\widetilde{\boldsymbol{a}}^\circ = [(\boldsymbol{a}^\circ)^\top, (\boldsymbol{a}^\circ)^\top]^\top$. Then

$$f(\boldsymbol{\tau}, \widetilde{\boldsymbol{a}}^\circ, \mathbf{x}) = f^\circ(\mathbf{x}), \qquad \|\widetilde{\boldsymbol{a}}^\circ\|_2^2 = 2\|\boldsymbol{a}^\circ\|_2^2, \tag{107}$$

where f° is given by (101). Form $\boldsymbol{\psi}(\mathbf{x}) = \sqrt{2/m}\mathbf{A}_{\boldsymbol{\tau}}(\mathbf{x})\mathbf{h}(\mathbf{x})$ and $\boldsymbol{\Sigma} = \int \boldsymbol{\psi}\boldsymbol{\psi}^\top \mathrm{d}\mu$ with this same attention vector $\boldsymbol{\tau}$. Then

$$\|\boldsymbol{\psi}(\mathbf{x})\|_2^2 \le \frac{2}{m}\sum_{j=1}^{m/2}\phi_j(\mathbf{x})^2 \le 2(d+1), \qquad \mathrm{tr}\,(\boldsymbol{\Sigma}) \le 5/4. \tag{108}$$

Proof. **Outline.** We use the fact that the two attention weights sum to one to preserve the comparison function. We then bound the feature vector pointwise and integrate this bound to control its population covariance.

Step 1. Exact preservation. Fix $\boldsymbol{\tau}$ and recall $f^\circ = \sqrt{2/m}\sum_j a_j^\circ \phi_j$. Substituting the duplicated coefficients into (29) gives $f(\boldsymbol{\tau}, \widetilde{\boldsymbol{a}}^\circ, \mathbf{x}) \overset{①}{=} f^\circ(\mathbf{x})$, where ① uses $g_{\boldsymbol{\tau}} + (1 - g_{\boldsymbol{\tau}}) = 1$. Duplication gives $\|\widetilde{\boldsymbol{a}}^\circ\|_2^2 = 2\|\boldsymbol{a}^\circ\|_2^2$, proving (107).

Step 2. Uniform feature bounds. Bounded gates control feature energy independently of the attention rule:

$$\|\boldsymbol{\psi}(\mathbf{x})\|_2^2 = \frac{2[g_{\boldsymbol{\tau}}(\mathbf{x})^2 + (1 - g_{\boldsymbol{\tau}}(\mathbf{x}))^2]}{m}\sum_{j=1}^{m/2}\phi_j(\mathbf{x})^2 \overset{②}{\le} \frac{2}{m}\sum_{j=1}^{m/2}\phi_j(\mathbf{x})^2.$$

Inequality ② uses $g_{\boldsymbol{\tau}}^2 + (1 - g_{\boldsymbol{\tau}})^2 \le 1$. Since both $\vec{\mathbf{q}}_j$ and $\mathbf{x}$ have norm one, each $\phi_j(\mathbf{x})^2 \le (\sqrt{d} + 1/2)^2 \le 2(d+1)$. This proves the pointwise bound in (108). The trace identity $\mathrm{tr}\,(\boldsymbol{\Sigma}) = \int \|\boldsymbol{\psi}(\mathbf{x})\|_2^2 \mathrm{d}\mu(\mathbf{x})$ and Lemma F.1 give $\mathrm{tr}\,(\boldsymbol{\Sigma}) \overset{③}{\le} (2/m)\sum_{j=1}^{m/2}\|\phi_j\|_{L^2}^2 = \kappa_d \le 5/4$. Inequality ③ integrates ② and uses the trace identity. Every bound holds for every finite $\boldsymbol{\tau}$, including a data-dependent choice. □

We next define an event on the sampled directions that gives a uniform bound over all input points:

$$\mathcal{E}_Q := \left\{ \frac{2d}{m}\sum_{j=1}^{m/2}\vec{\mathbf{q}}_j\vec{\mathbf{q}}_j^\top \preceq 2\mathbf{I}_d \right\}. \tag{109}$$

The event $\mathcal{E}_Q \in \mathscr{F}_1$ bounds the average squared scaled projection by 2, uniformly over unit inputs.

Lemma F.5. For $d \ge 3$ and $m \ge 4$ divisible by four, the direction event $\mathcal{E}_Q$ defined in (109) satisfies

$$\Pr\left[\mathcal{E}_Q^c\right] \le \min\left\{1, \exp(d\log 21 - m/64)\right\}. \tag{110}$$

On $\mathcal{E}_Q$, for every finite attention vector,

$$\sup_{\mathbf{x}\in\mathcal{X}}\|\boldsymbol{\psi}(\mathbf{x})\|_2^2 \le 9/4. \tag{111}$$

The lower bound in (49), with $C_0 = 512$, implies $\Pr\left[\mathcal{E}_Q^c\right] \le n^{-8}$ without using the upper width bound.

Proof. **Outline.** Step 1 uses spherical-coordinate moments to bound the upper tail of a quadratic form in each fixed direction. Step 2 constructs a finite covering and transfers these bounds to the largest eigenvalue of the direction covariance. Step 3 obtains the uniform feature bound and verifies the numerical constant.

Step 1. A fixed direction. Put $k = m/4$. Pairing each independent direction with its negative gives the direction covariance

$$\mathbf{A} = \frac{2d}{m}\sum_{j=1}^{m/2}\vec{\mathbf{q}}_j\vec{\mathbf{q}}_j^\top \overset{①}{=} \frac{d}{k}\sum_{j=1}^{k}\vec{\mathbf{q}}_j\vec{\mathbf{q}}_j^\top.$$

Equality ① pairs each direction with its negative, whose outer product is identical, and uses $k = m/4$. Fix a deterministic unit vector $\mathbf{u} \in \mathbb{R}^d$. For $j \in [k]$, let $X_j = d(\mathbf{u}^\top\vec{\mathbf{q}}_j)^2$. These variables are independent and nonnegative. Rotation invariance and (100) give $\mathbb{E}\left[X_j^r\right] \le (2r-1)!!$ for every integer $r \ge 1$. The zeroth moment is one. Expanding the exponential into its nonnegative series and

applying monotone convergence yield, for $0 \le t < 1/2$,

$$\mathbb{E}\left[\exp(tX_j)\right] \overset{②}{\le} 1 + \sum_{r=1}^{\infty} \frac{(2r-1)!!}{r!} t^r \overset{③}{=} (1-2t)^{-1/2}.$$

Inequality ② applies the moment bounds term by term under monotone convergence. Equality ③ is the binomial series. Markov's inequality, independence, and the choice $t = 3/16$ now give

$$\Pr\left[\mathbf{u}^\top \mathbf{A}\mathbf{u} \ge 8/5\right] \overset{④}{\le} \exp(-3k/10)(8/5)^{k/2} \overset{⑤}{\le} \exp(-k/16).$$

Inequality ④ applies exponential Markov and multiplies the independent moment bounds. For ⑤, $\log(8/5) < 19/40$ follows from $\exp(19/40) > 1 + 19/40 + (19/40)^2/2 + (19/40)^3/6 > 8/5$. As a result, $-3/10 + \frac{1}{2}\log(8/5) < -1/16$.

Step 2. A finite covering. We pass from fixed directions to the operator norm using a deterministic ε-net: a set $\mathcal{V} \subset \mathcal{X}$ within distance ε of every unit vector. Take $\varepsilon = 1/10$ and choose a maximal set with pairwise distances greater than ε. The open balls of radius $\varepsilon/2$ about its points are disjoint and lie in the ball of radius $1 + \varepsilon/2$. Comparing volumes gives $|\mathcal{V}| \le (1 + 2/\varepsilon)^d = 21^d$. This bound ensures that a greedy construction terminates, and maximality gives the covering property.

For a positive semidefinite matrix $\mathbf{A}$, choose a unit vector $\mathbf{u}$ attaining $\|\mathbf{A}\|_2 = \mathbf{u}^\top \mathbf{A}\mathbf{u}$ and choose $\mathbf{v} \in \mathcal{V}$ with $\|\mathbf{u} - \mathbf{v}\|_2 \le \varepsilon$. Then

$$\left|\mathbf{u}^\top \mathbf{A}\mathbf{u} - \mathbf{v}^\top \mathbf{A}\mathbf{v}\right| \overset{⑥}{\le} \left|(\mathbf{u}-\mathbf{v})^\top \mathbf{A}\mathbf{u}\right| + \left|\mathbf{v}^\top \mathbf{A}(\mathbf{u}-\mathbf{v})\right| \overset{⑦}{\le} 2\varepsilon\|\mathbf{A}\|_2.$$

Inequality ⑥ adds and subtracts $\mathbf{v}^\top \mathbf{A}\mathbf{u}$, and ⑦ uses the operator norm and the net distance. Hence $\|\mathbf{A}\|_2 > 2$ forces $\mathbf{v}^\top \mathbf{A}\mathbf{v} > 8/5$ at some net point. A union bound over Step 1 gives $\Pr\left[\mathcal{E}_Q^c\right] \overset{⑧}{\le} 21^d \exp(-k/16) = \exp(d\log 21 - m/64)$. Inequality ⑧ sums over at most 21^d deterministic net points. The probability bound of one completes (110). Only the $m/4$ independent directions enter this argument.

Step 3. Features and the stated probability. Pairing opposite directions bounds ReLU energy by a quadratic form, so the direction event controls every input simultaneously. Recall that $\sigma(t) = \max\{t, 0\}$ and that each sampled direction is paired with its negative. For every real t,

$$\frac{\sigma(t+1/2)^2 + \sigma(-t+1/2)^2}{2} \overset{⑨}{=} t^2 + \frac{1}{4} - \frac{1}{2}\sigma(|t| - 1/2)^2 \le t^2 + \frac{1}{4}. \tag{112}$$

For ⑨, both ReLU arguments are nonnegative when $|t| \le 1/2$. Otherwise only one is active, and expanding $(|t| + 1/2)^2/2$ gives the same identity. Summing over the $k = m/4$ pairs gives

$$\frac{2}{m}\sum_{j=1}^{m/2} \phi_j(\mathbf{x})^2 \overset{①}{\le} \frac{d}{k}\sum_{j=1}^{k} (\vec{\mathbf{q}}_j^\top \mathbf{x})^2 + \frac{1}{4} = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \frac{1}{4} \overset{②}{\le} \frac{9}{4} \qquad \text{on } \mathcal{E}_Q.$$

Inequality ① sums (112), and ② uses $\|\mathbf{A}\|_2 \le 2$ on $\mathcal{E}_Q$. The pointwise comparison in Step 2 of Proposition F.4 now proves (111) for every input and finite attention vector.

Finally, $512/64 = 8$ and $\log 21 < 8$. If $m \ge 512(d + H_n)$ with $H_n = \log(en)$, then $d\log 21 - m/64 \le d(\log 21 - 8) - 8H_n \le -8H_n$. Equation (110) therefore gives $\Pr\left[\mathcal{E}_Q^c\right] \le (en)^{-8} \le n^{-8}$. No relation between d and n is needed for this last implication. □

F.4 Concentration of the Second-Stage Covariance

We next compare the population and empirical covariances of the learned features. We condition on the first stage so that the feature map is fixed and the inputs in the second training sample remain independent. The resulting covariance inequality converts bounds on empirical squared predictions into bounds on their population squared norms.

Covariances and effective dimension. Define the learned population covariance by

$$\mathbf{\Sigma} := \int_{\mathcal{X}} \boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^\top \mathrm{d}\mu(\mathbf{x}). \tag{113}$$

The matrix in (113) is an uncentered second-moment matrix, with $\mathbf{c}^\top \mathbf{\Sigma}\mathbf{c} = \left\|\mathbf{c}^\top \boldsymbol{\psi}\right\|_{L^2}^2$ for $\mathbf{c} \in \mathbb{R}^m$. Subtracting the feature means would measure prediction variance and omit the squared prediction

mean. The empirical counterpart and effective dimension at scale $\lambda > 0$ are

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n_2} \sum_{i \in I_2} \boldsymbol{\psi}(\vec{\mathbf{x}}_i)\boldsymbol{\psi}(\vec{\mathbf{x}}_i)^\top, \qquad \widehat{\mathcal{N}}(\lambda) = \operatorname{tr}\left(\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}} + \lambda \mathbf{I}_m)^{-1}\right).$$

Each eigenvalue $u \geq 0$ contributes $u/(u+\lambda)$, so $0 \leq \widehat{\mathcal{N}}(\lambda) \leq \operatorname{rank}(\widehat{\mathbf{\Sigma}}) \leq m$. The filter comparison (126) relates these weights to fitted noise. We define $\widehat{\mathcal{N}}_0(\lambda)$ using $\boldsymbol{\psi}_0$ on the same inputs. These quantities are used only in the analysis.

Recall the sigma-fields from Appendix C.1: $\mathscr{F}_1$ fixes the learned map, and $\mathscr{F}_2$ also fixes the inputs in the second training sample. Conditional on $\mathscr{F}_1$, those inputs remain independent with law P. Both covariance matrices are $\mathscr{F}_2$-measurable.

We normalize by $\mathbf{\Sigma} + \lambda \mathbf{I}_m$ at the GD scale $\lambda = m^{-1}$. The positive shift permits singular feature covariances and controls low-energy directions through coefficient cost. Define

$$\mathcal{E}_S := \left\{ \left\| (\mathbf{\Sigma} + \lambda \mathbf{I}_m)^{-1/2} (\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})(\mathbf{\Sigma} + \lambda \mathbf{I}_m)^{-1/2} \right\|_2 \leq \frac{1}{2} \right\}. \tag{114}$$

This event belongs to $\mathscr{F}_2$, since the learned feature map and the inputs in the second training sample are measurable with respect to that sigma-field. Equivalently, $\mathcal{E}_S$ requires $\left|\mathbf{v}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{v}\right| \leq \frac{1}{2}\left(\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} + \lambda \|\mathbf{v}\|_2^2\right)$ for every $\mathbf{v} \in \mathbb{R}^m$. This equivalence follows by substituting $(\mathbf{\Sigma} + \lambda \mathbf{I}_m)^{1/2}\mathbf{v}$ in the norm bound. The added λ term allows error in directions of small population energy.

Lemma F.6. For every deterministic $m \geq 4$ divisible by four, on $\mathcal{E}_Q$ one has $\Pr[\mathcal{E}_S^c \mid \mathscr{F}_1] \leq 2m \exp(-n_2/(21m))$. As a result, $\Pr[(\mathcal{E}_Q \cap \mathcal{E}_S)^c] \leq \delta_{d,n,m}$, where (55) defines $\delta_{d,n,m}$. Under (49) with $C_0 = 512$, the bound on $\Pr[\mathcal{E}_S^c \mid \mathscr{F}_1]$ is at most n^{-8} on $\mathcal{E}_Q$. Consequently,

$$\mathcal{E} = \mathcal{E}_Q \cap \mathcal{E}_S, \qquad \Pr[\mathcal{E}^c] \leq 2n^{-8}. \tag{115}$$

On $\mathcal{E}_S$,

$$\mathbf{\Sigma} \preceq 2\widehat{\mathbf{\Sigma}} + \lambda \mathbf{I}_m. \tag{116}$$

In particular, every $\mathbf{c} \in \mathbb{R}^m$ with $\widehat{\mathbf{\Sigma}}\mathbf{c} = \mathbf{0}$ satisfies, on $\mathcal{E}_S$,

$$\left\|\mathbf{c}^\top \boldsymbol{\psi}\right\|_{L^2}^2 = \mathbf{c}^\top \mathbf{\Sigma} \mathbf{c} \leq m^{-1}\|\mathbf{c}\|_2^2. \tag{117}$$

Proof. **Outline.** We fix a first-stage outcome in the direction event $\mathcal{E}_Q$ from (109). We bound the norm and variance of the normalized summands from the second training sample, apply matrix Bernstein, and recover the unnormalized covariance comparison. The same quadratic-form inequality controls predictions whose evaluations on the second training sample vanish.

Step 1. Conditional summands. Recall $\lambda = m^{-1}$. Fix a first-stage outcome in $\mathcal{E}_Q$ and set $\mathbf{C} = \mathbf{\Sigma} + \lambda \mathbf{I}_m \succ 0$. For $i \in I_2$, define

$$\mathbf{Z}_i = \mathbf{C}^{-1/2}\boldsymbol{\psi}(\vec{\mathbf{x}}_i)\boldsymbol{\psi}(\vec{\mathbf{x}}_i)^\top \mathbf{C}^{-1/2}, \qquad \mathbf{B} = \mathbb{E}[\mathbf{Z}_i \mid \mathscr{F}_1] \overset{①}{=} \mathbf{C}^{-1/2}\mathbf{\Sigma}\mathbf{C}^{-1/2}.$$

Equality ① uses the conditional input law. The $\mathbf{Z}_i$ are conditionally independent with common mean $0 \preceq \mathbf{B} \preceq \mathbf{I}_m$. Equation (111) and $\mathbf{C} \succeq m^{-1}\mathbf{I}_m$ give $0 \preceq \mathbf{Z}_i \preceq K\mathbf{I}_m$, where $K = (9/4)m$.

Step 2. Norm and variance bounds. The normalized positive summands have a variance bound proportional to their mean, which is sharper than squaring their norm bound. Since $K \geq 1$, the inequalities $-\mathbf{I}_m \preceq \mathbf{Z}_i - \mathbf{B} \preceq K\mathbf{I}_m$ give $\|\mathbf{Z}_i - \mathbf{B}\|_2 \leq K$. The matrix inequality $\mathbf{Z}_i^2 \preceq K\mathbf{Z}_i$ follows by diagonalizing $\mathbf{Z}_i$. Because $\mathbf{B}$ is fixed under the conditioning, expansion of the centered square gives

$$\mathbb{E}\left[(\mathbf{Z}_i - \mathbf{B})^2 \mid \mathscr{F}_1\right] \overset{②}{=} \mathbb{E}\left[\mathbf{Z}_i^2 \mid \mathscr{F}_1\right] - \mathbf{B}^2 \overset{③}{\preceq} K\mathbf{B} \overset{④}{\preceq} K\mathbf{I}_m.$$

Equality ② expands the centered square using $\mathbb{E}[\mathbf{Z}_i \mid \mathscr{F}_1] = \mathbf{B}$. Inequality ③ uses $\mathbf{Z}_i^2 \preceq K\mathbf{Z}_i$ and drops $\mathbf{B}^2 \succeq 0$, and ④ uses $\mathbf{B} \preceq \mathbf{I}_m$. As a result, the variance parameter for the sum of n_2 centered matrices is at most $n_2 K$. This calculation does not require an individual $\mathbf{Z}_i$ to commute with $\mathbf{B}$.

Step 3. Recovering the covariance comparison. After concentration in normalized coordinates, a congruence returns the population-to-empirical comparison needed for prediction error. Theorem E.5, at $t = n_2/2$, gives

$$\Pr[\mathcal{E}_S^c \mid \mathscr{F}_1] \overset{⑤}{\leq} 2m \exp\left(-\frac{3n_2}{28K}\right) = 2m \exp\left(-\frac{n_2}{21m}\right). \tag{118}$$

Inequality ⑤ substitutes summand bound K, variance bound n_2K, and threshold $n_2/2$ into matrix Bernstein. The derivation of (118) uses no width condition. The decomposition $(\mathcal{E}_Q \cap \mathcal{E}_S)^c = \mathcal{E}_Q^c \cup (\mathcal{E}_Q \cap \mathcal{E}_S^c)$, conditional expectation, and (110) give the claimed bound by $\delta_{d,n,m}$. For $C_0 \geq 512$, the upper width condition and $n_2 \geq n/2$ imply $n_2/(21m) \geq 10\log(en)$. Since $m \leq n$, the last expression in (118) is at most $2n(en)^{-10} \leq n^{-8}$. Averaging the conditional probability on $\mathcal{E}_Q$ and adding $\Pr\left[\mathcal{E}_Q^c\right] \leq n^{-8}$ proves (115).

On $\mathcal{E}_S$, the defining norm bound (114) means that every eigenvalue of its symmetric normalized matrix lies in $[-1/2, 1/2]$. In particular, congruence by the positive definite matrix $\mathbf{C}^{1/2}$ gives $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} \overset{⑥}{\succeq} -\mathbf{C}/2 = -(\mathbf{\Sigma} + \lambda \mathbf{I}_m)/2$. Rearranging yields $\widehat{\mathbf{\Sigma}} \succeq \mathbf{\Sigma}/2 - \lambda \mathbf{I}_m/2$, which is (116). In ⑥, congruence preserves the positive semidefinite order. This argument does not require the two covariances to commute.

Applying (116) to $\mathbf{c} \in \ker \widehat{\mathbf{\Sigma}}$ gives (117). Finally, $\mathcal{E}_Q \in \mathscr{F}_1$ and $\mathcal{E}_S \in \mathscr{F}_2$, so $\mathcal{E} \in \mathscr{F}_2$ is independent of the noise in the second training sample. □

F.5 Exact Analysis of Discrete Gradient Descent

Output fitting makes the retained features and the learned products separately available to the predictor. The estimates below show how the finite GD horizon controls their coefficient cost and fitted noise. Combined with the preserved polynomial approximation in Appendix F.6, they give the risk bound for the learned feature kernel.

Let $\mathbf{Z} \in \mathbb{R}^{n_2 \times m}$ have rows $\boldsymbol{\psi}(\vec{\mathbf{x}}_i)^\top$ for $i \in I_2$, in increasing order. Set $\mathbf{D} = \mathbf{Z}/\sqrt{n_2}$ and let $\mathbf{y}_2 = [y_i]_{i \in I_2}$ in the same order. The normalized sample Gram matrix is $\mathbf{G} = \mathbf{D}\mathbf{D}^\top \in \mathbb{R}^{n_2 \times n_2}$, with entries $\widehat{K}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)/n_2$. As a result, $\widehat{\mathbf{\Sigma}} = \mathbf{D}^\top \mathbf{D}$ acts on coefficient vectors, while $\mathbf{G}$ acts on observation vectors. Their positive eigenvalues agree with multiplicities by singular value decomposition.

The output updates in stage two of Algorithm 1 use $\eta = 1/[4(d+1)]$ and run for $T = 4(d+1)m$ iterations, so $\eta T = m$. The deterministic stability identities below hold for every finite frozen attention vector. For $u \geq 0$, define the GD polynomials

$$q_T(u) = \eta \sum_{t=0}^{T-1} (1 - \eta u)^t, \qquad s_T(u) = u q_T(u). \tag{119}$$

For $u > 0$, summing the geometric series gives $q_T(u) = [1-(1-\eta u)^T]/u$, while $q_T(0) = \eta T = m$. The matrix $q_T(\widehat{\mathbf{\Sigma}})$ converts $\mathbf{Z}^\top \mathbf{y}_2/n_2$ into output coefficients. On an eigendirection of the sample Gram matrix $\mathbf{G}$ with eigenvalue u, the multiplier $s_T(u)$ is the fraction of the response fitted after T steps. The following lemma proves both identities and compares this fraction with $u/(u + m^{-1})$, making the training scale m^{-1} explicit.

Lemma F.7 (GD stability and exact representation)**.** Use the feature map (39) and the output updates (41)–(42), initialized at $\boldsymbol{a}(0) = \mathbf{0}$. For every realization of the directions, every finite attention vector, and every second training sample with finite responses,

$$\eta \left\| \widehat{\mathbf{\Sigma}} \right\|_2 \leq \eta \mathrm{tr}\left(\widehat{\mathbf{\Sigma}}\right) \leq \frac{1}{2}. \tag{120}$$

With ∇L_2 denoting the gradient of (40) with respect to $\boldsymbol{a}$, the loss satisfies

$$L_2(\boldsymbol{a}(t+1)) \leq L_2(\boldsymbol{a}(t)) - \frac{3\eta}{4} \|\nabla L_2(\boldsymbol{a}(t))\|_2^2, \qquad 0 \leq t < T. \tag{121}$$

Every iterate lies in the linear span of the feature vectors for the second training sample,

$$\boldsymbol{a}(t) \in \mathrm{span}\left\{ \boldsymbol{\psi}(\vec{\mathbf{x}}_i) : i \in I_2 \right\}, \qquad 0 \leq t \leq T. \tag{122}$$

The final iterate has the exact representation

$$\boldsymbol{a}(T) = q_T(\widehat{\mathbf{\Sigma}}) \mathbf{Z}^\top \mathbf{y}_2/n_2. \tag{123}$$

The vector of fitted responses is $\mathbf{Z}\boldsymbol{a}(T)$, and

$$\mathbf{Z}\boldsymbol{a}(T) = s_T(\mathbf{G})\mathbf{y}_2, \qquad 0 \preceq s_T(\mathbf{G}) \preceq \mathbf{I}_{n_2}. \tag{124}$$

As a result, fitting the responses does not increase their Euclidean norm. For every eigenvalue u of $\widehat{\mathbf{\Sigma}}$,

$$0 \le s_T(u) \le \min\{1, mu\}, \qquad uq_T(u)^2 \le m, \qquad u(1-\eta u)^{2T} \le \frac{1}{2em}. \tag{125}$$

For every $u \ge 0$ with $\eta u \le 1/2$, one also has

$$\frac{u}{u+m^{-1}} \le s_T(u) \le \min\{1, mu\} \le \frac{2u}{u+m^{-1}}. \tag{126}$$

These conclusions require no concentration event, lower bound on a positive eigenvalue, or width condition (49).

Proof. **Outline.** We first use the global feature bound to prove stability and training loss decrease. We then identify the span containing the iterates and expand the GD recurrence to obtain the fitted-response formula. Finally, we bound the scalar polynomial on the empirical spectrum, identify the comparison scale in (126), and prove that the fitted-response map does not increase Euclidean norms, treating zero eigenvalues explicitly.

Step 1. Stability and loss decrease. Stability is a deterministic property of the frozen feature matrix, so the loss bound can hold on every sample. We recall that $\boldsymbol{\psi}$ is the feature vector frozen after attention training, $\widehat{\mathbf{\Sigma}} = n_2^{-1}\sum_{i\in I_2}\boldsymbol{\psi}(\vec{\mathbf{x}}_i)\boldsymbol{\psi}(\vec{\mathbf{x}}_i)^\top$ is its positive semidefinite empirical covariance, and $\eta = 1/[4(d+1)]$. By (108), its trace is at most $2(d+1)$. Multiplication by $\eta = 1/[4(d+1)]$ proves (120). Fix an integer t with $0 \le t < T$ and, locally in this step, put $\mathbf{g} = \nabla L_2(\boldsymbol{a}(t)) = \widehat{\mathbf{\Sigma}}\boldsymbol{a}(t) - \mathbf{Z}^\top\mathbf{y}_2/n_2$. Here $\mathbf{g}$ denotes the current output gradient, rather than the first-stage gradient used earlier. The loss in (40) is quadratic, with Hessian $\widehat{\mathbf{\Sigma}}$. Its exact expansion and (120) give

$$L_2(\boldsymbol{a}(t)-\eta\mathbf{g}) \overset{①}{=} L_2(\boldsymbol{a}(t)) - \eta\|\mathbf{g}\|_2^2 + \frac{\eta^2}{2}\mathbf{g}^\top\widehat{\mathbf{\Sigma}}\mathbf{g} \overset{②}{\le} L_2(\boldsymbol{a}(t)) - \frac{3\eta}{4}\|\mathbf{g}\|_2^2.$$

Equality ① is the exact quadratic expansion, and ② uses $\eta\left\|\widehat{\mathbf{\Sigma}}\right\|_2 \le 1/2$. The update (42) identifies the left side with $L_2(\boldsymbol{a}(t+1))$, proving (121).

Step 2. Span and exact iterate. Each update is a linear combination of the feature vectors for the second training sample. Zero initialization therefore gives (122), placing every iterate in $(\ker\widehat{\mathbf{\Sigma}})^\perp$.

The recurrence $\boldsymbol{a}(t+1) = (\mathbf{I}_m - \eta\widehat{\mathbf{\Sigma}})\boldsymbol{a}(t) + \eta\mathbf{Z}^\top\mathbf{y}_2/n_2$ yields $\boldsymbol{a}(T) = \eta\sum_{t=0}^{T-1}(\mathbf{I}_m - \eta\widehat{\mathbf{\Sigma}})^t\mathbf{Z}^\top\mathbf{y}_2/n_2$, proving (123). This polynomial needs no inverse. Although $q_T(0) = m$, its input $\mathbf{Z}^\top\mathbf{y}_2$ has no component in $\ker\widehat{\mathbf{\Sigma}}$.

To pass from coefficients to fitted responses, use $\widehat{\mathbf{\Sigma}} = \mathbf{D}^\top\mathbf{D}$ and $\mathbf{G} = \mathbf{D}\mathbf{D}^\top$ in the polynomial q_T:

$$\mathbf{D}q_T(\widehat{\mathbf{\Sigma}})\mathbf{D}^\top \overset{③}{=} \mathbf{G}q_T(\mathbf{G}) \overset{④}{=} s_T(\mathbf{G}). \tag{127}$$

Equality ③ applies $\mathbf{D}\widehat{\mathbf{\Sigma}}^k\mathbf{D}^\top = \mathbf{G}^{k+1}$ termwise, and ④ uses $s_T(u) = uq_T(u)$. Multiplication of (123) by $\mathbf{Z}$ proves (124) on the whole observation space, without a matrix inverse.

Step 3. Scalar bounds. Scalar bounds on each eigenvalue separate optimization bias from fitted noise and identify the effective training scale m^{-1}. At $u = 0$, we have $q_T(0) = m$, $s_T(0) = 0$, and the last two left-hand sides of (125) are zero. For a positive eigenvalue, (120) gives $0 < \eta u \le 1/2$. Hence $s_T(u) = 1-(1-\eta u)^T$ lies in $[0,1]$. Every summand in (119) lies in $[0,1]$, so $s_T(u) \le \eta T u = mu$.

To compare this filter with the scale in (126), take $u > 0$ and bound its unfitted fraction:

$$(1-\eta u)^T \overset{⑤}{\le} \exp(-mu) \overset{⑥}{\le} \frac{1}{1+mu}, \qquad s_T(u) \overset{⑦}{\ge} \frac{mu}{1+mu} = \frac{u}{u+m^{-1}}.$$

Inequality ⑤ uses $1-\eta u \le \exp(-\eta u)$ and $\eta T = m$. Inequality ⑥ uses $\exp(mu) \ge 1+mu$, and ⑦ subtracts the resulting bound from one. If $mu \le 1$, then $mu \le 2mu/(1+mu)$. If $mu \ge 1$, then $1 \le 2mu/(1+mu)$. These two cases give the last inequality in (126). All its terms vanish at $u = 0$.

The upper bounds for s_T also give $uq_T(u)^2 = s_T(u)^2/u \overset{⑧}{\le} \min\{m^2u, 1/u\} \overset{⑨}{\le} m$ when $u > 0$. In ⑧, we use $s_T(u) \le \min\{mu, 1\}$. Inequality ⑨ treats $mu \le 1$ and $mu \ge 1$ separately. Also $(1-\eta u)^{2T} \le \exp(-2\eta T u) = \exp(-2mu)$. The function $u \mapsto u\exp(-2mu)$ on $[0,\infty)$ has maximum $1/(2em)$. This proves the last inequality in (125).

The common positive spectrum of $\mathbf{G}$ and $\widehat{\mathbf{\Sigma}}$, together with $s_T(0) = 0$, gives $0 \preceq s_T(\mathbf{G}) \preceq \mathbf{I}_{n_2}$. This proves (124), including norm contraction and annihilation of the Gram-matrix null space. □

Coefficient norms and zero functions. The finite-feature RKHS norm identifies the least coefficient cost of a prediction function. We also record the distinction between coefficients representing zero on the sphere and coefficients giving zero sample predictions.

Verification of the coefficient-norm identities. **Outline.** Step 1 identifies the RKHS norm with the minimum coefficient norm. Step 2 relates the population and empirical null spaces.

Step 1. The function norm. The RKHS of $\widehat{K}(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(\mathbf{x}')$ consists of the functions $h = \boldsymbol{a}^\top \boldsymbol{\psi}$, with

$$\|h\|^2_{\mathcal{H}_{\widehat{K}}} \overset{①}{=} \min \left\{ \|\boldsymbol{a}\|_2^2 : \boldsymbol{a}^\top \boldsymbol{\psi} = h \text{ on } \mathcal{X} \right\}.$$

For ①, restrict coefficients to the orthogonal complement of those representing the zero function, and transport the Euclidean inner product to their images. Each $\boldsymbol{\psi}(\mathbf{x})$ lies in this complement and represents evaluation at $\mathbf{x}$, giving the kernel $\widehat{K}$. Orthogonal projection attains the minimum.

Step 2. The null spaces. Recall $\mathbf{\Sigma} = \int \boldsymbol{\psi}\boldsymbol{\psi}^\top \mathrm{d}\mu$, the matrix $\mathbf{Z}$ of feature rows for the second training sample, and $\widehat{\mathbf{\Sigma}} = \mathbf{Z}^\top \mathbf{Z}/n_2$. The identity $\mathbf{c}^\top \mathbf{\Sigma} \mathbf{c} = \left\| \mathbf{c}^\top \boldsymbol{\psi} \right\|^2_{L^2}$, continuity of the features, and full support of μ give

$$\left\{ \mathbf{c} \in \mathbb{R}^m : \mathbf{c}^\top \boldsymbol{\psi} = 0 \text{ on } \mathcal{X} \right\} \overset{②}{=} \ker \mathbf{\Sigma} \overset{③}{\subseteq} \ker \mathbf{Z} \overset{④}{=} \ker \widehat{\mathbf{\Sigma}}.$$

Equality ② uses continuity and full support, ③ evaluates the zero function on the sample, and ④ uses $\widehat{\mathbf{\Sigma}} = \mathbf{Z}^\top \mathbf{Z}/n_2$. In particular, coefficients in $(\ker \mathbf{\Sigma})^\perp$ have minimum norm among all representations of their prediction function. □

The scale selected by output GD. Equation (126) places the transition of $s_T(u)$ at $u = m^{-1}$, within a factor of two of $u/(u + m^{-1})$. This identifies the GD scale without identifying GD with a penalized minimizer. Proposition B.1 retains the same scale.

For the noise calculation, assume the observation model (22)–(23) and an almost surely finite $\mathscr{F}_1$-measurable attention vector. Write $\mathbf{w}_2 = [w_i]_{i \in I_2}$ in increasing index order and recall $v_w = \mathbb{E}\left[w_1^2\right]$. Conditional on $\mathscr{F}_2$, this vector has mean zero and covariance $v_w \mathbf{I}_{n_2}$. Its fitted response vector is $s_T(\mathbf{G})\mathbf{w}_2$, and

$$\mathbb{E}\left[\frac{1}{n_2} \|s_T(\mathbf{G})\mathbf{w}_2\|_2^2 \mid \mathscr{F}_2 \right] \overset{①}{=} \frac{v_w}{n_2} \mathrm{tr}\left(s_T(\mathbf{G})^2 \right) \overset{②}{\leq} \frac{2v_w}{n_2} \widehat{\mathcal{N}}(m^{-1}). \tag{128}$$

Equality ① in (128) uses the conditional noise covariance. Inequality ② uses $s_T(u)^2 \leq s_T(u) \leq 2u/(u + m^{-1})$ and the common positive spectra of $\mathbf{G}$ and $\widehat{\mathbf{\Sigma}}$. This bounds fitted noise on the observed inputs. Lemma F.9 uses (116) to pass to fresh inputs, while Lemma F.8 controls outcomes where that comparison fails.

Lemma F.8. Let f^* be a deterministic measurable target in $L^2(P)$ and put $B_* = \|f^*\|_{L^2}$. Use the output updates (41)–(42) from zero. For every finite attention vector and every second training sample with finite responses,

$$\|\boldsymbol{a}(T)\|_2^2 \leq m \|\mathbf{y}_2/\sqrt{n_2}\|_2^2, \qquad \mathcal{R}(\widehat{f}) \leq \frac{5}{2} m \frac{1}{n_2} \sum_{i \in I_2} y_i^2 + 2B_*^2. \tag{129}$$

For the following expectation bounds, assume the observation model (22)–(23), directions independent of the observations, and an almost surely finite $\mathscr{F}_1$-measurable attention rule. The conditional mean $\bar{f} := \mathbb{E}\left[\widehat{f} \mid \mathscr{F}_2 \right]$ exists in $L^2(P)$ and satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{f}) \mid \mathscr{F}_2 \right] = \left\| \bar{f} - f^* \right\|^2_{L^2} + \frac{v_w}{n_2} \mathrm{tr}\left(\mathbf{\Sigma} q_T(\widehat{\mathbf{\Sigma}})^2 \widehat{\mathbf{\Sigma}} \right). \tag{130}$$

Here q_T is the GD polynomial in (119) and $v_w = \mathbb{E}\left[w_1^2\right]$. The conditional variance obeys, almost surely,

$$\mathbb{E}\left[\left\| \widehat{f} - \bar{f} \right\|^2_{L^2} \mid \mathscr{F}_2 \right] = \frac{v_w}{n_2} \mathrm{tr}\left(\mathbf{\Sigma} q_T(\widehat{\mathbf{\Sigma}})^2 \widehat{\mathbf{\Sigma}} \right) \leq \frac{5 m v_w}{4 n_2}. \tag{131}$$

If $f^* \in L^4(P)$, let $A \in \mathscr{F}_2$ and put $p = \Pr[A]$. The variance of $f^*(\mathbf{x})^2$ for $\mathbf{x} \sim P$ is $\|f^*\|_{L^4}^4 - B_*^4$, and

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{A\}}\right] \le 3mB_*^2 p + \frac{5m}{2}\sqrt{\frac{(\|f^*\|_{L^4}^4 - B_*^4)p(1-p)}{n_2}} + \frac{5mv_w}{4n_2}p. \tag{132}$$

Under the polynomial-target and width assumptions of the main theorem,

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{\mathcal{E}^c\}}\right] \le C_L B_*^2 n^{-7/2} + 5v_w n^{-8} \le C_L(B_*^2 + \sigma_0^2)n^{-7/2}. \tag{133}$$

The expectation bounds do not require a fourth noise moment. The restriction $A \in \mathscr{F}_2$ in (132) allows dependence on responses in the first training sample and inputs in the second training sample, but excludes dependence on the noises in the second training sample. The mean $\bar{f}$ holds the possibly noisy first-stage outcome fixed.

Proof. **Outline.** Step 1 obtains deterministic coefficient and risk bounds and verifies square integrability of the predictor. Step 2 separates the conditional mean from the fitted output noise and bounds that noise on all covariance outcomes. Step 3 centers the empirical target square and applies Cauchy–Schwarz to its correlation with the event indicator, retaining the averaging factor from independent inputs. Step 4 applies polynomial hypercontractivity and the covariance failure probability.

Step 1. A bound on every outcome. The GD linear map controls exceptional outcomes and ensures integrability. Recall $\mathbf{D} = \mathbf{Z}/\sqrt{n_2}$ and $\mathbf{y}_2 = [y_i]_{i\in I_2}$. Its squared operator norm satisfies $\left\|q_T(\widehat{\mathbf{\Sigma}})\mathbf{D}^\top\right\|_2^2 = \left\|q_T(\widehat{\mathbf{\Sigma}})^2\widehat{\mathbf{\Sigma}}\right\|_2 \le m$ by (125), also for singular $\widehat{\mathbf{\Sigma}}$. Hence (123) gives $\|\boldsymbol{a}(T)\|_2^2 \overset{①}{\le} m\|\mathbf{y}_2/\sqrt{n_2}\|_2^2$, where ① applies this operator bound. Also $\left\|\widehat{f}\right\|_{L^2}^2 \overset{②}{\le} (5m/4)n_2^{-1}\sum_{i\in I_2} y_i^2$, where ② uses $\|\mathbf{\Sigma}\|_2 \le \operatorname{tr}(\mathbf{\Sigma}) \le 5/4$ from Proposition F.4. The inequality $\left\|\widehat{f} - f^*\right\|_{L^2}^2 \le 2\left\|\widehat{f}\right\|_{L^2}^2 + 2B_*^2$ proves (129). Taking expectations gives $\mathbb{E}\left[\left\|\widehat{f}\right\|_{L^2}^2\right] \le (5m/4)(B_*^2 + v_w) < \infty$.

Joint continuity and (108) make the features continuous in $L^2(P)$ as functions of directions and attention. Formula (123) therefore makes $\widehat{f}$ measurable in this separable Hilbert space. The preceding second-moment bound ensures existence of its conditional mean $\bar{f}$.

Step 2. Exact conditional bias and variance. Conditional on $\mathscr{F}_2$, the vector $\mathbf{w}_2 = [w_i]_{i\in I_2}$ has mean zero and covariance $v_w\mathbf{I}_{n_2}$. Equation (123) therefore identifies $\bar{f}$ with the predictor fitted to $f^*(\vec{\mathbf{x}}_i)$ using the same learned representation. The coefficient vector of $\widehat{f} - \bar{f}$ is $q_T(\widehat{\mathbf{\Sigma}})\mathbf{Z}^\top\mathbf{w}_2/n_2$, whose covariance follows from

$$\mathbb{E}\left[(\mathbf{Z}^\top\mathbf{w}_2/n_2)(\mathbf{Z}^\top\mathbf{w}_2/n_2)^\top \mid \mathscr{F}_2\right] \overset{③}{=} \frac{v_w}{n_2^2}\mathbf{Z}^\top\mathbf{Z} = \frac{v_w}{n_2}\widehat{\mathbf{\Sigma}}.$$

Equality ③ uses the conditional covariance $v_w\mathbf{I}_{n_2}$. Since $\left\|\boldsymbol{a}^\top\boldsymbol{\psi}\right\|_{L^2}^2 = \boldsymbol{a}^\top\mathbf{\Sigma}\boldsymbol{a}$, multiplying this covariance on both sides by $q_T(\widehat{\mathbf{\Sigma}})$ and taking the trace against $\mathbf{\Sigma}$ gives (131). Its upper bound follows from $q_T(\widehat{\mathbf{\Sigma}})^2\widehat{\mathbf{\Sigma}} \preceq m\mathbf{I}_m$ and $\operatorname{tr}(\mathbf{\Sigma}) \le 5/4$, without a covariance event or common eigenbasis.

Finally, $\widehat{f} - \bar{f}$ is conditionally centered with finite second moment in $L^2(P)$. Its conditional expected inner product with $\bar{f} - f^*$ is zero. Expanding the squared error proves (130).

Step 3. Centering the empirical target square. Centering retains the factor $n_2^{-1/2}$ even when the exceptional event depends on the inputs. Fix $A \in \mathscr{F}_2$, put $p = \Pr[A]$, and let $S = n_2^{-1}\sum_{i\in I_2} f^*(\vec{\mathbf{x}}_i)^2$. Applying (129) to the noiseless responses gives $\left\|\bar{f} - f^*\right\|_{L^2}^2 \le (5m/2)S + 2B_*^2$. The moments needed to average this bound are

$$\mathbb{E}[S] = B_*^2, \qquad \mathbb{E}\left[(S - B_*^2)^2\right] \overset{④}{=} \frac{\|f^*\|_{L^4}^4 - B_*^4}{n_2}, \qquad \mathbb{E}\left[(\mathbb{1}_{\{A\}} - p)^2\right] = p(1-p).$$

Equality ④ is the variance formula for the average of the independent target squares, each with variance $\|f^*\|_{L^4}^4 - B_*^4$. Centering both S and the event indicator gives

$$\mathbb{E}\left[S\mathbb{1}_{\{A\}}\right] \overset{⑤}{=} B_*^2 p + \mathbb{E}\left[(S - B_*^2)(\mathbb{1}_{\{A\}} - p)\right] \overset{⑥}{\le} B_*^2 p + \sqrt{\frac{(\|f^*\|_{L^4}^4 - B_*^4)p(1-p)}{n_2}}. \tag{134}$$

Inequality ⑥ is Cauchy–Schwarz with the two variances just computed. In ⑤, the subtraction of p inside the second expectation is valid because $\mathbb{E}\left[S - B_*^2\right] = 0$. Multiplying the deterministic bias bound by $\mathbb{1}_{\{A\}}$, averaging, and using (134) now give

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_{L^2}^2 \mathbb{1}_{\{A\}}\right] \overset{⑦}{\le} 3mB_*^2 p + \frac{5m}{2}\sqrt{\frac{(\|f^*\|_{L^4}^4 - B_*^4)p(1-p)}{n_2}}.$$

Inequality ⑦ combines the deterministic bias bound with ⑥. Here $(5m/2 + 2)B_*^2 p \le 3mB_*^2 p$ because $m \ge 4$. Since $A \in \mathscr{F}_2$, multiplying (130) by its indicator and using (131) adds at most $5mv_w p/(4n_2)$. This proves (132), including $p = 0, 1$. If the target square is constant almost everywhere, its variance is zero and the square-root term vanishes. The proof uses a fourth moment only for the target. The noise still requires only its second moment.

Step 4. Polynomial targets and covariance failure. The polynomial fourth-moment bound and the small covariance-failure probability now make the exceptional contribution negligible. The event $\mathcal{E}^c$ belongs to $\mathscr{F}_2$ by Lemma F.6. Its probability is at most $2n^{-8}$ under the width conditions. Lemma E.6 gives $\|f^*\|_{L^4}^2 \le 3^L B_*^2$. Substitution in (132) gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\mathbb{1}_{\{\mathcal{E}^c\}}\right] \overset{⑧}{\le} 6mB_*^2 n^{-8} + 5 \cdot 3^L mB_*^2 n^{-9/2} + \frac{5mv_w}{2n_2} n^{-8}.$$

Inequality ⑧ substitutes the failure probability and target fourth moment into (132). To obtain the square-root term, we used $1 - \Pr[\mathcal{E}^c] \le 1$ and $n_2 \ge n/2$. Since $m \le n$, $n^{-7} \le n^{-7/2}$, and $v_w \le \sigma_0^2$, the first two terms sum to at most $(6+5\cdot 3^L)B_*^2 n^{-7/2}$ and the last is at most $5v_w n^{-8}$. Combining the target and noise estimates proves both inequalities in (133) after choosing C_L large enough. Every bound is uniform over the almost surely finite first-stage rules. The argument uses their measurability and the uniform feature bounds, with no moment bound on their attention parameters. □

Lemma F.9. Assume the observation model (22)–(23) with a fixed measurable target $f^* \in L^2(P)$. Let the frozen attention be an almost surely finite $\mathscr{F}_1$-measurable vector. Set $B_* = \|f^*\|_{L^2}$. Let $\boldsymbol{a}^\dagger \in \mathbb{R}^m$ be an $\mathscr{F}_1$-measurable comparison vector with $\mathbb{E}\left[\left\|\boldsymbol{a}^\dagger\right\|_2^2\right] < \infty$, and set $r = f^* - (\boldsymbol{a}^\dagger)^\top \boldsymbol{\psi}$. The vector may depend on the directions, the entire first training sample, and the fixed target, but not on the second training sample. The bound $\mathrm{tr}(\boldsymbol{\Sigma}) \le 5/4$ implies $\mathbb{E}\left[\|r\|_{L^2}^2\right] \le 2B_*^2 + (5/2)\mathbb{E}\left[\left\|\boldsymbol{a}^\dagger\right\|_2^2\right] < \infty$. The comparison vector is used only for analysis. For the training parameters (41), on the event $\mathcal{E}$ from (115), the conditional risk satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{f}) \mid \mathscr{F}_2\right] \le C\left(\|r\|_{L^2}^2 + \frac{\left\|\boldsymbol{a}^\dagger\right\|_2^2}{m} + \|r\|_{L^2(P_{n_2})}^2 + \frac{\sigma_0^2}{n_2}\widehat{\mathcal{N}}(m^{-1})\right). \tag{135}$$

The conditioning fixes the directions, observations in the first training sample, and inputs in the second training sample. The expectation averages only the remaining noises in the second training sample. A numerical constant $C = 9$ is sufficient.

Proof. **Outline.** Step 1 decomposes the coefficient error into optimization, fitted-residual, and fitted-noise terms. Step 2 bounds the optimization term, including coefficient directions absent from the second training sample. Step 3 controls the fitted residual by its empirical squared norm. Step 4 expands the conditional noise contribution in empirical eigenvectors and bounds it by the effective dimension. Step 5 combines the three bias terms and the variance through the preceding conditional bias–variance identity.

Step 1. Exact decomposition. Separating optimization, target residual, and noise allows each source of error to use the assumption that controls it. We condition on $\mathscr{F}_2$ and work on $\mathcal{E}$. We recall $\widehat{\boldsymbol{\Sigma}} = \mathbf{Z}^\top\mathbf{Z}/n_2$, $\mathbf{D} = \mathbf{Z}/\sqrt{n_2}$, $\boldsymbol{\Sigma} = \int \boldsymbol{\psi}\boldsymbol{\psi}^\top \mathrm{d}\mu$, and $r = f^* - (\boldsymbol{a}^\dagger)^\top\boldsymbol{\psi}$. We let $\mathbf{e}_2 = [r(\vec{\mathbf{x}}_i)]_{i\in I_2}$ and $\mathbf{w}_2 = [w_i]_{i\in I_2}$, with entries in increasing index order. The comparison vector and residual are

fixed under this conditioning. Their defining identity gives $\mathbf{y}_2 = \mathbf{Z}\boldsymbol{a}^\dagger + \mathbf{e}_2 + \mathbf{w}_2$. Formula (123) therefore yields

$$\boldsymbol{a}(T) - \boldsymbol{a}^\dagger \overset{①}{=} -(\mathbf{I}_m - \eta\widehat{\boldsymbol{\Sigma}})^T\boldsymbol{a}^\dagger + q_T(\widehat{\boldsymbol{\Sigma}})\mathbf{Z}^\top\mathbf{e}_2/n_2 + q_T(\widehat{\boldsymbol{\Sigma}})\mathbf{Z}^\top\mathbf{w}_2/n_2. \tag{136}$$

Equality ① substitutes the response decomposition and uses $q_T(\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}} = \mathbf{I}_m - (\mathbf{I}_m - \eta\widehat{\boldsymbol{\Sigma}})^T$. For every coefficient vector $\mathbf{c}$, the population squared norm of $\mathbf{c}^\top\boldsymbol{\psi}$ is $\mathbf{c}^\top\boldsymbol{\Sigma}\mathbf{c}$.

Step 2. Optimization. The covariance comparison converts an empirical spectral contraction into a population prediction bound, including empirical null directions. We bound the first term of (136). We set $\mathbf{v} = (\mathbf{I}_m - \eta\widehat{\boldsymbol{\Sigma}})^T\boldsymbol{a}^\dagger$. The comparison (116) gives $\mathbf{v}^\top\boldsymbol{\Sigma}\mathbf{v} \overset{②}{\le} 2\mathbf{v}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{v} + m^{-1}\|\mathbf{v}\|_2^2$, where ② evaluates the covariance order on $\mathbf{v}$. For each eigenvalue $u \ge 0$ of $\widehat{\boldsymbol{\Sigma}}$, stability gives $(1-\eta u)^{2T} \le 1$, and (125) gives $2u(1-\eta u)^{2T} \le 1/(em)$. As a result, $(2u+m^{-1})(1-\eta u)^{2T} \le (1+\exp(-1))/m$, including $u = 0$. Diagonalizing $\widehat{\boldsymbol{\Sigma}}$ in the preceding quadratic-form bound yields

$$\left\|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_m - \eta\widehat{\boldsymbol{\Sigma}})^T\boldsymbol{a}^\dagger\right\|_2^2 \overset{③}{\le} \frac{1+\exp(-1)}{m}\left\|\boldsymbol{a}^\dagger\right\|_2^2. \tag{137}$$

Inequality ③ sums the scalar bound over the empirical eigenvectors after ②. Only polynomials in $\widehat{\boldsymbol{\Sigma}}$ are diagonalized together. On $\ker\widehat{\boldsymbol{\Sigma}}$, the GD power is the identity, but (117) controls its population cost by $m^{-1}\left\|\boldsymbol{a}^\dagger\right\|_2^2$. The full quadratic-form comparison also controls cross terms, without commutation of $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$ or population orthogonality between empirical eigenspaces.

Step 3. Approximation residual. The residual need not be orthogonal to the features. We therefore control what GD fits from it by its empirical energy. We put $\mathbf{e} = \mathbf{e}_2/\sqrt{n_2}$ and $\mathbf{c} = q_T(\widehat{\boldsymbol{\Sigma}})\mathbf{D}^\top\mathbf{e}$. We recall that $\mathbf{G} = \mathbf{D}\mathbf{D}^\top$ is the normalized sample Gram matrix. The operator-norm argument in Step 1 of Lemma F.8 gives $\|\mathbf{c}\|_2^2 \le m\|\mathbf{e}\|_2^2$. Equation (127) also gives $\mathbf{D}\mathbf{c} = s_T(\mathbf{G})\mathbf{e}$. The norm bound in (124) therefore yields $\mathbf{c}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{c} = \|\mathbf{D}\mathbf{c}\|_2^2 \le \|\mathbf{e}\|_2^2$. Both conclusions hold when the empirical covariance is singular. Substitution into (116) proves

$$\mathbf{c}^\top\boldsymbol{\Sigma}\mathbf{c} \overset{④}{\le} 3\|\mathbf{e}\|_2^2 = 3\|r\|_{L^2(P_{n_2})}^2. \tag{138}$$

Inequality ④ combines (116) with the two preceding bounds. No residual orthogonality is needed. Indeed, for the minimizer of (48) at $\lambda = m^{-1}$, the normal equation gives $\int \boldsymbol{\psi} r\,\mathrm{d}\mu = m^{-1}\boldsymbol{a}^\dagger$, so the fitted-residual term in (136) generally remains nonzero.

Step 4. Conditional noise. We refine the all-outcome variance (131) using effective dimension. Recall $\bar{f} = \mathbb{E}\left[\widehat{f} \mid \mathscr{F}_2\right]$ from Lemma F.8. For a fixed conditioning outcome, choose orthonormal eigenvectors $\mathbf{v}_j$ of $\widehat{\boldsymbol{\Sigma}}$ with eigenvalues $u_j \ge 0$, $j \in [m]$. Expanding the variance trace gives

$$\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^{1/2}q_T(\widehat{\boldsymbol{\Sigma}})\mathbf{Z}^\top\mathbf{w}_2/n_2\right\|_2^2 \mid \mathscr{F}_2\right] \overset{⑤}{=} \frac{v_w}{n_2}\sum_{j=1}^m u_j q_T(u_j)^2\mathbf{v}_j^\top\boldsymbol{\Sigma}\mathbf{v}_j. \tag{139}$$

Equality ⑤ expands the trace in empirical eigenvectors. Here $\mathbf{v}_j^\top\boldsymbol{\Sigma}\mathbf{v}_j = \int(\mathbf{v}_j^\top\boldsymbol{\psi})^2\mathrm{d}\mu$ is their population second moment. No common eigenbasis is assumed, and a zero empirical eigenvalue contributes zero because $q_T(0) = m$.

The sum in (139) equals $\mathrm{tr}\left(\boldsymbol{\Sigma}q_T(\widehat{\boldsymbol{\Sigma}})^2\widehat{\boldsymbol{\Sigma}}\right)$ for any chosen eigenbasis, so no measurable choice of eigenvectors is needed. The covariance calculation and this exact identity hold almost surely without restricting to $\mathcal{E}$.

Using $v_w \le \sigma_0^2$ in (131) gives $\mathbb{E}\left[\left\|\widehat{f} - \bar{f}\right\|_{L^2}^2 \mid \mathscr{F}_2\right] \overset{⑥}{\le} (\sigma_0^2/n_2)\mathrm{tr}\left(\boldsymbol{\Sigma}q_T(\widehat{\boldsymbol{\Sigma}})^2\widehat{\boldsymbol{\Sigma}}\right)$. On $\mathcal{E}$, the covariance comparison bounds this trace by

$$\mathrm{tr}\left(\boldsymbol{\Sigma}q_T(\widehat{\boldsymbol{\Sigma}})^2\widehat{\boldsymbol{\Sigma}}\right) \overset{⑦}{\le} 2\mathrm{tr}\left(s_T(\widehat{\boldsymbol{\Sigma}})^2\right) + m^{-1}\mathrm{tr}\left(q_T(\widehat{\boldsymbol{\Sigma}})^2\widehat{\boldsymbol{\Sigma}}\right) \overset{⑧}{\le} \min\left\{3m, 6\widehat{\mathcal{N}}(m^{-1})\right\}. \tag{140}$$

For ⑦, we apply (116) inside the trace against the positive semidefinite matrix $q_T(\widehat{\boldsymbol{\Sigma}})^2\widehat{\boldsymbol{\Sigma}}$. In general, $\mathbf{A} \preceq \mathbf{B}$ and $\mathbf{C} \succeq 0$ imply $\mathrm{tr}(\mathbf{C}\mathbf{A}) \le \mathrm{tr}(\mathbf{C}\mathbf{B})$ by congruence with $\mathbf{C}^{1/2}$. This step does not assume that $\boldsymbol{\Sigma}$ commutes with $\widehat{\boldsymbol{\Sigma}}$.

To verify ⑧, each eigenvalue contributes at most three to the middle expression in (140), by (125). To obtain the effective-dimension bound, we fix $u > 0$ in the spectrum and put $x = mu$. Then

$$2s_T(u)^2 + m^{-1}uq_T(u)^2 \overset{⑨}{\leq} (2 + x^{-1}) \min\{1, x^2\} \overset{①}{\leq} \frac{6x}{1+x}. \tag{141}$$

Inequality ⑨ uses $s_T(u) \leq \min\{1, mu\}$ and $uq_T(u) = s_T(u)$. For $0 < x \leq 1$, inequality ① is equivalent to $(2x+1)(1+x) \leq 6$. For $x \geq 1$, it follows from $(2 + x^{-1})(1 + x^{-1}) \leq 6$. The zero eigenvalues contribute zero, so summing (141) proves (140).

Step 5. Combining the errors. We combine the three bias terms separately from the noise variance to retain the sharper noise constant. With $\bar{f}$ as in Lemma F.8, the conditional bias on $\mathcal{E}$ is $\bar{f} - f^* = -\mathbf{v}^\top \boldsymbol{\psi} + \mathbf{c}^\top \boldsymbol{\psi} - r$, where $\mathbf{v}$ and $\mathbf{c}$ are from Steps 2 and 3. Applying (130) together with (137), (138), and (140) gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f}) \mid \mathscr{F}_2\right] \overset{②}{\leq} 3\|r\|_{L^2}^2 + \frac{3(1+\exp(-1))}{m}\left\|\boldsymbol{a}^\dagger\right\|_2^2 + 9\|r\|_{L^2(P_{n_2})}^2 + \frac{6\sigma_0^2}{n_2}\widehat{\mathcal{N}}(m^{-1}). \tag{142}$$

Inequality ② uses the exact conditional bias–variance identity and bounds the squared sum of the three bias terms by three times their separate squared norms. Since $3(1 + \exp(-1)) < 9$, inequality (142) proves (135) with $C = 9$. The averaging argument uses the separate coefficients in (142) to obtain the numerical constants in (50). □

F.6 Preserved Polynomial Approximation

The next proposition supplies the measurable comparison coefficients and the approximation bound used in the expected-risk proof. Recall that $\mathcal{A}_\lambda$ in (48) combines regression risk with a coefficient penalty of size $\lambda > 0$.

Proposition F.10. Fix the sampled directions, a finite attention vector $\widehat{\boldsymbol{\tau}}$, and a target $f^* \in L^2(P)$. For every $\lambda > 0$,

$$\mathcal{A}_\lambda(\widehat{K}) \leq \mathcal{A}_\lambda(K_0). \tag{143}$$

Under the target and dimension assumptions of Theorem C.1, every measurable finite first-stage attention rule also satisfies

$$\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] \leq \mathbb{E}_{\mathbf{Q}}\left[\mathcal{A}_{m^{-1}}(K_0)\right] \leq \frac{C_L\Gamma_*^2}{m} \leq \frac{C_L B_*^2 r_0}{m}. \tag{144}$$

The first inequality is pointwise before expectation. As a result, its constant is independent of the attention rule, although the value of $\mathcal{A}_{m^{-1}}(\widehat{K})$ can depend on the responses in the first training sample.

Proof. **Outline.** Step 1 constructs the measurable population minimizers. Step 2 uses orthogonal feature coordinates to prove preservation. Step 3 inserts the ReLU comparison predictor to obtain the expected approximation bound.

Step 1. Measurable population minimizers. The coefficient penalty makes the minimizer unique and measurable, as required when averaging the conditional risk bound. We recall that $\boldsymbol{\psi} : \mathcal{X} \to \mathbb{R}^m$ is the learned feature map, $\boldsymbol{\Sigma} = \int \boldsymbol{\psi}\boldsymbol{\psi}^\top \mathrm{d}\mu$ is its population covariance, and $\lambda > 0$. The integrals are finite by (108) and $f^* \in L^2(P)$. Expanding the objective in (48) gives a quadratic function of $\boldsymbol{a}$ with Hessian $2(\boldsymbol{\Sigma} + \lambda \mathbf{I}_m)$, which is positive definite. Its unique minimizer is

$$\boldsymbol{a}_\lambda := (\boldsymbol{\Sigma} + \lambda \mathbf{I}_m)^{-1} \int_{\mathcal{X}} \boldsymbol{\psi}(\mathbf{x}) f^*(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}), \qquad \left\|\boldsymbol{a}_\lambda^\top \boldsymbol{\psi} - f^*\right\|_{L^2}^2 + \lambda\|\boldsymbol{a}_\lambda\|_2^2 = \mathcal{A}_\lambda(\widehat{K}) \overset{①}{\leq} B_*^2. \tag{145}$$

Inequality ① compares with the zero coefficient vector, recalling $B_* = \|f^*\|_{L^2}$. The same formula gives $\boldsymbol{a}_{\lambda,0}$ using $\boldsymbol{\psi}_0$ and its covariance, including when that covariance is singular.

Joint continuity and (108) give integrable envelopes $2(d+1)$ and $\sqrt{2(d+1)}|f^*|$ for the covariance and target-feature integrands. Here $\int |f^*| \mathrm{d}\mu \leq B_*$ by Cauchy–Schwarz. Dominated convergence makes both integrals continuous in the directions and finite attention vector. Inversion of $\boldsymbol{\Sigma} + \lambda \mathbf{I}_m$ is continuous, so (145) defines a measurable minimizer.

Equation (145) bounds $\|\boldsymbol{a}_\lambda\|_2^2$ by B_*^2/λ and the residual norm by B_*. These bounds supply all moments used below without a moment assumption on the attention parameters. The coefficient-

norm identity in Appendix F.5 also shows that $\mathcal{A}_\lambda(\widehat{K})$ depends only on the kernel: minimizing the squared coefficient norm over all representations of a function gives its squared RKHS norm.

Step 2. Learned features and nonincrease. Use the norm-preserving coordinates (78), with retained block $\boldsymbol{\psi}_+ = [\phi_j]_{j=1}^{m/2}/\sqrt{m}$ and added block $\boldsymbol{\psi}_- = 2a_g\boldsymbol{\psi}_+$. For coefficient blocks $\mathbf{b}, \mathbf{c} \in \mathbb{R}^{m/2}$, the learned objective is

$$J_\lambda(\mathbf{b}, \mathbf{c}) := \left\| \mathbf{b}^\top \boldsymbol{\psi}_+ + \mathbf{c}^\top \boldsymbol{\psi}_- - f^* \right\|_{L^2}^2 + \lambda \left(\|\mathbf{b}\|_2^2 + \|\mathbf{c}\|_2^2 \right).$$

The initial minimum is $\min_{\mathbf{b}} J_\lambda(\mathbf{b}, \mathbf{0})$. Therefore $\min_{\mathbf{b},\mathbf{c}} J_\lambda(\mathbf{b}, \mathbf{c}) \overset{②}{\leq} \min_{\mathbf{b}} J_\lambda(\mathbf{b}, \mathbf{0})$ by inclusion of feasible coefficients, proving (143) at the same penalty. These coordinates are used only in the proof.

Step 3. The expected approximation bound. The preserved ReLU comparison gives a quantitative bound at the finite-GD scale. We now use the polynomial target assumptions and set $\lambda = m^{-1}$. Lemma F.2 supplies f° and $\boldsymbol{a}^\circ \in \mathbb{R}^{m/2}$ with the bounds (102). Proposition F.4, applied also at $\widehat{\boldsymbol{\tau}} = \mathbf{0}$, represents f° using $\widetilde{\boldsymbol{a}}^\circ = [(\boldsymbol{a}^\circ)^\top, (\boldsymbol{a}^\circ)^\top]^\top$ and gives $\|\widetilde{\boldsymbol{a}}^\circ\|_2^2 = 2\|\boldsymbol{a}^\circ\|_2^2$. Evaluation at these coefficients and then expectation over $\mathbf{Q}$ yield

$$\mathbb{E}_{\mathbf{Q}}\left[\mathcal{A}_{m^{-1}}(K_0)\right] \overset{③}{\leq} \mathbb{E}_{\mathbf{Q}}\left[\|f^\circ - f^*\|_{L^2}^2 + \frac{2\|\boldsymbol{a}^\circ\|_2^2}{m}\right] \overset{④}{\leq} \frac{C_L \Gamma_*^2}{m}.$$

Inequality ③ uses the duplicated comparison vector, and ④ uses (102), absorbing numerical factors into C_L. The initial quantity depends only on directions and target. Averaging the pointwise inequality (143) therefore gives the first two inequalities in (144), including the learned quantity's dependence on the first training sample. The bound $\Gamma_*^2 \leq r_0 B_*^2$ in (31) gives the last. □

F.7 Polynomial Approximation and Proof of Theorem C.1

For the expected-risk conclusions of Theorem C.1, we combine Proposition C.2 with the preserved polynomial approximation. Steps 5–9 of the joint proof below establish its additional high-probability conclusion. Recall B_*, Γ_*, r_0 from (31) and the comparison predictor f° with coefficients $\boldsymbol{a}^\circ$ from Lemma F.2.

For $d \geq d_L$ and fixed $L \geq \ell_0$, that lemma gives, with expectation over the sampled directions alone,

$$\mathbb{E}_{\mathbf{Q}}\left[\|f^\circ - f^*\|_{L^2}^2\right] \leq 2C_L\Gamma_*^2/m, \qquad \mathbb{E}_{\mathbf{Q}}\left[\|\boldsymbol{a}^\circ\|_2^2\right] \leq C_L\Gamma_*^2. \tag{146}$$

The bounds in (146) control both approximation error and coefficient cost. The comparison predictor is used only in the proof.

For the learned feature map, we duplicate these coefficients as $\widetilde{\boldsymbol{a}}^\circ = [(\boldsymbol{a}^\circ)^\top, (\boldsymbol{a}^\circ)^\top]^\top \in \mathbb{R}^m$. Proposition F.4 gives $(\widetilde{\boldsymbol{a}}^\circ)^\top \boldsymbol{\psi} = f^\circ$ at every input and $\|\widetilde{\boldsymbol{a}}^\circ\|_2^2 = 2\|\boldsymbol{a}^\circ\|_2^2$. The same vector represents f° at initialization with exactly that squared norm. Evaluating (48) at this vector and using (146) therefore gives, after enlarging C_L,

$$\mathbb{E}\left[\mathcal{A}_{m^{-1}}(\widehat{K})\right] \overset{①}{\leq} \mathbb{E}_{\mathbf{Q}}\left[\|f^\circ - f^*\|_{L^2}^2 + \frac{2}{m}\|\boldsymbol{a}^\circ\|_2^2\right] \overset{②}{\leq} \frac{C_L\Gamma_*^2}{m} \overset{③}{\leq} \frac{C_L r_0 B_*^2}{m}. \tag{147}$$

Inequality ① evaluates the minimum at the preserved comparison coefficients. Inequality ② applies (146), and inequality ③ uses $N(d, \ell) \leq r_0$ in the weighted coefficient sum. The comparison cost in (147) depends only on the target and directions. Since preservation holds for every gate, averaging this cost over the directions bounds the learned objective even after averaging over the first training sample.

Proof of Theorems C.1 and 4.1. **Outline.** Step 1 fixes constants, invokes the deterministic representation result, and verifies measurability. Step 2 checks the hypotheses of Proposition C.2 and bounds its exceptional term under the theorem's width conditions. Step 3 uses the bound on regression risk plus coefficient penalty and the rank bound to obtain the stated target-class inequalities. Step 4 proves the consistency limits, verifies the logarithmic width example, and identifies the main-text formulation. For the additional high-probability assertion, Step 5 establishes a concentration inequality and the required input events. Step 6 constructs a preserved comparison predictor with concentrated error. Step 7 identifies the learned feature kernel's local complexity. Step 8 bounds the

finite-GD noise and compares empirical and population effective dimensions. Step 9 combines the errors and their failure probabilities.

Step 1. Constants and well-defined quantities. Recall that $\mathscr{F}_1$ records the directions and first training sample, whereas $\mathscr{F}_2$ also records the inputs in the second training sample. We take $C_0 = 512$, which suffices for Lemmas F.5 and F.6, and take d_L from Lemma F.1. The selection between the finitely many candidates in (36) is measurable under the specified tie rule. The learning rate (37) has a finite positive denominator, so the first-stage update is a measurable finite vector. Proposition D.1 proves the nonzero gradient, strict descent, strict enlargement with witness (77), and kernel evolution. Step 6 of its proof gives the population bound (75), including strict positivity and the uniform upper bound c_d.

The covariance measurability follows from the dominated-convergence argument in Step 1 of Proposition F.10. The empirical covariance, event $\mathcal{E} \in \mathscr{F}_2$, and finite GD iterates are measurable in their specified data. Lemma F.8 gives risk integrability.

Step 2. Applying the explicit-tail risk bound. We isolate the exceptional contribution before simplifying the rate. The observation and initialization assumptions, Step 1, and the bound $\|f^*\|_{L^4}^2 \le 3^L B_*^2$ from Lemma E.6 verify every hypothesis of Proposition C.2. Its bound is already averaged over all outcomes.

Recall $\delta_{d,n,m}$ from (55). Under (49) with $C_0 = 512$, Lemmas F.5 and F.6 bound each exponential term in its definition by n^{-8}. Hence $\delta_{d,n,m} \le 2n^{-8}$. Since $m \le n$ and $n_2 \ge n/2$,

$$m\left(\delta_{d,n,m} + \sqrt{\frac{\delta_{d,n,m}}{n_2}}\right) \overset{①}{\le} 2n^{-7} + 2n^{-7/2} \overset{②}{\le} 4n^{-7/2}, \qquad \frac{5m\sigma_0^2}{4n_2}\delta_{d,n,m} \overset{③}{\le} 5\sigma_0^2 n^{-8}.$$

Inequality ① substitutes $\delta_{d,n,m} \le 2n^{-8}$, $m \le n$, and $n_2 \ge n/2$. Inequality ② uses $n \ge 1$, and inequality ③ uses $m/n_2 \le 2$. Substitution in (57), followed by enlargement of C_L, proves (50). The constants are uniform over the admissible first-stage rules because the concentration and conditional risk bounds use feature bounds valid for every finite attention vector.

Step 3. Approximation, effective dimension, and target classes. We remove the remaining dependence on the learned covariance only after obtaining the sharper kernel-dependent bound. Proposition F.10 bounds the approximation term in (50) by $C_L\Gamma_*^2/m$ through (144). This proves (51). The deterministic bound $\widehat{\mathcal{N}}(m^{-1}) \overset{④}{\le} m$ gives (52). Inequality ④ bounds each of the at most m effective-dimension summands by one. Finally, $N(d,\ell) \ge 1$ implies $B_*^2 \le \Gamma_*^2$, so the restriction $\Gamma_* \le \gamma_0$ gives (53). All constants are independent of the target coefficients and the noise law beyond its stated variance bound. The expectation bound holds with the same constants for every target in the class. This uniform expectation bound does not assert a single event on which the inequality holds simultaneously for all targets.

Step 4. Joint consistency. We check the bias, fitted-noise, and exceptional-event terms separately to make the sufficient resource conditions explicit. The width bounds (49) give $m \ge C_0 d \to \infty$, $n \to \infty$, and $m/n \le 1/(C_0 \log(en)) \to 0$. As a result, (53) vanishes for fixed γ_0, σ_0. This proves the weighted-class assertion of part (iii). The elementary inequality $\Gamma_*^2 \le r_0 B_*^2$ in (31) gives

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \overset{⑤}{\le} C_L\left(\frac{r_0 B_*^2}{m} + \frac{\sigma_0^2 m}{n} + \frac{B_*^2 + \sigma_0^2}{n^{7/2}}\right).$$

Inequality ⑤ substitutes $\Gamma_*^2 \le r_0 B_*^2$ into (52). When B_*, σ_0 are bounded, $r_0/m \to 0$, and $m/n \to 0$, every term on the right tends to zero. In particular, $r_0 \asymp d^{\ell_0} \to \infty$ by Lemma E.4, so these assumptions imply $m, n \to \infty$ as needed for the remainder.

For the concrete example, $n \asymp d^{\ell_0}(\log d)^4$ gives $H_n = \log(en) \asymp \log d$. Rounding in (33) gives $n/H_n^2 \le m < n/H_n^2 + 4$, so $m \sim n/H_n^2 \asymp d^{\ell_0}(\log d)^2$. Here $H_n = O(\log d) = o(d)$, so $m/(d+H_n) \asymp d^{\ell_0 - 1}(\log d)^2 \to \infty$ and $mH_n/n \asymp 1/\log d \to 0$. The two finite-sample width conditions hold eventually, as does $d \ge d_L$. Also $r_0/m = O_L((\log d)^{-2})$ and $m/n = O((\log d)^{-2})$. Substitution into the displayed inequality proves (54). The ratios in that equation diverge and are $o(d)$ because every fixed power of $\log d$ is $o(d)$. The architecture fixes m before training, so this width applies at every iteration.

For every $\varepsilon > 0$, $\Pr\left[\mathcal{R}(\widehat{f}) > \varepsilon\right] \overset{⑥}{\leq} \mathbb{E}\left[\mathcal{R}(\widehat{f})\right]/\varepsilon \to 0$, where ⑥ is Markov's inequality. In the logarithmic example, this gives the uniform bound $C/[\varepsilon(\log d)^2]$. Each probability uses the training law for its own target and noise distribution, with no common event or coupling across dimensions required.

To recover Theorem 4.1, use the equivalent model and target definitions in Appendix A and the identical updates in Appendix B. The choice $C_0 = 512$ makes (49) exactly (12). Part (i), together with the coordinate identity in Appendix D.1, gives (9)–(11) and strict enlargement. Equation (52) is (13), and Step 4 gives (14), including uniformity and convergence in probability. This proves the deterministic and expected-risk assertions of both formulations. Steps 5–9 prove the additional high-probability assertion of Theorem 4.1.

Step 5. A concentration inequality and the input events. The expected-risk proof above does not control the comparison error on individual realizations. We obtain that control jointly in population and empirical norms, then bound the noise in the second training sample separately. For the rest of the proof, assume (15), fix $\delta \in (0,1)$, and abbreviate $t = t_\delta = \log(8/\delta) > 1$ and $\lambda = m^{-1}$. The confidence parameter occurs only in the proof.

We first record a dimension-independent concentration inequality and its proof. If $X_1, \dots, X_k$ are independent centered random elements of a real separable Hilbert space and numbers $b, V \geq 0$ satisfy $\|X_j\| \leq b$ almost surely and $\sum_j \mathbb{E}\left[\|X_j\|^2\right] \leq V$, then

$$\Pr\left[\left\|\sum_{j=1}^{k} X_j\right\| > \sqrt{2Vt} + 2bt/3\right] \leq 2e^{-t}. \tag{148}$$

The norm in this paragraph is the Hilbert-space norm. If $b = 0$ or $V = 0$, every summand vanishes almost surely and (148) is immediate. Otherwise, put $F_\theta(x) = \cosh(\theta\|x\|)$ for $\theta > 0$ and a vector x in the Hilbert space. Its second derivative in a direction u satisfies $D^2F_\theta(x)[u,u] \leq \theta^2 F_\theta(x)\|u\|^2$, also at $x = 0$ by continuity. This follows by differentiating the radial function and using $\sinh z \leq z\cosh z$ for $z \geq 0$. The triangle inequality also gives $F_\theta(x + sX_j) \leq e^{\theta s b}F_\theta(x)$ for $s \in [0,1]$. Taylor's formula, centering, and independence therefore give

$$\mathbb{E}\left[F_\theta(x + X_j)\right] \overset{⑦}{\leq} F_\theta(x)\left[1 + \frac{e^{\theta b} - 1 - \theta b}{b^2}\mathbb{E}\left[\|X_j\|^2\right]\right].$$

Inequality ⑦ integrates the second-derivative bound against $1 - s$ and removes the linear term by centering. Iteration, $1 + u \leq e^u$, and the exponential series yield, for $0 < \theta b < 3$,

$$\mathbb{E}\left[F_\theta\left(\sum_j X_j\right)\right] \overset{⑧}{\leq} \exp\left(\frac{\theta^2 V}{2(1 - \theta b/3)}\right).$$

For ⑧, the series estimate uses $j! \geq 2 \cdot 3^{j-2}$ for $j \geq 2$. Since $\cosh z \geq e^z/2$, exponential Markov gives a tail bound $2\exp[-z^2/(2(V + bz/3))]$ at level $z > 0$, by choosing $\theta = z/(V + bz/3)$. Substituting $z = \sqrt{2Vt} + 2bt/3$ proves (148).

Recall the direction and learned-covariance events $\mathcal{E}_Q, \mathcal{E}_S$ from (109) and (114). Their intersection has failure probability at most $2n^{-8}$. We additionally use the event determined by the inputs in the second training sample

$$\mathcal{E}_X := \left\{\frac{d}{n_2}\sum_{i \in I_2} \vec{\mathbf{x}}_i \vec{\mathbf{x}}_i^\top \preceq 2\mathbf{I}_d\right\}, \qquad \Pr\left[\mathcal{E}_X^c\right] \leq \exp(d\log 21 - n_2/16) \leq n^{-8}. \tag{149}$$

The first probability bound in (149) is exactly the fixed-direction exponential-moment and $1/10$-net argument in Steps 1–2 of Lemma F.5, with the n_2 independent inputs replacing the $m/4$ independent directions. For the last bound, $n_2 \geq m/4$ follows from (12), and hence $d\log 21 - n_2/16 \leq d\log 21 - m/64 \leq -8H_n$. These events require no additional sample or width restriction.

Step 6. A preserved comparison predictor with concentrated error. We construct a comparison predictor that simultaneously controls the two residual terms in the finite-GD decomposition. Recall

$\phi_{\mathbf{q}}(\mathbf{x}) = \sigma(\sqrt{d}\,\mathbf{q}^\top\mathbf{x} + 1/2)$ and the integral representation (103). Its coefficient function is

$$v(\mathbf{q}) = \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)} \frac{\beta_{\ell,j}}{\xi_{\ell,d}} Y_{\ell,j}(\mathbf{q}), \qquad S := \|v\|_{L^2}, \qquad S^2 \le C_L\Gamma_*^2.$$

Here $\xi_{\ell,d}$ is the nonzero ReLU coefficient from (97). If $S = 0$, then $f^* = 0$, and the zero comparison vector gives all the bounds below deterministically. We now treat $S > 0$. Put $k = m/4$, $b_\delta = S\sqrt{k/t}$, and $v_\delta(\mathbf{q}) = v(\mathbf{q})\mathbb{1}_{\{|v(\mathbf{q})|\le b_\delta\}}$. Define

$$f^\dagger(\mathbf{x}) := \frac{1}{k}\sum_{j=1}^{k} \frac{v_\delta(\vec{\mathbf{q}}_j)\phi_{\vec{\mathbf{q}}_j}(\mathbf{x}) + v_\delta(-\vec{\mathbf{q}}_j)\phi_{-\vec{\mathbf{q}}_j}(\mathbf{x})}{2}. \tag{150}$$

Only the analytical comparison function (150) uses truncated coefficients. The network, its observations, and all GD updates remain unchanged.

Fix the inputs in the second training sample on the event $\mathcal{E}_X$. Let P_{n_2} be their empirical probability measure. Embed each continuous function h as

$$Jh := \left(h, n_2^{-1/2}[h(\vec{\mathbf{x}}_i)]_{i\in I_2}\right) \in \mathcal{H} := L^2(P)\oplus\mathbb{R}^{n_2}, \qquad \|Jh\|_{\mathcal{H}}^2 = \|h\|_{L^2}^2 + \|h\|_{L^2(P_{n_2})}^2.$$

The entries indexed by I_2 are ordered increasingly. The continuous representatives make every evaluation well defined. The space $\mathcal{H}$ is a real separable Hilbert space. For every unit direction $\mathbf{q}$, Lemma F.1 and $\sigma(u+1/2)^2 \le 2u^2 + 1/2$ give

$$\|J\phi_{\mathbf{q}}\|_{\mathcal{H}}^2 \overset{⑨}{\le} \frac{5}{4} + \frac{2d}{n_2}\sum_{i\in I_2}(\mathbf{q}^\top\vec{\mathbf{x}}_i)^2 + \frac{1}{2} \overset{①}{\le} \frac{23}{4}.$$

Inequality ⑨ combines the population feature moment with the pointwise squared bound, and ① uses (149). Write $c_\phi = \sqrt{23/4}$ and, for a uniform direction $\mathbf{q}$, put $U_\delta(\mathbf{q}) := \frac{1}{2}J\big(v_\delta(\mathbf{q})\phi_{\mathbf{q}} + v_\delta(-\mathbf{q})\phi_{-\mathbf{q}}\big)$. These random elements are strongly measurable and bounded in $\mathcal{H}$, since v_δ is Borel measurable and $\mathbf{q}\mapsto J\phi_{\mathbf{q}}$ is continuous. The directions remain independent and uniform under the present conditioning. The triangle inequality and convexity give

$$\|U_\delta(\mathbf{q})\|_{\mathcal{H}} \le c_\phi b_\delta, \qquad \mathbb{E}\left[\|U_\delta(\mathbf{q})\|_{\mathcal{H}}^2\right] \overset{②}{\le} c_\phi^2 S^2, \qquad \|\mathbb{E}\,[U_\delta] - Jf^*\|_{\mathcal{H}} \overset{③}{\le} c_\phi S^2/b_\delta.$$

Inequality ② uses the identical uniform laws of $\mathbf{q}$ and $-\mathbf{q}$. For ③, the integral representation gives the difference as $J\int(v_\delta - v)\phi_{\mathbf{q}}\mathrm{d}\mu(\mathbf{q})$. Its norm is at most $c_\phi\int|v|\,\mathbb{1}_{\{|v|>b_\delta\}}\mathrm{d}\mu \le c_\phi S^2/b_\delta$. These bounds justify the Hilbert-space integrals, and dominated convergence supplies continuous representatives for the integrated functions.

Apply (148) to the centered independent elements $U_\delta(\vec{\mathbf{q}}_j) - \mathbb{E}\,[U_\delta]$, whose norm is at most $2c_\phi b_\delta$ and whose summed second moments are at most $kc_\phi^2S^2$. With conditional failure probability at most $2e^{-t}$,

$$\|J(f^\dagger - f^*)\|_{\mathcal{H}} \overset{④}{\le} c_\phi\left(S\sqrt{2t/k} + \frac{4b_\delta t}{3k} + \frac{S^2}{b_\delta}\right) \overset{⑤}{\le} CS\sqrt{t/k}. \tag{151}$$

Inequality ④ divides the concentration bound by k and adds the truncation bias. Inequality ⑤ substitutes $b_\delta = S\sqrt{k/t}$. The constant C is numerical.

We also need a coefficient bound that does not grow with the confidence factor. Put $a_j^\circ = v_\delta(\vec{\mathbf{q}}_j)/\sqrt{m/2}$ for $j\in[m/2]$ and $\boldsymbol{a}^\dagger = [(\boldsymbol{a}^\circ)^\top, (\boldsymbol{a}^\circ)^\top]^\top$. Proposition F.4 gives $(\boldsymbol{a}^\dagger)^\top\boldsymbol{\psi} = f^\dagger$ for the learned gate on every outcome. Define $\omega_j = (v_\delta(\vec{\mathbf{q}}_j)^2 + v_\delta(-\vec{\mathbf{q}}_j)^2)/2$ for $j\in[k]$. Then $0\le\omega_j\le b_\delta^2$, $\mathbb{E}\,[\omega_j]\le S^2$, and $\mathrm{Var}\,[\omega_j]\le b_\delta^2S^2$. Applying (148) in $\mathbb{R}$ to their centered sum gives, with failure probability at most $2e^{-t}$,

$$\left\|\boldsymbol{a}^\dagger\right\|_2^2 \overset{⑥}{=} \frac{2}{k}\sum_{j=1}^{k}\omega_j \overset{⑦}{\le} 2\left(S^2 + \sqrt{2b_\delta^2S^2t/k} + \frac{2b_\delta^2t}{3k}\right) \overset{⑧}{\le} 8S^2.$$

Equality ⑥ accounts for both opposite directions and both duplicated channel blocks. Inequality ⑦ is the scalar concentration bound, and ⑧ substitutes $b_\delta^2 = S^2k/t$. Combining this bound with the square of (151), $k = m/4$, and $t > 1$ yields

$$\left\|f^\dagger - f^*\right\|_{L^2}^2 + \left\|f^\dagger - f^*\right\|_{L^2(P_{n_2})}^2 + \lambda\left\|\boldsymbol{a}^\dagger\right\|_2^2 \le C_L\Gamma_*^2t/m. \tag{152}$$

Conditional on any realization of the inputs in the second training sample satisfying $\mathcal{E}_X$, the failure probability in (152) is at most $4e^{-t}$. The formula for $\boldsymbol{a}^\dagger$ depends only on the target, directions, and δ, even though its error was analyzed conditionally on the inputs in the second training sample. Preservation holds for every first-stage outcome, so this conditional analysis does not restrict the computed attention update.

Step 7. Local complexity of the learned feature kernel. We identify the stochastic scale before controlling the actual GD noise. Fix the first-stage sigma-field $\mathscr{F}_1$. Recall the population covariance $\boldsymbol{\Sigma}$ and effective dimension $\mathcal{N}_{\widehat{K}}(\lambda)$ from (16). Put $\mathbf{C} = \boldsymbol{\Sigma} + \lambda \mathbf{I}_m \succ 0$. For $r \geq 0$, define the localized class

$$\mathcal{F}_{\lambda,r} := \left\{ h_{\boldsymbol{a}} = \boldsymbol{a}^\top \boldsymbol{\psi} : \|h_{\boldsymbol{a}}\|_{L^2}^2 + \lambda \|\boldsymbol{a}\|_2^2 \leq r \right\}.$$

The constraint is $\boldsymbol{a}^\top \mathbf{C} \boldsymbol{a} \leq r$. Let ϵ_i, $i \in I_2$, be independent random signs, each taking values 1 and -1 with probability $1/2$, independently of the data. The conditional local Rademacher complexity is the expectation, over these signs and the independent inputs in the second training sample, of the supremum below. Direct maximization over the coefficient ellipsoid gives

$$\begin{aligned} \mathbb{E}\left[\sup_{h \in \mathcal{F}_{\lambda,r}} \frac{1}{n_2} \sum_{i \in I_2} \epsilon_i h(\vec{\mathbf{x}}_i) \mid \mathscr{F}_1 \right] &\overset{⑨}{=} \sqrt{r}\, \mathbb{E}\left[\left\| \mathbf{C}^{-1/2} \frac{1}{n_2} \sum_{i \in I_2} \epsilon_i \boldsymbol{\psi}(\vec{\mathbf{x}}_i) \right\|_2 \mid \mathscr{F}_1 \right] \\ &\overset{①}{\leq} \sqrt{r \mathcal{N}_{\widehat{K}}(\lambda)/n_2}. \end{aligned} \tag{153}$$

Equality ⑨ is Euclidean duality after multiplication by $\mathbf{C}^{1/2}$. Inequality ① uses Cauchy–Schwarz for the expectation and cancellation of cross terms between independent centered signs. The remaining second moment is $n_2^{-1} \mathrm{tr}\left(\mathbf{C}^{-1} \boldsymbol{\Sigma}\right)$. The supremum is measurable because the coefficient ellipsoid is compact and the objective is continuous in the coefficients.

The nonnegative function $r \mapsto \sigma_0 \sqrt{r \mathcal{N}_{\widehat{K}}(\lambda)/n_2}$ is nondecreasing, and its ratio to $\sqrt{r}$ is constant on $(0, \infty)$. Its positive fixed point, when $\sigma_0^2 \mathcal{N}_{\widehat{K}}(\lambda) > 0$, is $\sigma_0^2 \mathcal{N}_{\widehat{K}}(\lambda)/n_2$. If that product vanishes, the stochastic scale is zero. This identifies the local-complexity term in (17) at the actual stopping scale $\lambda = m^{-1}$. It does not replace the approximation and optimization terms in (152).

Step 8. Sub-Gaussian noise for the finite GD iterate. We use the exact filter to turn the localized noise bound into a prediction bound without requiring GD to minimize a penalized objective. Recall $\mathbf{Z}$, $\widehat{\boldsymbol{\Sigma}} = \mathbf{Z}^\top \mathbf{Z}/n_2$, and the polynomial q_T from (119). Let $\mathbf{w}_2 = [w_i]_{i \in I_2}$ in increasing index order, put $\boldsymbol{\zeta} = \mathbf{Z}^\top \mathbf{w}_2/n_2$, and define the fitted-noise coefficients $\boldsymbol{a}_w = q_T(\widehat{\boldsymbol{\Sigma}})\boldsymbol{\zeta}$. Conditional on $\mathscr{F}_2$, the feature map and inputs in the second training sample are fixed and the entries of $\mathbf{w}_2$ are independent centered sub-Gaussian variables with the same proxy σ_0^2.

On $\mathcal{E}_S$, the covariance comparison (116) and the scalar GD filter bounds give

$$\left\| \boldsymbol{a}_w^\top \boldsymbol{\psi} \right\|_{L^2}^2 \overset{②}{\leq} \boldsymbol{\zeta}^\top q_T(\widehat{\boldsymbol{\Sigma}})^2 (2\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}_m) \boldsymbol{\zeta} \overset{③}{\leq} 6 \left\| (\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}_m)^{-1/2} \boldsymbol{\zeta} \right\|_2^2. \tag{154}$$

Inequality ② applies the covariance order to $\boldsymbol{a}_w$. For ③, $q_T(u) \leq \min\{\lambda^{-1}, u^{-1}\}$ for $u > 0$, and $q_T(0) = \lambda^{-1}$. With $x = u/\lambda$, the needed bound is $(2x+1)(x+1)\min\{1, x^{-2}\} \leq 6$, checked separately on $0 < x \leq 1$ and $x \geq 1$. At $u = 0$ the corresponding product equals one. This argument diagonalizes only functions of $\widehat{\boldsymbol{\Sigma}}$.

For completeness, we establish the required quadratic-form tail directly from the noise moment generating function. Let $\mathbf{A} = (\widehat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}_m)^{-1/2} \mathbf{Z}^\top / n_2$ and $\mathbf{H} = \mathbf{A}^\top \mathbf{A} \succeq 0$, all fixed under $\mathscr{F}_2$. If $\sigma_0 = 0$ or $\mathbf{H} = 0$, then $\mathbf{A}\mathbf{w}_2 = 0$ almost surely and the claimed tail is immediate. Otherwise, an independent standard Gaussian vector $\boldsymbol{\gamma} \in \mathbb{R}^m$ and (15) give

$$\mathbb{E}\left[e^{\theta \|\mathbf{A}\mathbf{w}_2\|_2^2} \mid \mathscr{F}_2 \right] \overset{④}{=} \mathbb{E}\left[e^{\sqrt{2\theta} \boldsymbol{\gamma}^\top \mathbf{A}\mathbf{w}_2} \mid \mathscr{F}_2 \right] \overset{⑤}{\leq} \mathbb{E}\left[e^{\sigma_0^2 \theta \|\mathbf{A}^\top \boldsymbol{\gamma}\|_2^2} \mid \mathscr{F}_2 \right].$$

Equality ④ integrates the Gaussian moment generating function first. Inequality ⑤ conditions on $\boldsymbol{\gamma}$ and multiplies the independent noise moment generating functions. Nonnegativity justifies both orders of integration. Gaussian integration in an eigenbasis, followed by the series for $-\log(1-u)$, yields

$$\log \mathbb{E}\left[e^{\theta(\|\mathbf{A}\mathbf{w}_2\|_2^2 - \sigma_0^2 \mathrm{tr}(\mathbf{H}))} \mid \mathscr{F}_2 \right] \overset{⑥}{\leq} \frac{\sigma_0^4 \theta^2 \mathrm{tr}\left(\mathbf{H}^2\right)}{1 - 2\sigma_0^2 \theta \|\mathbf{H}\|_2}, \qquad 0 < 2\sigma_0^2 \theta \|\mathbf{H}\|_2 < 1.$$

For ⑥, the Gaussian integral is the product of $(1-2\sigma_0^2\theta\rho)^{-1/2}$ over the nonzero eigenvalues ρ of $\mathbf{H}$, and $-\log(1-u)-u \le u^2/[2(1-u)]$. Exponential Markov with $\theta = \sqrt{t}/[\sigma_0^2(\sqrt{\operatorname{tr}(\mathbf{H}^2)} + 2\|\mathbf{H}\|_2\sqrt{t})]$ therefore bounds $\|\mathbf{A}\mathbf{w}_2\|_2^2$ by $\sigma_0^2[\operatorname{tr}(\mathbf{H}) + 2\sqrt{\operatorname{tr}(\mathbf{H}^2)\,t} + 2\|\mathbf{H}\|_2 t]$ with conditional failure probability at most e^{-t}.

The nonzero eigenvalues of $\mathbf{H}$ are $u/[n_2(u+\lambda)]$, where u ranges over those of $\widehat{\mathbf{\Sigma}}$. Recall $\widehat{\mathcal{N}}(\lambda)$ from (19). Consequently $\operatorname{tr}(\mathbf{H}) = \widehat{\mathcal{N}}(\lambda)/n_2$, $\|\mathbf{H}\|_2 \le n_2^{-1}$, and $\operatorname{tr}(\mathbf{H}^2) \le \widehat{\mathcal{N}}(\lambda)/n_2^2$. Substitution into (154) gives, on $\mathcal{E}_S$ and outside conditional probability e^{-t},

$$\left\|\boldsymbol{a}_w^\top \boldsymbol{\psi}\right\|_{L^2}^2 \le \frac{6\sigma_0^2}{n_2}\left[\widehat{\mathcal{N}}(\lambda) + 2\sqrt{\widehat{\mathcal{N}}(\lambda)t} + 2t\right]. \tag{155}$$

To express (155) through the population local complexity (153), fix a first-stage outcome in $\mathcal{E}_Q$ and define $\ell_\lambda(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})^\top \mathbf{C}^{-1}\boldsymbol{\psi}(\mathbf{x})$, where $\mathbf{C} = \mathbf{\Sigma} + \lambda \mathbf{I}_m$ as in Step 7. Let $K_b = 9m/4$ and $N_\lambda = \mathcal{N}_{\widehat{K}}(\lambda)$. The feature bound (111) gives $0 \le \ell_\lambda \le K_b$, $\mathbb{E}_P[\ell_\lambda] = N_\lambda$, and $\operatorname{Var}_P(\ell_\lambda) \le K_b N_\lambda$. Applying (148) in $\mathbb{R}$ to the centered independent evaluations of ℓ_λ gives, outside conditional probability $2e^{-t}$,

$$\frac{1}{n_2}\sum_{i\in I_2} \ell_\lambda(\vec{\mathbf{x}}_i) \overset{⑦}{\le} N_\lambda + \sqrt{2K_bN_\lambda t/n_2} + \frac{2K_b t}{3n_2} \overset{⑧}{\le} 2N_\lambda + \frac{7K_b t}{6n_2}.$$

Inequality ⑦ is the scalar concentration bound. Inequality ⑧ uses $\sqrt{2ab} \le a + b/2$ with $a = N_\lambda$ and $b = K_b t/n_2$. On $\mathcal{E}_S$, $(\widehat{\mathbf{\Sigma}} + \lambda \mathbf{I}_m) \succeq \mathbf{C}/2$, so inversion and a trace against $\widehat{\mathbf{\Sigma}} \succeq 0$ yield

$$\widehat{\mathcal{N}}(\lambda) \overset{⑨}{\le} 2\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\mathbf{C}^{-1}\right) \overset{①}{\le} 4N_\lambda + \frac{7K_b t}{3n_2} \overset{②}{\le} 4N_\lambda + 6t. \tag{156}$$

Inequality ⑨ uses inverse ordering, which follows by conjugating the covariance inequality by $\mathbf{C}^{-1/2}$ and diagonalizing the resulting positive definite matrix. The trace preserves the order by positive semidefiniteness of $\widehat{\mathbf{\Sigma}}$. Inequality ① uses the preceding empirical average, and ② uses $K_b = 9m/4$ and $m/n_2 \le 1$, a consequence of (12). Combining (155) and (156), with $2\sqrt{uv} \le u + v$, gives

$$\left\|\boldsymbol{a}_w^\top \boldsymbol{\psi}\right\|_{L^2}^2 \le \frac{C\sigma_0^2}{n_2}\left(\mathcal{N}_{\widehat{K}}(\lambda) + t\right). \tag{157}$$

All inverses include the positive shift λ, so zero covariance eigenvalues cause no exception.

Step 9. Combining the bounds and their probabilities. We now apply the pathwise GD decomposition to the comparison vector from Step 6. Put $r = f^* - f^\dagger$, $\mathbf{e}_2 = [r(\vec{\mathbf{x}}_i)]_{i\in I_2}$, $\mathbf{v} = (\mathbf{I}_m - \eta\widehat{\mathbf{\Sigma}})^T \boldsymbol{a}^\dagger$, and $\mathbf{c} = q_T(\widehat{\mathbf{\Sigma}})\mathbf{Z}^\top \mathbf{e}_2/n_2$. Equation (136) gives $\widehat{f} - f^* = -\mathbf{v}^\top\boldsymbol{\psi} + \mathbf{c}^\top\boldsymbol{\psi} - r + \boldsymbol{a}_w^\top\boldsymbol{\psi}$. On $\mathcal{E}_S$, the proofs of (137) and (138) are pathwise and give $\left\|\mathbf{v}^\top\boldsymbol{\psi}\right\|_{L^2}^2 \le (1+e^{-1})\lambda\left\|\boldsymbol{a}^\dagger\right\|_2^2$ and $\left\|\mathbf{c}^\top\boldsymbol{\psi}\right\|_{L^2}^2 \le 3\|r\|_{L^2(P_{n_2})}^2$. Hence

$$\mathcal{R}(\widehat{f}) \overset{③}{\le} 8\lambda\left\|\boldsymbol{a}^\dagger\right\|_2^2 + 12\|r\|_{L^2(P_{n_2})}^2 + 4\|r\|_{L^2}^2 + 4\left\|\boldsymbol{a}_w^\top\boldsymbol{\psi}\right\|_{L^2}^2. \tag{158}$$

Inequality ③ bounds the squared norm of the sum of four functions by four times the sum of their squared norms and uses $1 + e^{-1} < 2$. Substituting (152) and (157) into (158) proves the first inequality in (17). Since $0 \le \mathcal{N}_{\widehat{K}}(\lambda) \le m$ and $n_2 \ge n/2$, its second inequality follows with the displayed factor two. Assumption (15) implies the original variance bound by its second-order expansion at zero, so the expectation argument in Steps 1–4 remains applicable.

We finish by checking the joint probability without assuming independence between favorable events. The event $\mathcal{E}_Q \cap \mathcal{E}_S \cap \mathcal{E}_X$ has complement of probability at most $3n^{-8}$. On $\mathcal{E}_X$, Step 6 contributes at most $4e^{-t}$ after averaging its conditional direction probabilities. On $\mathcal{E}_Q$, the scalar effective-dimension comparison in Step 8 contributes at most $2e^{-t}$ after averaging over the first stage. The quadratic-form noise bound contributes at most e^{-t} after averaging over $\mathscr{F}_2$. The union bound gives total failure probability at most $3n^{-8} + 7e^{-t} \le 3n^{-8} + \delta$. These bounds hold under each fixed target's training law, with constants depending only on L, and make no simultaneous-event assertion over different targets.

Finally, $\Gamma_*^2 \le r_0 B_*^2$, $r_0 \asymp d^{\ell_0}$, and the resource calculations in Step 4 show that fixed δ retains the rate $O((\log d)^{-2})$. For $\delta = (\log d)^{-a}$, one has $t = \log 8 + a\log\log d$, proving the stated rate

and probability tending to one. For comparison, $\delta = d^{-a}$ gives risk $O((\log d)^{-1})$ with probability at least $1 - d^{-a} - 3n^{-8}$. The constants in these assertions also depend on the fixed target and noise bounds, budget comparison constants, and a when present. If a probability of at least $1 - \delta$ is desired and $\delta \geq 6n^{-8}$, apply the bound with $\delta/2$, replacing t_δ by $\log(16/\delta)$. This completes the high-probability assertion without altering any training parameter. □

G Full Comparison with Related Regression Results

This appendix gives the full comparison summarized in the introduction. Section G.2 explains the target-norm distinctions used in Table 2. Section G.4 derives the resource ratios, and Section G.5 analyzes the polynomial-ridge risk appearing in the table.

For the literature comparison, we write $q_n := n/d^{\ell_0}$ and let m_{lit} denote the neuron count in each cited neural-network paper. The letter L denotes the fixed supplied upper degree bound in the harmonic-channel method, and $\varepsilon_0 > 0$ is its channel-selection threshold. We take a fixed confidence parameter $\delta \in (0, 1/2)$ when displaying its orders. The constants θ_1, θ_2 are fixed and positive. A "logarithmic factor" in the table means a fixed power of $\log d$, with the power chosen as in the cited result. The neural tangent kernel (NTK) is the inner product of network-parameter gradients at initialization (Jacot et al., 2018). The symbol $\varepsilon_n > 0$ in the Yang–Li GD row denotes the population critical radius of that paper's NTK. If λ_j^{lit}, $j \geq 1$, are its positive population eigenvalues in nonincreasing order and $\sigma_0^2 > 0$ is the noise variance proxy used in that comparison, it is the smallest $u > 0$ satisfying $\sigma_0 \left(\frac{1}{n}\sum_{j\geq 1} \min\left\{\lambda_j^{\mathrm{lit}}, u^2\right\}\right)^{1/2} \leq u^2$. This definition is local to the comparison and does not identify that NTK with our target kernel or learned feature kernel. The two Yang–Li rows assume sub-Gaussian noise and fixed initialization-kernel RKHS norm bounds. Let $\mathcal{X}_{\mathrm{lit}}$ be the bounded input set in the ordinary-GD theorem, and set $u_0 := \sup_{\mathbf{x}\in\mathcal{X}_{\mathrm{lit}}} \left\|[\mathbf{x}^\top, 1]^\top\right\|_2$. That theorem also requires $n \geq C \max\left\{(\lambda_1^{\mathrm{lit}})^{-1}, \sigma_0^2 u_0^2/2\right\}$, with $C > 0$ as in that theorem. Its stopping rule uses the empirical critical radius, computed from the empirical kernel eigenvalues and the noise proxy. These spectral and stopping conditions accompany its width and risk entries.

"High probability" below refers to each cited theorem's stated probability, including any confidence parameter. It is distinct from averaging risk over all training outcomes.

Throughout Table 2, B bounds the target L^2 norm, σ_0^2 bounds noise variance, and $H_n = \log(en)$. All limits use $d \to \infty$ with degree, target and noise bounds, and structural parameters fixed. The positive constants $C, b_0, c, \delta_0, \delta_1$ are fixed, with $\delta_0 < 1/2$. Montanari–Zhong require sufficiently large fixed logarithmic exponents $c_0, c_1 > 0$. Their polynomial ridge risk $\mathcal{R}_{\mathrm{poly}}$ is defined in Section G.5. Our alternative width uses $C_0 = 512$, and O_L hides a constant depending only on the fixed degree bound L. For Lee et al., $0 < \varepsilon < 1$ is the requested accuracy. Expected bounds average data and initialization, except that Hu et al. also average random target coefficients. A vanishing bound at fixed confidence does not by itself imply convergence in probability.

G.1 Detailed Related Work

For general spherical polynomials, Ghorbani et al. (2021) identify approximation thresholds for random features and networks linearized at initialization. Their upper bounds treat population-optimal output fitting with unlimited data, or exact-kernel regression with infinite width. Mei et al. (2022) and Hu et al. (2024) already analyze finite samples and finite random features, with sharp risk characterizations in their stated regimes. Our distinction is the simultaneous logarithmic resource budgets, finite-variance expected-risk guarantee, and response-dependent representation. Their stated asymptotic hypotheses do not cover these logarithmic budgets. Table 1 summarizes the closest guarantees, and Table 2 records the full conditions. This comparison does not establish a limitation of those estimators.

Random-feature approximation, estimation, and early-stopped stochastic GD have finite-sample guarantees (Rudi & Rosasco, 2017; Carratino et al., 2018). Kernel early stopping also has statistical guarantees (Yao et al., 2007; Raskutti et al., 2014). Defilippis et al. (2024) analyze regularized random-feature risk under concentration assumptions. Our analysis combines these established approximation and spectral-GD principles with exact coefficient preservation, strict enlargement for

the performed attention step, and noise conditioning that requires only a second moment for the expected-risk bound. Appendix G.4 tracks the dimension-dependent feature and target constants needed for comparison.

Feature learning can exploit additional target structure. Damian et al. (2022) analyze a two-stage ReLU procedure for polynomials of a fixed number of unknown projections under Gaussian inputs, a nondegenerate expected target Hessian, and an absolute-loss criterion. Lee et al. (2024) obtain vanishing squared risk with nearly linear sample size for a polynomial of one unknown projection, using normalized stochastic GD, polynomial activations, and regularized output fitting. These settings do not cover our full spherical polynomial class. Abbe et al. (2022) also study structured targets, while Ba et al. (2022) and Moniri et al. (2024) quantify benefits of one-step feature learning followed by output fitting for targets depending on one projection. Table 2 retains the precise comparisons.

Yang (2026) uses an activation built from polynomial-degree components and selects a degree during training. Yang & Li (2026) study augmented ReLU features and GD with a projection constraint. Their sharper stated rates use different target normalizations and additional noise, selection, or training-size conditions, detailed in Appendix G. Our method uses biased ReLU channels without degree selection. Its attention block is motivated by squeeze-and-excitation (Hu et al., 2018), with shared weights keeping the parameter count linear in the channel count (Wang et al., 2020). The guarantees concern the complementary gate in Section 2.1, rather than arbitrary attention architectures.

G.2 Target Normalization and Kernel Norms

We first quantify the restriction imposed by a fixed weighted target radius and the resulting risk of constant prediction.

Proposition G.1 (Weighted target classes and constant prediction)**.** Assume the regression model (22) and noise condition (23), with $d \geq 3$ and the polynomial target (30). Recall the target norms B_* and Γ_* from (31), and suppose $\Gamma_* \leq \gamma_0$ for a fixed $\gamma_0 \geq 0$. Every nonconstant harmonic dimension satisfies $N(d,\ell) \geq d$. Consequently, the squared L^2 norm of the nonconstant part of f^* is at most γ_0^2/d. Let $c_* := \int f^* \mathrm{d}\mu$ and let $\bar{y} = n^{-1}\sum_{i=1}^n y_i$, regarded as a constant predictor. Then

$$\mathbb{E}\left[\mathcal{R}(\bar{y})\right] = \|f^* - c_*\|_{L^2}^2 + \frac{\|f^* - c_*\|_{L^2}^2 + \mathbb{E}\left[w_1^2\right]}{n} \leq \frac{\gamma_0^2}{d} + \frac{\gamma_0^2/d + \sigma_0^2}{n}. \tag{159}$$

The weighted-class consistency statement alone therefore does not demonstrate learning a nonvanishing nonconstant signal. For every fixed $B > 0$, the larger class $B_* \leq B$ includes pure-degree targets with L^2 norm B, independently of d.

Proof. **Outline.** We first bound the harmonic dimensions and the nonconstant energy. We then compute the risk of the empirical mean.

Step 1. Nonconstant energy. Recall that Γ_*^2 weights each squared degree-ℓ coefficient by $N(d,\ell)$. Equations (86) and (87) give $N(d,\ell) \geq d$ for $\ell \geq 1$. Orthogonality and $Y_{0,1} = 1$ give $c_* = \beta_{0,1}$ and $\|f^* - c_*\|_{L^2}^2 \overset{①}{\leq} \Gamma_*^2/d \leq \gamma_0^2/d$, where ① compares these nonconstant weights with d. By contrast, $BY_{\ell,1}$ has norm B and mean zero for every $1 \leq \ell \leq \ell_0$, proving the assertion about the larger class.

Step 2. Risk of the empirical mean. Orthogonality separates the nonconstant target from mean estimation: $\mathcal{R}(\bar{y}) \overset{②}{=} \|f^* - c_*\|_{L^2}^2 + (\bar{y} - c_*)^2$, where ② uses $\int (f^* - c_*)\mathrm{d}\mu = 0$. The responses have mean c_*, and centering and independence of the noise give $\mathbb{E}\left[(y_1 - c_*)^2\right] \overset{③}{=} \|f^* - c_*\|_{L^2}^2 + \mathbb{E}\left[w_1^2\right]$ by cancellation of the signal-noise cross term in ③. Independence across observations gives $\mathbb{E}\left[(\bar{y} - c_*)^2\right] \overset{④}{=} \mathbb{E}\left[(y_1 - c_*)^2\right]/n$. This variance formula justifies ④. Averaging the risk identity and applying Step 1 proves (159). □

The following lemma proves (164) using the diagonal bound of a continuous rotation-invariant kernel.

Lemma G.2 (Target norm for a normalized spherical kernel)**.** Let $d \geq 3$ and let $K(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}^\top \mathbf{x}')$ be a positive semidefinite kernel on $\mathcal{X} = \mathbb{S}^{d-1}$, where $k : [-1, 1] \to \mathbb{R}$ is continuous and $k(1) \leq 1$. Let $\mathcal{H}_K$ be its RKHS. For each integer $\ell \geq 0$, we define $\lambda_\ell := \int_{-1}^{1} k(t) P_\ell^{(d)}(t) \mathrm{d}\nu_d(t)$. These are the eigenvalues on the harmonic spaces $\mathcal{H}_\ell$. They satisfy $0 \leq \lambda_\ell \leq N(d, \ell)^{-1}$. If the polynomial target in (30) belongs to $\mathcal{H}_K$, then every coefficient with $\lambda_\ell = 0$ vanishes and

$$\|f^*\|_{\mathcal{H}_K}^2 = \sum_{\substack{0 \leq \ell \leq \ell_0 \\ \lambda_\ell > 0}} \sum_{j=1}^{N(d,\ell)} \frac{\beta_{\ell,j}^2}{\lambda_\ell} \geq \Gamma_*^2. \tag{160}$$

As a result, $\|f^*\|_{\mathcal{H}_K} \leq R$ implies $\|f^* - \beta_{0,1}\|_{L^2}^2 \leq R^2/d$ for every $R \geq 0$.

Proof. **Outline.** Step 1 constructs the harmonic eigenfunctions inside the RKHS and treats zero eigenvalues. Step 2 applies Bessel's inequality to kernel sections, bounds each eigenvalue by its multiplicity, and computes the polynomial target norm.

Step 1. Harmonic functions in the RKHS. We construct the relevant RKHS elements directly so that the norm formula also accounts for degrees with zero eigenvalue. We recall that μ is uniform probability measure and that $Y_{\ell,j}$ are orthonormal in $L^2(P)$. Write $K_{\mathbf{x}} = K(\mathbf{x}, \cdot)$ for the kernel section, so the reproducing property is $\langle h, K_{\mathbf{x}} \rangle_{\mathcal{H}_K} = h(\mathbf{x})$ for $h \in \mathcal{H}_K$. The map $\mathbf{x} \mapsto K_{\mathbf{x}}$ is continuous in RKHS norm, since $\|K_{\mathbf{x}} - K_{\mathbf{x}'}\|_{\mathcal{H}_K}^2 \overset{①}{=} 2[k(1) - k(\mathbf{x}^\top \mathbf{x}')]$. Equality ① expands the RKHS norm using the reproducing property. Multiplying this map by a harmonic gives a continuous Hilbert-space-valued function on the compact sphere, whose integral is therefore well defined. We define $h_{\ell,j} := \int_{\mathcal{X}} Y_{\ell,j}(\mathbf{x}) K_{\mathbf{x}} \mathrm{d}\mu(\mathbf{x})$. Bounded linear functionals commute with this integral. The reproducing property and the Funk–Hecke formula (92) give $h_{\ell,j}(\mathbf{x}') \overset{②}{=} \lambda_\ell Y_{\ell,j}(\mathbf{x}')$. Equality ② is the Funk–Hecke evaluation of the integral. Taking its RKHS inner product with itself and using the reproducing property gives $\|h_{\ell,j}\|_{\mathcal{H}_K}^2 \overset{③}{=} \int Y_{\ell,j}(\mathbf{x}) h_{\ell,j}(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \overset{④}{=} \lambda_\ell$. Equality ③ passes the bounded inner-product functional through the integral. Equality ④ uses ② and $\int Y_{\ell,j}^2 \mathrm{d}\mu = 1$. Hence $\lambda_\ell \geq 0$. If $\lambda_\ell > 0$, then $Y_{\ell,j} = h_{\ell,j}/\lambda_\ell$ belongs to $\mathcal{H}_K$. For two harmonic indices (ℓ, j) and (r, i), the same integral identity gives $\langle h_{\ell,j}, h_{r,i} \rangle_{\mathcal{H}_K} \overset{⑤}{=} \lambda_r \delta_{\ell r} \delta_{ji}$. Equality ⑤ uses orthogonality after evaluating $h_{r,i}$. As a result, the functions $\sqrt{\lambda_\ell} Y_{\ell,j}$, over positive eigenvalues and all corresponding indices, are orthonormal in $\mathcal{H}_K$.

If $\lambda_\ell = 0$, then $h_{\ell,j} = 0$. For every $f^* \in \mathcal{H}_K$, this implies $\beta_{\ell,j} = \int f^* Y_{\ell,j} \mathrm{d}\mu \overset{⑥}{=} \langle f^*, h_{\ell,j} \rangle_{\mathcal{H}_K} = 0$. Equality ⑥ again passes the inner product through the defining integral.

Step 2. Multiplicity and the target norm. The addition formula turns a bound on kernel sections into a bound on each harmonic eigenvalue, including its multiplicity. We fix a degree with $\lambda_\ell > 0$ and an input $\mathbf{x}$. Bessel's inequality for the finite orthonormal family from Step 1, applied to $K_{\mathbf{x}}$, and the addition formula (88) give

$$N(d, \ell) \lambda_\ell \overset{⑦}{=} \sum_{j=1}^{N(d,\ell)} \left| \langle K_{\mathbf{x}}, \sqrt{\lambda_\ell} Y_{\ell,j} \rangle_{\mathcal{H}_K} \right|^2 \overset{⑧}{\leq} \|K_{\mathbf{x}}\|_{\mathcal{H}_K}^2 \overset{⑨}{=} k(1) \leq 1.$$

Equality ⑦ uses reproduction and the addition formula, ⑧ is Bessel's inequality, and ⑨ uses $\|K_{\mathbf{x}}\|_{\mathcal{H}_K}^2 = K(\mathbf{x}, \mathbf{x})$. The eigenvalue bound also holds for $\lambda_\ell = 0$. Step 1 eliminates zero-eigenvalue terms from the target's finite harmonic expansion. Its coordinates in the remaining RKHS-orthonormal family are $\beta_{\ell,j}/\sqrt{\lambda_\ell}$, proving the equality in (160). The eigenvalue bound proves its inequality. Finally, $\|f^* - \beta_{0,1}\|_{L^2}^2 \overset{①}{\leq} \Gamma_*^2/d$, where ① applies the multiplicity comparison from Step 1 of Proposition G.1. This proves the last assertion. □

The projected kernel displayed in Section G.4 is continuous, positive semidefinite, and has diagonal one. The lemma therefore proves (164), including the bound for a pure-degree target. A fixed RKHS radius consequently restricts the target class as dimension increases. It cannot be replaced by a fixed L^2 radius while keeping the source's norm-dependent constants fixed.

Table 2: Complete comparison of regression guarantees and sufficient resources. The sample ratio is $q_n \coloneqq n/d^{\ell_0}$, where n is the sample size, d the input dimension, and ℓ_0 the degree bound. Feature ratios divide the stated count by d^{ℓ_0}. Here N_{RF} counts random features, m_{lit} source neurons, m our channels, and r_0 harmonic coordinates. Risk is squared error against the noiseless target, except for the explicitly marked absolute-loss row. Population fitting and exact kernels are distinguished from finite-sample training. Constants, target normalization, noise assumptions, and asymptotic regimes are specified in the text. The displayed resource conditions are sufficient guarantees, not necessary lower bounds.

Method	Target and assumptions	Sample and feature scales	Risk guarantee	Observations and training
Ghorbani et al. (2021, Theorem 1(b) and its proof)	Any deterministic spherical polynomial of degree at most ℓ_0, with $\|f^*\|_{L^2} \le B$. No low-dimensional structure or minimum coefficient size.	$n = \infty$ (population fit). $N_{\mathrm{RF}}/d^{\ell_0} \to \infty$.	Expected minimum regression risk under population fitting $O(B^2 d^{\ell_0}/N_{\mathrm{RF}})$, hence it tends to zero as $d \to \infty$ with $N_{\mathrm{RF}}/d^{\ell_0} \to \infty$.	Uniform spherical directions. Assumption 1(b) supplies all target degrees. This is the best population fit, without finite-sample output training.
Ghorbani et al. (2021, Theorem 4)	Any deterministic spherical polynomial of degree at most ℓ_0, with $\|f^*\|_{L^2} \le B$. No low-dimensional structure or minimum coefficient size.	$m_{\mathrm{lit}} = \infty$ (exact kernel). $q_n/\log d \to \infty$, $n \le C d^{\ell_0+1-\delta_0}$.	$o_{\mathbb{P}}(1)$ as $d \to \infty$ in the displayed sample and feature regime.	Uniform spherical inputs and Gaussian noise. The kernel retains degrees $0, \dots, \ell_0$, with the source's spectral and regularization conditions.
Mei et al. (2022, Theorem 2)	Any deterministic spherical polynomial of degree at most ℓ_0, with $\|f^*\|_{L^2} \le B$. No low-dimensional structure or minimum coefficient size.	$d^{\ell_0+\delta_0} \le n, N_{\mathrm{RF}} \le d^{\ell_0+1-\delta_0}$. Either $N_{\mathrm{RF}} \ge n d^{\delta_1}$ or $n \ge N_{\mathrm{RF}} d^{\delta_1}$.	$o_{\mathbb{P}}(1)$ as $d \to \infty$ in the displayed sample and feature regime.	Uniform spherical inputs and Gaussian noise. Fixed random features, source Assumption 3 with both degree thresholds ℓ_0, and the stated ridge-penalty range.

Table 2 continues in the next part.

TABLE 2 (continued)

Method	Target and assumptions	Sample and feature scales	Risk guarantee	Observations and training
Montanari & Zhong (2022, Assumptions 3.1–3.2 and Theorem 3.3)	Any deterministic spherical polynomial of degree at most ℓ_0, with $\|f^*\|_{L^2} \le B$. No low-dimensional structure is imposed.	If $\ell_0 > 1$: $d^{\ell_0}(\log d)^{c_0} \le n \le d^{\ell_0+1}/(\log d)^{c_0}$. If $\ell_0 = 1$: $b_0 d \le n \le d^2/(\log d)^{c_0}$. $m_{\mathrm{lit}} d/n \ge \log^{c_1}(m_{\mathrm{lit}} d)$.	$\mathcal{R}_{\mathrm{poly}} + \widetilde{O}_{\mathbb{P}}(\sqrt{n/(m_{\mathrm{lit}} d)} + \sqrt{n/d^{\ell_0+1}})$. Sufficient for convergence to zero in probability as $d \to \infty$: $q_n/\log d \to \infty$, bounded effective penalty, eigenvalues at least $cd^{-\ell_0}$ in every degree $0, \dots, \ell_0$ for fixed $c > 0$, and (172).	Uniform spherical inputs and centered finite-variance noise. Ridge regression fits $m_{\mathrm{lit}} d$ coefficients of fixed parameter-derivative features. The activation must cover the target degrees. Nonlinear-network transfer needs additional assumptions, as explained in Section G.5.
Zandieh et al. (2023, Theorem 5 and Remark 1)	Any spherical polynomial of degree at most the supplied ℓ_0. No low-dimensional structure or coefficient lower bound.	$n \ge Cr_0 \log r_0$ for sufficiently large C. Space dimension r_0. The algorithm uses n kernel centers.	Exact values: 0 with probability at least $1 - 10^{-4}$. Gaussian noise: $O(r_0/n)$ with probability at least 0.98. This bound tends to zero as $d \to \infty$ with $r_0/n \to 0$.	The degree is supplied. Harmonic interpolation uses exact function values. Remark 1 treats independent standard Gaussian response noise.
Yang (2026, Theorems 1–2)	Degree-ℓ_0 spherical polynomial with a fixed weighted harmonic norm bound, stronger than a fixed L^2 bound, and a lower bound on every normalized coefficient. See (162).	For fixed failure parameter $\delta \in (0, 1/2)$, $m_{\mathrm{lit}} \ge Cn^4 d^{-2\ell_0} \log(2n/\delta)$. Selection also requires $n, m_{\mathrm{lit}} \ge Cd^L \varepsilon_0^{-2} \log d$ at fixed norm, noise proxy, and confidence. Equations (162)–(163) retain the threshold dependence.	$O(d^{\ell_0}/n) = O(q_n^{-1})$ with probability at least $1 - 2\delta - o(1)$. The bound tends to zero as $d \to \infty$ with $q_n \to \infty$ and all selection conditions met.	Uniform spherical inputs and sub-Gaussian noise. Harmonic activation, supplied degree cap, attention-based degree selection, and output GD.

Table 2 continues in the next part.

TABLE 2 (continued)

Method	Target and assumptions	Sample and feature scales	Risk guarantee	Observations and training
Hu et al. (2024, Theorem 1)	Critical exponents both ℓ_0, random high-degree coefficients.	$q_n \to \theta_2$ $N_{\mathrm{RF}}/d^{\ell_0} \to \theta_1$	Expected $O(1)$, explicit bias–variance limit	Fixed random features and a fixed positive ridge penalty.
Yang & Li (2026, Theorem 3.1)	Supplied degree, bounded initialization-kernel RKHS norm, and sub-Gaussian noise.	$q_n \geq Cd^{\ell_0}\log(4/\delta)$ $m_{\mathrm{lit}}/d^{\ell_0} \geq Cq_n^{25/2}d^{5/2-\ell_0}$	$O(\log(4/\delta)/q_n)$, high probability	Augmented ReLU and projected GD. Adaptive degree selection adds conditions. See (164) for the target restriction.
Yang & Li (2024, Theorem V.1), version 6	Bounded inputs and initialization-NTK RKHS norm, with sub-Gaussian noise.	Not degree-specific $m_{\mathrm{lit}}/d^{\ell_0} \geq Cd^{5/2-\ell_0}\varepsilon_n^{-25}$	$O(\varepsilon_n^2)$, high probability	Input-weight GD with empirical early stopping. The text defines ε_n and states the accompanying spectral and stopping conditions.

Table 2 continues in the next part.

G.3 Complete Source Guarantees

The table's fixed-confidence guarantees retain their stated probabilities. For Damian et al., $n \asymp d^{2+\kappa}$ and $m_{\mathrm{lit}} \asymp d^{\kappa}$, with fixed $\kappa > 0$ and fixed structural parameters, make the displayed excess absolute-loss bound vanish. This does not imply squared-risk consistency.

For Lee et al., $\varepsilon = (\log d)^{-2}$ gives vanishing squared risk with n of order d times a fixed logarithmic power and logarithmic width, under their single-projection target restriction and degree-dependent activation. The expectation in the Hu et al. row also averages random target coefficients and need not vanish. The calculations below specify the normalization and resource conditions behind the spherical comparisons.

G.4 Resource Comparisons

The sample lower bound in Proposition H.1 requires $n/d^{\ell_0} \to \infty$ for uniform consistency under fixed positive Gaussian noise variance. It does not make the fourth logarithmic power necessary. The width lower bounds in Appendix H concern target-independent features and do not prove an optimal width for learned attention. Accordingly, the main theorem gives explicit sufficient budgets, without asserting optimal logarithmic overheads or a minimax risk rate.

The following resource comparisons distinguish population approximation, finite-sample regression, and reconstruction from exact values, as does the table. Appendix H treats necessary counts separately.

The closest target-class comparisons in Table 2 are Ghorbani et al., Mei et al., and Montanari–Zhong. Their results allow deterministic targets in the full bounded-L^2 spherical polynomial class, without restricting the relevant directions or imposing coefficient lower bounds. Changing the sphere's radius from $\sqrt{d}$ to one preserves the polynomial degree and the L^2 norm when the function and the uniform probability measure are rescaled together. As a result, the benefit over these baselines is not a larger polynomial target class. It is the combination of learned finite features, the specified finite GD procedure, explicit sufficient resource factors, and an unconditional expected-risk guarantee. The cited asymptotic probability statements hold along deterministic target sequences, and should not be interpreted as a single event holding simultaneously for every target.

The two Ghorbani et al. rows in Table 2 allow our logarithmic resource orders separately. The sample choice in (54) gives $n/(d^{\ell_0}\log d) \asymp (\log d)^3 \to \infty$ and $n/d^{\ell_0+1-\delta} \to 0$ for every fixed

TABLE 2 (continued)

Method	Target and assumptions	Sample and feature scales	Risk guarantee	Observations and training
Damian et al. (2022, Theorem 1)	Unit-L^2 polynomial of a fixed number of projections. Gaussian inputs, symmetric fixed-magnitude noise, full-rank expected Hessian on the relevant subspace with controlled conditioning.	$q_n = \widetilde{\Omega}(d^{2-\ell_0})$ is sufficient at fixed structural parameters. $m_{\mathrm{lit}}/d^{\ell_0}$, with the loss dependence shown at right.	Excess absolute loss $\widetilde{O}(\sqrt{d/n} + m_{\mathrm{lit}}^{-1/2} + n^{-1/4})$ with probability at least 0.99.	One ReLU representation step and regularized output training. This row bounds excess absolute loss, not squared risk.
Lee et al. (2024, Section 3.2)	Fixed univariate degree-ℓ_0 polynomial of one unknown unit-direction projection. Gaussian inputs and source noise assumptions.	$q_n = \widetilde{O}(d^{1-\ell_0}\varepsilon^{-2} + d^{-\ell_0}\varepsilon^{-8})$ suffices. $m_{\mathrm{lit}}/d^{\ell_0} = \widetilde{O}(d^{-\ell_0}\varepsilon^{-1})$ suffices.	Squared risk $O(\varepsilon)$ with probability tending to one as $d \to \infty$.	Degree-dependent randomized polynomial activations, normalized stochastic GD with sample reuse, and regularized output fitting.
Learned attention Theorem C.1	Every spherical polynomial of degree at most unknown ℓ_0, with $\|f^*\|_{L^2} \le B$. No low-dimensional structure or minimum coefficient size.	$n/(d^{\ell_0}H_n) \to \infty$. $m \sim n/(C_0 H_n)$, as in (66).	Expected $O_L(B^2 d^{\ell_0} H_n/n + \sigma_0^2/H_n) \to 0$ as $d \to \infty$. High-probability bound: (17) under (15).	Uniform spherical inputs and finite-variance noise. Biased ReLU channels and two-stage GD with learnable channel attention.
Supplied harmonics Proposition G.3	The same full bounded-L^2 spherical polynomial class, with degree bound ℓ_0 supplied to the estimator.	The same sample regime $n/(d^{\ell_0}H_n) \to \infty$. $r_0/d^{\ell_0} \asymp 1$.	Expected $O_L((B^2 + \sigma_0^2)d^{\ell_0}/n) \to 0$ as $d \to \infty$.	Uniform spherical inputs and finite-variance noise. Empirical harmonic coefficients use the same second training sample.

$\delta \in (0,1)$, while $m/d^{\ell_0} \asymp (\log d)^2 \to \infty$. Counting our $m/4$ independent directions preserves the latter limit. However, their population-fitting theorem and exact-kernel theorem do not together establish a guarantee with both resources finite. Ghorbani et al. (2021, Section 3.5) discuss that joint random-feature regime as a conjecture. Our theorem controls finite n, m jointly with learned features and finite-variance noise. This compares available guarantees and does not establish necessary logarithmic powers or an impossibility for their estimators.

Other spherical comparisons have different asymptotic or observation regimes. Mei & Montanari (2022) take n/d and N_{RF}/d to finite positive limits. Xiao et al. (2022) take n/d^{ℓ_0} to a finite positive limit and average over random target coefficients whose covariance is proportional to the identity within each degree. Their limiting-risk formulas therefore do not assert uniform consistency for our deterministic class at logarithmically diverging resource ratios.

Zandieh et al. (2023, Theorem 5) reconstruct a supplied-degree harmonic projection. For an exactly degree-bounded target, their relative-error guarantee gives zero error on the reconstruction event using $n \geq Cr_0 \log r_0$ exact evaluations. Their Remark 1 separately gives an $O(r_0/n)$ squared-risk bound under standard Gaussian response noise with probability at least 0.98. Their returned kernel expansion uses the sampled inputs as centers in a space of dimension r_0. At our sample scale this noisy rate is $O((\log d)^{-4})$, sharper than our uniform network bound. Our advantage in this comparison is a finite learned network trained without the degree, rather than a smaller risk rate. The cited Gaussian-noise statement is distinct from our unconditional bound over all centered finite-variance noise laws.

Montanari & Zhong (2022, Section 3.4 and Theorem 3.3) also allow centered finite-variance noise and the full polynomial target class. Their fitted tangent model uses fixed parameter-derivative features and does not receive the degree. Table 2 records its sample window, width requirement, and polynomial-risk approximation. Proposition G.4 verifies when that comparison risk vanishes, and (172) controls the remaining terms. As a result, the source can give consistency on the full class. Its positive effective penalty is not an obstruction when empirical eigenvalues grow sufficiently quickly. The source's unspecified logarithmic exponents cannot be identified with our explicit powers four and two.

Our theorem instead gives the explicit expected bound (59) for the biased ReLU features in Section A.1, and establishes the powers four and two in (54). Its width rule has $m \asymp n/(\log n)^2$, so the number of fitted output coefficients is smaller than the sample size, whereas the cited tangent model requires $m_{\mathrm{lit}} d \geq n[\log(m_{\mathrm{lit}} d)]^{c_1}$. This comparison concerns fitted scalar coefficients and does not claim that our network has fewer neurons. Our representation also changes strictly during training. Neither this structural conclusion nor the different upper bounds establishes a uniform ordering of the two predictors' actual risks.

The broader comparisons require assumptions that are not captured by resource ratios to d^{ℓ_0} alone. Defilippis et al. (2024) require feature concentration and spectral conditions for their risk approximation. The constrained or regularized shallow-network analyses of Bach (2017) and Yang & Zhou (2024) also retain their representation-norm assumptions and optimization formulations. The Gaussian or binary-input target restrictions in this broader feature-learning literature are substantive. For example, dependence on one unknown linear projection permits a much smaller model class than all degree-ℓ_0 spherical polynomials. Fu et al. (2025) study compositions of a few quadratic features using three layers. These results allow us to compare learning mechanisms, without implying that our method has a smaller risk or sample requirement on their structured classes.

We use the same notation N_{RF} for the feature counts in the following asymptotic comparisons. The spherical result of Mei et al. (2022, Theorem 2) imposes fixed positive gaps between n, N_{RF} and adjacent polynomial-degree thresholds, together with a power separation between n and N_{RF}. Its higher-moment remainder concerns only the target components above the fitted degree. Those components vanish for every degree-ℓ_0 polynomial when both fitted degree thresholds are ℓ_0, which justifies the full bounded-L^2 target class in Table 2. The cited theorem includes zero ridge penalty in both sample–feature regimes, so our advantage over that result is not the absence of an explicit penalty. It does not cover $n \asymp d^{\ell_0}(\log d)^p$ and $N_{\mathrm{RF}} \asymp d^{\ell_0}(\log d)^{p-2}$ for fixed $p > 2$. More quantitatively, for the fixed positive gap δ_0 in Table 2, the lower sample and feature scales in that theorem exceed our displayed choices by factors of orders $d^{\delta_0}/(\log d)^4$ and $d^{\delta_0}/(\log d)^2$, respectively. Both

factors diverge. The comparison concerns sufficient guarantees for the same full polynomial class, without a direction or subspace restriction.

The framework of Hu et al. (2024, Section 2.2) uses fixed positive exponents κ_1, κ_2 and finite positive constants θ_1, θ_2, with $N_{\mathrm{RF}}/d^{\kappa_1} \to \theta_1$ and $n/d^{\kappa_2} \to \theta_2$. The logarithmic factors in the displayed sequence also preclude those fixed positive limits. These comparisons concern the hypotheses of the cited statements, not impossibility results for their estimators or an exhaustive claim about all random-feature results.

The harmonic-channel and projected-GD results in Table 2 report sharper risk rates under larger sufficient sample-size and width conditions. We derive their resource ratios below, retaining the definitions of q_n, m_{lit}, and the fixed confidence parameter δ from the start of Appendix G.

The reference paper and its selection threshold. The reference assumes sub-Gaussian noise with variance proxy σ_0^2. This means $\mathbb{E}\left[\exp(tw_i)\right] \le \exp(t^2\sigma_0^2/2)$ for all $t \in \mathbb{R}$. We retain that stronger condition in this literature comparison. The reference method (Yang, 2026, Theorems 1–2) uses the degree cap $L \ge \ell_0$, the weighted radius $\gamma_0 > 0$ from (31), and a threshold $0 < \varepsilon_0 \le \beta_0^2/3$, where $\beta_0 > 0$ is its lower bound on every normalized coefficient. In our notation that condition is $|\beta_{\ell,j}|/\sqrt{\mu_\ell} \ge \beta_0$ for $0 \le \ell \le \ell_0$ and $j \in [N(d,\ell)]$, where $\mu_\ell = N(d,\ell)^{-1}$. Its displayed regression guarantee is

$$m_{\mathrm{lit}} \ge C\frac{n^4 \log(2n/\delta)}{d^{2\ell_0}} \quad \Longleftrightarrow \quad \frac{m_{\mathrm{lit}}}{d^{\ell_0}} \ge C d^{\ell_0} q_n^4 \log(2n/\delta). \tag{161}$$

Equation (161) explains the reference width entry. It is only one part of that method's conditions. For fixed γ_0, noise proxy, L, and δ, the explicit selection conditions have sufficient sample and width orders $d^L \varepsilon_0^{-2} \log d$, with their constants retained. As a result, its sample ratio has the order displayed in Table 2. For the fixed δ in Table 2, the success probability in the reference is at least $1-2\delta-o(1)$ as $d \to \infty$. Obtaining a probability tending to one requires $\delta \to 0$ and the corresponding δ-dependent sample and width conditions.

The threshold dependence is essential for comparison along a sequence of dimensions. Summing the squared coefficient lower bounds and using (31) gives

$$r_0\beta_0^2 \overset{①}{\le} \Gamma_*^2 \le \gamma_0^2, \qquad \varepsilon_0 \overset{②}{\le} \frac{\gamma_0^2}{3r_0}. \tag{162}$$

Inequality ① sums the lower bounds on all weighted squared coefficients. Inequality ② uses $\varepsilon_0 \le \beta_0^2/3$ and the resulting bound $\beta_0^2 \le \gamma_0^2/r_0$. Equation (162) rules out dimension-independent positive β_0 when γ_0 is fixed and $r_0 \to \infty$. Let $r_L := \sum_{\ell=0}^{L} N(d,\ell)$. Even with shrinking β_0, the reference's Theorem 1 requires $n \ge 400\gamma_0^4\varepsilon_0^{-2} r_L \log(12r_L/\delta)$. Substitution of (162) gives

$$n \overset{①}{\ge} 3600 r_0^2 r_L \log(12r_L/\delta), \qquad \frac{n}{d^{\ell_0}} \overset{②}{\ge} c_L d^{L+\ell_0} \log d \quad \text{for all sufficiently large } d. \tag{163}$$

Inequality ① substitutes $\varepsilon_0^{-2} \ge 9r_0^2/\gamma_0^4$ into the source condition. Inequality ② uses $r_0 \asymp d^{\ell_0}$ and $r_L \asymp d^L$, with the fixed parameters in this comparison. Equation (163) is a consequence of that sufficient condition, not a necessary sample lower bound for degree selection or for regression. It also shows why substituting our logarithmic sample size into the reference risk rate while ignoring its selection assumptions would not give a valid competing theorem. Our procedure imposes no coefficient lower bound and does not recover the degree.

Projected GD and ordinary input-weight GD. The projected-GD result (Yang & Li, 2026, Section 2.2) assumes sub-Gaussian noise and a fixed RKHS norm bound for its initialization kernel. We denote that kernel by K_{proj} and its RKHS by $\mathcal{H}_{K_{\mathrm{proj}}}$. On the unit sphere it is

$$K_{\mathrm{proj}}(\mathbf{x},\mathbf{x}') = \frac{(1+\mathbf{x}^\top\mathbf{x}')(\pi - \arccos(\mathbf{x}^\top\mathbf{x}'))}{2\pi}, \qquad K_{\mathrm{proj}}(\mathbf{x},\mathbf{x}) = 1.$$

Lemma G.2 proves the following implication for every fixed radius $R \ge 0$:

$$\|f^*\|_{\mathcal{H}_{K_{\mathrm{proj}}}} \le R \quad \Longrightarrow \quad \Gamma_*^2 \le R^2, \qquad \|f^* - \beta_{0,1}\|_{L^2}^2 \le R^2/d. \tag{164}$$

In particular, a pure degree-ℓ_0 target with L^2 norm $B > 0$ has squared RKHS norm at least $N(d,\ell_0)B^2$. Equation (164) shows that the fixed-radius projected-GD guarantee is not uniform over our full bounded-L^2 class. Allowing its RKHS radius to grow would also require tracking that radius in its constants.

Yang & Li (2026, Theorem 3.1) give, for a supplied degree, the sample requirement $n \geq C\log(4/\delta)d^{2\ell_0}$ and width requirement $m_{\mathrm{lit}} \geq C(n/d^{\ell_0})^{25/2}d^{5/2}$. Dividing by d^{ℓ_0} gives the two ratios in Table 2. The sample requirement alone exceeds the sample size in our logarithmic example for every fixed $\ell_0 \geq 1$. Their adaptive Theorem 3.2 adds a coefficient lower bound and threshold-dependent selection conditions, so the supplied-degree row must not be read as an unknown-degree guarantee at those sample orders alone. Their sharper reported risk and our smaller sufficient sample and width range address different guarantees.

For the ordinary input-weight GD guarantee of Yang & Li (2024, Theorem V.1), recall the population critical radius ε_n defined at the start of Appendix G. Dividing the width $d^{5/2}\varepsilon_n^{-25}$ by d^{ℓ_0} gives the corresponding entry in Table 2. Substituting $\varepsilon_n^2 = d^{\ell_0}/n$ without verifying its defining spectral equation and the target RKHS norm would be unjustified. Equation (54) supplies our ratios explicitly. These comparisons distinguish sufficient guarantees and do not order the exact risks of the trained predictors.

A benchmark when the degree is supplied. For the next comparison, we retain the same input law, target, noise, and second training sample indexed by I_2. We supply the degree bound ℓ_0 to an estimator based on explicit harmonic features. This estimator serves only as a benchmark for the risk bound and is not used by the attention algorithm.

Proposition G.3. Under (22)–(23) and (30), define empirical harmonic coefficients and their predictor by

$$\widehat{\beta}_{\ell,j} := \frac{1}{n_2}\sum_{i\in I_2} y_i Y_{\ell,j}(\vec{\mathbf{x}}_i), \qquad \widehat{f}_{\mathrm{pol}} := \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)} \widehat{\beta}_{\ell,j}Y_{\ell,j}. \tag{165}$$

Then the exact expected risk is

$$\mathbb{E}\left[\mathcal{R}(\widehat{f}_{\mathrm{pol}})\right] = \frac{(r_0-1)B_*^2 + r_0\mathbb{E}\left[w_1^2\right]}{n_2} \leq \frac{2r_0(B_*^2+\sigma_0^2)}{n}. \tag{166}$$

For a fixed noise law and $B \geq 0$, the worst-case risk over the targets in (30) with $B_* \leq B$ is

$$\sup_{f^*\in\bigoplus_{\ell=0}^{\ell_0}\mathcal{H}_\ell,\ \|f^*\|_{L^2}\leq B} \mathbb{E}\left[\mathcal{R}(\widehat{f}_{\mathrm{pol}})\right] = \frac{(r_0-1)B^2 + r_0\mathbb{E}\left[w_1^2\right]}{n_2}. \tag{167}$$

These assertions require no width condition, matrix inversion, or knowledge of the noise variance.

Proof. **Outline.** We first verify unbiasedness of each estimated coefficient. We then sum the coefficient variances using orthogonality and the addition formula, and maximize the resulting expression over the norm ball.

Step 1. Coefficient variances. Recall that the functions $Y_{\ell,j}$ form an orthonormal family in $L^2(P)$ and $\beta_{\ell,j}$ is the corresponding target coefficient. Then $\mathbb{E}\left[y_i Y_{\ell,j}(\vec{\mathbf{x}}_i)\right] \overset{①}{=} \beta_{\ell,j}$, where ① uses centered independent noise and the coefficient definition. Boundedness of the harmonics and $\mathbb{E}\left[y_i^2\right] < \infty$ ensure finite variances. Independence across observations in (165) gives

$$\mathbb{E}\left[(\widehat{\beta}_{\ell,j}-\beta_{\ell,j})^2\right] \overset{②}{=} \frac{1}{n_2}\left(\mathbb{E}\left[y_1^2 Y_{\ell,j}(\vec{\mathbf{x}}_1)^2\right] - \beta_{\ell,j}^2\right).$$

Equality ② is the variance of the average of the independent response-harmonic products.

Step 2. Summation. Both the target and estimator lie in $\bigoplus_{\ell=0}^{\ell_0}\mathcal{H}_\ell$. Hence $\mathcal{R}(\widehat{f}_{\mathrm{pol}}) \overset{③}{=} \sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)}(\widehat{\beta}_{\ell,j}-\beta_{\ell,j})^2$, where ③ is orthonormality. The addition formula (88) gives $\sum_{\ell=0}^{\ell_0}\sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x})^2 = r_0$. Summing the coefficient variances therefore yields

$$\mathbb{E}\left[\mathcal{R}(\widehat{f}_{\mathrm{pol}})\right] \overset{④}{=} \frac{r_0\mathbb{E}\left[y_1^2\right] - B_*^2}{n_2}, \qquad \mathbb{E}\left[y_1^2\right] \overset{⑤}{=} B_*^2 + \mathbb{E}\left[w_1^2\right].$$

Equality ④ sums ②, using the addition formula and $\sum_{\ell,j}\beta_{\ell,j}^2 = B_*^2$. Equality ⑤ uses the centered independent noise. This proves the identity in (166). The bounds $n_2 \geq n/2$ and $\mathbb{E}\left[w_1^2\right] \leq \sigma_0^2$ prove its inequality. No independence between different estimated harmonic coefficients is needed, since orthogonality eliminates cross terms in the integrated prediction error.

For the supremum in (167), the noise law is fixed and the only target-dependent quantity in (166) is B_*^2. Its coefficient $r_0 - 1$ is nonnegative. The largest permitted value is B^2, and the target $f^* = BY_{\ell_0,1}$ attains it. This proves (167), including $B = 0$. □

For $n \asymp d^{\ell_0}(\log d)^p$ with fixed $p > 2$, the uniform risk in (167) is $O_{L,p}((B^2 + \sigma_0^2)(\log d)^{-p})$. Here the implicit constant also depends on the comparison constants in the choice of n. This is a comparison with the uniform bounded-L^2 bound (70), which gives a slower stated rate. It does not order the exact risks for individual targets, or compare the finer weighted-class bound (53) uniformly across noise levels. The distinction is that (165) uses the supplied degree and explicit harmonics, whereas Algorithm 1 uses ReLU channels and chooses its optimization parameters from the observed samples and d. Consistency on nonvanishing targets distinguishes the trained network from constant prediction.

General random-feature bounds. The basic random-feature ridge theorem of Rudi & Rosasco (2017, Theorem 1) assumes bounded features and, in its displayed form, bounded responses. Carratino et al. (2018, Assumptions 2–4 and Theorem 1) analyze stochastic output GD with bounded features and a conditional response-moment assumption of every even order. Their finite-sample results already establish that finite random features and gradient fitting can yield consistency. The latter moment assumption is stronger than the finite noise variance assumed here. Neither source should be summarized as requiring infinite width.

For the present activation, the deterministic individual-feature supremum is $\sqrt{d} + 1/2$. The integral representation in Lemma F.2 has squared coefficient cost at most $C_L\Gamma_*^2$, which can have order d^{ℓ_0} on the full bounded-L^2 polynomial class. As a result, feature and target constants must be tracked when applying a general fixed-distribution bound to a sequence with growing d. Rescaling features to bound their supremum changes the coefficient cost and the training scale. Our paired-feature concentration and preservation arguments keep these dependences explicit while allowing finite-variance noise.

These observations distinguish the proved assumptions and calculations. They do not prove that every refinement of the earlier methods fails at our resource scales. Accordingly, we make no general priority claim for consistency with finite random features or for the optimality of the logarithmic overheads.

G.5 When the Polynomial Ridge Risk Vanishes

The term $\mathcal{R}_{\mathrm{poly}}$ in Table 2 is a regression risk, not a negligible remainder. Montanari & Zhong (2022, Theorem 3.3) identify it as the risk of a polynomial kernel truncated at an analytical degree s, with an increased ridge penalty. Their sample–dimension regime determines which s their comparison theorem permits. We specialize to $s = \ell_0$ under the explicit sample assumption in Proposition G.4. Their algorithm fits the full neural-tangent model without receiving the target degree. The truncated kernel and its increased penalty are comparison objects used to analyze that fitted model.

We first give a sufficient consistency criterion for this term and then state the separate requirements on the neural-tangent approximation. All limits in this subsection are as $d \to \infty$, with fixed $\ell_0 \geq 1$, uniformly bounded target norm and noise variance, and dimension-dependent sample size and width.

We rescale the source's sphere to the unit sphere with uniform probability law μ, as in Section G.4. Recall that $Y_{k,j}$ is an orthonormal degree-k harmonic and that $r_0 = \sum_{k=0}^{\ell_0} N(d,k)$. Let $\rho_{k,d} \geq 0$ be the eigenvalue of the source's truncated kernel on each degree-k harmonic. This defines the kernel by

$$K_{\mathrm{poly}}(\mathbf{x}, \mathbf{x}') = \sum_{k=0}^{\ell_0} \rho_{k,d} \sum_{j=1}^{N(d,k)} Y_{k,j}(\mathbf{x}) Y_{k,j}(\mathbf{x}'). \tag{168}$$

In the source's Gegenbauer notation, $\rho_{k,d}$ equals its degree-k coefficient divided by $N(d,k)$. As a result, positivity must be checked also at degree zero if the comparison includes arbitrary constant components.

Let σ_{lit} be the source activation and G a standard Gaussian variable. For each integer $k \geq 0$, set $H_k = \mathrm{He}_k / \sqrt{k!}$, using the probabilists' Hermite polynomials defined in the proof of Lemma F.1.

As a result, H_k is orthonormal for the standard Gaussian law and has positive leading coefficient. Define $v_{\ell_0} = \sum_{k\geq\ell_0}(\mathbb{E}\,[\sigma'_{\mathrm{lit}}(G)H_k(G)])^2$. The source's activation assumptions make this sum finite and positive. Let $\lambda_d \geq 0$ be the source's ridge penalty for the sum of squared training errors, without division by n. The effective penalty in Theorem 3.3 is $\tau_d := \lambda_d + v_{\ell_0} > 0$.

For independent observations $(\vec{\mathbf{x}}_i, y_i)_{i=1}^n$, define the matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ by $\mathbf{K} = (K_{\mathrm{poly}}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j))_{i,j=1}^n$, the column vector $\mathbf{k}(\mathbf{x}) \in \mathbb{R}^n$ by $\mathbf{k}(\mathbf{x}) = (K_{\mathrm{poly}}(\vec{\mathbf{x}}_i, \mathbf{x}))_{i=1}^n$, and the response vector $\mathbf{y} = (y_i)_{i=1}^n \in \mathbb{R}^n$. The predictor is $\widehat{f}_{\mathrm{poly}}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top(\mathbf{K} + \tau_d\mathbf{I}_n)^{-1}\mathbf{y}$, and $\mathcal{R}_{\mathrm{poly}} = \left\|\widehat{f}_{\mathrm{poly}} - f^*\right\|^2_{L^2(\mu)}$. The positivity of τ_d makes this definition valid on every sample.

Proposition G.4 (A sufficient consistency criterion for polynomial ridge regression)**.** Fix the analytical cutoff $s = \ell_0 \geq 1$. Let $c_0 > 1$ be the fixed sufficiently large logarithmic exponent in Table 2, and let $b_0 > 0$ be fixed. This source exponent c_0 is distinct from our numerical width constant $C_0 = 512$. Assume that $n = n(d)$ satisfies the sample regime of Montanari & Zhong (2022, Assumption 3.1), specialized to this cutoff, for all sufficiently large d:

$$\begin{cases} d^{\ell_0}(\log d)^{c_0} \leq n \leq d^{\ell_0+1}/(\log d)^{c_0}, & \ell_0 > 1, \\ b_0 d \leq n \leq d^2/(\log d)^{c_0}, & \ell_0 = 1. \end{cases} \tag{169}$$

This assumption selects a regime for the analysis and does not supply ℓ_0 to training. Suppose f^* has degree at most ℓ_0 and $\|f^*\|_{L^2(\mu)} \leq B$. Assume that the inputs are independent with law μ and that $y_i = f^*(\vec{\mathbf{x}}_i) + w_i$, where the noises are independent of the inputs and of each other, centered, and have variance at most σ_0^2. For the kernel in (168), suppose $\rho_{\min,d} := \min_{0\leq k\leq\ell_0}\rho_{k,d} > 0$. Then, for every $t > 0$,

$$\Pr\left[\mathcal{R}_{\mathrm{poly}} > t\right] \leq 2r_0\exp\left(-\frac{3n}{28r_0}\right) + \frac{4B^2(\tau_d/(n\rho_{\min,d}))^2 + 2\sigma_0^2 r_0/n}{t}. \tag{170}$$

Consequently, for fixed target-norm and noise bounds B, σ_0, the sufficient conditions

$$\frac{r_0\log(r_0+1)}{n} \longrightarrow 0, \qquad \frac{\tau_d}{n\rho_{\min,d}} \longrightarrow 0 \quad \text{as } d\to\infty \tag{171}$$

imply $\mathcal{R}_{\mathrm{poly}} \to 0$ in probability. The first condition controls the number of estimated harmonic coefficients, and the second makes shrinkage negligible even in the least strongly represented target degree. For $\ell_0 > 1$, the first condition follows from (169), since $c_0 > 1$ and $r_0 \asymp d^{\ell_0}$. For $\ell_0 = 1$, it remains an additional sufficient requirement.

Proof. **Outline.** Step 1 concentrates the empirical Gram matrix of the orthonormal harmonics. Step 2 writes the ridge estimator in these coordinates and bounds its conditional squared bias and noise variance. Step 3 combines the good-design probability with Markov's inequality.

Step 1. The harmonic design. Concentrating the harmonic Gram matrix permits coefficientwise control of the ridge estimator. The finite-dimensional bound below does not require (169), which specifies its application at $s = \ell_0$. Recall that r_0 counts harmonics through degree ℓ_0. All matrices in this proof refer to polynomial ridge regression.

We stack the harmonics in a fixed order as $\mathbf{Y}(\mathbf{x}) \in \mathbb{R}^{r_0}$, and let $\mathbf{Z} \in \mathbb{R}^{n\times r_0}$ have rows $\mathbf{Y}(\vec{\mathbf{x}}_i)^\top$. Orthonormality and (88) give $\mathbb{E}\left[\mathbf{Y}(\vec{\mathbf{x}}_i)\mathbf{Y}(\vec{\mathbf{x}}_i)^\top\right] = \mathbf{I}_{r_0}$ and $\|\mathbf{Y}(\mathbf{x})\|_2^2 = r_0$. For $\mathbf{X}_i = \mathbf{Y}(\vec{\mathbf{x}}_i)\mathbf{Y}(\vec{\mathbf{x}}_i)^\top - \mathbf{I}_{r_0}$, the rank-one eigenvalue r_0 gives $\|\mathbf{X}_i\|_2 \leq r_0$ and $\mathbb{E}\left[\mathbf{X}_i^2\right] \overset{①}{=} (r_0-1)\mathbf{I}_{r_0}$. Equality ① expands the square using $(\mathbf{Y}\mathbf{Y}^\top)^2 = r_0\mathbf{Y}\mathbf{Y}^\top$ and orthonormality. Put $\mathbf{G} = \mathbf{Z}^\top\mathbf{Z}/n$ and $\mathcal{E}_{\mathrm{pol}} = \{\|\mathbf{G} - \mathbf{I}_{r_0}\|_2 \leq 1/2\}$. Then $\Pr\left[\mathcal{E}_{\mathrm{pol}}^c\right] \overset{②}{\leq} 2r_0\exp(-3n/(28r_0))$, where ② applies (95) with summand bound r_0, variance bound nr_0, and threshold $n/2$. On $\mathcal{E}_{\mathrm{pol}}$, one has $\mathbf{G} \succeq \mathbf{I}_{r_0}/2$.

Step 2. Conditional bias and variance. The harmonic coordinates separate the deterministic shrinkage error from the fitted noise without requiring an inverse of the unregularized design. We write $f^*(\mathbf{x}) = \mathbf{b}^\top\mathbf{Y}(\mathbf{x})$ for $\mathbf{b} \in \mathbb{R}^{r_0}$, so $\|\mathbf{b}\|_2 \leq B$. Let $\mathbf{D} \in \mathbb{R}^{r_0\times r_0}$ be the diagonal matrix that repeats $\rho_{k,d}$ exactly $N(d,k)$ times in the chosen harmonic order. Then $\mathbf{K} = \mathbf{Z}\mathbf{D}\mathbf{Z}^\top$. Define $\mathbf{A} := (\tau_d/n)\mathbf{D}^{-1} \in \mathbb{R}^{r_0\times r_0}$. The normal equations give the fitted harmonic coefficient vector

$\widehat{\mathbf{b}} \in \mathbb{R}^{r_0}$ and predictor as

$$\widehat{\mathbf{b}} \overset{③}{=} (\mathbf{G}+\mathbf{A})^{-1}\mathbf{Z}^\top \mathbf{y}/n, \qquad \widehat{f}_{\mathrm{poly}}(\mathbf{x}) = \widehat{\mathbf{b}}^\top \mathbf{Y}(\mathbf{x}).$$

Equality ③ follows from the normal equations obtained by minimizing $\|\mathbf{y}-\mathbf{Z}\mathbf{b}'\|_2^2 + \tau_d(\mathbf{b}')^\top \mathbf{D}^{-1}\mathbf{b}'$ over $\mathbf{b}' \in \mathbb{R}^{r_0}$. Its unique minimizer agrees with the kernel formula because $(\mathbf{Z}^\top\mathbf{Z}+\tau_d\mathbf{D}^{-1})\mathbf{D}\mathbf{Z}^\top = \mathbf{Z}^\top(\mathbf{K}+\tau_d\mathbf{I}_n)$. In particular, $\mathcal{R}_{\mathrm{poly}} = \left\|\widehat{\mathbf{b}}-\mathbf{b}\right\|_2^2$.

On $\mathcal{E}_{\mathrm{pol}}$, we put $\mathbf{C} = (\mathbf{G}+\mathbf{A})^{-1}$. Then $\|\mathbf{C}\|_2 \le 2$ and $\|\mathbf{A}\|_2 = \tau_d/(n\rho_{\min,d})$. Conditional on the inputs, with $\mathbf{w} = (w_i)_{i=1}^n$, the coefficient error is $\widehat{\mathbf{b}}-\mathbf{b} \overset{④}{=} -\mathbf{C}\mathbf{A}\mathbf{b} + \mathbf{C}\mathbf{Z}^\top\mathbf{w}/n$. Equality ④ substitutes $\mathbf{y} = \mathbf{Z}\mathbf{b}+\mathbf{w}$ and uses $\mathbf{C}\mathbf{G}-\mathbf{I}_{r_0} = -\mathbf{C}\mathbf{A}$. The noise term is conditionally centered. Its conditional covariance is bounded above by $(\sigma_0^2/n)\mathbf{C}\mathbf{G}\mathbf{C}$. Indeed, independence and the variance bounds give $\mathrm{Cov}(\mathbf{w} \mid \mathbf{Z}) \preceq \sigma_0^2\mathbf{I}_n$, and congruence by $\mathbf{C}\mathbf{Z}^\top/n$ gives this inequality. Moreover, $\mathbf{C}\mathbf{G}\mathbf{C} \overset{⑤}{\preceq} \mathbf{C}(\mathbf{G}+\mathbf{A})\mathbf{C} \overset{⑥}{=} \mathbf{C}$. Inequality ⑤ uses $\mathbf{A} \succeq 0$ and congruence by $\mathbf{C}$, and equality ⑥ uses the definition of $\mathbf{C}$. Hence the conditional squared bias is at most $4B^2(\tau_d/(n\rho_{\min,d}))^2$, and the conditional variance trace is at most $2\sigma_0^2 r_0/n$. Their sum bounds $\mathbb{E}\left[\mathcal{R}_{\mathrm{poly}} \mid \vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_n\right]$ on $\mathcal{E}_{\mathrm{pol}}$.

Step 3. The probability statement. Splitting off the exceptional design event and applying Markov's inequality gives $\Pr[\mathcal{R}_{\mathrm{poly}} > t] \overset{⑦}{\le} \Pr\left[\mathcal{E}_{\mathrm{pol}}^c\right] + t^{-1}\mathbb{E}\left[\mathcal{R}_{\mathrm{poly}}\mathbb{I}_{\{\mathcal{E}_{\mathrm{pol}}\}}\right]$, justifying ⑦. Steps 1 and 2 bound these terms and prove (170). The first condition in (171) implies $\log(2r_0) - 3n/(28r_0) \to -\infty$ and $r_0/n \to 0$. The second makes the squared-bias bound vanish. For fixed B, σ_0, applying (170) at every fixed $t > 0$ proves convergence in probability. When $\ell_0 > 1$, the sample lower bound in (169) gives $r_0 \log(r_0+1)/n = O((\log d)^{1-c_0}) \to 0$, verifying the statement's final observation. This argument does not bound unconditional expected risk on the exceptional event. □

For a concrete interpretation, suppose $\tau_d = O(1)$ and $\rho_{\min,d} \ge cd^{-\ell_0}$ for a fixed $c > 0$. Since $r_0 \asymp d^{\ell_0}$, the condition $q_n/\log d \to \infty$, where $q_n = n/d^{\ell_0}$, implies (171). The same proof gives $\mathcal{R}_{\mathrm{poly}} = O_{\mathbb{P}}(B^2 q_n^{-2} + \sigma_0^2 q_n^{-1})$. As a result, the polynomial ridge term does vanish on this full bounded-L^2 target class. A positive constant effective penalty does not prevent consistency, because the relevant empirical eigenvalues grow with n. This criterion is sufficient.

In the source's linear Gaussian-noise example, Corollary 3.2 and Remark 4 give, under their hypotheses, the sharper condition $n/d \to \infty$ for the linear ridge term with fixed positive effective penalty and a nonzero linear kernel coefficient. When $n/d \to \kappa \in (0,\infty)$, its bias or variance generally remains nonzero.

The source's activation assumption alone does not ensure the quantitative lower bound on $\rho_{\min,d}$. If $\rho_{k,d} = 0$ for a degree k carrying target energy, every fitted function from (168) omits that component. Orthogonality then gives $\mathcal{R}_{\mathrm{poly}} \ge \sum_{j=1}^{N(d,k)} \beta_{k,j}^2$, independently of n and width. This obstruction is relevant only to omitted components. When all target degrees are represented sufficiently strongly, the preceding criterion establishes consistency.

Finally, let $C, C' > 0$ denote the fixed logarithmic exponents in Montanari–Zhong's Theorem 3.3. In addition to that theorem's sample and width hypotheses, a sufficient condition for its two comparison remainders to vanish is

$$\frac{n[\log(m_{\mathrm{lit}}d)]^{C'}}{m_{\mathrm{lit}}d} \longrightarrow 0, \qquad \frac{n(\log n)^C}{d^{\ell_0+1}} \longrightarrow 0 \quad \text{as } d \to \infty. \tag{172}$$

Under the source's activation and width hypotheses, the sample regime (169) together with (171) and (172) implies vanishing neural-tangent risk in probability. Under the concrete spectral lower bound and the assumption $\tau_d = O(1)$ above, one admissible choice is $n = d^{\ell_0}(\log d)^a$ with $a > \max\{1, c_0\}$, and $m_{\mathrm{lit}}d = n(\log d)^b$ with $b > \max\{c_1, C'\}$, after integer rounding. Here c_0, c_1 are the source exponents displayed in Table 2. The upper sample condition and the second limit in (172) then hold because a fixed power of $\log d$ is negligible relative to d. This explicitly gives consistency for the same unrestricted polynomial class in that source's parameter regime. It does not identify the unspecified source exponents with the powers four and two in our theorem.

The neural-tangent row concerns the fitted tangent model. The transfer to a nonlinear network in Montanari & Zhong (2022, Theorem 5.1) additionally requires that theorem's smooth activation, sub-Gaussian responses, symmetric initialization, scaling, and gradient-flow conditions.

H Necessary Sample Size and the Scope of Width Lower Bounds

This appendix separates a sample-size lower bound that applies to every estimator from a width restriction for fixed random features. The distinction is necessary because our first-stage attention features depend on the observed responses. The structural result for our attention network is strict enlargement of the prediction space at the chosen width. The lower bounds below specify the scope of the resource comparisons. They do not imply that strict feature enlargement reduces the necessary sample size or width.

H.1 Sample Size for the Full Polynomial Class

A full finite-dimensional L^2 ball under Gaussian noise admits a minimax lower bound involving the smaller of its squared radius and the noise variance times the dimension-to-sample-size ratio. See Zhang et al. (2015, Section 3.3.1, equation (17)) and the finite-rank lower-bound argument in Raskutti et al. (2012, Theorem 2(a)). We give the elementary argument directly for our spherical model. This makes the dependence on the radius and noise explicit and avoids identifying the different norm constraints in the cited results.

Recall from (31) that $r_0 = \sum_{\ell=0}^{\ell_0} N(d,\ell)$ is the dimension of the degree-at-most-ℓ_0 polynomial space in $L^2(P)$. For a radius $B > 0$, define

$$\mathcal{F}_{d,\ell_0}(B) := \left\{ f \in \bigoplus_{\ell=0}^{\ell_0} \mathcal{H}_\ell : \|f\|_{L^2} \le B \right\}. \tag{173}$$

The ball in (173) is the bounded-L^2 class covered by the main consistency result. In the next proposition, the Gaussian noise standard deviation is $\sigma_0 > 0$, so its variance equals σ_0^2. This is a submodel of our finite-variance observation model. Uniform consistency means that the supremum of expected regression risk over this ball tends to zero.

Proposition H.1. Let $d \ge 3$, $n \ge 1$, and $\ell_0 \ge 1$ be integers. Use the observation equation (22), allowing any positive sample size for this lower bound, and assume independent noise $w_i \sim \mathcal{N}(0,\sigma_0^2)$ with $\sigma_0 > 0$. For every $B > 0$,

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}_{d,\ell_0}(B)} \mathbb{E}\left[\left\| \widehat{f} - f^* \right\|_{L^2}^2 \right] \ge \frac{3}{128} \min\left\{ B^2, \frac{\sigma_0^2 r_0}{n} \right\}. \tag{174}$$

The infimum ranges over measurable estimators from the n observations into $L^2(P)$ with its norm Borel sigma-field, including randomized estimators and estimators supplied with ℓ_0, B, and σ_0.

Proof. **Outline.** We construct a finite family by independently changing the signs of small harmonic coefficients. We bound the statistical distinguishability of two targets differing in one coefficient. We then convert every coefficient-sign error into squared prediction error and sum the resulting testing bounds.

Step 1. A finite family inside the target ball. We choose coefficient amplitudes small enough both to stay inside the norm ball and to make neighboring targets statistically hard to distinguish. We relabel the orthonormal harmonics of degrees at most ℓ_0 as $Y_1, \dots, Y_{r_0}$. This relabeling is local to the proof. Set

$$a^2 := \min\left\{ \frac{B^2}{r_0}, \frac{\sigma_0^2}{16n} \right\}, \qquad f_{\boldsymbol{\theta}} := a \sum_{j=1}^{r_0} \theta_j Y_j, \quad \boldsymbol{\theta} = (\theta_1, \dots, \theta_{r_0}) \in \{-1,1\}^{r_0}, \quad a > 0.$$

Orthonormality gives $\|f_{\boldsymbol{\theta}}\|_{L^2}^2 \overset{①}{=} r_0 a^2 \le B^2$, so every target belongs to (173). Equality ① uses orthonormality and $\theta_j^2 = 1$. Let $\mathsf{P}_{\boldsymbol{\theta}}$ be the joint law of all n observations under $f_{\boldsymbol{\theta}}$, and let $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ denote expectation under this law. An estimator's independent random seed, when used, is included in this law with the same distribution for all $\boldsymbol{\theta}$.

Step 2. Adjacent targets are difficult to distinguish. Bounding the distance between the observation laws limits how reliably any estimator can recover one coefficient sign. For probability laws P, Q with densities p, q relative to a common measure ν, write $D(\mathsf{P}, \mathsf{Q}) := \int p \log(p/q) \mathrm{d}\nu$ for their relative entropy and $\mathrm{TV}(\mathsf{P}, \mathsf{Q}) := \sup_A |\mathsf{P}(A) - \mathsf{Q}(A)|$ for their total variation distance, where the supremum is over measurable events. Let $\boldsymbol{\theta}^{(j)}$ be $\boldsymbol{\theta}$ with its jth sign reversed. Conditional on the inputs, the observations have independent Gaussian responses with the same variance. The relative entropy between Gaussian laws with means u, v and common variance σ_0^2 is $(u-v)^2/(2\sigma_0^2)$. Averaging over the common input law and using orthonormality therefore gives

$$D(\mathsf{P}_{\boldsymbol{\theta}}, \mathsf{P}_{\boldsymbol{\theta}^{(j)}}) \overset{②}{=} \frac{n}{2\sigma_0^2} \|f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta}^{(j)}}\|_{L^2}^2 \overset{③}{=} \frac{2na^2}{\sigma_0^2} \overset{④}{\leq} \frac{1}{8}. \tag{175}$$

Equality ② adds the conditional Gaussian relative entropies and averages over the common input law. Equality ③ uses $f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta}^{(j)}} = 2a\theta_j Y_j$. Inequality ④ uses $a^2 \leq \sigma_0^2/(16n)$. An independent random seed contributes zero to this relative entropy. Pinsker's inequality bounds total variation by the square root of half the relative entropy. Consequently, (175) gives $\mathrm{TV}(\mathsf{P}_{\boldsymbol{\theta}}, \mathsf{P}_{\boldsymbol{\theta}^{(j)}}) \overset{⑤}{\leq} 1/4$. Inequality ⑤ substitutes the relative-entropy bound into Pinsker's inequality.

Step 3. Testing errors imply prediction error. A wrong coefficient sign costs at least a^2 in squared error, so the testing difficulty can be summed across all coordinates. We fix any estimator with finite expected risk at every target in this family. Estimators with infinite risk at one of them already satisfy (174). Define its projected coefficients $\widehat{\beta}_j := \int \widehat{f} Y_j \mathrm{d}\mu$ and set $\widehat{\theta}_j = 1$ when $\widehat{\beta}_j \geq 0$, and $\widehat{\theta}_j = -1$ otherwise. Each coefficient is measurable because integration against Y_j is a continuous linear functional on $L^2(P)$. These coefficients are well defined almost surely under every law in the finite family. Projection onto the harmonic span and the definition of $\widehat{\theta}_j$ give

$$\left\|\widehat{f} - f_{\boldsymbol{\theta}}\right\|_{L^2}^2 \overset{⑥}{\geq} \sum_{j=1}^{r_0} (\widehat{\beta}_j - a\theta_j)^2 \overset{⑦}{\geq} a^2 \sum_{j=1}^{r_0} \mathbb{1}_{\{\widehat{\theta}_j \neq \theta_j\}}.$$

Inequality ⑥ is Bessel's inequality for the error, and inequality ⑦ uses the distance of a coefficient with the wrong sign from $a\theta_j$. For fixed j and fixed remaining signs, we pair the targets with $\theta_j = 1$ and $\theta_j = -1$. If A is the event $\widehat{\theta}_j = 1$, the sum of their sign-error probabilities is $\mathsf{P}_{\boldsymbol{\theta}}(A^c) + \mathsf{P}_{\boldsymbol{\theta}^{(j)}}(A) \overset{⑧}{\geq} 1 - \mathrm{TV}(\mathsf{P}_{\boldsymbol{\theta}}, \mathsf{P}_{\boldsymbol{\theta}^{(j)}}) \geq 3/4$, where $\theta_j = 1$. Inequality ⑧ is the definition of total variation applied to A. As a result, averaging over all sign vectors gives at least $3/8$ error probability for every coordinate. Consequently,

$$\sup_{f^* \in \mathcal{F}_{d,\ell_0}(B)} \mathbb{E}\left[\left\|\widehat{f} - f^*\right\|_{L^2}^2\right] \overset{⑨}{\geq} 2^{-r_0} \sum_{\boldsymbol{\theta} \in \{-1,1\}^{r_0}} \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\widehat{f} - f_{\boldsymbol{\theta}}\right\|_{L^2}^2\right] \overset{①}{\geq} \frac{3r_0 a^2}{8}.$$

Inequality ⑨ bounds a supremum by the average over the finite subfamily. Inequality ① combines ⑦ with the average sign-error bound $3/8$. The definition of a gives $3r_0a^2/8 \overset{②}{\geq} (3/128)\min\left\{B^2, \sigma_0^2 r_0/n\right\}$. Inequality ② uses $\min\{x, y/16\} \geq \min\{x, y\}/16$ for nonnegative x, y. Taking the infimum over estimators proves (174). For fixed positive B, σ_0, its right side can tend to zero only if $r_0/n \to 0$. Lemma E.4 gives $r_0 \asymp d^{\ell_0}$ for fixed ℓ_0. As a result, uniform expected-risk consistency requires $n/d^{\ell_0} \to \infty$. □

More quantitatively, for any desired worst-case expected risk ε with $0 < \varepsilon < 3B^2/128$, inequality (174) requires $n \geq 3\sigma_0^2 r_0/(128\varepsilon)$. Conversely, the supplied-harmonic estimator of Proposition G.3 is uniformly consistent when $r_0/n \to 0$ and B, σ_0 are fixed. As a result, the sample-size condition is necessary and sufficient for uniform consistency when no architectural or computational restrictions are imposed and the degree bound is supplied. Our attention theorem achieves consistency with the stated logarithmic sample overhead and without supplying the degree.

The lower bound uses the full fixed-radius L^2 ball. It does not impose a fixed upper bound on the weighted norm Γ_* in (31). Under that latter restriction, nonconstant target energy can vanish with d, as explained in Proposition G.1. Positive fixed noise variance is also essential to the stated necessity argument. The proposition does not assert this sample requirement for each individual structured polynomial or for noiseless regression.

H.2 Width Bounds for Fixed Representations

The next proposition specializes the approximation-dimension bound of Kamath et al. (2020, Theorem 19) to spherical polynomials. We give a direct proof that makes the target class, feature count, and target-independent feature law explicit. In this proposition, m counts all scalar features available to the output layer. It equals the neuron count when each fixed neuron supplies one feature and there is no separately fitted output constant. Such a constant can instead be included as a constant feature.

Proposition H.2 (Width bound for fixed representations)**.** Let $d \geq 3$, $\ell_0 \geq 1$, and $m \geq 1$ be integers, and let $0 < \varepsilon < 1$. Recall that P is the uniform probability measure on $\mathcal{X}$ and that $\mathcal{H}_{\ell_0}$ is the degree-ℓ_0 spherical harmonic space. Put $r = N(d, \ell_0) = \dim \mathcal{H}_{\ell_0}$. Define the normalized target class $\mathcal{U}_{\ell_0} = \{f \in \mathcal{H}_{\ell_0} : \|f\|_{L^2} = 1\}$. Targets are evaluated through their continuous polynomial representatives. Auxiliary randomness ω has law ν on a probability space $(\Omega, \mathcal{A}, \nu)$. This space, its law, and the feature map are fixed for the entire target class. They may depend on d, ℓ_0, and P, but not on the chosen f. All random functions below are jointly measurable in their random arguments and inputs. We suppress ω in the notation for a realized map and write $\mathbb{E}_\nu[\cdot]$ for expectation over this auxiliary randomness.

Let $\mathbf{\Phi} = (\Phi_1, \dots, \Phi_m)^\top : \Omega \times \mathcal{X} \to \mathbb{R}^m$ be a feature map with $\Phi_j(\omega, \cdot) \in L^2(P)$ for ν-almost every ω and every j. Its coordinates need not be independent. Then

$$\sup_{f \in \mathcal{U}_{\ell_0}} \mathbb{E}_\nu \left[\inf_{\boldsymbol{a} \in \mathbb{R}^m} \left\| \boldsymbol{a}^\top \mathbf{\Phi} - f \right\|_{L^2}^2 \right] \leq \varepsilon \quad \Longrightarrow \quad m \geq (1 - \varepsilon) r. \tag{176}$$

The expectation in (176) is over the feature map. The fitted coefficients may have full knowledge of the target.

For fixed ℓ_0 and fixed $\varepsilon \in (0, 1)$, the right-hand side of (176) has order d^{ℓ_0} as $d \to \infty$, since $r \sim d^{\ell_0}/\ell_0!$. The bound also applies to a uniform guarantee over the full degree-at-most-ℓ_0 unit L^2 ball, which contains $\mathcal{U}_{\ell_0}$.

Proof. **Outline.** Step 1 bounds how much of an orthonormal family a finite-dimensional prediction space can represent. Step 2 verifies measurability and averages over a common feature law to prove the width claim.

Step 1. Orthogonal targets and prediction dimension. We first bound the total energy that a space of dimension s can capture from r orthogonal unit targets. Recall the orthonormal basis $Y_{\ell_0,1}, \dots, Y_{\ell_0,r}$ of $\mathcal{H}_{\ell_0}$ in $L^2(P)$. Each basis function belongs to $\mathcal{U}_{\ell_0}$. For any finite-dimensional subspace $\mathcal{S} \subset L^2(P)$, let $\Pi_{\mathcal{S}}$ be its orthogonal projection and let $e_1, \dots, e_s$ be an orthonormal basis of $\mathcal{S}$, where $s = \dim \mathcal{S}$. Bessel's inequality gives

$$\sum_{j=1}^{r} \|\Pi_{\mathcal{S}} Y_{\ell_0,j}\|_{L^2}^2 \overset{①}{=} \sum_{u=1}^{s} \sum_{j=1}^{r} \left| \langle e_u, Y_{\ell_0,j} \rangle_{L^2(P)} \right|^2 \overset{②}{\leq} s.$$

Equality ① expands each projection in the orthonormal basis of $\mathcal{S}$, and inequality ② applies Bessel's inequality to each e_u. Since every $Y_{\ell_0,j}$ has norm one, the Pythagorean identity yields

$$\frac{1}{r} \sum_{j=1}^{r} \inf_{h \in \mathcal{S}} \|h - Y_{\ell_0,j}\|_{L^2}^2 \overset{③}{=} 1 - \frac{1}{r} \sum_{j=1}^{r} \|\Pi_{\mathcal{S}} Y_{\ell_0,j}\|_{L^2}^2 \overset{④}{\geq} 1 - \frac{s}{r}.$$

Equality ③ is the Pythagorean projection identity, and inequality ④ substitutes ②.

Step 2. Random fixed features. We must average over the feature law without choosing the target after observing the features. Measurability of the minimum error makes this averaging valid. The event on which all coordinates of $\mathbf{\Phi}$ belong to $L^2(P)$ is measurable and has ν-probability one, because each squared coordinate norm is the integral of a jointly measurable nonnegative function. We may set $\mathbf{\Phi} = \mathbf{0}$ outside this event without changing any expectation. For each remaining realization, put $\mathcal{S}_{\mathbf{\Phi}} = \operatorname{span}\{\Phi_1, \dots, \Phi_m\}$. This is a closed subspace of $L^2(P)$ with dimension at most m.

For fixed f and coefficients, Tonelli's theorem makes the integrated squared error measurable in ω. Its dependence on the coefficients is continuous, since Cauchy–Schwarz gives

$$\left\|\boldsymbol{a}^\top \boldsymbol{\Phi} - \mathbf{b}^\top \boldsymbol{\Phi}\right\|_{L^2} \overset{⑤}{\leq} \|\boldsymbol{a} - \mathbf{b}\|_2 \left(\sum_{j=1}^{m} \|\Phi_j\|_{L^2}^2\right)^{1/2}, \qquad \boldsymbol{a}, \mathbf{b} \in \mathbb{R}^m.$$

Inequality ⑤ is the integrated Cauchy–Schwarz bound. Continuity makes the infimum over $\mathbb{R}^m$ equal to the infimum over the countable dense set $\mathbb{Q}^m$, proving measurability. The zero predictor bounds this minimum between zero and one for $f \in \mathcal{U}_{\ell_0}$. We can therefore average Step 1 over the common law ν, without choosing a measurable basis of the random span:

$$\sup_{f \in \mathcal{U}_{\ell_0}} \mathbb{E}_\nu \left[\inf_{\boldsymbol{a} \in \mathbb{R}^m} \left\|\boldsymbol{a}^\top \boldsymbol{\Phi} - f\right\|_{L^2}^2\right] \overset{⑥}{\geq} \frac{1}{r} \sum_{j=1}^{r} \mathbb{E}_\nu \left[\inf_{h \in \mathcal{S}_{\boldsymbol{\Phi}}} \|h - Y_{\ell_0,j}\|_{L^2}^2\right] \overset{⑦}{\geq} \max\{0, 1 - m/r\}.$$

Inequality ⑥ averages the fixed orthonormal targets, and ⑦ combines ④, $s \leq m$, and nonnegativity. An upper bound ε forces $m \geq (1-\varepsilon)r$, proving (176). Some fixed basis target attains at least this average expected error. Its index depends only on the feature law, not on the realized features.

Finally, Lemma E.4 gives $N(d, \ell_0) \sim d^{\ell_0}/\ell_0!$ for fixed ℓ_0, proving the stated order. □

To relate this proof to Kamath et al. (2020, Theorem 19), we note that the Gram matrix of the normalized targets $Y_{\ell_0,1}, \ldots, Y_{\ell_0,r}$ is $\mathbf{I}_r$, with smallest eigenvalue 1. We take that theorem's eigenvalue threshold to be 1 and its error parameter to be $\varepsilon/2$, since its loss is half the squared error. The condition $\varepsilon < 1$ makes this choice admissible and yields (176).

Proposition H.2 gives a worst-case necessary width without requiring independent feature coordinates or rotational invariance. Its common-law hypothesis is essential: an unrestricted target-dependent map $\boldsymbol{\Phi}_f(\mathbf{x}) = [f(\mathbf{x})] \in \mathbb{R}^1$ represents each target exactly with one coordinate. A representation learned from responses need not have a common law across targets, even if it is subsequently frozen. The proposition therefore does not apply directly to our learned attention. Proposition H.1 supplies the separate sample requirement under fixed positive noise variance and applies to every estimator. Neither result makes the logarithmic overheads in our sufficient guarantee necessary.

H.3 A Finite-Sample Width Obstruction for Fixed Features

Proposition H.2 requires a uniform guarantee over the target class. We next use rotational invariance to obtain an obstruction for every fixed target under the stated random-feature law.

A neuron count need not equal the number of scalar features. The neural-tangent model of Ghorbani et al. (2021, Theorem 2(a)) uses derivatives of the network at initialization as fixed features and fits md coefficients for m neurons. For a deterministic unit-norm degree-ℓ_0 polynomial orthogonal to all lower degrees, with $\ell_0 \geq 2$, its minimum regression risk tends to one in probability when $m/d^{\ell_0 - 1} \to 0$ as $d \to \infty$. This conclusion uses their Assumptions 2(a)–(b) at level $\ell_0 - 2$, uniform spherical inputs, and independent uniform initialization directions. As a result, the relevant scalar-feature count remains below order d^{ℓ_0} even though the neuron threshold differs. Neither this fixed-linearization bound nor the target-independent-map bound applies directly to our response-dependent attention features.

Let $N_{\mathrm{RF}} \geq 1$ be a deterministic integer, let $\vec{\mathbf{q}}_1^{\,\mathrm{RF}}, \ldots, \vec{\mathbf{q}}_{N_{\mathrm{RF}}}^{\,\mathrm{RF}}$ be independent uniform directions, and define the random prediction space $\mathcal{S}_{\mathrm{RF}} = \operatorname{span}\left\{\mathbf{x} \mapsto \sigma\big(\sqrt{d}\,(\vec{\mathbf{q}}_j^{\,\mathrm{RF}})^\top \mathbf{x} + 1/2\big) : j \in [N_{\mathrm{RF}}]\right\}$. The directions are independent of the deterministic target. The following bound recovers the fixed-feature obstruction of Ghorbani et al. (2021). It also applies to our paired baseline and requires no asymptotic assumption on the activation coefficients.

Proposition H.3. Let $d \geq 3$, $\ell \geq 1$, and $f \in \mathcal{H}_\ell$ be deterministic with $\|f\|_{L^2} = 1$. Then

$$\mathbb{E}\left[\inf_{h \in \mathcal{S}_{\mathrm{RF}}} \|h - f\|_{L^2}^2\right] \geq \max\left\{0, 1 - \frac{N_{\mathrm{RF}}}{N(d,\ell)}\right\}. \tag{177}$$

For $0 < \varepsilon < 1$,

$$\Pr\left[\inf_{h \in \mathcal{S}_{\mathrm{RF}}} \|h - f\|_{L^2}^2 \le 1 - \varepsilon\right] \le \min\left\{1, \frac{N_{\mathrm{RF}}}{\varepsilon N(d,\ell)}\right\}. \tag{178}$$

If $\mathcal{S}_{\mathrm{RF}}$ is instead the initial feature span of our network, both bounds hold with $m/4$ in place of N_{RF}, where m is its channel count. All expectations and probabilities in this proposition are over the directions only.

Proof. **Outline.** Step 1 projects the approximation problem onto the target harmonic space. Step 2 uses rotation invariance and the addition formula to compute the expected coefficient-space projection. Step 3 obtains the expectation and probability bounds and verifies the improvement for opposite pairs.

Step 1. Projection onto one degree. Projecting onto the target degree removes irrelevant components of a candidate fit and reduces the obstruction to a Euclidean dimension calculation. Recall the orthonormal basis $Y_{\ell,1}, \dots, Y_{\ell,N(d,\ell)}$ of $\mathcal{H}_\ell$. In this proof, we put $r = N(d,\ell)$ and define the column vector $\mathbf{Y}_\ell(\mathbf{x}) := [Y_{\ell,1}(\mathbf{x}), \dots, Y_{\ell,r}(\mathbf{x})]^\top \in \mathbb{R}^r$. We write $f = \boldsymbol{\beta}^\top \mathbf{Y}_\ell$ with $\|\boldsymbol{\beta}\|_2 = 1$. The orthogonal projection onto $\mathcal{H}_\ell$ is denoted by Π_ℓ. By the Funk–Hecke formula (98), $\Pi_\ell \phi_{\mathbf{q}} \overset{①}{=} \xi_{\ell,d} \mathbf{Y}_\ell(\mathbf{q})^\top \mathbf{Y}_\ell$. Equality ① computes each harmonic coefficient by the Funk–Hecke formula. Let V be the subspace of $\mathbb{R}^r$ spanned by $\mathbf{Y}_\ell(\overrightarrow{\mathbf{q}}_j^{\mathrm{RF}})$, and let $\mathbf{P}_V$ be its Euclidean orthogonal projection. The coefficients of $\Pi_\ell h$ lie in V for every $h \in \mathcal{S}_{\mathrm{RF}}$, including when $\xi_{\ell,d} = 0$. Orthogonal projection cannot increase a norm, so

$$\inf_{h \in \mathcal{S}_{\mathrm{RF}}} \|h - f\|_{L^2}^2 \overset{②}{\ge} \inf_{\mathbf{v} \in V} \|\mathbf{v} - \boldsymbol{\beta}\|_2^2 \overset{③}{=} 1 - \boldsymbol{\beta}^\top \mathbf{P}_V \boldsymbol{\beta}.$$

Inequality ② is projection contraction and coefficient inclusion, and ③ is Pythagoras with $\|\boldsymbol{\beta}\|_2 = 1$. Step 2 of Proposition H.2 gives measurability of the minimum. For the projector, put $\mathbf{A} = \sum_j \mathbf{Y}_\ell(\overrightarrow{\mathbf{q}}_j^{\mathrm{RF}}) \mathbf{Y}_\ell(\overrightarrow{\mathbf{q}}_j^{\mathrm{RF}})^\top$. Spectral decomposition gives $\mathbf{P}_V = \lim_{k \to \infty} \mathbf{A}(\mathbf{A} + k^{-1}\mathbf{I}_r)^{-1}$, a limit of measurable matrices with norm at most one.

Step 2. Expected projection. Rotation invariance makes the expected captured energy identical in every direction of the target harmonic space. Let $\mathbf{R}$ be an orthogonal transformation of $\mathbb{R}^d$. Rotation preserves $\mathcal{H}_\ell$ and its L^2 inner product, so an orthogonal matrix $\mathbf{O}_{\mathbf{R}}$ of order r satisfies $\mathbf{Y}_\ell(\mathbf{R}\mathbf{x}) = \mathbf{O}_{\mathbf{R}} \mathbf{Y}_\ell(\mathbf{x})$. Rotating all directions sends $\mathbf{P}_V$ to $\mathbf{O}_{\mathbf{R}} \mathbf{P}_V \mathbf{O}_{\mathbf{R}}^\top$. The joint law of the directions is unchanged. Consequently the kernel $J(\mathbf{x}, \mathbf{x}') = \mathbf{Y}_\ell(\mathbf{x})^\top \mathbb{E}[\mathbf{P}_V] \mathbf{Y}_\ell(\mathbf{x}')$ is invariant under simultaneous rotations of $\mathbf{x}, \mathbf{x}'$. For fixed $\mathbf{x}$, it is a degree-ℓ harmonic in $\mathbf{x}'$ invariant under every rotation fixing $\mathbf{x}$. Step 1 of Proposition E.3 makes it a scalar multiple of $P_\ell^{(d)}(\mathbf{x}^\top \mathbf{x}')$. The diagonal $J(\mathbf{x}, \mathbf{x})$ is constant by rotation invariance. Its integral is $\mathrm{tr}(\mathbb{E}[\mathbf{P}_V]) = \mathbb{E}[\dim V]$ by orthonormality. Since $\mu(\mathcal{X}) = 1$ and $P_\ell^{(d)}(1) = 1$, the addition formula gives

$$J(\mathbf{x}, \mathbf{x}') \overset{④}{=} \frac{\mathbb{E}[\dim V]}{r} \mathbf{Y}_\ell(\mathbf{x})^\top \mathbf{Y}_\ell(\mathbf{x}'), \qquad \mathbb{E}[\mathbf{P}_V] \overset{⑤}{=} \frac{\mathbb{E}[\dim V]}{r} \mathbf{I}_r.$$

Equality ④ is the addition formula with the computed diagonal. Integrating against each $Y_{\ell,j}(\mathbf{x}) Y_{\ell,k}(\mathbf{x}')$ gives ⑤, with all interchanges justified by the bounded projector and finite basis. In particular, $\mathbb{E}[\boldsymbol{\beta}^\top \mathbf{P}_V \boldsymbol{\beta}] \overset{⑥}{=} \mathbb{E}[\dim V]/r \overset{⑦}{\le} \min\{1, N_{\mathrm{RF}}/r\}$. Equality ⑥ uses ⑤ and $\|\boldsymbol{\beta}\|_2 = 1$, while inequality ⑦ uses $\dim V \le \min\{r, N_{\mathrm{RF}}\}$.

Step 3. Bounds and opposite pairs. The expected projection controls both mean error and the probability of capturing a fixed amount of target energy. Taking expectations in the projection bound from Step 1 and using the expected projection from Step 2 proves (177). Its left side also lies in $[0, 1]$ because the zero function is available. If the minimum regression risk is at most $1 - \varepsilon$, Step 1 forces $\boldsymbol{\beta}^\top \mathbf{P}_V \boldsymbol{\beta} \ge \varepsilon$. Markov's inequality gives $\Pr[\boldsymbol{\beta}^\top \mathbf{P}_V \boldsymbol{\beta} \ge \varepsilon] \overset{⑧}{\le} \mathbb{E}[\boldsymbol{\beta}^\top \mathbf{P}_V \boldsymbol{\beta}]/\varepsilon$. Inequality ⑧ and Step 2 prove (178).

For the initial span of our network, the joint direction law is again invariant under simultaneous rotations. Homogeneity gives $\mathbf{Y}_\ell(-\mathbf{q}) \overset{⑨}{=} (-1)^\ell \mathbf{Y}_\ell(\mathbf{q})$. Equality ⑨ evaluates a homogeneous polynomial of degree ℓ at $-\mathbf{q}$. Each opposite pair therefore adds at most one dimension to V, and duplication adds none. As a result, $\dim V \le m/4$. Steps 1 and 2 remain valid and give the stated

replacements. They use rotation invariance of the full collection, not independence between the two members of a pair. □

For a deterministic sequence $f_d \in \mathcal{H}_{\ell_0}$ of unit norm and fixed $\ell_0 \geq 1$, Lemma E.4 and (178) yield

$$N_{\mathrm{RF}} = o(d^{\ell_0}) \quad \Longrightarrow \quad \inf_{h \in \mathcal{S}_{\mathrm{RF}}} \|h - f_d\|_{L^2}^2 \longrightarrow 1 \quad \text{in probability.} \tag{179}$$

Equation (177) also gives expectation convergence in (179).

I Simulation Study

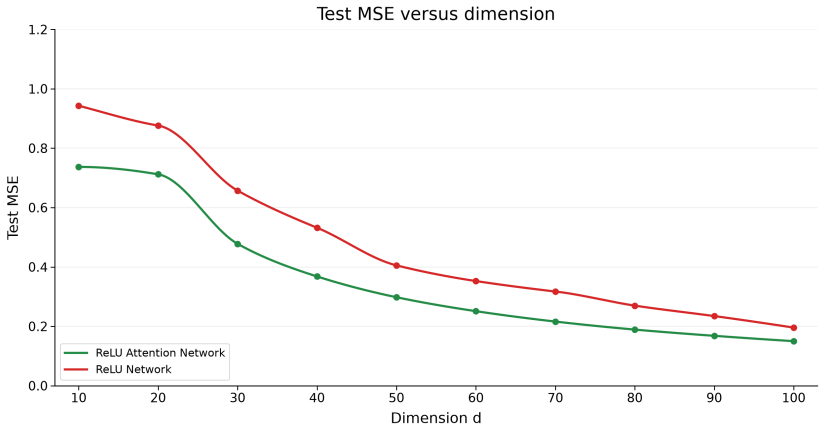


Figure 1: Illustration of the test loss by the ReLU attention network (5) and its non-attention version, ReLU network, with $d \in [10 : 100]$.

We demonstrate the simulation results by training the ReLU attention network (5) with Algorithm 1.

Target Function and Training Data. We use the cubic target $f^*(\mathbf{x}) = \sqrt{d(d+2)(d+4)}\, x_1 x_2 x_3$, where x_j is the jth coordinate. As a result, the true degree is $\ell_0 = 3$. This homogeneous harmonic polynomial has $B_*^2 = 1$ and $\Gamma_*^2 = N(d,3) = d(d-1)(d+4)/6$. The normalization follows from $\mathbb{E}\left[x_1^2 x_2^2 x_3^2\right] = 1/[d(d+2)(d+4)]$. We retain the resource schedule in (180) for the training data size n and network width m:

$$n(d) \coloneqq \lceil \alpha_{\mathrm{sim}} d^3 (\log d)^4 \rceil, \qquad m(d) \coloneqq 4 \left\lceil \frac{n}{4H_n^2} \right\rceil, \qquad H_n = \log(en), \tag{180}$$

which satisfies Theorem 4.1. For each dimension d in [10,100] with an increment of 10, we randomly sample $n(d)$ points $\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^{n(d)}$ as an i.i.d. sample of random variables distributed uniformly on the unit sphere $\mathbb{S}^{d-1}$. For each d, we also uniformly and independently sample 1000 points on the unit sphere $\mathbb{S}^{d-1}$ as the test data. We train the ReLU attention network (5) using GD by Algorithm 1 on an NVIDIA A100 GPU card with a learning rate specified by Algorithm 1, and report the test loss as test MSE in Figure 1. Figure 1 also illustrates the test loss of a ReLU network, whose architecture is the same as the ReLU attention network (5) without channel attention.

We report the average performance of 10 random initializations of the network parameters at each d in Figure 1, where the ReLU attention network and the ReLU network share the same random initial biased ReLU features at the same non-attention layer. It can be observed that the ReLU attention network always exhibits lower test loss than the ReLU network at each dimension, demonstrating the benefit of learnable channel attention.