FEATURE LEARNING, EFFECTIVE DIMENSION, AND ARCHITECTURE-SPECIFIC CRITICAL DEPTH

LIGHTMAN CHANG

ABSTRACT. The minimax rate for nonparametric regression of C^s functions on $\mathbb{R}^d$ is $n^{-2s/(2s+d)}$, which deteriorates rapidly with the ambient dimension. Practical neural networks trained on high-dimensional inputs nonetheless generalize at rates that do not exhibit such severe dimension dependence. We give a precise mechanistic account of this gap for the *single-index* target class $f^*(\mathbf{x}) = \rho(\langle \mathbf{v}^*, \mathbf{x}\rangle)$ under Gaussian inputs, working within the noise-propagation framework of [1]. Our principal contribution (Theorem 3.1) shows that, conditional on the alignment hypothesis established in [2], the test risk of two-layer ReLU networks transitions from the NTK-regime rate $\Theta(n^{-2s/(2s+d)})$ to $\Theta(n^{-2s/(2s+1)})$ in the feature-learning regime; the proof routes the NTK rate through spherical-harmonic multiplicities $N_l \asymp l^{d-2}$ and the feature-learning rate through the rank-one effective dimension induced by the collapsed equivalent kernel. We then introduce an architecture-specific critical depth L^* (Theorem 6.2) defined as the smallest depth at which the effective rank of the noise propagation operator falls below $n^{2s/(2s+1)}$, and we derive $L^*_{\mathrm{FC}} = \lfloor (d-1)/2 \rfloor + 1 = \lceil d/2 \rceil$ for fully-connected ReLU networks (with the per-layer exponent $\alpha_0 = (d{+}1)/(2(d{-}1))$ from spherical-harmonic NTK eigenvalues; the rate carries a depth-dependent prefactor $L^{L\alpha_0} = \exp(\Theta(d \log d))$ for $L = \Theta(d)$, tracked explicitly in Proposition 7.4) and a depth-decoupled formula for residual networks. The analogous result for transformer architectures is stated as a heuristic conjecture (Conjecture 9.1) pending a rigorous attention-head spectral computation. The cross-entropy extension via a Fisher-corrected operator $\mathbf{M}_{\mathrm{CE}}$ and the rank-$(K-1)$ effective rank (K = number of classes) at the Neural Collapse fixed point are treated in the companion paper [3].

Date: May 7, 2026.

2020 *Mathematics Subject Classification.* Primary 68T07; Secondary 62G05, 41A25, 60J60.

Key words and phrases. Feature learning, neural tangent kernel, effective dimension, curse of dimensionality, single-index models, critical depth, ResNet, Transformer.

1. Introduction

1.1. **The dimension dependence of the kernel rate.** A long-standing tension in the theory of overparameterized learning concerns the role of the ambient input dimension d. Classical nonparametric estimation gives a sharp lower bound: for the Hölder class $C^s(\mathbb{R}^d)$, no estimator can attain a rate better than $n^{-2s/(2s+d)}$, and kernel ridge regression with a smooth rotation-invariant kernel attains exactly this rate. The exponent is dominated by d for any fixed smoothness s, so that even modest inputs (e.g. $d = 100$) require an exponentially large number of samples for moderate accuracy. This is the *curse of dimensionality*.

Empirical practice presents a different picture. Networks trained on inputs of dimension d in the hundreds or thousands typically attain test errors that scale far better than $n^{-2s/(2s+d)}$. Two complementary explanations are usually offered. The first is structural: real targets are well approximated by low-complexity composites (single-index, multi-index, or sparse functions of low intrinsic dimension), so that the relevant rate is governed by an intrinsic dimension $k \ll d$. The second is algorithmic: gradient-based training does not produce an NTK-regime estimator; rather, it performs feature learning, and the post-training Jacobian concentrates spectral mass on directions that are aligned with the target structure.

The first explanation is essentially structural and rests on assumptions about the data-generating process. The second, however, makes a more delicate prediction: even when the input is genuinely d-dimensional, an overparameterized network can attain a rate that scales as if the input were one-dimensional, provided that feature learning succeeds at aligning the weights with the target direction. The present paper makes this prediction quantitative and unconditional, given the alignment guarantees already established in [2].

1.2. **Setting and prior work.** We work within the framework introduced in [1], which models the test mean squared error of an interpolating predictor through the *noise propagation operator* $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\,\mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma}$ is the diagonal matrix of singular values of the training Jacobian and $\mathbf{\Gamma}$ is the test–train cross-correlation matrix of gradient features. The trace $\mathrm{tr}(\mathbf{M})$ governs the noise contribution to test error, and the benign overfitting trichotomy ($\beta < \alpha + 1/2$ catastrophic, $\beta = \alpha + 1/2$ tempered, $\beta > \alpha + 1/2$ benign) follows from the power-law decay of singular values $\sigma_j \asymp j^{-\alpha}$ and gradient-feature norms $\varrho_j \asymp j^{-\beta}$. The companion paper [2] establishes the alignment hypothesis: in the

rich regime, gradient flow drives the weights of every active neuron of a two-layer ReLU network into the span of the teacher direction $\mathbf{v}^*$.

We refer to [1] for the unified generalization bound and the benign/tempered/catastrophic dichotomy, and to [2] for the alignment lemma and its proof via ReLU directional independence. The present paper does *not* reprove these results: we cite them as black-box inputs and develop their consequences.

The literature on dimension-free rates for neural networks is now substantial. Bach [6] gave a convex/variational analysis of two-layer networks and derived dimension-adaptive approximation rates. Bietti and Bruna [9] studied spherical harmonic decompositions of the NTK and gave matching upper and lower bounds on kernel ridge regression. Damian, Lee, and Soltanolkotabi [12] proved that gradient descent learns single-index targets via a feature-learning step that escapes the NTK regime, attaining the dimension-free rate $n^{-2s/(2s+1)}$. Ben Arous, Gheissari, and Jagannath [8] characterized the sample complexity of online SGD for single-index recovery in terms of the information exponent κ, identifying the $\Theta(d^\kappa)$ sample threshold for signal detection. The two results are complementary: the latter analyzes the signal-detection phase (number of samples needed to escape from the NTK regime), while the former analyzes the rate once detection has occurred. Rakhlin and Zhai [17] showed that interpolation by Laplace kernels is consistent only in high dimension, isolating the dimension dependence of benign overfitting in the NTK regime. Papyan, Han, and Donoho [16] introduced the Neural Collapse phenomenon, in which last-layer features and class means collapse onto a simplex equiangular tight frame at the terminal phase of training; the cross-entropy implications of Neural Collapse for our framework are treated in the companion paper [3].

1.3. **Contributions.** This paper makes two main contributions, both conditional on the alignment hypothesis of [2], plus a heuristic conjecture for transformers.

(i) **Feature-learning dimension collapse** (Theorem 3.1). For single-index targets with smoothness s, we show that the test risk transitions from the NTK-regime rate $\Theta(n^{-2s/(2s+d)})$ to the dimension-free rate $\Theta(n^{-2s/(2s+1)})$ when feature learning succeeds. The proof identifies the spherical-harmonic multiplicity $N_l \asymp l^{d-2}$ as the source of the curse of dimensionality in the NTK regime, and the multiplicity collapse to a one-dimensional kernel (at the level of the population equivalent kernel) as the

mechanism responsible for the dimension-free rate after alignment.

(ii) **Architecture-specific critical depth for FC and ResNet** (Theorem 6.2). We define L_A^* as the smallest depth at which the effective rank $r_{\text{eff}}(\mathbf{M}_A)$ falls below $n^{2s/(2s+1)}$. Using the per-layer singular-value exponent $\alpha_0 = (d+1)/(2(d-1))$ from Lemma 7.1, we derive $L_{\text{FC}}^* = \lfloor (d-1)/2 \rfloor + 1 = \lceil d/2 \rceil$ for fully-connected ReLU stacks (via depth-multiplicative spectral decay; the depth-dependent constant $L^{L\alpha_0}$ is tracked explicitly in Proposition 7.4), and $L_{\text{ResNet}}^* \in \{1, \infty\}$ decoupled from depth (via the identity-bypass structure of skip connections). For transformers, we record the analogous heuristic prediction as an explicit conjecture (Conjecture 9.1) and discuss what would be required to make it rigorous.

The cross-entropy extension via a Fisher-corrected operator $\mathbf{M}_{\text{CE}}$ and the rank-$(K-1)$ effective rank at the Neural Collapse fixed point are treated separately in the companion paper [3].

1.4. **Delta from prior work.** The kernel-regime rate $n^{-2s/(2s+d)}$ for kernel ridge regression in the NTK regime is folklore (see Caponnetto–De Vito [10] for the abstract minimax-rate theory and the references therein for spherical NTK derivations). The dimension-free single-index rate $n^{-2s/(2s+1)}$ is established for one-pass SGD by Damian–Lee–Soltanolkotabi [12] and for one-step gradient by Ba–Erdogdu–Suzuki–Wu–Zhang [7]. The new contribution of this paper is the systematic identification of the spherical-harmonic multiplicity $N(d,l) \asymp l^{d-2}$ as the precise source of the curse in the NTK regime, the rank-one effective dimension as the mechanism of the rate under alignment, and the depth-multiplicative chain that gives $L_{\text{FC}}^* = \Theta(d)$ for plain ReLU stacks. The ResNet decoupling and the transformer conjecture are, to our knowledge, not in the prior literature in this exact form.

The multi-index extension lies adjacent to the staircase-learning literature of Abbe–Adsera–Misiakiewicz [4] and Mei–Misiakiewicz–Montanari [15]; we discuss this in Section 10.

1.5. **Organization.** Section 2 recalls the noise propagation framework of [1] and the alignment hypothesis of [2]. Section 3 states Theorem 3.1, the feature-learning dimension collapse. Sections 4 and 5 prove the NTK-regime and feature-learning halves of Theorem 3.1 respectively. Section 6 states Theorem 6.2; Sections 7 and 8 prove the FC and ResNet cases. Section 9 states Conjecture 9.1 for transformers. Section 10 discusses limitations and open directions.

2. Preliminaries

2.1. **Recap of the noise propagation framework.** We follow the notation of [1]. Let $f(\cdot\,;\theta)\colon \mathbb{R}^d \to \mathbb{R}$ be a parametric predictor with parameters $\theta \in \mathbb{R}^p$, and let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a training set with $y_i = g^*(\mathbf{x}_i) + \xi_i$, ξ_i i.i.d. sub-Gaussian with variance proxy σ^2. Let θ^* be a minimum-norm interpolating parameter and let

$$\mathbf{J} \in \mathbb{R}^{n\times p}, \qquad \mathbf{J}_{i,:} = \nabla_\theta f(\mathbf{x}_i;\theta^*)^\top,$$

be the training Jacobian, with SVD $\mathbf{J} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. The j-th gradient feature is $\psi_j(\mathbf{x}) = \nabla_\theta f(\mathbf{x};\theta^*)^\top \mathbf{v}_j$, the cross-correlation matrix is $\mathbf{\Gamma}_{jl} = \mathbb{E}_{\mathbf{x}}[\psi_j(\mathbf{x})\psi_l(\mathbf{x})]$, and the noise propagation operator is

$$\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\,\mathbf{\Sigma}^{-1} \in \mathbb{R}^{n\times n}.$$

Theorem 1 of [1] states that for any $\delta \in (0,1)$, with probability $1-\delta$,

$$\text{(1)} \qquad \mathrm{MSE}_{\text{test}} \le C\big(B_{\text{signal}}^2 + \sigma^2\,\mathrm{tr}(\mathbf{M}) + \sigma^2\,\|\mathbf{M}\|_F\,\sqrt{\log(1/\delta)} + \delta_{\text{lin}}^2\big).$$

Theorem 2 of [1] states that under power-law decay $\sigma_j \asymp j^{-\alpha}$, $\varrho_j \asymp j^{-\beta}$, with $\varrho_j := \sqrt{\mathbf{\Gamma}_{jj}}$ (the diagonal feature-norm sequence),

$$\mathrm{tr}(\mathbf{M}) < \infty \iff \beta > \alpha + \tfrac{1}{2}.$$

Definition 2.1 (Effective rank)**.** For a positive semidefinite operator $\mathbf{A}$ with non-negative eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots$, the *effective rank* is

$$r_{\text{eff}}(\mathbf{A}) := \frac{(\sum_j \lambda_j)^2}{\sum_j \lambda_j^2} = \frac{(\mathrm{tr}\,\mathbf{A})^2}{\|\mathbf{A}\|_F^2}.$$

For $\mathbf{M}$, the effective rank quantifies the number of significant noise-propagation directions; benign overfitting requires $r_{\text{eff}}(\mathbf{M})$ to be at most polylogarithmic in n.

2.2. **Single-index targets and the alignment hypothesis.**

Definition 2.2 (Single-index target)**.** A function $g^*\colon \mathbb{R}^d \to \mathbb{R}$ is a *single-index target* with link ρ and direction $\mathbf{v}^*$ if there exist $\mathbf{v}^* \in \mathbb{S}^{d-1}$ and $\rho\colon \mathbb{R} \to \mathbb{R}$ such that

$$g^*(\mathbf{x}) = \rho(\langle \mathbf{v}^*, \mathbf{x}\rangle) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

We say $g^* \in \mathcal{H}^s$ if $\rho \in C^s(\mathbb{R})$ with bounded derivatives up to order s.

We work throughout under the standard Gaussian input distribution $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$, with link function ρ that is non-affine and has information exponent $\kappa = 1$ in the sense of [8]; that is, the first non-trivial Hermite coefficient of ρ is non-zero. This is the regime in which the alignment hypothesis of [2] applies.

Assumption 2.3 (Alignment, after [2])**.** Let $f_\theta(\mathbf{x}) = p^{-1/2}\sum_{j=1}^{p} a_j \operatorname{ReLU}(\mathbf{w}_j^\top \mathbf{x})$ be a two-layer ReLU student. Suppose the inputs are $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$ and the target is single-index $g^*(\mathbf{x}) = \rho(\langle \mathbf{v}^*, \mathbf{x}\rangle)$ with $\kappa = 1$. Suppose f_θ is trained by population gradient flow from a small-scale initialization. Then at any limit point θ^* of the gradient flow, the active neurons satisfy $\mathbf{w}_j \in \operatorname{span}(\mathbf{v}^*)$, and the dead neurons satisfy $\|\mathbf{w}_j\| = O(\alpha_{\text{init}})$.

This is exactly the conclusion of Theorem 5a–c of [2]; we use it as a black box. The point of Theorem 3.1 below is to quantify the consequence of Assumption 2.3 for the test risk.

2.3. **Equivalent kernels in the NTK and feature-learning regimes.**

Definition 2.4 (Equivalent kernel)**.** The *equivalent kernel* of an interpolating predictor at θ^* is the population kernel

$$K_{\theta^*}(\mathbf{x}, \mathbf{x}') := \mathbb{E}_{\theta^*}\left[\nabla_\theta f(\mathbf{x}; \theta^*)^\top \nabla_\theta f(\mathbf{x}'; \theta^*)\right],$$

where the expectation is taken over the randomness in the trained parameters (e.g. over initializations or training noise). When f_θ is in the NTK regime, $K_{\theta^*} \approx K_{\text{NTK}}$ is the rotation-invariant Neural Tangent Kernel of [14]. After feature learning, K_{θ^*} is no longer rotation-invariant.

Lemma 2.5 (Spherical harmonic decomposition of the NTK; [9])**.** *Let K_{NTK} be the NTK of a two-layer ReLU network with input dimension d and Gaussian inputs. Then K_{NTK} admits the spherical-harmonic eigendecomposition*

$$K_{\text{NTK}}(\mathbf{x}, \mathbf{x}') = \sum_{l=0}^{\infty} \lambda_l^{(d)} \sum_{m=1}^{N(d,l)} Y_{l,m}(\mathbf{x}) Y_{l,m}(\mathbf{x}'),$$

where $\{Y_{l,m}\}$ is the orthonormal basis of spherical harmonics of degree l on $\mathbb{S}^{d-1}$, $N(d,l) = \binom{d+l-1}{d-1} - \binom{d+l-3}{d-1}$ is the multiplicity, and there exist constants $c_l, C_l > 0$ such that, for $l \geq 1$,

$$c_l\, l^{-(d+1)} \leq \lambda_l^{(d)} \leq C_l\, l^{-(d+1)}.$$

The asymptotic multiplicity is $N(d,l) = \Theta(l^{d-2})$ as $l \to \infty$ with d fixed.

Remark 2.6. We will use Lemma 2.5 as a black box; the eigenvalue rate $l^{-(d+1)}$ and the multiplicity l^{d-2} are standard consequences of the homogeneity properties of the ReLU activation and the structure of the spherical Funk–Hecke formula. See [9], Theorem 3, for a complete proof.

3. Theorem 1: Feature Learning Dimension Collapse

Theorem 3.1 (Feature Learning Dimension Collapse)**.** *Let* $g^*(\mathbf{x}) = \rho(\langle \mathbf{v}^*, \mathbf{x}\rangle)$ *be a single-index target with* $\rho \in \mathcal{H}^s$ *for some* $s > 1/2$, $\mathbf{v}^* \in \mathbb{S}^{d-1}$, *and let* $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$. *Let* f_θ *be a two-layer ReLU network of width* p *trained on* $S \sim (g^* + \mathcal{N}(0, \sigma^2))^n$ *to a minimum-norm interpolant* θ^*.

(NTK regime). *If the network is trained in the NTK regime so that the equivalent kernel coincides with the population NTK in the sense that* $\mathbf{M} \approx \mathbf{M}_{\mathrm{NTK}}$ *to leading order, the optimally ridge-regularized test risk satisfies*

$$\mathbb{E}_S \, \mathrm{MSE}^{\mathrm{NTK}}_{\mathrm{test}}(n) = \Theta(n^{-2s/(2s+d)}). \tag{2}$$

(Feature-learning regime). *If the network is trained in the rich regime so that Assumption 2.3 holds at* θ^*, *the optimally ridge-regularized test risk satisfies*

$$\mathbb{E}_S \, \mathrm{MSE}^{\mathrm{FL}}_{\mathrm{test}}(n) = \Theta(n^{-2s/(2s+1)}). \tag{3}$$

The improvement factor is

$$\frac{\mathrm{MSE}^{\mathrm{NTK}}}{\mathrm{MSE}^{\mathrm{FL}}} = \Theta\big(n^{2s(d-1)/((2s+1)(2s+d))}\big),$$

which grows polynomially in n *for any fixed* $d > 1$.

Remark 3.2 (Status of the alignment hypothesis). The condition under which the feature-learning conclusion applies is exactly that of Assumption 2.3, which is in turn established by Theorem 5a–c of [2]. The present theorem is thus a conditional statement of the form "alignment ⇒ dimension-free rate". Removing the alignment hypothesis (e.g. proving the same rate without an alignment guarantee) is open; see Section 10.

The proof of Theorem 3.1 occupies the next two sections. Section 4 proves the NTK rate (2); Section 5 proves the feature-learning rate (3).

4. Proof of Theorem 3.1: NTK rate via spherical harmonic multiplicity

We prove (2) by reducing the test risk to that of optimally ridge-regularized regression with respect to the population NTK and applying classical kernel-rate arguments.

4.1. **Reduction to kernel ridge regression.** In the NTK regime, the linearization residual δ_{lin} in (1) is $o(1)$ and the gradient features $\psi_j(\mathbf{x}) = \nabla_\theta f(\mathbf{x}; \theta^*)^\top \mathbf{v}_j$ depend only on the kernel structure. Specifically, the columns of $\mathbf{V}$ diagonalize $\mathbf{J}^\top \mathbf{J}$, which converges in the wide limit to the Gram matrix of K_{NTK} on the training set. The unified bound (1) therefore reduces, up to $o(1)$ corrections, to the test risk of kernel ridge regression with kernel K_{NTK}.

4.2. **Source-condition formulation.** Let $\{(\lambda_{l,m}, e_{l,m})\}$ enumerate the eigenpairs of K_{NTK} in the spherical-harmonic basis of Lemma 2.5. The target $g^*(\mathbf{x}) = \rho(\langle \mathbf{v}^*, \mathbf{x} \rangle)$, viewed as a function on $\mathbb{R}^d$ with Gaussian measure, has Hermite expansion

$$\rho(\langle \mathbf{v}^*, \mathbf{x} \rangle) = \sum_{l=0}^{\infty} c_l \, H_l(\langle \mathbf{v}^*, \mathbf{x} \rangle),$$

where H_l is the l-th (probabilist's) Hermite polynomial. For a C^s link function ρ, the standard Hermite-decay result on the Gaussian line is $|c_l|^2 = O(l^{-2s})$ in the Hermite-normalized basis (this is the standard Sobolev source condition for the Ornstein–Uhlenbeck operator, see e.g. [5, Theorem 4.6]). In particular, $\sum_l c_l^2 \, l^{2s} < \infty$ for $\rho \in C^s(\mathbb{R})$, which is the source condition we use in the bias bound below.

Each Hermite mode $H_l(\langle \mathbf{v}^*, \mathbf{x} \rangle)$ is supported on the spherical-harmonic shell of degree l in the eigenbasis of K_{NTK}; indeed, $H_l(\langle \mathbf{v}^*, \mathbf{x} \rangle)$ is a degree-l polynomial that depends only on $\langle \mathbf{v}^*, \mathbf{x} \rangle$, hence is a degree-$l$ zonal harmonic with respect to $\mathbf{v}^*$. By rotational invariance of K_{NTK}, this zonal harmonic is a single eigenfunction of K_{NTK} with eigenvalue $\lambda_l^{(d)}$. Consequently, in the eigenbasis of K_{NTK}, the coefficients of g^* are supported on $N(d, l) \asymp l^{d-2}$ many shells of equal magnitude up to the zonal/non-zonal split.

4.3. **Bias–variance tradeoff and the optimal rate.** For ridge regression with regularization $\eta > 0$ and i.i.d. Gaussian inputs and noise, the standard bias–variance bound (see, e.g., [17, Section 2]) gives

$$\mathbb{E}_S \, \text{MSE}_{\text{test}} \lesssim \underbrace{\sum_{l,m} \frac{\eta^2 c_{l,m}^2}{(\lambda_l^{(d)} + \eta)^2}}_{\text{bias}} + \underbrace{\frac{\sigma^2}{n} \sum_{l,m} \frac{\lambda_l^{(d)}}{\lambda_l^{(d)} + \eta}}_{\text{variance}}, \tag{4}$$

where $c_{l,m}$ are the eigen-coefficients of g^*.

Bias term. By the source condition $\sum_l c_l^2\, l^{2s} < \infty$ and the upper bound $\lambda_l^{(d)} \le C l^{-(d+1)}$ from Lemma 2.5,

$$\sum_{l,m} \frac{\eta^2 c_{l,m}^2}{(\lambda_l^{(d)} + \eta)^2} \asymp \eta^{2s/(d+1)},$$

where the equivalence follows by partitioning the sum at the index l_η for which $\lambda_{l_\eta}^{(d)} = \eta$, namely $l_\eta \asymp \eta^{-1/(d+1)}$, and integrating both contributions.

Variance term. The effective dimension is

$$\mathrm{df}(\eta) := \sum_{l,m} \frac{\lambda_l^{(d)}}{\lambda_l^{(d)} + \eta} \asymp \sum_{l \le l_\eta} N(d,l) \asymp l_\eta^{d-1} \asymp \eta^{-(d-1)/(d+1)},$$

where we used $N(d,l) \asymp l^{d-2}$ from Lemma 2.5 and summed via $\sum_{l \le L} l^{d-2} \asymp L^{d-1}$. The variance term is therefore $\asymp \sigma^2\, \eta^{-(d-1)/(d+1)}/n$.

Combining. Setting bias and variance equal, $\eta^{2s/(d+1)} \asymp \sigma^2\, \eta^{-(d-1)/(d+1)}/n$, yields $\eta^* \asymp n^{-(d+1)/(2s+d)}$ and

$$\mathbb{E}_S\, \mathrm{MSE}_{\mathrm{test}}^{\mathrm{NTK}}(n) \asymp (\eta^*)^{2s/(d+1)} \asymp n^{-2s/(2s+d)}.$$

The matching lower bound follows from the standard nonparametric minimax lower bound on $\mathcal{H}^s(\mathbb{S}^{d-1})$, since the NTK at this optimally tuned ridge attains the minimax rate; see [17]. This proves (2). □

4.4. **Where the dimension dependence enters.** The exponent d in $n^{-2s/(2s+d)}$ comes from two compounding sources in the variance term: the eigenvalue rate $l^{-(d+1)}$ and the multiplicity $N(d,l) = \Theta(l^{d-2})$. The product $\lambda_l^{(d)} N(d,l) \asymp l^{-3}$ (after summation) is dimension-independent, but the threshold l_η scales as $\eta^{-1/(d+1)}$ and is summed against $N(d,l)$, producing $\mathrm{df}(\eta) \asymp \eta^{-(d-1)/(d+1)}$. The exponent $(d-1)/(d+1)$ is the source of the curse: when d is large, the effective dimension grows almost linearly with η^{-1}, and the variance dominates.

This identifies the precise mechanism that feature learning must defeat: to escape the curse, the equivalent kernel must lose the multiplicity factor l^{d-2}. The next section shows how alignment achieves this.

5. Proof of Theorem 3.1: Feature-learning rate via rank-one effective dimension

We now prove (3). The argument proceeds in three steps: (i) characterize the equivalent kernel under alignment as a one-dimensional ridge kernel along $\mathbf{v}^*$; (ii) compute the resulting eigenvalue sequence and effective dimension; (iii) optimize the bias–variance tradeoff.

5.1. **Equivalent kernel under alignment.**

Lemma 5.1 (Population kernel collapse)**.** *Suppose Assumption 2.3 holds at θ^*, and let $\bar{K}(t,t') := K^{(1)}_{\mathrm{NTK}}(t,t')$ denote the one-dimensional ReLU NTK on $\mathbb{R}$. Define the* population equivalent kernel $\bar{K}^{\mathrm{pop}}_{\theta^*}(\mathbf{x},\mathbf{x}')$ *by integrating out the component orthogonal to $\mathbf{v}^*$:*

$$\bar{K}^{\mathrm{pop}}_{\theta^*}(\mathbf{x},\mathbf{x}') := \mathbb{E}_{\mathbf{x}_\perp,\mathbf{x}'_\perp}\big[\, K_{\theta^*}(\mathbf{x},\mathbf{x}')\big],$$

where $\mathbf{x} = \langle \mathbf{v}^*, \mathbf{x}\rangle\, \mathbf{v}^* + \mathbf{x}_\perp$ and likewise for $\mathbf{x}'$, with $\mathbf{x}_\perp, \mathbf{x}'_\perp$ drawn from the Gaussian marginal on the orthogonal complement of $\mathbf{v}^*$. Then

$$\bar{K}^{\mathrm{pop}}_{\theta^*}(\mathbf{x},\mathbf{x}') = \bar{K}\big(\langle \mathbf{v}^*, \mathbf{x}\rangle, \langle \mathbf{v}^*, \mathbf{x}'\rangle\big) + c_0 + O(\alpha_{\mathrm{init}}^2),$$

where c_0 is a constant that does not depend on $\mathbf{x}, \mathbf{x}'$ (and is absorbed into the constant "bias" direction of the equivalent kernel).

Proof. Under Assumption 2.3, the active neurons satisfy $\mathbf{w}_j = w_j \mathbf{v}^*$ for some scalar $w_j \in \mathbb{R}$, and dead neurons have norm $O(\alpha_{\mathrm{init}})$. Write $t := \langle \mathbf{v}^*, \mathbf{x}\rangle$, $t' := \langle \mathbf{v}^*, \mathbf{x}'\rangle$ for the projected coordinates, and decompose $\mathbf{x} = t\,\mathbf{v}^* + \mathbf{x}_\perp$, $\mathbf{x}' = t'\,\mathbf{v}^* + \mathbf{x}'_\perp$ with $\mathbf{x}_\perp, \mathbf{x}'_\perp$ Gaussian on the $(d-1)$-dimensional orthogonal complement of $\mathbf{v}^*$.

The gradient of the predictor at $\mathbf{x}$ has two pieces

$$\nabla_{a_j} f_{\theta^*}(\mathbf{x}) = p^{-1/2}\,\mathrm{ReLU}(w_j t),$$
$$\nabla_{\mathbf{w}_j} f_{\theta^*}(\mathbf{x}) = p^{-1/2} a_j \mathbf{1}\{w_j t > 0\}\,\mathbf{x},$$

for active j (both $O(\alpha_{\mathrm{init}})$ for dead j). Hence the pointwise kernel decomposes as

$$K_{\theta^*}(\mathbf{x},\mathbf{x}') = A(t,t') + B(t,t')\,\mathbf{x}^\top \mathbf{x}',$$

where A, B are deterministic functions of the projected coordinates arising from summing over j. Decomposing $\mathbf{x}^\top\mathbf{x}' = t\,t' + \mathbf{x}_\perp^\top \mathbf{x}'_\perp$, the population equivalent kernel is

$$\bar{K}^{\mathrm{pop}}_{\theta^*}(\mathbf{x},\mathbf{x}') = \mathbb{E}_{\mathbf{x}_\perp,\mathbf{x}'_\perp}\big[K_{\theta^*}(\mathbf{x},\mathbf{x}')\big] = A(t,t') + B(t,t')\,t\,t' + B(t,t')\,\mathbb{E}[\mathbf{x}_\perp^\top \mathbf{x}'_\perp].$$

Since $\mathbf{x}_\perp, \mathbf{x}'_\perp$ are independent mean-zero Gaussian, $\mathbb{E}[\mathbf{x}_\perp^\top \mathbf{x}'_\perp] = 0$, and the perpendicular term *vanishes* in expectation. Hence

$$\bar{K}^{\mathrm{pop}}_{\theta^*}(\mathbf{x},\mathbf{x}') = A(t,t') + B(t,t')\,t\,t' =: \tilde{K}(t,t'),$$

which depends on $\mathbf{x}, \mathbf{x}'$ only through (t,t'). The function $\tilde{K}(t,t')$ is precisely the one-dimensional ReLU NTK $\bar{K}(t,t')$ on the projected inputs, up to an additive constant c_0 contributed by neurons whose ReLU-gate is identically activated (formally, $c_0 = \mathbb{E}_{w_j,a_j}[\,\cdot\,]$ aggregated across j, modifying $\bar{K}$ only on its bias direction). The dead-neuron contribution contributes a residual of squared norm $O(\alpha_{\mathrm{init}}^2)$. □

Remark 5.2 (Pointwise vs. population kernel). The collapse to the one-dimensional kernel holds at the level of the *population* kernel (averaged over the perpendicular components of the inputs), not the pointwise kernel: the term $\mathbf{x}_\perp^\top \mathbf{x}'_\perp$ is a random quantity, not a constant, and the perpendicular components contribute fluctuations of order $O(\sqrt{d-1})$ at any single $(\mathbf{x}, \mathbf{x}')$. The bias–variance analysis in Section 5 works at the population level (i.e. uses $\bar{K}_{\theta^*}^{\mathrm{pop}}$ in the integrated risk), so the population statement is the relevant one. A pointwise version would require additional concentration of $\mathbf{x}_\perp^\top \mathbf{x}'_\perp$ and would only contribute $O((d-1)/p)$ corrections in the wide limit; we do not pursue it here.

5.2. **Eigenvalue sequence of the one-dimensional ReLU NTK.**

Lemma 5.3 (1D ReLU NTK eigenvalues)**.** *Let $\bar{K}$ denote the 1D ReLU NTK on $\mathbb{R}$ with Gaussian measure $\mathcal{N}(0,1)$. Then $\bar{K}$ has an eigenbasis given by the Hermite polynomials $\{H_l\}_{l\geq 0}$, with eigenvalues $\bar{\lambda}_l = \Theta(l^{-3})$ for $l \geq 1$.*

Proof sketch. The 1D ReLU NTK is rotation-invariant on $\mathbb{R}$ (i.e. depends only on $|t|, |t'|, tt'$). Its eigenfunctions in the Gaussian Hilbert space are the Hermite polynomials H_l, with eigenvalues given by the spectral decomposition computed e.g. in [9, Appendix A]. The asymptotic rate $\bar{\lambda}_l \asymp l^{-3}$ follows from the Hermite-coefficient calculation for ReLU. □

The crucial difference between the 1D and d-dimensional cases is the multiplicity: for the 1D kernel, each eigenvalue $\bar{\lambda}_l$ has multiplicity exactly 1, whereas in d dimensions the corresponding eigenvalue has multiplicity $N(d,l) \asymp l^{d-2}$.

5.3. **Effective dimension and the optimal rate.**

Proposition 5.4 (1D effective dimension)**.** *For the 1D ReLU NTK with eigenvalues $\bar{\lambda}_l \asymp l^{-3}$, the effective dimension at ridge $\eta > 0$ satisfies*

$$\mathrm{df}_{\mathrm{FL}}(\eta) := \sum_{l\geq 1} \frac{\bar{\lambda}_l}{\bar{\lambda}_l + \eta} \asymp \eta^{-1/3}.$$

Proof. The threshold l_η at which $\bar{\lambda}_{l_\eta} = \eta$ is $l_\eta \asymp \eta^{-1/3}$. The effective dimension is $\sum_{l\leq l_\eta} 1 + \sum_{l>l_\eta} \bar{\lambda}_l/\eta$, which equals $l_\eta + O(l_\eta)$ by direct integration, hence $\asymp \eta^{-1/3}$. □

By Lemma 5.1, the population equivalent kernel of f_{θ^*} in the feature-learning regime coincides with $\bar{K}$ up to a constant and $O(\alpha_{\mathrm{init}}^2)$, both of which are negligible for the leading-order rate (the constant contributes

only to the bias direction; cf. Remark 5.2). Substituting Proposition 5.4 into the bias–variance decomposition (4) but with the 1D eigenvalues $\bar{\lambda}_l$ in place of $\lambda_l^{(d)}$:

- the bias term is $\Theta(\eta^{2s/3})$. To see this, fix the threshold l_η at which $\bar{\lambda}_{l_\eta} = \eta$, namely $l_\eta \asymp \eta^{-1/3}$ (cf. Proposition 5.4). Splitting the bias sum at l_η,

$$\sum_{l\geq 1} \frac{\eta^2\, c_l^2}{(\bar{\lambda}_l + \eta)^2} \asymp \sum_{l\leq l_\eta} c_l^2 \;+\; \eta^2 \sum_{l>l_\eta} \frac{c_l^2}{\bar{\lambda}_l^2}.$$

  Under the source condition $\sum_l c_l^2 l^{2s} < \infty$, the second sum is bounded by $\eta^2\, l_\eta^{6-2s} \asymp \eta^2\, \eta^{-(6-2s)/3} = \eta^{2s/3}$ (for $s < 3$, with the obvious adaptation for higher s), and the first sum is $O(l_\eta^{-2s}) = O(\eta^{2s/3})$. Hence the bias is $\Theta(\eta^{2s/3})$;
- the variance term is $\sigma^2\eta^{-1/3}/n$ by Proposition 5.4.

Equating bias and variance, $\eta^{2s/3} \asymp \sigma^2\eta^{-1/3}/n$, gives $\eta^* \asymp n^{-3/(2s+1)}$ and

$$\mathbb{E}_S\, \mathrm{MSE}_{\mathrm{test}}^{\mathrm{FL}}(n) \asymp (\eta^*)^{2s/3} \asymp n^{-2s/(2s+1)}.$$

This proves (3). □

5.4. **Comparison with the kernel rate.** The exponent improvement is

$$\frac{2s}{2s+1} - \frac{2s}{2s+d} = \frac{2s(d-1)}{(2s+1)(2s+d)},$$

which is positive for any $d > 1$ and any $s > 0$, and approaches 1 as $d \to \infty$ for fixed s. The improvement factor in test risk is

$$\frac{\mathrm{MSE}^{\mathrm{NTK}}}{\mathrm{MSE}^{\mathrm{FL}}} = n^{2s/(2s+1) - 2s/(2s+d)} = n^{2s(d-1)/((2s+1)(2s+d))},$$

which grows polynomially in n.

Remark 5.5. The rate $n^{-2s/(2s+1)}$ is exactly the minimax rate for non-parametric regression of a C^s function on $\mathbb{R}$. Theorem 3.1 therefore states that, in the feature-learning regime, the two-layer ReLU network attains the minimax rate of the *intrinsic* one-dimensional problem, not the ambient d-dimensional problem. The role of the network is to learn the hidden direction $\mathbf{v}^*$, after which the estimation problem becomes one-dimensional.

6. Theorem 2: Architecture-Specific Critical Depth

We now turn to the depth dependence of the feature-learning rate. The key observation is that, even when the alignment hypothesis holds at the final layer, the depth of the network controls the rate of spectral decay α that can be attained by composition. We define the critical depth as the smallest L for which the network can attain the benign overfitting condition $\beta > \alpha + 1/2$ for the *single-index* target.

Definition 6.1 (Architecture, critical depth)**.** Let A denote an architecture family parametrized by depth L and width (taken to infinity). Fix smoothness $s = 1$ throughout this definition (matching the regularity at which Theorem 6.2 applies). For each L, let $\mathbf{M}_{A,L}$ denote the noise propagation operator of the trained network on a single-index target, with effective rank $r_{\text{eff}}(\mathbf{M}_{A,L})$ as in Definition 2.1. The *critical depth* for A is

$$L_A^*(n,d) := \min\{L \in \mathbb{N} : r_{\text{eff}}(\mathbf{M}_{A,L}) \le n^{2s/(2s+1)}\big|_{s=1} = n^{2/3}\},$$

or ∞ if no such L exists. Equivalently, L_A^* is the smallest depth at which the test risk attains the dimension-free rate $n^{-2/3}$ in the trichotomy of [1].

Theorem 6.2 (Architecture-Specific Critical Depth)**.** *Under the alignment hypothesis (Assumption 2.3) and Gaussian inputs $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$, the critical depths for the single-index target with $s = 1$ satisfy:*

(FC ReLU stacks). *For fully-connected ReLU networks with depth L and width going to infinity,*

$$L_{\text{FC}}^*(n,d) = \lfloor (d-1)/2 \rfloor + 1 = \lceil d/2 \rceil. \tag{5}$$

The implied rate carries a depth-dependent prefactor of $L^{L\alpha_0} = \exp(\Theta(d \log d))$ for $L = \Theta(d)$, tracked explicitly in Proposition 7.4 and Remark 7.5.

(ResNet, identity skip). *For residual networks $h_{\ell+1} = h_\ell + g_\ell(h_\ell)$ with each g_ℓ a two-layer ReLU block,*

$$L_{\text{ResNet}}^*(n,d) \in \{1, \infty\}, \tag{6}$$

where the value depends only on whether the per-block exponents (α_0, β_0) already satisfy $\beta_0 > \alpha_0 + 1/2$, independent *of L.*

The Transformer case is treated separately as a heuristic conjecture (Conjecture 9.1 in Section 9), since a complete proof requires an explicit attention-head spectral computation that we do not establish here.

The proof of Theorem 6.2 is split into two sections: Section 7 treats the FC case via a depth-multiplicative spectral decay argument, and

Section 8 treats the ResNet case via the identity-bypass structure. Section 9 states the analogous heuristic for transformer architectures as a conjecture and discusses what would be required to make it rigorous.

7. Proof of Theorem 6.2: FC ReLU case

7.1. **Per-layer exponents and depth-multiplicative bound.** We first derive the per-layer singular-value exponent α_0 from the spherical-harmonic NTK eigenvalues.

Lemma 7.1 (Per-layer singular-value exponent)**.** *For a single-layer ReLU computation on Gaussian inputs of dimension d, let σ_j denote the singular values of the per-sample gradient features ordered by magnitude, and let j be the corresponding flattened index across all spherical-harmonic shells. Then*

$$\sigma_j \asymp j^{-\alpha_0}, \qquad \alpha_0 = \frac{d+1}{2(d-1)}.$$

Proof. By Lemma 2.5, the NTK eigenvalues on the l-th spherical-harmonic shell satisfy $\lambda_l^{(d)} \asymp l^{-(d+1)}$, and the multiplicity of shell l is $N(d,l) \asymp l^{d-2}$. The flattened index is $j(l) \asymp \sum_{l' \le l} N(d,l') \asymp l^{d-1}$, so $l \asymp j^{1/(d-1)}$. The eigenvalue at flattened index j is therefore $\lambda_j = \lambda_{l(j)}^{(d)} \asymp j^{-(d+1)/(d-1)}$, and the corresponding singular value is $\sigma_j = \sqrt{\lambda_j} \asymp j^{-(d+1)/(2(d-1))}$, giving the claimed α_0. □

Assumption 7.2 (One-step rich-regime gain; labeled `(B0)`)**.** For a single-index target with $s = 1$ and Gaussian inputs, after one gradient step in the rich regime starting from a small-scale random initialization, the per-layer cross-correlation diagonal $\varrho_j := \sqrt{\mathbf{\Gamma}_{jj}}$ satisfies the power-law decay

$$\varrho_j \asymp j^{-\beta_0}, \qquad \beta_0 = \alpha_0 + \frac{1}{d-1} = \frac{d+3}{2(d-1)},$$

i.e. a $+1/(d-1)$ exponent gain over α_0 in the flattened-shell index.

Remark 7.3 (Status of Assumption 7.2). We treat this as an explicit assumption rather than a fully proved lemma. The supporting evidence is twofold. *(i) Spiked-RMT analysis.* The one-step gradient feature analysis of Ba–Erdogdu–Suzuki–Wu–Zhang [7] shows that, at information exponent $\kappa = 1$, the post-step feature kernel exhibits a rank-one spiked deformation of the kernel-regime spectrum, with the spike concentrated along $\mathbf{v}^*$. *(ii) DLS spike strength.* Damian–Lee–Soltanolkotabi [12] establish that the spike contributes a multiplicative factor of order $1/(d-1)$ to the rate exponent in the flattened-shell

index for $\kappa = 1$. Combining (i) and (ii) is consistent with the formula $\beta_0 = \alpha_0 + 1/(d-1)$, but a direct derivation in the precise singular-value-on-Jacobian framework of [1] is not given in either reference; we therefore record (B0) as an assumption and use it as a black-box input to the rest of this section. A first-principles derivation via the spherical-harmonic decomposition of the post-one-step Jacobian is left to future work.

The single-layer rates do not by themselves satisfy $\beta_0 > \alpha_0 + 1/2$:

$$\beta_0 - \alpha_0 = \frac{1}{d-1} < \frac{1}{2} \quad \text{for } d \geq 3.$$

Composition is therefore necessary to reach the benign regime.

Proposition 7.4 (Depth multiplicative upper bound)**.** *Let $\mathbf{J}_{\text{total}}$ be the end-to-end Jacobian of an L-layer FC ReLU stack at a trained interpolant satisfying Assumption 2.3 at the final layer. Then there exists a depth-dependent constant $C_L > 0$ such that*

$$\sigma_j(\mathbf{J}_{\text{total}}) \leq C_L\, j^{-L\alpha_0}, \qquad C_L = C_0^L \cdot L^{L\alpha_0},$$

where C_0 is the per-layer constant from Lemma 7.1.

Proof. The end-to-end Jacobian factors as $\mathbf{J}_{\text{total}} = \mathbf{J}_L \mathbf{J}_{L-1} \cdots \mathbf{J}_1$ where each $\mathbf{J}_\ell$ is a per-layer Jacobian. By Lemma 7.1, each $\mathbf{J}_\ell$ has spectral decay

$$\sigma_j(\mathbf{J}_\ell) \leq C_0\, j^{-\alpha_0}, \qquad \alpha_0 = \frac{d+1}{2(d-1)}, \tag{7}$$

on the spherical-harmonic basis (the dependence on ℓ in C_0 is suppressed since intermediate layers receive Gaussian-like inputs by the NTK propagation; we take $C_0 = \max_\ell C_\ell$ uniformly).

Apply Weyl's product inequality [13, Theorem 3.3.14]: for any matrices $A_1, \ldots, A_L$ and indices $j_1, \ldots, j_L \geq 1$,

$$\sigma_{j_1 + \cdots + j_L - L + 1}(A_L \cdots A_1) \leq \prod_{\ell=1}^{L} \sigma_{j_\ell}(A_\ell).$$

Setting all $j_\ell = j$, we obtain

$$\sigma_{Lj-L+1}(\mathbf{J}_{\text{total}}) \leq \prod_{\ell=1}^{L} \sigma_j(\mathbf{J}_\ell) \leq C_0^L\, j^{-L\alpha_0}.$$

Re-index by $j' = Lj - L + 1$, equivalently $j = (j' + L - 1)/L \leq j'$, and note that $j \geq j'/L$ for $j' \geq L$. Hence

$$\sigma_{j'}(\mathbf{J}_{\text{total}}) \leq C_0^L\, (j'/L)^{-L\alpha_0} = C_0^L\, L^{L\alpha_0}\, (j')^{-L\alpha_0}.$$

Setting $C_L = C_0^L L^{L\alpha_0}$ gives the stated bound. The $L^{L\alpha_0}$ factor is depth-dependent and explicitly tracked; it does not affect the leading $j^{-L\alpha_0}$ rate but is a factor in the constant. □

Remark 7.5 (Tracking the $L^{L\alpha_0}$ constant). The depth-dependent constant $C_L = C_0^L L^{L\alpha_0}$ enters the bias and variance constants in (4) but does not alter the j-exponent $L\alpha_0$. For the qualitative critical-depth result $L^*_{\mathrm{FC}}(n,d) = \lceil d/2 \rceil$, what matters is the exponent comparison $\beta > \alpha + 1/2$, which is unaffected by C_L. For sample-complexity bounds with explicit constants, one must absorb C_L into the $\Theta(\cdot)$ in the rate; we do not pursue this quantification here.

Remark 7.6 (Limit of the multiplicative inequality). Proposition 7.4 provides only an upper bound on the spectral decay of $\mathbf{J}_{\text{total}}$. Beyond the leading $j^{-L\alpha_0}$ rate, no *equality* $\sigma_j(\mathbf{J}_{\text{total}}) = \Theta(\sigma_j(\mathbf{J}_L) \cdots \sigma_j(\mathbf{J}_1))$ holds in general: such a per-index multiplicative identity would require the right singular vectors of $\mathbf{J}_{\ell+1}$ to coincide with the left singular vectors of $\mathbf{J}_\ell$ across every ℓ, which is a measure-zero condition. A simple 2×2 example illustrates this gap. Let

$$\mathbf{J}_1 = \mathrm{diag}(1, \epsilon), \qquad \mathbf{J}_2 = \mathrm{diag}(\epsilon, 1).$$

Both factors have singular values $\{1, \epsilon\}$, so $\sigma_1(\mathbf{J}_\ell)\sigma_2(\mathbf{J}_\ell) = \epsilon$ for each ℓ, and the naive per-index multiplicative *equality* would predict $\sigma_2(\mathbf{J}_2\mathbf{J}_1) = \sigma_2(\mathbf{J}_2)\sigma_2(\mathbf{J}_1) = \epsilon^2$. Direct computation gives

$$\mathbf{J}_2\mathbf{J}_1 = \mathrm{diag}(\epsilon, \epsilon) = \epsilon\,\mathbf{I},$$

so $\sigma_2(\mathbf{J}_2\mathbf{J}_1) = \epsilon$, not ϵ^2. This is consistent with Weyl's product *inequality*, which only asserts $\sigma_2(\mathbf{J}_2\mathbf{J}_1) \le \sigma_1(\mathbf{J}_2)\sigma_2(\mathbf{J}_1) = 1 \cdot \epsilon = \epsilon$, and which is saturated here. The mechanism is that $\mathbf{J}_2$'s *top* singular direction e_2 coincides with $\mathbf{J}_1$'s *bottom* singular direction e_2, so the small-singular-value direction of $\mathbf{J}_1$ is amplified by the top singular value of $\mathbf{J}_2$. This shows the multiplicative bound on σ_j can be loose by a factor as large as $\sigma_1(\mathbf{J}_{\ell+1})/\sigma_j(\mathbf{J}_{\ell+1})$ per layer when intermediate singular vectors misalign, and motivates our use of an upper-bound argument throughout: only the leading $j^{-L\alpha_0}$ rate is universal, while sharper per-index constants depend on the alignment of consecutive singular bases.

7.2. **The β_0 exponent and depth scaling.** A symmetric argument applied to the cross-correlation matrix $\mathbf{\Gamma}_{\text{total}}$ via Assumption 7.2 gives the per-depth bound $\varrho_j(\mathbf{\Gamma}_{\text{total}}) \le C_L j^{-L\beta_0}$ with $\beta_0 = (d+3)/(2(d-1))$ and C_L a depth-dependent constant analogous to Proposition 7.4.

Combined with Proposition 7.4 and Lemma 7.1, the per-depth exponents satisfy

$$\alpha = L\alpha_0 = \frac{L(d+1)}{2(d-1)}, \qquad \beta = L\beta_0 = \frac{L(d+3)}{2(d-1)}.$$

The benign overfitting condition $\beta > \alpha + 1/2$ becomes

$$\frac{L(d+3)}{2(d-1)} > \frac{L(d+1)}{2(d-1)} + \frac{1}{2},$$

which simplifies to

$$\frac{L \cdot 2}{2(d-1)} > \frac{1}{2} \iff \frac{L}{d-1} > \frac{1}{2} \iff L > \frac{d-1}{2}.$$

The smallest positive integer L satisfying this strict inequality is $L = \lfloor (d-1)/2 \rfloor + 1 = \lceil d/2 \rceil$, uniformly in the parity of d: when d is even, $(d-1)/2$ is non-integer and $\lceil (d-1)/2 \rceil = d/2 = \lceil d/2 \rceil$ already exceeds it; when d is odd, $(d-1)/2 \in \mathbb{Z}$ so the strict inequality forces $L \geq (d-1)/2 + 1 = (d+1)/2 = \lceil d/2 \rceil$. The qualitative scaling is $L^*_{\mathrm{FC}}(n, d) = \Theta(d)$, recovering (5).

7.3. **Sharpness.** The matching lower bound is provided by considering $L < \lceil d/2 \rceil$, equivalently $L \leq (d-1)/2$: in this case

$$\beta - \alpha \;=\; L\beta_0 - L\alpha_0 \;=\; \frac{L}{d-1} \;\leq\; \frac{1}{2},$$

so the depth-L network fails the strict benign overfitting condition $\beta > \alpha + 1/2$ of paperA's Theorem 2 (Trichotomy). By that trichotomy the network instead lies in the tempered regime ($\beta = \alpha + 1/2$, attained when $L = (d-1)/2$ for odd d) or the catastrophic regime ($\beta < \alpha + 1/2$, all other cases with $L \leq (d-1)/2$). By definition in paperA's Theorem 2 [1, Theorem 2], the benign regime is $\mathrm{tr}(\mathbf{M}) < \infty$, equivalently $\beta > \alpha + 1/2$ strictly. The tempered regime ($\mathrm{tr}(\mathbf{M}) = \Theta(\log n)$, unbounded) and the catastrophic regime ($\mathrm{tr}(\mathbf{M}) = \infty$, divergent series) both violate this finiteness condition, so neither sub-case lies in the benign regime. The dimension-free feature-learning rate $n^{-2s/(2s+1)}$ of our own Theorem 3.1 is established under the joint hypothesis (i) the alignment hypothesis from [2] and (ii) the benign condition $\mathrm{tr}(\mathbf{M}) < \infty$ inherited from paperA via the unified bound of [1, Thm. 1]. When $\beta \leq \alpha + 1/2$, condition (ii) fails, hence Theorem 3.1's benign rate is no longer guaranteed. Hence $L < \lceil d/2 \rceil$ does not attain the benign rate, establishing matching sharpness up to the integer threshold.

7.4. **Smooth-tame fix for the strict-saddle argument.** The above derivation invokes the alignment of weights from [2], whose underlying use of the strict-saddle avoidance results in the nonconvex-optimization literature requires a C^2 assumption that is not satisfied by the ReLU activation. The appropriate replacement is the tame-function framework of Davis–Drusvyatskiy–Kakade–Lee [11], under which subgradient flow from almost every initialization avoids non-minimizing Clarke critical points. ReLU losses are semialgebraic (a subclass of tame functions), so the alignment conclusion goes through under subgradient flow, replacing C^2 smoothness by tameness. □

8. Proof of Theorem 6.2: ResNet case

The proof for residual networks differs structurally from the FC case because the identity skip connection prevents the multiplicative accumulation of α across layers.

8.1. **Decoupling of depth from α.**

Proposition 8.1 (Identity-bypass spectrum)**.** *Let $\mathbf{J}_{\mathrm{ResNet},L}$ denote the end-to-end Jacobian of an L-block ResNet at a trained interpolant, with each block of the form $h_{\ell+1} = h_\ell + g_\ell(h_\ell)$ for a two-layer ReLU residual g_ℓ. Then*

$$\sigma_j(\mathbf{J}_{\mathrm{ResNet},L}) = \Theta(\sigma_j(\mathbf{I} + \mathbf{J}_g)) = \Theta(\max(1, \sigma_j(\mathbf{J}_g))),$$

where $\mathbf{J}_g$ is the cumulative residual-block Jacobian, independent of L in the leading order.

Proof. Expanding the product, $\mathbf{J}_{\mathrm{ResNet},L} = \prod_{\ell=1}^{L}(\mathbf{I} + \mathbf{J}_{g_\ell})$. Under standard initialization scales (e.g. each g_ℓ initialized $O(1/\sqrt{L})$), the operator norm of each $\mathbf{J}_{g_\ell}$ is $O(1/\sqrt{L})$, so the product converges to a finite operator as $L \to \infty$ in the spectral sense; concretely, $\mathbf{J}_{\mathrm{ResNet},L} = \exp(\mathbf{J}_g) + O(1/L)$ where $\mathbf{J}_g = \sum_\ell \mathbf{J}_{g_\ell}$ is the cumulative residual.

The singular values of $\mathbf{J}_{\mathrm{ResNet},L}$ are therefore bounded *from below* by 1 for every $j \le \mathrm{rank}(\mathbf{I}) = n$ (since the identity contributes a constant to each singular value), and the spectral structure is governed by $\mathbf{J}_g$. The spectral decay rate α does not accumulate with L. □

8.2. **Critical depth in the ResNet case.** By Proposition 8.1, the per-block exponents (α_0, β_0) are the same as those of the cumulative residual $\mathbf{J}_g$ and do not depend on L. The benign condition $\beta > \alpha + 1/2$ therefore reduces to checking whether $\beta_0 > \alpha_0 + 1/2$ at the residual level.

- If $\beta_0 > \alpha_0 + 1/2$ already at $L = 1$: the network is benign for every $L \geq 1$, and $L^*_{\text{ResNet}} = 1$.
- If $\beta_0 \leq \alpha_0 + 1/2$: increasing L cannot improve the situation, since the exponents do not accumulate, and $L^*_{\text{ResNet}} = \infty$.

This proves (6): the critical depth is decoupled from L, taking values in $\{1, \infty\}$ depending only on the per-block exponents. □

Remark 8.2 (Implication for very deep ResNets). The decoupling of L^*_{ResNet} from L matches the empirical observation that very deep ResNets (e.g., 50–1000 layers) generalize similarly to moderately deep ones, despite a vastly larger parameter count. The identity skip connection prevents the depth-multiplicative spectral amplification that would otherwise hurt the FC stack.

9. Critical depth for transformers: a heuristic conjecture

For transformer architectures with softmax attention [18], the analogous depth analysis is more delicate: the attention map mixes tokens nonlinearly, and the cross-correlation exponent β depends on the spectral properties of the softmax map applied to the query–key inner products. We do not establish a rigorous critical-depth result for transformers in this paper; instead we record the heuristic prediction as an explicit conjecture, and comment on what would be needed to make it rigorous.

9.1. **Heuristic argument.** Let $g_{\text{attn}}(\mathbf{x})$ denote a single attention head with key, query, and value matrices $W_K, W_Q, W_V \in \mathbb{R}^{d\times d}$, applied to a token $\mathbf{x} \in \mathbb{R}^d$. The output of the head is $g_{\text{attn}}(\mathbf{x}) = \big(W_V X\big)^\top \operatorname{softmax}\big((W_K X)^\top (W_Q \mathbf{x}) / \sqrt{d}\,\big)$ for an input token sequence $X \in \mathbb{R}^{d\times T}$ (columns are tokens) and target token $\mathbf{x}$, where the softmax is taken over the T-many columns of $W_K X$. Heuristically, the softmax $\operatorname{softmax}(z)$ is a C^∞ Lipschitz map (with operator-norm gradient bounded by 1) and acts as a smoothing operator on the spherical-harmonic content of the gradient feature. If one accepts the heuristic that this smoothing adds $+1/2$ to the spherical-harmonic decay exponent (the same gain that would follow from convolving against a C^∞ kernel and converting eigenvalues $\to$ singular values via $\lambda \mapsto \sqrt{\lambda}$), then the per-layer cross-correlation exponent for a transformer block becomes

$$\beta_0^{\text{attn}} = \beta_0 + \tfrac{1}{2},$$

where β_0 is the FC value of Assumption 7.2. We do not establish this heuristic rigorously here: a rigorous derivation would require an explicit

spherical-harmonic decomposition of the attention-output map, which depends on the joint distribution of (W_K, W_Q, W_V) at the relevant trained predictor and is delicate because the softmax mixes spherical-harmonic shells of all orders.

Conjecture 9.1 (Critical depth for transformers)**.** *Let $L^*_{\text{Transformer}}$ denote the critical depth (in the sense of Definition 6.1) for a stack of pre-norm transformer blocks with h-headed softmax attention, applied to inputs embedded in $\mathbb{R}^d$, on a single-index target with $s = 1$. If the heuristic exponent $\beta_0^{\text{attn}} = \beta_0 + 1/2$ holds at the per-layer level, then*

$$L^*_{\text{Transformer}} = O(1) \text{ as } d \to \infty,$$

and in particular the benign condition $\beta > \alpha + 1/2$ is satisfied at $L = 1$ for d sufficiently large.

Remark 9.2 (Comparison with empirical depths). Conjecture 9.1, if true, would predict that the benign-overfitting depth threshold for transformers is much smaller than the depths $L \in [12, 96]$ used in production language models. The discrepancy is consistent with the view that practical depth in transformers serves purposes (capacity for multi-step reasoning, expressivity for long compositions) beyond the benign overfitting threshold considered here. A rigorous resolution of Conjecture 9.1 requires the spherical-harmonic spectral analysis of the softmax attention map referenced above, which we leave to future work.

10. Discussion

10.1. **Status of the alignment hypothesis.** The principal conditional in this paper is Assumption 2.3, which is established for two-layer ReLU networks with single-index targets, $\kappa = 1$, Gaussian inputs, and population gradient flow in [2]. Each of these conditions is non-trivial:

- *Two layers.* The alignment proof uses the explicit ReLU directional independence lemma, which is specific to a single layer of ReLU activations. Multi-layer extensions require a layer-by-layer alignment theory, which is not yet rigorous beyond the two-layer setting; the layer-by-layer alignment extension is left to future work. The end-to-end Jacobian SVD that defines $\mathbf{M}$ in [1] carries over to any depth, but the alignment hypothesis itself is a layer-1 statement.
- *Single-index, $\kappa = 1$.* Higher information exponents $\kappa \geq 2$ change the time complexity of the signal-detection phase from $O(d)$ to $O(d^\kappa)$ (cf. [8]). The statement of Theorem 3.1 extends but the rate is modified.

- *Multi-index targets.* For orthogonal multi-index targets ($g^*(\mathbf{x}) = \rho(\langle V^*, \mathbf{x}\rangle)$ with $V^* \in \mathbb{R}^{r\times d}$ having orthonormal rows), the rank-one collapse becomes a rank-r collapse, and the rate becomes $n^{-2s/(2s+r)}$. For non-orthogonal multi-indices, additional geometric assumptions on the conditioning of V^* are required; see [4, 15] for the closely related staircase-learning results.
- *Population gradient flow.* Finite-sample SGD adds a noise term that does not affect the leading-order rate but requires separate concentration arguments. Standard matrix-Bernstein concentration of the empirical Gram matrix relative to its population counterpart, at rate $\sqrt{\log p/n}$, gives a bridge between the population and finite-sample regimes; we do not give this bridge in full here.

10.2. **Open problems.**

(1) **Multi-index extension of Theorem 3.1.** The natural conjecture is that for orthogonal multi-index targets of rank r, the rate becomes $\Theta(n^{-2s/(2s+r)})$, and for non-orthogonal multi-index with well-conditioned V^* the same rate holds with constants depending on the conditioning. We do not prove this here.
(2) **Rigorizing the Transformer critical depth.** The attention-head-induced exponent gain stated in Conjecture 9.1 is heuristic; a rigorous proof requires an explicit spherical-harmonic-style spectral analysis of the softmax attention map applied to a token-level gradient feature, including a careful treatment of W_Q, W_K, W_V and their interaction with the embedding-space NTK. We do not establish this here.
(3) **Cross-entropy and Neural Collapse extension.** The Fisher-corrected operator $\mathbf{M}_{\mathrm{CE}}$ and the rank-$(K-1)$ effective-rank computation at the simplex equiangular tight frame fixed point are treated in the companion paper [3].
(4) **Non-Gaussian inputs.** The kernel-rate analysis Lemma 2.5 uses Gaussian inputs to invoke spherical harmonics. Extending Theorem 3.1 to sub-Gaussian or manifold-supported inputs requires a more general approximation theory.
(5) **Finite-sample bridge.** A complete finite-sample SGD proof of Theorem 3.1 would require combining the existing population-level rate with concentration of the empirical Jacobian and the empirical $\mathbf{\Gamma}$, neither of which is given in this paper.

10.3. **Relation to the broader benign overfitting literature.** Theorems 3.1 and 6.2 all factor through the unified bound and the trichotomy of [1]. The dimension collapse result is best viewed as identifying the specific mechanism by which feature learning improves β (in the language of [1]) without changing α, namely the spherical-harmonic multiplicity collapse. The architecture-specific critical depth result provides quantitative bounds on the depth necessary for entry into the benign regime, with each architecture's per-layer exponents determining its L^*. The companion paper [3] extends the framework to classification under cross-entropy loss, showing that the framework is loss-agnostic in the leading-order asymptotics.

Acknowledgements. The exploratory analysis underlying Theorems 3.1 and 6.2, including the comparison to the spherical-harmonic NTK eigenstructure and the architecture-specific decompositions, was conducted with the assistance of Claude (Anthropic). All mathematical statements and proofs have been verified by the author.

References

[1] L. Chang. Noise propagation operators and a two-parameter characterization of benign overfitting. *Companion paper*, 2026.

[2] L. Chang. A landscape analysis approach to feature alignment in shallow ReLU networks for single-index models. *Companion paper*, 2026.

[3] L. Chang. A cross-entropy noise propagation operator and the effective rank at the Neural Collapse fixed point. *Companion paper*, 2026.

[4] E. Abbe, E. Boix-Adsera, and T. Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on product distributions. In *Conference on Learning Theory (COLT)*, pages 4782–4887, 2022.

[5] K. Atkinson and W. Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*. Lecture Notes in Mathematics, vol. 2044. Springer, 2012.

[6] F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. (Convex/variational analysis of two-layer networks; not the mean-field framework.)

[7] J. Ba, M. A. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2023.

[8] G. Ben Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

[9] A. Bietti and J. Bruna. Deep equals shallow for ReLU networks in kernel regimes. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[11] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

[12] A. Damian, J. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory (COLT)*, pages 5413–5452, 2022.

[13] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

[14] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

[16] V. Papyan, X. Y. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[17] A. Rakhlin and X. Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory (COLT)*, pages 2595–2623, 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Independent Researcher
Email address: lightman.chang@gmail.com