Low rank is enough for the neural tangent kernel: depth and rank commutes for lazy training

Janis (Heran) Aiad[1]* Haizhao Yang[2]
Shijun Zhang[3]
[1] École Polytechnique, École Normale Supérieure Paris-Saclay
[2] University of Maryland, Department of Mathematics; Department of Computer Science
[3] The Hong Kong Polytechnic University

Abstract

In the lazy regime, training deep neural networks reduces to kernel regression, and the spectrum of the neural tangent kernel (NTK) controls convergence and stability. To reduce parameters, one can use low-rank structure and random features: deep networks with low-rank bottlenecks (RF-LR) freeze random feature maps and train only narrow readouts of dimension $r \ll N$ per layer. We study how depth L and bottleneck rank r shape this kernel: we take the sequential infinite-width limit $N \to \infty$, then analyze how the remaining randomness concentrates as $r \to \infty$.

We derive an explicit *new* NTK recursion for RF-LR with a visible $1/r$ factor at each bottleneck layer and a closed-form expansion at any depth. For the deterministic mean (proxy) kernel we prove sharp depth scaling—correlations align to 1 at rate $O(k^{-2})$ (same as for MLPs at the edge of chaos), the kernel magnitude saturates, and the diagonal–off-diagonal gap decays as $\asymp 1/(rk)$—and we give condition-number bounds: $\kappa \geq \Omega(r \cdot L)$ in general, and $\kappa_\perp = 1$ or $\kappa_\perp = 1 + o(1)$ for equicorrelated or high-dimensional spherical data (full proof in the appendix). Under a fixed parameter budget $O(NLr)$, depth and rank trade off, and from a conditioning perspective they commute. A main message is that *low rank does not shrink the kernel function class*: for the three-layer network, the mean RF-LR kernel induces the same RKHS as the shallow ReLU kernel, so expressivity is preserved while trainable parameters drop from $O(LN^2)$ to $O(LrN)$. We prove a rigorous proxy–empirical bound for equicorrelated data (Appendix C.4); for general (non-equicorrelated) data the concentration is sketched and left open. We also link the $I(r) \sim 1/\sqrt{r}$ scaling to $1/\sqrt{r}$ sub-Gaussian concentration of the empirical kernel around the proxy. Numerical experiments (Appendix E) confirm the depth scaling, conditioning bounds, and proxy–empirical concentration. All condition-number statements refer to the proxy kernel; we see strong experimental agreement with our theory and proxy predictions, and exact RKHS equivalence holds for three layers, with extension to depth $L \geq 4$ left to future work.

***K*eywords** Neural Tangent Kernel · Random Features · Low Rank · Scaling Laws · Infinite Width · Marchenko–Pastur

Contents

1 Introduction **4**

1.1 Related work 4

2 Network definition and EOC parameterization **5**

2.1 RF-LR architecture 5

2.2 Assumptions 5

*This work, including all theoretical derivations, calculations, and manuscript, was completed by J.H-A. during a Visiting Assistant Researcher at University of Maryland.

2.3 Parameterization, signal propagation, and kernels . . . 6

3 NTK recursion **6**

3.1 Large-depth regime: probabilistic kernel and exponential decay . . . 7

4 Depth scaling of the RF-LR NTK **8**

4.1 Main results . . . 8

5 Low rank is enough for the NTK RKHS (three-layer mean kernel) **10**

5.1 Microscopic distributions: Fisher–Kibble decoupling . . . 10

5.2 Endpoint expansion via hypergeometric analysis . . . 11

5.3 RKHS equivalence via endpoint behavior . . . 11

6 Conclusion and discussion **12**

A Summary of main results **15**

B Notation and standing assumptions **16**

B.1 Standing assumptions (expanded) . . . 16

B.2 Finite-width empirical NTK and limiting gradient flow . . . 16

B.3 EOC scaling and the origin of $1/r$ . . . 16

B.4 Finite-width effects . . . 17

B.5 Proofs of NTK Recursions . . . 17

B.5.1 Proof of Theorem 3.1: Infinite-width NTK recursion . . . 17

B.5.2 Proof of Theorem 2.1: Gaussian process composition . . . 18

B.5.3 Proof of three-layer NTK Corollary . . . 18

B.5.4 Proof of General L-layer NTK Corollary . . . 18

B.5.5 Reference: full-width MLP limiting NTK at the edge of chaos . . . 19

B.6 Comparison with the classical fully-trained NTK recursion . . . 20

B.7 Probabilistic recursion, GP concatenation, and bottleneck scaling . . . 20

B.8 Comparison table: depth/rank effects in the kernel regime . . . 22

B.9 Techniques and proof sketch . . . 22

C Proofs of low-rank NTK RKHS **22**

C.1 Background and statements for Section 5 . . . 22

C.2 Discussion on Fisher and Kibble Distributions . . . 23

C.3 Open directions: deep mean NTK recursion and effective depth . . . 25

C.4 Concentration of the random Gram matrix around the proxy . . . 26

C.5 Rank-driven concentration . . . 29

C.6 Proof of Corollary C.2: Concentration bound for three-layer NTK . . . 30

C.7 Proof of Corollary C.1: Mean NTK over Fisher–Kibble has same RKHS . . . 32

C.8 What extends to general depth vs. what remains open . . . 34

C.9 Asymptotic scaling of $I(r)$ . . . 35

D Proofs for depth scaling **35**

D.1 Correlation propagation and inverse cosine distances . . . 35
D.2 RF-LR NTK: recursion-driven depth scaling (not the full-MLP closed form) . . . 36
D.3 Inverse cosine distance matrices and spectral bounds . . . 36
D.4 Preliminaries and full statements for Section 4.1 . . . 37
D.5 Proof of Proposition 4.1 . . . 37
D.6 Proof of Corollary 4.1 . . . 38
D.7 Proof of Theorem 4.1 . . . 39
D.8 Remark: proxy lower bounds vs positivity (smallest eigenvalue) . . . 40
D.9 Remark: infinite-depth limit after centering . . . 41

E Experiments **41**

E.1 Correlation alignment and equicorrelated spectrum (Theorem 4.1, Corollary 4.1) . . . 41
E.2 Product decay and entry variance . . . 42
E.3 Smallest eigenvalue (finite-width Monte Carlo) . . . 43
E.4 Condition number κ (proxy vs empirical; equicorrelated, high-dim, non-equicorrelated) . . . 43
E.5 Non-equicorrelated κ and kernel regression risk . . . 44
E.6 RKHS Puiseux exponent vs depth (Corollary 5.2, extension to $L \geq 4$) . . . 44

1 Introduction

In the lazy/NTK regime, training deep neural networks linearizes and reduces to kernel regression with an (approximately) fixed Gram matrix [4]. To reduce parameters, one can factorize weights as $W = LR^\top$ with $r \ll \min\{n, m\}$; a central question is whether such factorizations preserve favorable optimization and convergence [1]. We combine this with the random-feature lazy regime: we freeze the feature (left) directions and train only the readout (right) factors into a bottleneck of dimension r per layer. That combination defines the RF-LR architecture studied in this paper: RF-LR freezes random feature directions and trains only linear readouts (and biases) into a bottleneck of dimension $r \ll N$ at each layer. This paper asks: *how do depth L and bottleneck dimension r control conditioning-relevant scales for optimization under the NTK perspective?* We adopt the standard NTK setup (ReLU activation, isotropic Gaussian initialization, sequential infinite-width limit) and analyze the induced kernel at initialization.

The architecture stacks high-then-low width layers with a bottleneck $r \ll n_1$; we train readouts and biases while keeping feature directions frozen. In the kernel regime, the Gram matrix spectrum controls optimization rates [4, 11]. We characterize limiting kernels and spectra under a composition-of-GPs view in the sequential infinite-width limit. Figure 1 illustrates the layout.

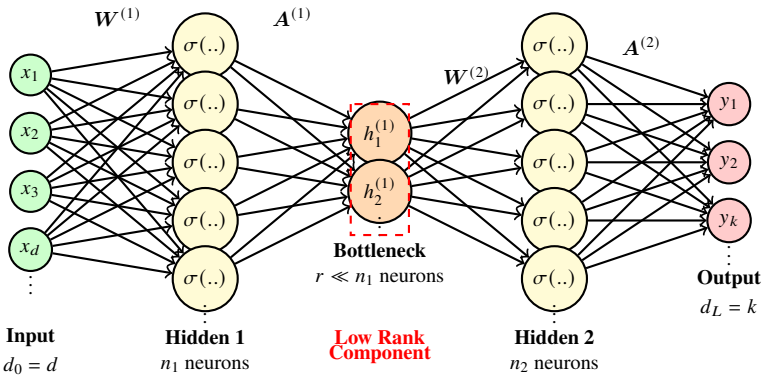

Figure 1: RF-LR: low-rank bottleneck (red dashed) reduces width from n_1 to r; readouts trained, feature directions frozen.

1.1 Related work

The NTK framework [4] characterizes infinitely wide networks as kernel methods. Extensions cover finite-width corrections [20, 6], depth scaling [22, 9], and RMT limits for Gram spectra in the extensive-width regime ($n \sim N^2$, layers in the RMT limit) [23]. Low-rank architectures reduce parameters and change optimization dynamics [29, 30, 31]; freezing features isolates r as the key control parameter. LoRA theory situates low-rank fine-tuning between kernel and feature-learning regimes [**?**, **?**]. For ReLU kernels, Bietti and Bach [17] established the "deep equals shallow" RKHS phenomenon via endpoint expansions; we combine Fisher–Kibble decoupling [15, 16] with a Puiseux analysis near $\rho = 1$.

Depth analyses for fully trained MLPs at the edge of chaos [**?**, **?**] show that increasing depth drives correlations toward 1, so layerwise kernels become nearly constant and the deep mean kernel approaches a rank-one structure. The same qualitative picture holds in RF-LR; the mean-kernel infinite-depth limit collapses after centering unless one rescales time or learning rate (Appendix D.9). Freezing the random features removes some cross-layer dependencies present in fully trained networks; together with Fisher–Kibble decoupling, this yields explicit formulas and suggests a path toward joint large-n regimes via the RMT extensive-width framework [23]. The recursion (Theorem 3.1) also suggests organizing finite-width corrections as an expansion in the bottleneck dimension r. Natural connections for future work include tensor programs [54] and depth-dependent NTK spectral analyses for MLPs at the edge of chaos [**?**], while retaining explicit control of the bottleneck scaling.

In the lazy/NTK regime, gradient flow reduces to kernel regression with an (approximately) fixed kernel [4]. Our contributions are:

- **Explicit NTK recursion and closed form.** We derive the infinite-width NTK recursion for RF-LR with an explicit $1/r$ factor at each bottleneck layer and a closed-form L-layer expansion (Theorem 3.1, Corollary 3.1).

- **Depth scaling and condition number.** For the deterministic proxy kernel we prove sharp depth scaling: correlation alignment $1 - \rho_k = \Theta(k^{-2})$ (the same as for full-width MLPs at the edge of chaos [**?**, **?**]), kernel saturation to a fixed point, and diagonal–off-diagonal gap $\asymp 1/(rk)$ (Theorem 4.1). We prove a condition-number *lower bound* $\kappa \geq \Omega(r \cdot L)$ for the proxy Gram in general (Proposition 4.1), and *exact* conditioning $\kappa_\perp = 1$ or $\kappa_\perp = 1 + o(1)$ for equicorrelated or high-dimensional spherical data (Corollary 4.1). Under a fixed parameter budget $O(NLr)$, depth and rank thus trade off from a conditioning perspective.
- **RKHS equivalence (three layers).** We show that the mean three-layer RF-LR kernel induces the same RKHS as the shallow ReLU kernel (Corollary 5.2): the bottleneck does not shrink the kernel-regime function class, while reducing trainable parameters from $O(LN^2)$ to $O(LrN)$ per layer. The proof uses Fisher–Kibble decoupling and a Puiseux expansion at the endpoint; extension to $L \geq 4$ remains open.
- **Numerical experiments.** Appendix E presents deterministic and finite-width experiments that confirm the depth scaling (Theorem 4.1), equicorrelated and high-dimensional conditioning (Corollary 4.1), and proxy–empirical concentration (Theorem 4.2); non-equicorrelated data illustrate the $\kappa \geq \Omega(r \cdot L)$ lower bound (Proposition 4.1).

All condition-number bounds refer to the proxy Gram matrix; we give a rigorous proxy–empirical bound on $\mathbf{1}^\perp$ for equicorrelated data (Theorem 4.2). Proof sketch: Appendix B.9.

We study the RF-LR NTK in the sequential infinite-width limit and its deterministic proxy; we do not analyze feature learning, adaptive optimizers, or continual learning. Regime: fix n, take $N \to \infty$, then study concentration as $r \to \infty$; we do not pursue $n \to \infty$ bulk spectral laws here. Sections 2–5 define the architecture, derive the recursion, state depth–rank scaling, and give the RKHS identification.

2 Network definition and EOC parameterization

2.1 RF-LR architecture

Let $\boldsymbol{h}^{(0)}(x) = x \in \mathbb{R}^{\boldsymbol{d}_0}$. For $\ell = 1, \ldots, L$,

$$\boldsymbol{h}^{(\ell)}(x) = \frac{1}{\sqrt{\boldsymbol{n}_\ell}} \sum_{j=1}^{\boldsymbol{n}_\ell} A_j^{(\ell)} \; \boldsymbol{\sigma}\left(w_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x)\right) \; + \; c^{(\ell)}. \tag{1}$$

Training policy: The readouts $A^{(\ell)}$ and layer biases $c^{(\ell)} \in \mathbb{R}^{\boldsymbol{d}_\ell}$ are trained; the feature directions $w^{(\ell)}$ are frozen random draws (i.i.d. Gaussian). The $1/\sqrt{\boldsymbol{n}_\ell}$ scaling (NTK parameterization) ensures a well-defined infinite-width limit.

Width vs bottleneck dimension. We distinguish the feature width $\boldsymbol{n}_\ell$ (number of random features) from the output dimension $\boldsymbol{d}_\ell = \dim(\boldsymbol{h}^{(\ell)})$. The trainable matrix $A^{(\ell)} \in \mathbb{R}^{\boldsymbol{d}_\ell \times \boldsymbol{n}_\ell}$ has columns $A_j^{(\ell)} \in \mathbb{R}^{\boldsymbol{d}_\ell}$. A *bottleneck* layer corresponds to $\boldsymbol{d}_\ell = r \ll \boldsymbol{n}_\ell$, yielding $O(r\boldsymbol{n}_\ell)$ trainable parameters and $\operatorname{rank}(A^{(\ell)}) \leq r$ automatically.

Training parameters, biases, and initialization. Initialization is

$$w_j^{(\ell)} \sim \mathcal{N}\big(0, I_{\boldsymbol{d}_{\ell-1}}/\boldsymbol{d}_{\ell-1}\big), \quad A_{ij}^{(\ell)} \sim \mathcal{N}(0, 2), \quad c^{(\ell)} = 0.$$

We use the *edge-of-chaos (EOC)* scaling for the weights so that $\operatorname{Cov}\big(w_j^{(\ell)}\big) = I_{\boldsymbol{d}_{\ell-1}}/\boldsymbol{d}_{\ell-1}$ [48]. In our bottleneck regime we take $\boldsymbol{d}_{\ell-1} = r$ for $\ell \geq 2$, so $w_j^{(\ell)} \in \mathbb{R}^r$ with covariance I_r/r.

We include a layer bias $c^{(\ell)}$ for completeness, but we avoid introducing extra scaling parameters for it. Bias scaling can be absorbed into the learning rate; the infinite-width and NTK structures remain qualitatively identical, and the kernel-regime predictions depend on kernel geometry, not absolute scales.

With this setup in place, we record the standing assumptions and conventions used throughout the paper.

2.2 Assumptions

We take the sequential infinite-width limit ($n_1 \to \infty, \ldots, n_L \to \infty$) with i.i.d. Gaussian random features frozen after initialization under NTK scaling $1/\sqrt{n_\ell}$ [4]; for $\ell \geq 2$ we assume the bottleneck regime $\boldsymbol{d}_\ell = r \ll n_\ell$ with edge-of-chaos scaling $\operatorname{Cov}(w_j^{(\ell)}) = I_r/r$. For RKHS and concentration we use normalized inputs (typically $x \in \mathbb{S}^{d-1}$) so kernels are

zonal in the cosine similarity ρ, and we assume positive 1-homogeneity of the activation (satisfied by $\boldsymbol{\sigma}$), which is used for Fisher–Kibble decoupling (Lemma 5.1) and radial/angular separation in the three-layer kernel. Full statements and discussion are in Appendix B (standing assumptions, EOC scaling, and Assumption B.1).

2.3 Parameterization, signal propagation, and kernels

We briefly record the scaling conventions used in the kernel calculations; details are in Appendix B.3 and Appendix B.5.1. *EOC scaling:* we choose the variance σ_A so that the forward variance remains stable across depth (edge of chaos), leading to a recursion for the layerwise variance $q^{(\ell)}$ and a corresponding condition on σ_A. *Bottleneck scaling and the* $1/r$ *factor:* in the bottleneck regime $\boldsymbol{d}_{\ell-1} = r$ for $\ell \geq 2$, the EOC scaling uses $\mathrm{Cov}(w) = I_r/r$, which keeps base and derivative kernels $O(1)$, and the RF-LR NTK recursion carries an explicit prefactor $1/r$ at bottleneck layers. With an abuse of notation, we denote by $\Sigma^{(\ell)}$ the base kernel and by $\dot{\Sigma}^{(\ell)}$ the derivative kernel.

Definition 2.1 (Base and derivative kernels)**.** In the sequential infinite-width limit, the layer-ℓ base kernel is

$$\Sigma^{(\ell)}(x, x') = \mathbb{E}_w\Big[\ \boldsymbol{\sigma}\left(w^\top \boldsymbol{h}^{(\ell-1)}(x)\right)\ \boldsymbol{\sigma}\left(w^\top \boldsymbol{h}^{(\ell-1)}(x')\right)\Big], \tag{2}$$

and the derivative kernel is

$$\dot{\Sigma}^{(\ell)}(x, x') = \mathbb{E}_w\Big[\dot{\boldsymbol{\sigma}}\big(w^\top \boldsymbol{h}^{(\ell-1)}(x)\big)\, \dot{\boldsymbol{\sigma}}\big(w^\top \boldsymbol{h}^{(\ell-1)}(x')\big)\ \|w\|^2\Big]. \tag{3}$$

For $\ell = 1$ and centered Gaussian w, $\Sigma^{(1)}$ is rotation-invariant with standard closed forms in terms of input cosine similarity $\rho = \langle x, x'\rangle/(\|x\|\|x'\|)$:

$$\Sigma^{(1)}(x, x') = \frac{\|x\|\|x'\|}{2\pi}\big((\pi - \theta)\cos\theta + \sin\theta\big), \quad \theta = \arccos(\rho).$$

We emphasize that these are not the NNGP kernel and the classical derivative kernel appearing in the NTK formulation. Rather, they are the base kernel and the derivative kernel of the Gaussian process $\boldsymbol{h}^{(\ell)}$ conditional on $\boldsymbol{h}^{(\ell-1)}$. By rotation invariance of Gaussian weights, these kernels depend on inputs only through the cosine similarity $\rho = \langle x, x'\rangle/(\|x\|\,\|x'\|)$. We will therefore write $\Sigma^{(\ell)}(\rho)$ and $\dot{\Sigma}^{(\ell)}(\rho)$. When no confusion arises (e.g., for $\ell = 1$), we will also abuse notation and write $\sigma(\rho)$ for the scalar kernel function; this σ denotes the kernel-as-a-function-of-ρ, not the activation $\boldsymbol{\sigma}$. We can now state the base and derivative kernels precisely.

Given these definitions, we can now state the following composition result.

Theorem 2.1 (Gaussian process composition)**.** *In the sequential infinite-width limit, the hidden states* $\boldsymbol{h}^{(\ell)}$ *form a composition of Gaussian processes:*

- *Conditionally on* $\boldsymbol{h}^{(\ell-1)}$*, for any finite collection of inputs* $\{x_1, \ldots, x_m\}$*, the vector* $(\boldsymbol{h}^{(\ell)}(x_1), \ldots, \boldsymbol{h}^{(\ell)}(x_m))$ *converges in distribution to a multivariate Gaussian.*
- *The conditional limiting Gaussian has mean zero and covariance matrix* $[\Sigma^{(\ell)}(x_i, x_j)]_{i,j=1}^m$*, where* $\Sigma^{(\ell)}$ *is the base kernel defined in Definition 2.1 and depends on* $\boldsymbol{h}^{(\ell-1)}$*.*

Proof. See Appendix B.5.2. □

3 NTK recursion

We now state the main result characterizing the NTK recursion for RF-LR. This recursion governs how the kernel evolves through layers and determines the training dynamics in the infinite-width limit.

Theorem 3.1 (Infinite-width NTK recursion)**.** *Under NTK parameterization $1/\sqrt{n_\ell}$ and a sequential limit $n_1 \to \infty, \ldots, n_L \to \infty$ with $w^{(\ell)} \sim \mathcal{N}(0, \mathbf{I}_{d_{\ell-1}}/d_{\ell-1})$ i.i.d. frozen after initialization, the layer-ℓ NTK satisfies*

$$\Theta^{(0)}(x,x') = 0, \tag{4}$$

$$\Theta^{(1)}(x,x') = 1 + \Sigma^{(1)}(x,x'), \tag{5}$$

$$\Theta^{(\ell)}(x,x') = 1 + \frac{1}{r}\Theta^{(\ell-1)}(x,x')\cdot\dot{\Sigma}^{(\ell)}(x,x') + \frac{1}{r}\Sigma^{(\ell)}(x,x'), \quad \ell \geq 2, \tag{6}$$

where the factor $1/r$ appears for bottleneck layers $\ell \geq 2$ (low-rank regime). For $\ell = 1$, $\Sigma^{(1)}$ is deterministic; for $\ell > 1$, $\Sigma^{(\ell)}$ and $\dot{\Sigma}^{(\ell)}$ are random fields whose fluctuations shrink with the rank r of bottlenecks. The additive constant 1 term comes from the trained layer biases $c^{(\ell)}$ and corresponds to the constant mode. For mean-zero targets (or after centering labels), the relevant spectrum is that of the kernel restricted to $\mathbf{1}^\perp$, so $\kappa_\perp$ is the quantity governing convergence; the constant mode then plays no role.

Furthermore, for a finite-width network initialized at random parameters θ_0, the empirical NTK $\Theta^{(L)}_{\theta_0}$ is a random *kernel (a Gram operator of gradients). Consequently the kernel gradient flow trajectory $t \mapsto f_t$ is also a random object (measurable with respect to θ_0) and satisfies*

$$\frac{d}{dt}f_t(x) = -\int \Theta^{(L)}_{\theta_0}(x,x')\big(f_t(x') - y(x')\big)d\mu(x'),$$

where μ is the data distribution and y is the target function. In the sequential infinite-width NTK limit, $\Theta^{(L)}_{\theta_0}$ concentrates and converges (in probability) to the deterministic kernel $\Theta^{(L)}$, yielding the corresponding deterministic limiting gradient flow.

Proof: See Appendix B.5.1, Proof B.5.1.

We can now derive an explicit closed-form expression for the NTK at any depth by expanding the recursion.

Corollary 3.1 (General L-layer NTK; closed form)**.** *Under the assumptions of Theorem 3.1, for any $L \geq 1$, the NTK admits the explicit form:*

$$\Theta^{(L)}(x,x') = \sum_{\ell=1}^{L}\frac{1}{r^{L-\ell}}\left(\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)}(x,x')\right) + \sum_{\ell=1}^{L}\frac{1}{r^{\max(1,L-\ell)}}\Sigma^{(\ell)}(x,x')\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)}(x,x'), \tag{7}$$

where the first sum has bias contributions (empty product = 1 when $\ell = L$) and the second has fresh-basis contributions; each bottleneck layer contributes a factor $1/r$. For $L = 2$: $\Theta^{(2)} = 1 + \frac{1}{r}\big(\dot{\Sigma}^{(2)} + \Sigma^{(1)}\dot{\Sigma}^{(2)} + \Sigma^{(2)}\big)$: the leading 1 is the bias (constant mode); $\frac{1}{r}\dot{\Sigma}^{(2)}$ is the bias-path term (derivative at layer 2 only); $\frac{1}{r}\Sigma^{(1)}\dot{\Sigma}^{(2)}$ and $\frac{1}{r}\Sigma^{(2)}$ are the fresh-basis terms (layers 1 and 2). This matches Theorem 3.1 by substituting $\Theta^{(1)} = 1 + \Sigma^{(1)}$ into the recursion for $\ell = 2$.

Proof. Expanding the recursion in Theorem 3.1 yields the two sums in (7). A complete proof by induction (including an explicit $L = 3$ expansion) is given in Appendix B.5.4; see also Appendix B.1. □

Finite-width corrections and comparison with the full-width MLP NTK at EOC are given in Appendix B.4 and Appendix B.6 (Proposition B.1).

3.1 Large-depth regime: probabilistic kernel and exponential decay

The mean (probabilistic) recursion and the resulting nested conditional expectations are essential for understanding the deep ($L \to \infty$) behavior, but the full discussion is lengthy. We therefore defer the probabilistic recursion, its nested-expectation form, and the exponential depth-suppression bound to Appendix B.7 (see also Appendix C.3 for related open problems).

The three-layer case admits a compact representation that makes the Fisher–Kibble structure explicit; we present it in Section 5.1.

4 Depth scaling of the RF-LR NTK

4.1 Main results

Random correlations and deterministic proxy. At each layer $\ell \geq 2$, the correlation $\rho_\ell = \cos \angle(\boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x'))$ is a *random variable*: given the population correlation from the previous layer, Fisher's law and Lemma C.2 give $\mathbb{E}[(\rho_\ell - \varrho(\rho_{\ell-1}))^2] \leq C/r$ and $\mathbb{P}(|\rho_\ell - \varrho(\rho_{\ell-1})| \geq t) \leq 6\exp(-c\, r\, t^2)$, so fluctuations are $O(1/\sqrt{r})$ sub-Gaussian. The NTK recursion passes through this chain of random correlations. The *deterministic proxy* replaces ρ_ℓ by the deterministic iterates $\rho_k = \varrho^{\circ(k-1)}(\rho_1)$ and analyzes the scalar recursion (84); for large r the random kernel concentrates around the proxy (Appendix B.7); a formal bound $\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} = o_P(1)$ as $r \to \infty$ and its implication for empirical conditioning are given in Appendix C.4. In the equicorrelated case, a full proof yields $\|(\hat{K} - K_{\text{proxy}})|_{\mathbf{1}^\perp}\|_{\text{op}} = O_P(L/r + 1/\sqrt{r})$ (Theorem 4.2). All condition-number statements in this section refer to the *proxy* Gram matrix; we give a lower bound on its condition number and, in one special case, exact conditioning with high probability.

Depth–rank scaling. The recursion contains an explicit $1/r$ bottleneck factor, which leads to saturation of the kernel magnitude as depth grows, while the diagonal–off-diagonal gap decays with depth in the proxy recursion (the same $1 - \rho_k = \Theta(k^{-2})$ correlation alignment as for full-width MLPs at the edge of chaos [**?**, **?**]). Theorem 4.1 analyzes this deterministic proxy, which approximates the mean path of the random recursion for large r. We refer to Appendix D.4 for notation and kernel expansions.

Theorem 4.1 (Depth scaling for the deterministic proxy RF-LR NTK)**.** *Fix $r > 1$ and $\rho_1 \in (-1, 1)$. Let $\Theta^{(k)}(\rho_1)$ be the deterministic proxy defined by* (84)*, with $\rho_k = \varrho^{\circ(k-1)}(\rho_1)$ deterministic. Then (as for MLPs at EOC [**?**, **?**]):*

- *(**Correlation alignment**) The correlation recursion aligns polynomially:*

$$1 - \rho_k \;=\; \Theta(k^{-2}).$$

- *(**Kernel saturation**) $\Theta^{(k)}(\rho_1) \to \Theta_\star(r)$ as $k \to \infty$, where $\Theta_\star(r)$ is the fixed point of the limiting recursion obtained by setting $\dot{s} = s = 1/2$, namely*

$$\Theta_\star(r) \;=\; \frac{2r+1}{2r-1}.$$

  Moreover,

$$\Theta^{(k)}(\rho_1) - \Theta_\star(r) \;=\; O(k^{-1}).$$

- *(**Depth-induced gap decay**) Let $\Theta^{(k)}_{\text{diag}} := \Theta^{(k)}(1)$ (on-diagonal value, i.e. identical inputs) and $\Theta^{(k)}_{\text{off}} := \Theta^{(k)}(\rho_1)$ (off-diagonal value for an input pair with initial cosine similarity ρ_1). Then*

$$\Theta^{(k)}_{\text{diag}} - \Theta^{(k)}_{\text{off}} \;=\; \Theta\Big(\frac{1}{r\,k}\Big),$$

  for large k.

Proof. See Appendix D.7. □

Proposition 4.1 (Lower bound on condition number for the proxy)**.** *In the setting of Theorem 4.1, consider the deterministic proxy kernel $\Theta^{(L)}(\rho)$ and form the $n \times n$ Gram matrix $\mathbf{M}$ with entries $M_{ij} = \Theta^{(L)}(\rho_{ij})$ for pairwise input cosine similarities $\rho_{ij} = \langle x_i, x_j\rangle/(\|x_i\|\|x_j\|) \in (-1, 1)$. Let $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum eigenvalues of $\mathbf{M}$ restricted to $\mathbf{1}^\perp$. Assume: (i) $n \geq 3$; (ii) there exists a pair (i, j) with ρ_{ij} bounded away from 1; (iii) the dataset is not equicorrelated (the ρ_{ij} for $i \neq j$ are not all equal). Then $\lambda_{\max} = \Theta(1)$, $\lambda_{\min} \leq O(1/(rL))$, and the condition number $\kappa = \lambda_{\max}/\lambda_{\min}$ satisfies $\kappa \geq \Omega(r \cdot L)$. When assumption (iii) is violated (equicorrelated data),*

the proposition does not apply; instead Corollary 4.1 gives the best-case $\kappa_\perp = 1$ (and similarly for approximately equicorrelated high-dimensional data).

Proof. See Appendix D.5. □

Theorem 4.2 (Operator norm on $\mathbf{1}^\perp$ in the equicorrelated case)**.** *Fix $n, L \geq 1$, $r \geq 2$, and the equicorrelated setup $\rho_{1,ii} = 1$, $\rho_{1,ij} = \rho_0$ for $i \neq j$ with $\rho_0 \in (-1, 1)$. Let $\hat{K}$ be the $n \times n$ empirical RF-LR NTK Gram matrix and K_{proxy} the deterministic proxy with entries $(K_{\text{proxy}})_{ij} = \Theta^{(L)}(\rho_{1,ij})$ (Definition D.4). Then there exist constants $C, c > 0$ (depending on L, ρ_0) such that for any $\epsilon > 0$,*

$$\mathbb{P}\Big(\big\|(\hat{K} - K_{\text{proxy}})\big|_{\mathbf{1}^\perp}\big\|_{\text{op}} \geq \epsilon\Big) \leq C \exp\Big(-c\,\frac{r\,\epsilon^2}{L^2}\Big) + C \exp\Big(-c\,\frac{r\,\epsilon}{L}\Big).$$

In particular, $\big\|(\hat{K} - K_{\text{proxy}})\big|_{\mathbf{1}^\perp}\big\|_{\text{op}} = O_P(L/r + 1/\sqrt{r})$ as $r \to \infty$ with n, L fixed. Hence if $r = O(L^2)$, then $\big\|(\hat{K} - K_{\text{proxy}})\big|_{\mathbf{1}^\perp}\big\|_{\text{op}} = O_P(L/r)$.

Proof. See Appendix C.4. □

Corollary 4.1 (Exact conditioning for equicorrelated and high-dimensional random data)**.** *In the setting of Proposition 4.1, let $\kappa_\perp = \lambda_{\max}(\mathbf{M}|_{\mathbf{1}^\perp})/\lambda_{\min}(\mathbf{M}|_{\mathbf{1}^\perp})$ denote the condition number on the mean-zero subspace. The following hold.*

- ***Equicorrelated data.*** *If $\rho_{ij} = \rho_0$ for all $i \neq j$ (and $\rho_{ii} = 1$), then on $\mathbf{1}^\perp$ all $n - 1$ eigenvalues of $\mathbf{M}$ equal $\lambda_\perp = \Theta^{(L)}(1) - \Theta^{(L)}(\rho_0) = \Theta(1/(rL))$. Hence*

$$\kappa_\perp = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\lambda_\perp}{\lambda_\perp} = 1.$$

- ***High-dimensional random data.*** *If the inputs $x_1, \ldots, x_n \in \mathbb{S}^{d-1}$ are drawn i.i.d. uniformly from the unit sphere, then as $d \to \infty$ (with n, r, L fixed) with high probability $\max_{i \neq j} |\rho_{ij}| = O(1/\sqrt{d})$, so the proxy Gram matrix is approximately equicorrelated with $\rho_0 = O(1/\sqrt{d})$. On that event, all eigenvalues on $\mathbf{1}^\perp$ are $\lambda_\perp = \Theta^{(L)}(1) - \Theta^{(L)}(\bar{\rho}) + o(1) = \Theta(1/(rL))(1 + o(1))$ with $\bar{\rho} = O(1/\sqrt{d})$, and they differ from each other by $o(\lambda_\perp)$. Hence*

$$\kappa_\perp = \frac{\lambda_{\max}}{\lambda_{\min}} = 1 + o(1) \qquad \text{as } d \to \infty.$$

  On that high-probability event, the empirical Gram matrix $\hat{K}$ concentrates around the proxy: $\|(\hat{K} - K_{\text{proxy}})|_{\mathbf{1}^\perp}\|_{\text{op}} = O_P(L/r + 1/\sqrt{r})$ (Theorem 4.2 with $\rho_0 = O(1/\sqrt{d})$).

Proof. See Appendix D.6. □

Limited scope of exact conditioning. Equicorrelated and high-dimensional i.i.d. spherical data are special; the general conditioning guarantee in this section is the proxy lower bound (Proposition 4.1). Summary of condition-number scaling:

- **Equicorrelated:** $\kappa_\perp = 1$ (exact; all eigenvalues on $\mathbf{1}^\perp$ equal $\lambda_\perp = \Theta(1/(rL))$).
- **High-dim i.i.d. spherical:** $\kappa_\perp = 1 + o(1)$ as $d \to \infty$ (approximately equicorrelated).
- **General (non-equicorrelated):** $\kappa \geq \Omega(r \cdot L)$; explicit: $\lambda_{\max} = \Theta(1)$, $\lambda_{\min} \leq O(1/(rL))$, hence $\kappa = \lambda_{\max}/\lambda_{\min} \geq \Omega(r \cdot L)$.

Proxy vs. optimization. Proposition 4.1 gives a *lower bound* on the condition number of the *proxy* Gram matrix; we see strong experimental agreement with our theory and proxy predictions, and exact RKHS equivalence holds for three layers, with extension to depth $L \geq 4$ left to future work. For the proxy and non-equicorrelated data, depth drives the kernel toward near-constant entries (gap $\Theta(1/(rL))$, as for MLPs at EOC) and $\kappa \geq \Omega(rL)$. The actual network uses random correlations with $O(1/\sqrt{r})$ sub-Gaussian fluctuations (Appendix B.7), so for large r the random kernel

concentrates around the proxy. The condition number is relevant for kernel regression (stability and convergence speed depend on $\lambda_{\max}/\lambda_{\min}$); under a fixed parameter budget $O(NLr)$, depth and rank trade off from a conditioning perspective, and the choice of r and L is a matter of approximation power (effective depth and RKHS expressivity).

5 Low rank is enough for the NTK RKHS (three-layer mean kernel)

We show that low rank does not shrink the kernel-regime function class: the mean three-layer RF-LR kernel induces the same RKHS as the shallow ReLU kernel.

Roadmap. We first derive the Fisher–Kibble decoupling (Section 5.1), which separates angular (correlation) and radial (norm) randomness and yields a compact three-layer empirical NTK form. We then obtain a Puiseux expansion of the mean kernel near $\rho = 1$ (Section 5.2), which controls the RKHS. The leading exponent matches the shallow ReLU kernel, so the RKHSs coincide (Corollary 5.2). The analysis is restricted to three layers because the Fisher–Kibble integral admits a closed form for a single bottleneck; for $L \geq 4$ (depth at least four layers), the nested Fisher-chain expectations require a Laplace-type expansion that remains open (Appendix C.3).

5.1 Microscopic distributions: Fisher–Kibble decoupling

We can now state a compact representation under a homogeneity assumption.

Lemma 5.1 (Fisher–Kibble decoupling)**.** *Let x, y be input vectors and let x_1, y_1 denote their rank-r random projections. Define the empirical correlation $\rho_1 = \cos\angle(x_1, y_1)$ and squared norms $u = \|x_1\|^2, v = \|y_1\|^2$. Then:*

- *ρ_1 follows Fisher's correlation distribution [15, 16] (centered at ρ). For $r > 2$:*

$$p_{\text{Fisher}}(\rho_1) \;=\; \frac{(r-2)\,\Gamma(r-1)\,(1-\rho^2)^{\frac{r-1}{2}}\,(1-\rho_1^2)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(1-\rho\rho_1)^{r-\frac{3}{2}}}\;{}_2F_1\!\left(\frac{1}{2},\frac{1}{2};r-\frac{1}{2};\frac{1+\rho\rho_1}{2}\right).$$

- *(u, v) follow Kibble's [49] bivariate chi-square law ($u, v \geq 0$; I_ν is the modified Bessel function):*

$$p_{\text{Kibble}}(u,v) \;=\; \frac{(uv)^{\frac{r}{4}-\frac{1}{2}}\,\exp\!\left(-\frac{u+v}{2(1-\rho^2)}\right)}{\Gamma(r/2)\,(2(1-\rho^2))^{\frac{r}{2}+1}\,\rho^{\frac{r}{4}-\frac{1}{2}}}\;I_{\frac{r}{2}-1}\!\left(\frac{\rho\sqrt{uv}}{1-\rho^2}\right).$$

- *Angular and radial parts are independent: $p(\rho_1, u, v) = p_{\text{Fisher}}(\rho_1)\,p_{\text{Kibble}}(u, v)$.*

Proof: See Appendix C.2, Proof C.2.

Homogeneity assumption enabling decoupling. By positive 1-homogeneity and rotational invariance, the base kernel factorizes and $(\rho_1, \|x_1\|, \|y_1\|)$ decouple into angular (Fisher) and radial (Kibble) components; see Appendix C.1 and Appendix C.2 for details, explicit densities, and the induced compact three-layer empirical NTK form.

Compact three-layer empirical NTK form. Under the assumptions of Theorem 3.1 with $L = 2$, the three-layer empirical RF-LR NTK admits the representation

$$\mathbf{\Theta}^{(2)}(x,x') \;=\; 1 \;+\; \frac{1}{r}\,\mathbf{\Theta}^{(1)}(x,x')\,\dot{\Sigma}^{(2)}(x,x') \;+\; \frac{1}{r}\,\Sigma^{(2)}(x,x'), \qquad \mathbf{\Theta}^{(1)}(x,x') = 1 + \Sigma^{(1)}(x,x'), \tag{8}$$

and, under the positive 1-homogeneity assumption (satisfied by $\boldsymbol{\sigma}$), the layer-2 fields can be written in terms of the Fisher angular variable ρ_1 and the Kibble radial factor $w_r := \sqrt{uv}/r$ as

$$\dot{\Sigma}^{(2)}(x,x') \;=\; 1 - \frac{\arccos(\rho_1)}{\pi}, \qquad \Sigma^{(2)}(x,x') \;=\; w_r\,\Sigma^{(1)}(\rho_1). \tag{9}$$

Combining (8)–(9) yields an explicit three-layer kernel in the decoupled microscopic variables (ρ_1, u, v); equivalently,

$$\mathbf{\Theta}^{(2)}(x,x') \;=\; 1 \;+\; \frac{1}{r}\,\mathbf{\Theta}^{(1)}(x,x')\left(1 - \frac{\arccos(\rho_1)}{\pi}\right) \;+\; \frac{1}{r}\,w_r\,\Sigma^{(1)}(\rho_1). \tag{10}$$

See Remark C.1 (Appendix C.2) for the full derivation, the explicit Fisher and Kibble densities, and the induced compact form (10).

5.2 Endpoint expansion via hypergeometric analysis

The RKHS of zonal kernels on the sphere is controlled by the Puiseux behavior near $\rho = \pm 1$ [17]. For the mean three-layer kernel, the key step is to analyze the Fisher expectation $\mathbb{E}[\arccos(\hat{\rho}_r)]$ as $\rho \to 1$.

Proposition 5.1 (Puiseux expansion of $\mathbb{E}(\arccos(\hat{\rho}_r))$ near $\rho = 1$)**.** *Let $r > 2$ and $\hat{\rho}_r \sim \mathrm{Fisher}(\rho, r)$. Write $\rho = 1 - t$ with $t \downarrow 0$. Then*

$$\mathbb{E}[\arccos(\hat{\rho}_r)] \;=\; \sqrt{2t}\, I(r) \;+\; O(t^{3/2}),$$

where $I(r) \in (0, \infty)$ admits the closed form

$$I(r) \;=\; \frac{(r-2)\, 2^{r-\frac{5}{2}}}{\sqrt{2\pi}}\, \frac{\Gamma\!\left(\frac{r-1}{2}\right)\Gamma\!\left(\frac{r}{2}-1\right)}{\Gamma(r-1)}\, C_1(r), \qquad C_1(r) \;=\; \frac{\Gamma\!\left(r-\frac{1}{2}\right)\Gamma\!\left(r-\frac{3}{2}\right)}{\Gamma(r-1)^2}. \tag{11}$$

Proof. Insert Fisher's density (Appendix C.1) into $\int_{-1}^{1} \arccos(s)\, p_{\mathrm{Fisher}}(s \mid 1-t, r)\, ds$ and change variables $s = 1 - v$. Near $t, v \to 0$, use $\arccos(1-v) = \sqrt{2v} + O(v^{3/2})$ and apply the hypergeometric connection formula (DLMF §15.8.2 [51]) to replace ${}_2F_1(\frac{1}{2}, \frac{1}{2}; r - \frac{1}{2}; 1 - \frac{t+v}{2})$ by the constant $C_1(r)$ up to $O((t+v)^{r-3/2})$. Then the leading integral reduces, after $w = v/t$, to a Beta integral and yields (11). See Appendix C.7 for the full proof. □

Proposition 5.2 ($I(r)$ decays as $1/\sqrt{r}$)**.** *Let $I(r)$ be defined by* (11)*. Then $I(r) = O(1/\sqrt{r})$ as $r \to \infty$.*

Proof sketch. Apply Stirling's formula to the Gamma ratios in (11). A detailed proof is given in Proposition C.2 (Appendix C.9). □

Corollary 5.1 (Puiseux expansion of the mean three-layer NTK near $\rho = 1$)**.** *Let $\tilde{\boldsymbol{\Theta}}^{(2)}(\rho)$ denote the mean three-layer RF-LR NTK and let $K_\infty(\rho)$ denote the deterministic $r \to \infty$ limit kernel,*

$$K_\infty(\rho) = \Theta^{(1)}(\rho)\left(1 - \frac{\arccos(\rho)}{\pi}\right) + \Sigma^{(1)}(\rho) + 1,$$

as in Appendix C.7. Writing $\rho = 1 - t$ with $t \downarrow 0$, one has

$$\tilde{\boldsymbol{\Theta}}^{(2)}(1-t) \;=\; 1 + \frac{1}{r}\big(K_\infty(1) - 1\big) \;-\; \frac{1}{r}\left[\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi}\, I(r)\right] t^{1/2} \;+\; O(t^{3/2}), \tag{12}$$

where $I(r)$ is as in Proposition 5.1. In particular, since $I(r) = O(1/\sqrt{r})$ (Proposition 5.2), the bracket equals $2/\pi + O(1/\sqrt{r})$, so the $t^{1/2}$ coefficient is $\frac{2}{\pi r} + O(r^{-3/2})$.

Proof sketch. Combine the three-layer mean-kernel identity (Appendix C.7) with Proposition 5.1 and the ReLU endpoint expansion $\arccos(1-t) = \sqrt{2t} + O(t^{3/2})$. A complete proof is given in Appendix C.7. □

Interpretation: $1/r$ vs. $I(r)$. The $1/r$ prefactor in (12) comes from the EOC initialization and the RF-LR recursion: each bottleneck layer contributes $1/r$ (Section 2, Appendix B.3). With unnormalized classical NTK (no bottleneck scaling), the Puiseux coefficient would be $\left[\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi} I(r)\right]$ and the $I(r) \sim 1/\sqrt{r}$ term would directly reflect the Fisher–Kibble randomness. Thus the $1/r$ decay in the expansion is due to EOC; $I(r) \sim 1/\sqrt{r}$ is the signature of the randomness-induced deviation from the deterministic limit.

5.3 RKHS equivalence via endpoint behavior

Main results. For zonal kernels on the sphere, the RKHS is determined by the Puiseux exponents at the endpoints $\rho = \pm 1$ [17]. Our main results in this section are:

- **Puiseux expansion:** The mean three-layer NTK $\tilde{\boldsymbol{\Theta}}^{(2)}(\rho)$ has a leading $t^{1/2}$ term near $\rho = 1 - t$ (Corollary 5.1). The coefficient involves $\mathbb{E}[\arccos(\hat{\rho}_r)]$, whose expansion uses the hypergeometric connection formula (DLMF §15.8.2 [51]) applied to the ${}_2F_1$ in Fisher's density (Proposition 5.1).

- $I(r)$ **scaling:** $I(r) = O(1/\sqrt{r})$ as $r \to \infty$ (Proposition 5.2), so the $t^{1/2}$ coefficient is an $O(1)$ perturbation of the shallow ReLU constant.
- **RKHS equivalence:** The three-layer mean kernel induces the same RKHS as the shallow ReLU kernel (Corollary 5.2 below), by the Bietti–Bach criterion [17, Theorem 1]: matching the exponent $1/2$ at $\rho = 1$ suffices.

See Appendix C.7 for the full proof of the Puiseux expansion and RKHS identification.

Corollary 5.2 (Low rank is enough for the NTK RKHS: three-layer mean kernel)**.** *Under the RF-LR setting with ReLU nonlinearity and isotropic random features on $\mathbb{S}^{d-1}$, the three-layer mean NTK $\tilde{\mathbf{\Theta}}^{(2)}$ induces the same RKHS as the shallow ReLU kernel. In particular, the corresponding RKHSs coincide as sets with equivalent norms.*

Proof sketch. The mean three-layer kernel is zonal. Proposition 5.1 gives a Puiseux expansion near $\rho = 1$ with leading exponent $t^{1/2}$; Proposition 5.2 shows the r-dependent coefficient is an $O(1)$ perturbation of the shallow ReLU coefficient. Since the shallow ReLU kernel has the same endpoint exponent $1/2$, the Bietti–Bach criterion [17] yields RKHS equivalence. See Appendix C.7 for a detailed argument. □

6 Conclusion and discussion

We gave an NTK analysis for low-rank random-feature architectures (RF-LR) under explicit assumptions, yielding a new explicit recursion, sharp depth scaling for the proxy kernel, condition-number bounds, and in the three-layer case RKHS equivalence with the shallow ReLU kernel and concentration in the bottleneck dimension. In the lazy/NTK regime with mean-zero targets, stability and convergence depend on the condition number $\lambda_{\max}/\lambda_{\min}$ of the centered Gram matrix. Our results give a lower bound on the proxy condition number and in one setting its exact value w.h.p. Numerical experiments (Appendix E) confirm the depth scaling, conditioning bounds, and proxy–empirical concentration. Exact RKHS equivalence holds for three layers, with extension to depth $L \geq 4$ left to future work; the link to GD convergence or sample complexity is also left to future work.

The empirical three-layer NTK concentrates around its deterministic limit with sub-Gaussian tails in r (Corollary C.2, Appendix C.6). The $I(r) = O(1/\sqrt{r})$ scaling (Proposition C.2) matches Fisher–Kibble corrections as the mean kernel's leading endpoint coefficient vanishes in r. Extending RKHS equivalence to depth $L \geq 4$ would require controlling endpoint expansions through the full Fisher chain (Appendix C.7). Conditioning holds for arbitrary L, whereas RKHS equivalence is proved only for three layers (Appendix C.8). Natural next steps are concentration bounds for general depth L, bulk spectral laws in the RMT limit $n \sim N^2$ by adapting [23] to low-rank depth $L \geq 2$, and the impact of data geometry (hypercube vs. spherical).

On the experimental side, confirming lazy-kernel predictions and testing $O(1/n) + O(1/r)$ finite-width corrections would strengthen the link between the asymptotic theory and practice. Our scope is the lazy/NTK regime, where tractability is greatest.

References

[1] Shijun Zhang, Hongkai Zhao, Yimin Zhong, and Haomin Zhou. Structured and balanced multi-component and multi-layer neural networks, 2025.

[2] Siddhartha Rao Kamalakara, Acyr Locatelli, Bharat Venkitesh, Jimmy Ba, Yarin Gal, and Aidan N. Gomez. Exploring low rank training of deep neural networks, 2022.

[3] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations, 2017.

[4] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[5] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy, 2019.

[6] Ethan Dyer and Guy Gur-Ari. Asymptotics of wide networks from feynman diagrams, 2019.

[7] Max Guillen, Philipp Misof, and Jan E. Gerken. Finite-width neural tangent kernels from feynman diagrams, 2025.

[8] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks, 2023.

[9] Dávid Terjék. The spectrum of the neural tangent kernel at initialization. *arXiv preprint arXiv:2211.10688*, 2022.

[10] Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, and Shafei Wang. Seeking consistent flat minima for better domain generalization via refining loss landscapes, 2025.

[11] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning, 2020.

[12] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[13] Stanislav Fort and Surya Ganguli. Large scale structure of neural network loss landscapes. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Sanghoon Na and Haizhao Yang. Curse of dimensionality in neural network optimization, 2025.

[15] Ronald Aylmer Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1915.

[16] Harold Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(2):193–225, 1953.

[17] Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes, 2021. ICLR 2021.

[18] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks, 2023.

[19] Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows, 2025.

[20] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 4825–4836, 2019.

[22] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[23] Lucas Benigni and Elliot Paquette. Eigenvalue distribution of the neural tangent kernel in the quadratic scaling, 2025.

[24] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

[25] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, pages 1313–1320, 2008.

[26] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[27] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in Neural Information Processing Systems*, 30, 2017.

[28] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 21(3):1759–1786, 2015.

[29] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6655–6659, 2013.

[30] Alexander Novikov, Dmitry Podoprikhin, Anton Osokin, and Dmitry Vetrov. Tensorizing neural networks. In *Advances in Neural Information Processing Systems*, pages 442–450, 2015.

[31] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *International Conference on Machine Learning*, pages 244–253, 2018.

[32] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310, 2019.

[33] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Training behavior of deep neural network in frequency domain. *International Conference on Neural Information Processing*, pages 264–274, 2019.

[34] Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 33:18126–18136, 2020.

[35] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pages 7537–7547, 2020.

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

[37] Alex Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. *Conference on Learning Theory*, pages 5413–5452, 2022.

[38] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. *Conference on Learning Theory*, pages 4782–4887, 2023.

[39] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[40] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[41] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[42] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[43] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[44] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[45] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[46] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *Conference on Learning Theory*, pages 1305–1338, 2020.

[47] Quynh Nguyen, Marco Mondelli, and Guido Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks, 2022.

[48] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the selection of initialization and activation function for deep neural networks, 2018.

[49] W. F. Kibble. A two-variate gamma type distribution. *Sankhyā: The Indian Journal of Statistics*, 5:137–150, 1941.

[50] Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models, 2021.

[51] F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain. Nist digital library of mathematical functions. `https://dlmf.nist.gov/`, 2024. Release 1.2.1 of 2024-06-15.

[52] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1964.

[53] Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1), feb 2010.

[54] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture, 2020.

Contents

A Summary of main results

This work develops an NTK analysis for low-rank random-feature architectures (RF-LR) under explicit assumptions. Appendix E reports numerical illustrations that accompany the theory.

Main: depth/rank scaling and condition number. The trainable parameter budget scales as $O(NLr)$. In the deterministic proxy mean recursion, depth drives correlations toward 1 with $1 - \rho_k = \Theta(k^{-2})$ (the same correlation alignment as for full-width MLPs at the edge of chaos [**?**, **?**]), the kernel magnitude saturates, and the diagonal–off-diagonal gap decays as $\asymp 1/(rk)$ (Theorem 4.1). We prove a lower bound on the condition number of the *proxy* Gram matrix, $\kappa \geq \Omega(r \cdot L)$ (Proposition 4.1), and exact conditioning $\kappa_\perp = 1$ or $\kappa_\perp = 1 + o(1)$ for equicorrelated or high-dimensional random data (Corollary 4.1). Under a fixed budget $O(NLr)$, depth and rank trade off from a conditioning perspective; the choice of r and L is a matter of approximation power (e.g. effective depth and RKHS expressivity).

Recursion. We provide an explicit NTK recursion for arbitrary depth in RF-LR and a closed-form expansion (Theorem 3.1, Corollary 3.1).

Low rank is enough for the NTK RKHS (three layers). In the NTK regime, the three-layer RF-LR mean kernel induces the same RKHS as the shallow ReLU kernel (Corollary 5.2), showing that the bottleneck does not reduce the kernel-regime function class while lowering the trainable parameter scaling from dense $O(LN^2)$ to $O(LrN)$.

Concentration in the bottleneck dimension. For the three-layer case, the empirical kernel concentrates around its deterministic limit $K_\infty(\rho)$ with tails controlled by the bottleneck dimension r (Theorem C.1, Corollary C.2). For equicorrelated data, a full proof in the appendix gives $\|(\hat{K} - K_{\text{proxy}})|_{\mathbf{1}^\perp}\|_{\text{op}} = O_P(L/r + 1/\sqrt{r})$ (Theorem 4.2), linking proxy and empirical conditioning on the mean-zero subspace.

B Notation and standing assumptions

B.1 Standing assumptions (expanded)

We summarize the conventions used throughout the main text.

- **Sequential infinite width.** Widths $n_1, \dots, n_L$ tend to infinity sequentially under NTK scaling $1/\sqrt{n_\ell}$. Random features $w^{(\ell)}$ are i.i.d. Gaussian and frozen after initialization.
- **Training policy.** Only the readouts $A^{(\ell)}$ (and layer biases $c^{(\ell)}$) are trained; all other parameters are frozen.
- **Bottleneck regime.** For $\ell \geq 2$, the representation dimension is $d_\ell = r \ll n_\ell$. This induces an explicit $1/r$ factor in the RF-LR NTK recursion at bottleneck layers.
- **Zonal kernels.** By rotational invariance of Gaussian features, base and derivative kernels depend on inputs only through the cosine similarity $\rho = \langle x, x' \rangle / (\|x\| \, \|x'\|)$. For RKHS results we typically take $x \in \mathbb{S}^{d-1}$.
- **Homogeneity.** We use Assumption B.1 (positive 1-homogeneity) for Fisher–Kibble decoupling and for separating radial and angular components in three-layer calculations.

B.2 Finite-width empirical NTK and limiting gradient flow

For a finite-width network initialized at random parameters θ_0, the empirical NTK $\Theta^{(L)}_{\theta_0}$ is a random kernel (the Gram operator of gradients $\langle \nabla_\theta f(x;\theta_0), \nabla_\theta f(x';\theta_0) \rangle$). The kernel gradient flow trajectory $t \mapsto f_t$ is therefore a random object (measurable with respect to θ_0) and satisfies

$$\frac{d}{dt} f_t(x) \;=\; -\int \Theta^{(L)}_{\theta_0}(x, x') \left(f_t(x') - y(x') \right) d\mu(x'),$$

where μ is the data distribution and y is the target function. In the sequential infinite-width NTK limit, $\Theta^{(L)}_{\theta_0}$ concentrates and converges in probability to the deterministic kernel $\Theta^{(L)}$, yielding the corresponding deterministic limiting gradient flow driven by $\Theta^{(L)}$ [4].

Assumption B.1 (Homogeneous activation)**.** The activation function σ is positively 1-homogeneous:

$$\sigma(\alpha u) \;=\; \alpha \, \sigma(u) \quad \text{for all } \alpha \geq 0, \; u \in \mathbb{R}.$$

In particular, $\boldsymbol{\sigma}$ satisfies this property. This assumption is used for Fisher–Kibble decoupling (Lemma 5.1) and for separating radial and angular fluctuations in the three-layer kernel.

B.3 EOC scaling and the origin of $1/r$

Let $\Sigma_w = \mathrm{Cov}(w_j^{(\ell)})$. Writing $q^{(\ell)} = \mathbb{E}\|h^{(\ell)}(x)\|^2$ for a fixed unit-norm input, weight independence yields the variance recursion

$$q^{(\ell)} = \frac{\sigma_A^2}{2} \, q^{(\ell-1)} \, \mathrm{Tr}(\Sigma_w),$$

The edge-of-chaos (EOC) condition chooses σ_A so that the linearized coefficient equals 1,

$$\frac{\sigma_A^2}{2} \, \mathrm{Tr}(\Sigma_w) = 1,$$

preventing exploding/vanishing signal propagation across depth. In particular, under $\Sigma_w = I_{d_{\ell-1}} / d_{\ell-1}$ one has $\mathrm{Tr}(\Sigma_w) = 1$ and $\sigma_A = \sqrt{2}$.

In the bottleneck regime $d_{\ell-1} = r$ for $\ell \geq 2$, EOC prescribes $\Sigma_w = I_r / r$. For ReLU, the (normalized) derivative kernel has the closed form

$$\dot{\bar{\Sigma}}^{(\ell)}(\rho) = \mathbb{P}(Z > 0, Z' > 0) = \frac{1}{2} - \frac{\arccos(\rho)}{2\pi} \in \left[0, \frac{1}{2}\right],$$

so $\dot{\Sigma}^{(\ell)}(x, x') = \dot{\bar{\Sigma}}^{(\ell)}(\rho_\ell)$ remains $O(1)$ under EOC and is uniformly bounded by $1/2$. The RF-LR NTK recursion (Theorem 3.1) carries an explicit prefactor $1/r$ at bottleneck layers because gradients propagate through an r-dimensional bottleneck and the corresponding Jacobian metric concentrates on an r-dimensional subspace; see Appendix B.5.1 for the detailed derivation.

B.4 Finite-width effects

Beyond the infinite-width limit, one can in principle study finite-width corrections to the kernel via cumulant/resolvent expansions or diagrammatic methods; see [6] for a Feynman-diagram approach in wide-network asymptotics. These corrections are deferred to future work.

B.5 Proofs of NTK Recursions

B.5.1 Proof of Theorem 3.1: Infinite-width NTK recursion

Proof sketch. Decompose the NTK at layer ℓ into (i) a bias contribution (constant +1), (ii) Term 2: gradient inner products for the layer-ℓ readouts $A^{(\ell)}$, and (iii) Term 1: backpropagated NTK from previous layers times the Jacobian metric. Apply the Law of Large Numbers conditionally on $\boldsymbol{h}^{(\ell-1)}$ as $n_\ell \to \infty$; the sequential limit ensures lower layers have already converged.

Proof. We prove the recursion by decomposing the NTK at layer ℓ into contributions of the trainable parameters up to layer ℓ, and invoking the Law of Large Numbers (LLN) under the sequential infinite-width limit. Let $f^{(\ell)}$ denote the network output truncated at layer ℓ. For finite width, the (random) NTK reads

$$\Theta^{(\ell)}_{\text{rand}}(x,x') \;=\; \underbrace{\big(\nabla_{\theta^{(\ell-1)}} f^{(\ell)}(x)\big)^\top \big(\nabla_{\theta^{(\ell-1)}} f^{(\ell)}(x')\big)}_{\text{Term 1}} \;+\; \underbrace{\sum_{j=1}^{n_\ell} \frac{\partial f^{(\ell)}(x)}{\partial A_j^{(\ell)}} \frac{\partial f^{(\ell)}(x')}{\partial A_j^{(\ell)}}}_{\text{Term 2}}. \tag{13}$$

Bias parameters. We include a trainable layer bias $c^{(\ell)}$ (initialized at zero). The gradient feature with respect to $c^{(\ell)}$ is constant, so it contributes an additive constant term to the NTK; under our normalization this contribution is 1. This yields the additive +1 term in the recursion.

Term 2 (layer-ℓ output weights $A^{(\ell)}$). By the forward pass, $\partial f^{(\ell)}(x)/\partial A_j^{(\ell)} = \frac{\sigma_A}{\sqrt{n_\ell}}\, \boldsymbol{\sigma}\big(\boldsymbol{w}_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x)\big)$. Therefore

$$\text{Term 2} \;=\; \frac{\sigma_A^2}{n_\ell} \sum_{j=1}^{n_\ell} \boldsymbol{\sigma}\big(\boldsymbol{w}_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x)\big)\; \boldsymbol{\sigma}\big(\boldsymbol{w}_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x')\big). \tag{14}$$

Conditionally on $\boldsymbol{h}^{(\ell-1)}$, the summands are i.i.d. over $\boldsymbol{w}_j^{(\ell)}$. As $n_\ell \to \infty$ (with lower layers already taken to their limits), LLN yields

$$\text{Term 2} \xrightarrow{p} \sigma_A^2\, \Sigma^{(\ell)}(x,x'). \tag{15}$$

Term 1 (previous layers). By the chain rule,

$$\nabla_{\theta^{(\ell-1)}} f^{(\ell)}(x) \;=\; \Big(\frac{\partial f^{(\ell)}(x)}{\partial \boldsymbol{h}^{(\ell-1)}(x)}\Big)\, \nabla_{\theta^{(\ell-1)}} \boldsymbol{h}^{(\ell-1)}(x). \tag{16}$$

Hence Term 1 factorizes as $\Theta^{(\ell-1)}_{\text{rand}}(x,x')$ multiplied by a Jacobian metric:

$$\text{Term 1} \;=\; \Theta^{(\ell-1)}_{\text{rand}}(x,x') \;\cdot\; \Big\langle \frac{\partial f^{(\ell)}}{\partial \boldsymbol{h}^{(\ell-1)}}(x), \frac{\partial f^{(\ell)}}{\partial \boldsymbol{h}^{(\ell-1)}}(x') \Big\rangle. \tag{17}$$

Taking the inner product and applying LLN as $n_\ell \to \infty$ yields

$$\Big\langle \frac{\partial f^{(\ell)}}{\partial \boldsymbol{h}^{(\ell-1)}}(x), \frac{\partial f^{(\ell)}}{\partial \boldsymbol{h}^{(\ell-1)}}(x') \Big\rangle \xrightarrow{p} \sigma_A^2\, \dot\Sigma^{(\ell)}(x,x'). \tag{18}$$

Therefore,

$$\text{Term 1} \xrightarrow{p} \Theta^{(\ell-1)}(x,x') \cdot \sigma_A^2\, \dot\Sigma^{(\ell)}(x,x'). \tag{19}$$

Combining the bias contribution with the two limits completes the recursion. For bottleneck layers $\ell \geq 2$, the Jacobian and output contributions scale with $1/r$ (from the r-dimensional bottleneck); we absorb σ_A^2 into the normalization. Thus

$$\Theta^{(1)}(x,x') \;=\; 1 \;+\; \Sigma^{(1)}(x,x'), \tag{20}$$

$$\Theta^{(\ell)}(x,x') \;=\; 1 \;+\; \frac{1}{r}\, \Theta^{(\ell-1)}(x,x') \cdot \sigma_A^2\, \dot\Sigma^{(\ell)}(x,x') \;+\; \frac{1}{r}\, \sigma_A^2\, \Sigma^{(\ell)}(x,x') \quad (\ell \geq 2). \tag{21}$$

The sequential limit assumption ensures that lower-layer random objects have already converged when taking the next-layer limit, justifying the conditional LLN applications. □

B.5.2 Proof of Theorem 2.1: Gaussian process composition

Proof sketch. By induction on ℓ: for $\ell = 0$ the output is deterministic; for $\ell \geq 1$, conditionally on $\boldsymbol{h}^{(\ell-1)}$ the forward pass is a sum of i.i.d. terms over j, so the multivariate CLT yields Gaussian convergence. The covariance matches the base kernel $\Sigma^{(\ell)}$ by definition.

Proof. We prove by induction on ℓ that $\boldsymbol{h}^{(\ell)}$ converges to a Gaussian process with covariance kernel $\Sigma^{(\ell)}$.

Base case $\ell = 0$. For $\ell = 0$, we have $\boldsymbol{h}^{(0)}(x) = x$, which is deterministic and trivially a Gaussian process.

Inductive step. Assume that $\boldsymbol{h}^{(\ell-1)}$ converges to a Gaussian process with covariance kernel $\Sigma^{(\ell-1)}$. Consider the forward pass at layer ℓ:

$$\boldsymbol{h}^{(\ell)}(x) = \frac{1}{\sqrt{\boldsymbol{n}_\ell}} \sum_{j=1}^{\boldsymbol{n}_\ell} A_j^{(\ell)} \ \boldsymbol{\sigma}\left(w_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x)\right) + c^{(\ell)}. \tag{22}$$

For any finite collection of inputs $\{x_1, \ldots, x_m\}$, conditionally on $\boldsymbol{h}^{(\ell-1)}$, the terms $A_j^{(\ell)} \ \boldsymbol{\sigma}\left(w_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x_i)\right)$ are independent across j and identically distributed. Since $A_j^{(\ell)} \sim \mathcal{N}(0, \sigma_A^2)$ and $w_j^{(\ell)}$ are frozen Gaussian draws, the conditional distribution of each term is centered with finite variance (and $c^{(\ell)}$ is deterministic at initialization).

By the multivariate Central Limit Theorem, as $\boldsymbol{n}_\ell \to \infty$, the vector $(\boldsymbol{h}^{(\ell)}(x_1), \ldots, \boldsymbol{h}^{(\ell)}(x_m))$ converges in distribution to a multivariate Gaussian. The covariance is given by

$$\mathbb{E}\Big[\boldsymbol{h}^{(\ell)}(x_i)\, \boldsymbol{h}^{(\ell)}(x_j)\Big] = \mathbb{E}_w\Big[\ \boldsymbol{\sigma}\left(w^\top \boldsymbol{h}^{(\ell-1)}(x_i)\right) \ \boldsymbol{\sigma}\left(w^\top \boldsymbol{h}^{(\ell-1)}(x_j)\right)\Big] = \Sigma^{(\ell)}(x_i, x_j), \tag{23}$$

where the expectation is taken over the frozen weights w and the limiting Gaussian process $\boldsymbol{h}^{(\ell-1)}$. The last equality follows from the definition of the base kernel in Definition 2.1.

The sequential infinite-width limit ensures that $\boldsymbol{h}^{(\ell-1)}$ has already converged to its Gaussian process limit when taking the limit for layer ℓ, justifying the application of the CLT conditionally on $\boldsymbol{h}^{(\ell-1)}$. □

B.5.3 Proof of three-layer NTK Corollary

Proof sketch. Apply the recursion (Theorem 3.1) for $\ell = 1$ and $\ell = 2$; for $\ell = 1$ use $\boldsymbol{\Theta}^{(0)} = 0$; for $\ell = 2$ substitute $\boldsymbol{\Theta}^{(1)} = 1 + \Sigma^{(1)}$. Under homogeneity, the mean form involves expectations over Fisher (angular) and Kibble (radial); as $r \to \infty$ these concentrate, recovering K_∞.

Proof. Apply Theorem 3.1 for $\ell = 1$ and $\ell = 2$. For $\ell = 1$, since $\boldsymbol{\Theta}^{(0)} = 0$, the recursion gives:

$$\boldsymbol{\Theta}^{(1)}(x, x') = 1 + \boldsymbol{\Theta}^{(0)}(x, x') \cdot \dot{\Sigma}^{(1)}(x, x') + \Sigma^{(1)}(x, x') = 1 + \Sigma^{(1)}(x, x'). \tag{24}$$

For $\ell = 2$, substitute $\boldsymbol{\Theta}^{(1)} = 1 + \Sigma^{(1)}$ into the recursion:

$$\boldsymbol{\Theta}^{(2)}(x, x') = 1 + \frac{1}{r}\, \boldsymbol{\Theta}^{(1)}(x, x') \cdot \dot{\Sigma}^{(2)}(x, x') + \frac{1}{r}\, \Sigma^{(2)}(x, x'). \tag{25}$$

This is the stated three-layer recursion with the bias contribution.

Mean explicit form (low-rank). Under the homogeneous ReLU setting on the sphere, the mean three-layer NTK (with $1/r$ factors) is

$$\tilde{\boldsymbol{\Theta}}^{(2)}(\rho) \ = \ 1 \ + \ \frac{1}{r}\, \boldsymbol{\Theta}^{(1)}(\rho)\, \mathbb{E}\left[1 - \frac{\arccos(\hat{\rho}_r)}{\pi}\right] \ + \ \frac{1}{r}\, \mathbb{E}\left[\Sigma^{(1)}(\hat{\rho}_r)\, \|x_1\| \|y_1\|\right], \tag{26}$$

where the expectations are over Fisher and Kibble. As $r \to \infty$, $\hat{\rho}_r \to \rho$ and $\|x_1\| \|y_1\| / r \to 1$, so the limit $K_\infty(\rho) = \Theta^{(1)}(\rho)(1 - \arccos(\rho)/\pi) + \Sigma^{(1)}(\rho) + 1$ is recovered (without the $1/r$ prefactors, since the low-rank scaling is absorbed in the limit). □

B.5.4 Proof of General L-layer NTK Corollary

Remark B.1 (Explicit $2L$-term unrolling). Equivalently, expanding the recursion yields an explicit *linear* number of terms (namely $2L$). For example, for $L = 3$:

$$\begin{aligned}\boldsymbol{\Theta}^{(3)}(x,x') = 1 + \frac{1}{r}\dot{\Sigma}^{(2)} \cdot \dot{\Sigma}^{(3)} + \frac{1}{r}\dot{\Sigma}^{(3)} \\ + \frac{1}{r^2}\Sigma^{(1)} \cdot \dot{\Sigma}^{(2)} \cdot \dot{\Sigma}^{(3)} + \frac{1}{r}\Sigma^{(2)} \cdot \dot{\Sigma}^{(3)} + \frac{1}{r}\Sigma^{(3)}.\end{aligned} \tag{27}$$

Each bottleneck layer contributes a factor $1/r$; layer 1 has no such factor. In total, there are $2L$ terms for depth L.

Proof sketch. By induction on L: for $L = 1$ the recursion gives $\boldsymbol{\Theta}^{(1)} = 1 + \Sigma^{(1)}$. For $L \geq 2$, expand $\boldsymbol{\Theta}^{(L)}$ via the recursion into a bias term and two paths (derivative-propagated and fresh-base); each path contributes L terms with the appropriate $1/r$ factors, yielding $2L$ terms in total.

Proof. We prove the explicit form by induction on L, accounting for the +1 bias terms in the recursion.

Base case $L = 1$. From Theorem 3.1, $\boldsymbol{\Theta}^{(1)} = 1 + \Sigma^{(1)}$. The formula with $L = 1$ gives:

$$\boldsymbol{\Theta}^{(1)}(x,x') = \sum_{\ell=1}^{1}\left(\prod_{k=\ell+1}^{1}\dot{\Sigma}^{(k)}(x,x')\right) + \sum_{\ell=1}^{1}\Sigma^{(\ell)}(x,x')\prod_{k=\ell+1}^{1}\dot{\Sigma}^{(k)}(x,x'). \tag{28}$$

For the first sum: when $\ell = 1$, $\prod_{k=2}^{1}\dot{\Sigma}^{(k)} = 1$ (empty product). For the second sum: when $\ell = 1$, $\Sigma^{(1)} \cdot \prod_{k=2}^{1}\dot{\Sigma}^{(k)} = \Sigma^{(1)} \cdot 1 = \Sigma^{(1)}$. Therefore, $\boldsymbol{\Theta}^{(1)} = 1 + \Sigma^{(1)}$, which matches the recursion.

Inductive step. Assume the formula holds for depth $L - 1$:

$$\boldsymbol{\Theta}^{(L-1)}(x,x') = \sum_{\ell=1}^{L-1}\left(\prod_{k=\ell+1}^{L-1}\dot{\Sigma}^{(k)}(x,x')\right) + \sum_{\ell=1}^{L-1}\Sigma^{(\ell)}(x,x')\prod_{k=\ell+1}^{L-1}\dot{\Sigma}^{(k)}(x,x'). \tag{29}$$

By Theorem 3.1, for $L \geq 2$,

$$\boldsymbol{\Theta}^{(L)}(x,x') = 1 + \frac{1}{r}\,\boldsymbol{\Theta}^{(L-1)}(x,x') \cdot \dot{\Sigma}^{(L)}(x,x') + \frac{1}{r}\,\Sigma^{(L)}(x,x'). \tag{30}$$

Substituting the inductive hypothesis and distributing the $1/r$ factor yields the stated closed form; the first sum gains a new term 1 (from $\ell = L$) and each existing term is multiplied by $(1/r)\dot{\Sigma}^{(L)}$, while the second sum gains $(1/r)\Sigma^{(L)}$ and each existing term is multiplied by $(1/r)\dot{\Sigma}^{(L)}$. The first term 1 corresponds to $\ell = L$ in the first sum (with empty product $\prod_{k=L+1}^{L}\dot{\Sigma}^{(k)} = 1$). The last term $\Sigma^{(L)}$ corresponds to $\ell = L$ in the second sum (with empty product $\prod_{k=L+1}^{L}\dot{\Sigma}^{(k)} = 1$). Therefore:

$$\boldsymbol{\Theta}^{(L)}(x,x') = \sum_{\ell=1}^{L}\left(\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)}(x,x')\right) + \sum_{\ell=1}^{L}\Sigma^{(\ell)}(x,x')\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)}(x,x'), \tag{31}$$

completing the induction. The total number of terms is $L + L = 2L$, which grows linearly with depth. □

B.5.5 Reference: full-width MLP limiting NTK at the edge of chaos

Proposition B.1 (Closed form for full-width MLP EOC NTK)**.** *Let $l \geq 1$ and consider an infinitely wide l-layer MLP with (a,b)-ReLU activation initialized at the edge of chaos as in [**?**]. Denote by ϱ the corresponding cosine map and by $\rho_1(x,x')$ the input cosine similarity. Then the limiting NTK satisfies*

$$K_\infty(x,x') \;=\; \|x\|\,\|x'\|\left(\sum_{k=1}^{l}\varrho^{\circ(k-1)}\big(\rho_1(x,x')\big)\prod_{k'=k}^{l-1}\varrho'\Big(\varrho^{\circ(k'-1)}\big(\rho_1(x,x')\big)\Big)\right) I_{m_l},$$

with the convention that the empty product (for $k = l$) equals 1*. See [**?**] for the derivation.*

B.6 Comparison with the classical fully-trained NTK recursion

For comparison, consider a standard fully-connected MLP where *all* weights (and biases) are trained under NTK parameterization. In the sequential infinite-width limit, the classical NTK recursion takes the form [4]

$$\Theta^{(0)}_{\mathrm{MLP}}(x,x') = 0, \qquad \Theta^{(\ell)}_{\mathrm{MLP}}(x,x') = \Theta^{(\ell-1)}_{\mathrm{MLP}}(x,x')\,\dot{K}^{(\ell)}(x,x') + K^{(\ell)}(x,x'), \quad \ell \geq 1, \tag{32}$$

where $K^{(\ell)}$ is the (deterministic) NNGP kernel at layer ℓ and $\dot{K}^{(\ell)}$ the associated derivative kernel (both obtained from the signal-propagation recursion). Unrolling (32) yields one contribution per layer (a sum of L terms, each propagated through subsequent $\dot{K}^{(k)}$ factors).

In contrast, the RF-LR recursion of Theorem 3.1 (i) carries explicit $1/r$ prefactors at bottleneck layers, and (ii) uses conditional base/derivative kernels $\Sigma^{(\ell)}, \dot{\Sigma}^{(\ell)}$ rather than the classical NNGP kernels $K^{(\ell)}, \dot{K}^{(\ell)}$. These differences lead to the $2L$-term expansion in Corollary 3.1, reflecting separate bias-path and fresh-basis contributions, each propagated through subsequent derivative kernels and weighted by bottleneck factors.

Relation to fully trained MLPs at the edge of chaos. For fully trained (full-width) MLPs at the edge of chaos, the limiting NTK admits a single-layer sum structure (one contribution per layer) without RF-LR bottleneck prefactors; see Proposition B.1 for a reference closed form [**?**]. In contrast, the RF-LR expansion (7) contains $2L$ terms because the recursion has both bias-path and fresh-basis-path contributions, each propagated through subsequent derivative kernels and weighted by explicit $1/r$ factors at bottlenecks.

B.7 Probabilistic recursion, GP concatenation, and bottleneck scaling

Random correlation chain and Fisher's law. At each layer $\ell \geq 2$, the pair $(\boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x'))$ is a $2r$-dimensional Gaussian conditional on lower-layer randomness. The *layer-ℓ sample correlation*

$$\rho_\ell \;=\; \cos\angle\big(\boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x')\big) \;=\; \frac{\langle \boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x')\rangle}{\|\boldsymbol{h}^{(\ell-1)}(x)\|\,\|\boldsymbol{h}^{(\ell-1)}(x')\|}$$

is a *random variable*. Given the population correlation $\rho = \varrho(\rho_{\ell-1})$ (the conditional mean under the ReLU cosine map), Fisher's law [15, 16] and Lemma C.2 give $\mathbb{E}[(\rho_\ell - \rho)^2] \leq C/r$ and $\mathbb{P}(|\rho_\ell - \rho| \geq t) \leq 6\exp(-c\,r\,t^2)$. Thus the chain $\rho_1 \to \rho_2 \to \cdots \to \rho_L$ is a *Markov chain of random variables* with $O(1/\sqrt{r})$ fluctuations at each transition. The NTK recursion passes through these random correlations successively: $\Sigma^{(\ell)} = \bar{\Sigma}^{(\ell)}(\rho_\ell)$, $\dot{\Sigma}^{(\ell)} = \bar{\dot{\Sigma}}^{(\ell)}(\rho_\ell)$ with random ρ_ℓ. The deterministic proxy (Definition D.4) replaces this random path by the deterministic iterates $\rho_k = \varrho^{\circ(k-1)}(\rho_1)$, which represent the mean/infinite-r limit path.

Kernel randomness and GP concatenation. For $\ell > 1$, $\Sigma^{(\ell)}$ and $\dot{\Sigma}^{(\ell)}$ are random fields because they depend on $\boldsymbol{h}^{(\ell-1)}$, which is the output of a (conditional) Gaussian process. By Theorem 2.1, the hidden states form a composition of GPs: each layer's output is conditionally Gaussian given the previous layer, but the unconditional law is a deep GP. In the bottleneck regime, the pair $(\boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x'))$ is a $2r$-dimensional Gaussian conditional on lower-layer randomness, with covariance determined by $\Sigma^{(\ell-1)}$. The base kernel $\Sigma^{(\ell)}$ and derivative kernel $\dot{\Sigma}^{(\ell)}$ are thus expectations over the frozen features $\boldsymbol{w}^{(\ell)}$, conditional on this Gaussian pair, and inherit randomness through $\Sigma^{(\ell-1)}$.

Bottleneck scaling and explicit ReLU bounds. At a bottleneck layer $\ell \geq 2$, the RF-LR NTK recursion carries an explicit prefactor $1/r$ (Theorem 3.1). For ReLU and isotropic weights with $\mathrm{Cov}(w) = I_r/r$, one has

$$\dot{\Sigma}^{(\ell)}(x,x') \;=\; \mathbb{E}_w\Big[\dot{\sigma}\big(w^\top \boldsymbol{h}^{(\ell-1)}(x)\big)\,\dot{\sigma}\big(w^\top \boldsymbol{h}^{(\ell-1)}(x')\big)\,\|w\|^2\Big] \;=\; \bar{\dot{\Sigma}}^{(\ell)}(\rho_\ell),$$

where $\rho_\ell = \cos\angle(\boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x'))$ and $\bar{\dot{\Sigma}}^{(\ell)}(\rho)$ denotes the corresponding scalar derivative kernel evaluated at cosine similarity $\rho \in [-1,1]$. Under the EOC normalization $\mathrm{Tr}(\mathrm{Cov}(w)) = 1$, the ReLU derivative kernel has the explicit form

$$\bar{\dot{\Sigma}}^{(\ell)}(\rho) = \mathbb{P}(Z > 0, Z' > 0) = \frac{1}{2} - \frac{\arccos(\rho)}{2\pi} \in \Big[0, \frac{1}{2}\Big],$$

where (Z, Z') is centered Gaussian with correlation ρ. In particular, $\sup_{\rho\in[-1,1]} \bar{\dot{\Sigma}}^{(\ell)}(\rho) = 1/2$, uniformly in ℓ. Similarly, $\Sigma^{(\ell)}(x,x') = \mathbb{E}_w[\sigma(w^\top \boldsymbol{h}^{(\ell-1)}(x))\,\sigma(w^\top \boldsymbol{h}^{(\ell-1)}(x'))]$ is an $O(1)$ scalar kernel, and the $1/r$ factors enter through the recursion.

Probabilistic recursion (nested expectations). The mean (probabilistic) NTK $\tilde{\boldsymbol{\Theta}}^{(L)} = \mathbb{E}[\boldsymbol{\Theta}^{(L)}]$ satisfies a recursion involving expectations of products:

$$\tilde{\boldsymbol{\Theta}}^{(1)} \;=\; 1 \;+\; \mathbb{E}\big[\Sigma^{(1)}\big], \qquad \tilde{\boldsymbol{\Theta}}^{(\ell)} \;=\; 1 \;+\; \frac{1}{r}\,\mathbb{E}\big[\boldsymbol{\Theta}^{(\ell-1)}\cdot\dot{\Sigma}^{(\ell)}\big] \;+\; \frac{1}{r}\,\mathbb{E}\big[\Sigma^{(\ell)}\big], \quad \ell \geq 2. \tag{33}$$

To emphasize the nested structure, write $\mathbb{E}_{\leq \ell-1}$ for expectation over all randomness up to depth $\ell-1$, and $\mathbb{E}_{w^{(\ell)}}[\cdot \mid \boldsymbol{h}^{(\ell-1)}]$ for expectation over the frozen feature direction $w^{(\ell)}$ at layer ℓ, conditional on the previous hidden states. Then the mean kernels are iterated conditional expectations:

$$\mathbb{E}\big[\Sigma^{(\ell)}(x,x')\big] \;=\; \mathbb{E}_{\leq \ell-1}\Big[\mathbb{E}_{w^{(\ell)}}\big[\sigma(w^\top \boldsymbol{h}^{(\ell-1)}(x))\,\sigma(w^\top \boldsymbol{h}^{(\ell-1)}(x')) \;\big|\; \boldsymbol{h}^{(\ell-1)}\big]\Big],$$

$$\mathbb{E}\big[\boldsymbol{\Theta}^{(\ell-1)}(x,x')\,\dot{\Sigma}^{(\ell)}(x,x')\big] \;=\; \mathbb{E}_{\leq \ell-1}\Big[\boldsymbol{\Theta}^{(\ell-1)}(x,x')\cdot \mathbb{E}_{w^{(\ell)}}\big[\dot\sigma(w^\top \boldsymbol{h}^{(\ell-1)}(x))\,\dot\sigma(w^\top \boldsymbol{h}^{(\ell-1)}(x'))\,\|w\|^2 \;\big|\; \boldsymbol{h}^{(\ell-1)}\big]\Big].$$

Since $\boldsymbol{\Theta}^{(\ell-1)}$ and $\Sigma^{(\ell)}, \dot{\Sigma}^{(\ell)}$ share the randomness of $\boldsymbol{h}^{(\ell-1)}$, one has $\mathbb{E}[\boldsymbol{\Theta}^{(\ell-1)}\cdot\dot{\Sigma}^{(\ell)}] \neq \tilde{\boldsymbol{\Theta}}^{(\ell-1)}\cdot\mathbb{E}[\dot{\Sigma}^{(\ell)}]$ in general; controlling this dependence (and the resulting nested conditional expectations) is the main technical point for a rigorous deep mean-kernel theory.

Exponential depth suppression (ReLU, bottlenecks). Assume for simplicity that all layers $\ell \geq 2$ are bottlenecks with the same rank r. For ReLU under EOC we have the explicit uniform bound

$$\bar{\dot{\Sigma}}^{(\ell)}(\rho) \leq c_0 := \frac{1}{2}, \qquad \forall \rho \in [-1,1],\ \forall \ell \geq 2.$$

From the closed form (7), the contribution from layer ℓ to $\boldsymbol{\Theta}^{(L)}$ is multiplied by $\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)}$. Writing $\dot{\Sigma}^{(k)} = \bar{\dot{\Sigma}}^{(k)}(\rho_k)$, we obtain

$$\frac{1}{r^{L-\ell}}\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)} \;\leq\; \frac{1}{r^{L-\ell}}\prod_{k=\ell+1}^{L}\bar{\dot{\Sigma}}^{(k)}(\rho_k) \;\leq\; \frac{c_0^{L-\ell}}{r^{L-\ell}}.$$

Hence, contributions from early layers ($\ell \ll L$) decay exponentially in $L-\ell$ when $c_0/r < 1$, and the NTK becomes increasingly localized to the top of the network as depth grows.

Proposition B.2 (Effective depth decomposition and a $\log r$ window)**.** *Assume the uniform derivative-kernel bound $\dot{\Sigma}^{(\ell)}(x,x') \leq c_0$ for all $\ell \geq 2$ and all input pairs (x,x'), and assume that for a given (x,x') one has a uniform base-kernel bound*

$$|\Sigma^{(\ell)}(x,x')| \leq C_\Sigma(x,x'), \qquad \forall \ell \in \{1,\dots,L\}.$$

Fix $L \geq 2$, $r > c_0$, and an integer $m \in \{1,\dots,L\}$. Define the top-m truncation $\boldsymbol{\Theta}^{(L)}_{\mathrm{top},m}(x,x')$ *by restricting both sums in the explicit expansion* (7) *to indices $\ell \in \{L-m+1,\dots,L\}$. Then*

$$\big|\boldsymbol{\Theta}^{(L)}(x,x') - \boldsymbol{\Theta}^{(L)}_{\mathrm{top},m}(x,x')\big| \;\leq\; \frac{1+C_\Sigma(x,x')}{1-c_0/r}\left(\frac{c_0}{r}\right)^m.$$

In particular, taking $m = \lceil \alpha \log r \rceil$ (for any $\alpha > 0$) yields a $\log r$ effective-depth window: the contribution of layers $\ell \leq L - \lceil \alpha \log r \rceil$ is at most

$$\big|\boldsymbol{\Theta}^{(L)}(x,x') - \boldsymbol{\Theta}^{(L)}_{\mathrm{top},\lceil \alpha \log r\rceil}(x,x')\big| \;\leq\; \frac{1+C_\Sigma(x,x')}{1-c_0/r}\left(\frac{c_0}{r}\right)^{\lceil \alpha \log r\rceil} \;=\; o(r^{-p})$$

for every fixed $p > 0$ as $r \to \infty$.

Proof sketch. The remainder is the sum of terms in (7) with $\ell \leq L-m$; each such term carries a factor $\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)} \leq c_0^{L-\ell}$ and hence $(c_0/r)^{L-\ell}$. Summing over $\ell \leq L-m$ yields a geometric-series tail; substituting $m = \lceil \alpha \log r \rceil$ gives the stated decay.

Proof. By definition of $\boldsymbol{\Theta}^{(L)}_{\mathrm{top},m}$, the remainder is the sum of the terms in (7) with $\ell \leq L-m$. For such ℓ, we have $L-\ell \geq m$, and by the derivative-kernel bound $\dot{\Sigma}^{(k)}(x,x') \leq c_0$ we obtain

$$\frac{1}{r^{L-\ell}}\prod_{k=\ell+1}^{L}\dot{\Sigma}^{(k)}(x,x') \;\leq\; \left(\frac{c_0}{r}\right)^{L-\ell}.$$

Moreover, for $\ell \le L - m \le L - 1$ the base-kernel term carries the same prefactor $r^{-(L-\ell)}$ in (7), hence

$$\frac{1}{r^{L-\ell}} \left|\Sigma^{(\ell)}(x,x')\right| \prod_{k=\ell+1}^{L} \dot{\Sigma}^{(k)}(x,x') \;\le\; C_\Sigma(x,x') \left(\frac{c_0}{r}\right)^{L-\ell}.$$

Summing these bounds over $\ell \le L - m$ and using a geometric-series tail gives

$$\left|\mathbf{\Theta}^{(L)}(x,x') - \mathbf{\Theta}^{(L)}_{\mathrm{top},m}(x,x')\right| \;\le\; \left(1 + C_\Sigma(x,x')\right) \sum_{j=m}^{\infty} \left(\frac{c_0}{r}\right)^{j} \;=\; \frac{1 + C_\Sigma(x,x')}{1 - c_0/r} \left(\frac{c_0}{r}\right)^{m},$$

as claimed. The $\log r$ specialization follows by substituting $m = \lceil \alpha \log r \rceil$ and observing that $(c_0/r)^{\lceil \alpha \log r \rceil} = o(r^{-p})$ for every fixed $p > 0$. □

B.8 Comparison table: depth/rank effects in the kernel regime

B.9 Techniques and proof sketch

We highlight the main proof ingredients and how they combine.

Recursion and closed form. The RF-LR NTK recursion (Theorem 3.1) follows from a layerwise decomposition of the gradient inner product into bias, readout, and backpropagated contributions, together with conditional law of large numbers arguments under the sequential infinite-width limit [4]. Iterating the recursion yields the explicit $2L$-term expansion (Corollary 3.1).

Microscopic decoupling and concentration. For three layers, the empirical kernel depends on the sample cosine and norm product of rank-r Gaussian projections. Rotational invariance yields Fisher–Kibble decoupling (Lemma 5.1), separating angular and radial fluctuations [15, 16, 49]. Concentration of Gaussian norms and sample correlations then yields sub-Gaussian deviations in r for the empirical kernel (Theorem C.1 and Corollary C.2).

RKHS identification via endpoint expansions. The RKHS of zonal kernels on the sphere is controlled by the Puiseux behavior near $\rho = \pm 1$ [17]. We compute the mean three-layer kernel's endpoint expansion by combining Fisher's density with a hypergeometric connection formula, showing that the leading $t^{1/2}$ exponent is preserved after taking expectations, hence the mean kernel induces the same RKHS as the shallow ReLU kernel (Corollary 5.2).

Depth scaling and effective depth. Depth dependence is controlled by correlation propagation and by products of derivative kernels. In the bottleneck regime, the RF-LR recursion carries explicit $1/r$ prefactors at each bottleneck layer, and for ReLU one has $\sup_\rho \bar{\dot{\Sigma}}^{(\ell)}(\rho) \le 1/2$, which yields exponential suppression of early-layer contributions as depth increases (Section 4). A refined mean-kernel statement is obtained by analyzing the deterministic proxy recursion (Theorem 4.1).

Future work: quadratic random-matrix scaling. Establishing bulk spectral laws for the *empirical* RF-LR NTK in a quadratic scaling regime $n \to \infty$ requires additional resolvent/local-law inputs. The Benigni–Paquette framework [23] establishes RMT limits for Gram spectra in the extensive-width regime ($n \sim N^2$) with layers taken in the RMT limit, via resolvent replacement and Gaussian-equivalence machinery for shallow two-layer NTKs. Adapting these methods to the present low-rank random-feature structure and to depth $L \ge 2$ is left for future work.

C Proofs of low-rank NTK RKHS

C.1 Background and statements for Section 5

RKHS viewpoint in the NTK regime. In the NTK regime, training dynamics linearize:

$$f_t(x) \approx f_0(x) + \langle \nabla_\theta f_0(x),\, \theta_t - \theta_0 \rangle.$$

The learned function stays in the RKHS induced by the (limiting) kernel Θ, with norm

$$\|f\|^2_{\mathcal{H}_\Theta} = \langle f, \Theta^{-1} f \rangle_{L^2}.$$

For RF-LR with frozen random features, universal approximation holds due to full support of the frozen Gaussian draws $w^{(\ell)}$, so any function in the Barron space [50] can be approximated with error $O(1/\sqrt{N})$. Classical NTK analysis suggests that feature weights move little during training; RF-LR enforces this by design by freezing the feature weights, which isolates kernel-regime behavior without an additional weight-deviation analysis.

Mean RKHS and random features. Each realization ω of the random features yields a kernel $K_\omega(x, x') = \Theta^{(2)}_\omega(x, x')$ and an induced RKHS $\mathcal{H}_\omega$. The mean kernel $\tilde{K}(x, x') = \mathbb{E}_\omega[K_\omega(x, x')]$ is deterministic; its RKHS $\mathcal{H}_{\tilde{K}}$ is the RKHS of $\tilde{K}$. It is not $\mathbb{E}[\mathcal{H}_\omega]$ (expectation of Hilbert spaces is undefined). For each ω, the feature map $\varphi_\omega(x) = K_\omega(\cdot, x) \in \mathcal{H}_\omega$ is random, whereas the canonical feature map of the mean kernel $\varphi_{\tilde{K}}(x) = \tilde{K}(\cdot, x) \in \mathcal{H}_{\tilde{K}}$ is deterministic. The mean RKHS describes the typical function space obtained after averaging over random feature realizations.

Remark (application of Bietti–Bach). Once the mean kernel's Puiseux endpoint behavior is established (Appendix C.7), the RKHS equivalence in Corollary 5.2 follows directly from the characterization theorem of [17].

Corollary C.1 (Mean NTK over Fisher–Kibble has same RKHS: three-layer case)**.** *Under Assumption B.1 and the RF-LR setting with ReLU nonlinearity and isotropic random features on $\mathbb{S}^{d-1}$, the mean three-layer NTK $\tilde{\mathbf{\Theta}}^{(2)}$ (obtained by taking expectations over Fisher and Kibble distributions) is a zonal kernel that induces the same RKHS as the shallow ReLU kernel. In particular, the RKHSs coincide as sets with equivalent norms.*

Proof: See Appendix C.7.

Main RKHS result (stated in the main text). Corollary 5.2 is stated in Section 5. It follows by combining the Puiseux endpoint analysis in Appendix C.7 with the RKHS characterization of [17].

C.2 Discussion on Fisher and Kibble Distributions

Lemma C.1 (Fisher–Kibble decoupling)**.** *Let x, y be input vectors and let x_1, y_1 denote their rank-r random projections. Define the empirical correlation $\rho_1 = \cos\angle(x_1, y_1)$ and squared norms $u = \|x_1\|^2, v = \|y_1\|^2$. Then:*

- *ρ_1 follows Fisher's correlation distribution [15] [16], centered at the true correlation ρ.*
- *(u, v) follow Kibble's [49] bivariate Gamma (chi-square) law.*
- *Angular and radial parts are independent: $p(\rho_1, u, v) = p_{Fisher}(\rho_1)\, p_{Kibble}(u, v)$.*

Remark C.1 (Fisher–Kibble decoupling and independence). By rotational invariance, the empirical correlation and squared norms satisfy

$$\rho_1 \sim \text{Fisher}(\rho, r), \qquad (u, v) \sim \text{Kibble}(r, \rho), \qquad \text{and} \qquad p(\rho_1, u, v) = p_{\text{Fisher}}(\rho_1)\, p_{\text{Kibble}}(u, v).$$

Homogeneity and factorization (why angular and radial parts decouple). For positively 1-homogeneous activations (Assumption B.1), the first-layer base kernel separates radial and angular dependence: for all $x, x' \neq 0$,

$$\Sigma^{(1)}(x, x') = \|x\|\,\|x'\|\,\tilde{\Sigma}^{(1)}(\rho), \qquad \rho = \frac{\langle x, x'\rangle}{\|x\|\,\|x'\|},$$

where $\tilde{\Sigma}^{(1)}$ is a scalar function on $[-1, 1]$ (for ReLU it is the standard arc-cosine kernel). Under isotropic Gaussian projections, rotational invariance implies that the empirical correlation $\rho_1 = \cos\angle(x_1, y_1)$ depends only on the angular part of the projected pair, while the norms $(\|x_1\|, \|y_1\|)$ depend only on the radial part; consequently ρ_1 is independent of $(u, v) = (\|x_1\|^2, \|y_1\|^2)$ and the joint law factorizes as stated above. The Fisher distribution density [15] is

$$p(\rho_1 \mid \rho, r) = \frac{(r-2)\,\Gamma(r-1)\,(1-\rho^2)^{\frac{r-1}{2}}\,(1-\rho_1^2)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(1-\rho\rho_1)^{r-\frac{3}{2}}}\,{}_2F_1\Big(\tfrac{1}{2}, \tfrac{1}{2};\, r-\tfrac{1}{2};\, \tfrac{1+\rho\rho_1}{2}\Big), \qquad r > 2.$$

For large r, the inverse Gudermann function applied to the complementary angle is approximately normal:

$$\text{gd}^{-1}\Big(\frac{\pi}{2} - \arccos(\rho_1)\Big) \approx \mathcal{N}\Big(\text{arctanh}(\rho),\, \frac{1}{r-3}\Big), \qquad \text{hence} \qquad \text{Var}(\rho_1) = O(1/r).$$

Kibble's [49] bivariate chi-square law for $(u, v) = (\|x_1\|^2, \|y_1\|^2)$ has density

$$f(u,v) = \frac{(uv)^{\frac{r/2-1}{2}} \exp\left(-\frac{u+v}{2(1-\rho^2)}\right)}{\Gamma(r/2)\left(2(1-\rho^2)\right)^{\frac{r}{2}+1} \rho^{\frac{r/2-1}{2}}} I_{\frac{r}{2}-1}\left(\frac{\rho\sqrt{uv}}{1-\rho^2}\right), \qquad u, v \geq 0,$$

where I_ν is the modified Bessel function of the first kind. The moments are

$$\mathbb{E}[u] = \mathbb{E}[v] = r, \qquad \mathrm{Var}(u) = \mathrm{Var}(v) = 2r, \qquad \mathrm{Cov}(u,v) = 2r\,\rho^2.$$

Defining $w_r = \sqrt{uv}/r$, concentration results imply $\mathbb{E}[w_r] \to 1$ and $\mathrm{Var}(w_r) = O(1/r)$.

Compact three-layer empirical NTK representation. In the three-layer low-rank setting, denote

$$\rho_1 = \cos\angle(\boldsymbol{h}^{(1)}(x), \boldsymbol{h}^{(1)}(x')), \qquad w_r = \frac{\|\boldsymbol{h}^{(1)}(x)\|\,\|\boldsymbol{h}^{(1)}(x')\|}{r}.$$

Then the empirical three-layer NTK admits the compact form

$$\boldsymbol{\Theta}^{(2)}(x,x') = 1 + \frac{1}{r}\,\boldsymbol{\Theta}^{(1)}(\rho_1)\left(1 - \frac{\arccos(\rho_1)}{\pi}\right) + \frac{1}{r}\,w_r\,\Sigma^{(1)}(\rho_1),$$

with $\Sigma^{(1)}(u) = \frac{1}{\pi}\left(\sqrt{1-u^2} + u(1-\arccos u)\right)$.

Proof sketch. The rank-r projections $x_1 = Px$, $y_1 = Py$ form a ρ-correlated Gaussian pair. By rotational invariance, the empirical correlation $\rho_1 = \cos\angle(x_1, y_1)$ (angular part) and the squared norms $(u, v) = (\|x_1\|^2, \|y_1\|^2)$ (radial part) are independent. The angular part follows Fisher's distribution; the radial part follows Kibble's bivariate chi-square law.

Proof. Let $x, y \in \mathbb{R}^d$ be fixed input vectors with correlation $\rho = \langle x, y\rangle/(\|x\|\|y\|)$. Consider a rank-$r$ random projection matrix $P \in \mathbb{R}^{r\times d}$ with entries $P_{ij} \sim \mathcal{N}(0, 1/d)$ i.i.d., so that $x_1 = Px$ and $y_1 = Py$ are the projected vectors.

Since P has i.i.d. Gaussian entries, the projected vectors $x_1, y_1 \in \mathbb{R}^r$ form a ρ-correlated bivariate Gaussian pair. Specifically, for each component $i = 1, \ldots, r$, the pairs (x_{1i}, y_{1i}) are i.i.d. with joint distribution

$$\begin{pmatrix} x_{1i} \\ y_{1i} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \|x\|^2/d & \rho\|x\|\|y\|/d \\ \rho\|x\|\|y\|/d & \|y\|^2/d \end{pmatrix}\right). \tag{34}$$

By rotational invariance of the Gaussian projection and the fact that the correlation ρ is preserved under orthogonal transformations, we can assume without loss of generality that the covariance structure factors into angular and radial components.

Angular part (Fisher distribution). The empirical correlation is defined as

$$\rho_1 = \frac{\langle x_1, y_1\rangle}{\|x_1\|\|y_1\|} = \frac{\sum_{i=1}^r x_{1i}y_{1i}}{\sqrt{\sum_{i=1}^r x_{1i}^2}\sqrt{\sum_{i=1}^r y_{1i}^2}}. \tag{35}$$

This is the sample correlation coefficient of r pairs of correlated Gaussian random variables. Fisher [15, 16] showed that the distribution of ρ_1 depends only on the true correlation ρ and the sample size r, with density given in Remark C.1. The key observation is that ρ_1 is a function purely of the angles between the projected vectors, independent of their magnitudes.

Radial part (Kibble distribution). The squared norms $u = \|x_1\|^2 = \sum_{i=1}^r x_{1i}^2$ and $v = \|y_1\|^2 = \sum_{i=1}^r y_{1i}^2$ are sums of correlated chi-square variables. Since (x_{1i}, y_{1i}) are ρ-correlated Gaussians, the bivariate distribution of (u, v) follows Kibble's [49] bivariate chi-square law. The joint density involves the modified Bessel function $I_\nu(z)$, which appears because the dot product $\langle x_1, y_1\rangle = \sum_{i=1}^r x_{1i}y_{1i}$ can be expressed as a weighted sum of products of correlated normals, whose distribution relates to Bessel functions through the generating function of the bivariate chi-square distribution.

Specifically, the joint characteristic function of (u, v) is

$$\phi(s,t) = \mathbb{E}[e^{i(su+tv)}] = \left(1 - 2i(1-\rho^2)(s+t) - 4\rho^2 st\right)^{-r/2}, \tag{36}$$

and the inverse Fourier transform yields Kibble's density, which contains the modified Bessel function $I_{\frac{r}{2}-1}$ as stated in Remark C.1. The Bessel function arises from the integral representation of the bivariate chi-square density when the correlation $\rho \neq 0$.

Independence of angular and radial parts. Rotational invariance of isotropic Gaussian projections implies that $\rho_1 = \cos \angle(x_1, y_1)$ and the norms $(\|x_1\|, \|y_1\|)$ are independent, yielding the factorization $p(\rho_1, u, v) = p_{\text{Fisher}}(\rho_1) \cdot p_{\text{Kibble}}(u, v)$. This is the classical independence of the sample correlation from sample variances in bivariate normal data. □

C.3 Open directions: deep mean NTK recursion and effective depth

The deterministic proxy (Definition D.4) analyzes the scalar recursion along the mean path $\rho_k = \varrho^{\circ(k-1)}(\rho_1)$. The actual network uses the *random* correlation chain ρ_ℓ with $O(1/\sqrt{r})$ per-layer fluctuations (Appendix B.7). A rigorous link requires: (i) propagating the Fisher-type concentration $\mathbb{E}[(\rho_\ell - \varrho^{\circ(\ell-1)}(\rho_1))^2] \le C/r$ through depth to show $\Theta^{(L)}_{\text{random}} - \Theta^{(L)}_{\text{proxy}} = O_P(\sqrt{L/r})$ or similar; (ii) a rigorous recursion for $\tilde{\mathbf{\Theta}}^{(\ell)}$ with explicit covariance bounds; (iii) extension of Fisher–Kibble decoupling to nested layers to control mixed terms such as $\mathbb{E}[\mathbf{\Theta}^{(\ell-1)}\dot{\Sigma}^{(\ell)}]$ (where dependence through the deep GP chain prevents naive factorization). The exponential suppression of early-layer contributions suggests that very deep RF-LR may exhibit an "effective depth" phenomenon similar to ResNets, where only the top layers meaningfully contribute to the NTK.

Lemma C.2 (Nonasymptotic concentration of Gaussian sample cosine)**.** *Let $r \ge 1$ and let $(X, Y) \in \mathbb{R}^r \times \mathbb{R}^r$ have i.i.d. coordinates $\{(X_i, Y_i)\}_{i=1}^r$, each distributed as a centered bivariate normal with $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2] = 1$ and $\mathbb{E}[X_i Y_i] = \rho \in (-1, 1)$. Define the empirical cosine (sample correlation without centering)*

$$\widehat{\rho}_r \;=\; \frac{\langle X, Y\rangle}{\|X\|\,\|Y\|}.$$

Then there exist absolute constants $c, C > 0$ such that for all $t \in (0, 1)$,

$$\mathbb{P}\big(|\widehat{\rho}_r - \rho| \ge t\big) \;\le\; 6\exp(-c\, r\, t^2), \qquad \mathbb{E}\big[(\widehat{\rho}_r - \rho)^2\big] \;\le\; \frac{C}{r}.$$

Proof sketch. Write $\widehat{\rho}_r = S_{xy}/\sqrt{S_{xx}S_{yy}}$ where S_{xx}, S_{yy}, S_{xy} are sample averages. Each of $S_{xx} - 1$, $S_{yy} - 1$, $S_{xy} - \rho$ is sub-exponential; Bernstein's inequality yields tail bounds. On the event that S_{xx}, S_{yy} are close to 1, bound $|\widehat{\rho}_r - \rho|$ via a ratio expansion; union bound over the three tails gives the stated rate.

Proof. Write

$$S_{xx} = \frac{1}{r}\|X\|^2, \qquad S_{yy} = \frac{1}{r}\|Y\|^2, \qquad S_{xy} = \frac{1}{r}\langle X, Y\rangle, \qquad \text{so that} \qquad \widehat{\rho}_r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Each of $S_{xx} - 1$, $S_{yy} - 1$, and $S_{xy} - \rho$ is an average of i.i.d. centered sub-exponential random variables (since $X_i^2 - 1$, $Y_i^2 - 1$, and $X_iY_i - \rho$ have finite ψ_1 norms for Gaussian data). Therefore, by Bernstein's inequality, there exist absolute constants $c_0 > 0$ and $C_0 < \infty$ such that for all $t \in (0, 1)$,

$$\mathbb{P}\big(|S_{xx} - 1| \ge t\big) \le 2e^{-c_0 r t^2}, \quad \mathbb{P}\big(|S_{yy} - 1| \ge t\big) \le 2e^{-c_0 r t^2}, \quad \mathbb{P}\big(|S_{xy} - \rho| \ge t\big) \le 2e^{-c_0 r t^2}.$$

On the event $E_t = \{|S_{xx} - 1| \le t,\ |S_{yy} - 1| \le t\}$ with $t \le 1/2$, we have $S_{xx}, S_{yy} \in [1/2, 3/2]$, hence

$$\left|\frac{1}{\sqrt{S_{xx}S_{yy}}} - 1\right| \le C_0\big(|S_{xx} - 1| + |S_{yy} - 1|\big) \le 2C_0 t$$

for an absolute constant C_0. Using

$$\widehat{\rho}_r - \rho = \frac{S_{xy} - \rho}{\sqrt{S_{xx}S_{yy}}} \;+\; \rho\Big(\frac{1}{\sqrt{S_{xx}S_{yy}}} - 1\Big),$$

we obtain on E_t the bound $|\widehat{\rho}_r - \rho| \le C_1(|S_{xy} - \rho| + t)$ for an absolute constant C_1. Taking $t \asymp t'$ and applying a union bound over the three Bernstein tails yields $\mathbb{P}(|\widehat{\rho}_r - \rho| \ge t') \le 6e^{-crt'^2}$ for $t' \in (0, 1)$ and absolute $c > 0$. The second-moment bound follows by integrating the tail bound: $\mathbb{E}[(\widehat{\rho}_r - \rho)^2] = \int_0^\infty 2u\,\mathbb{P}(|\widehat{\rho}_r - \rho| \ge u)\,du \le C/r$. □

C.4 Concentration of the random Gram matrix around the proxy

Let $\hat{K}$ be the $n \times n$ random Gram matrix with entries $\hat{K}_{ij} = \hat{\Theta}_r^{(L)}(x_i, x_j)$ (the empirical RF-LR NTK over the random correlation chain), and let K_{proxy} be the deterministic proxy Gram matrix with entries $(K_{\text{proxy}})_{ij} = \Theta^{(L)}(\rho_{1,ij})$ (Definition D.4). Under the same per-layer sub-Gaussian concentration as in Appendix B.7,

$$\mathbb{E}\big[(\rho_\ell - \varrho^{\circ(\ell-1)}(\rho_1))^2\big] \le C/r, \qquad \mathbb{P}\big(|\rho_\ell - \varrho^{\circ(\ell-1)}(\rho_1)| \ge t\big) \le 6\exp(-crt^2),$$

the kernel recursion is Lipschitz in the correlation path. Propagating the per-step bounds through L layers (e.g. by a union bound over the L steps and the fact that $\Theta^{(k)}$ is Lipschitz in ρ with constant $O(1)$), one obtains that for each pair (i, j),

$$\big|\hat{K}_{ij} - (K_{\text{proxy}})_{ij}\big| = O_P(\sqrt{L/r}).$$

A union over the n^2 pairs and the inequality $\|A\|_{\text{op}} \le \|A\|_{\text{F}} \le n \max_{i,j} |A_{ij}|$ yield the following. The bound in Proposition C.1 is *sketch-based*; a full proof with explicit constants and rate is given only for the equicorrelated case (Theorem 4.2).

Proposition C.1 (Proxy–empirical Gram matrix concentration)**.** *Fix $n, L \ge 1$, $r > 1$, and inputs $x_1, \dots, x_n$ with pairwise cosine similarities $\rho_{1,ij} \in (-1, 1)$. Let $\hat{K}$ and K_{proxy} be as above. There exist constants $C, c > 0$ (depending on L, n and the $\rho_{1,ij}$) such that for any $\epsilon \in (0, 1)$,*

$$\mathbb{P}\big(\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} \ \ge \ \epsilon\big) \ \le \ C \, \exp\!\left(-c \, \frac{r\,\epsilon^2}{L\,n^2}\right).$$

In particular, for $r \to \infty$ with n, L fixed, $\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} = o_P(1)$. If in addition the proxy Gram matrix has minimum eigenvalue $\lambda_{\min}(K_{\text{proxy}}|_{\mathbf{1}^\perp}) \ge \gamma > 0$, then for r large enough so that $\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} < \gamma/2$ with high probability, Weyl's inequality gives

$$\big|\lambda_{\min}(\hat{K}|_{\mathbf{1}^\perp}) - \lambda_{\min}(K_{\text{proxy}}|_{\mathbf{1}^\perp})\big| \le \gamma/2,$$

so the condition number of $\hat{K}$ restricted to $\mathbf{1}^\perp$ is within a constant factor of that of K_{proxy}.

Proof sketch. Per-layer concentration (Lemma C.2) and Lipschitz dependence of the scalar recursion on ρ imply that each entry $\hat{K}_{ij}$ differs from $(K_{\text{proxy}})_{ij}$ by $O_P(\sqrt{L/r})$. Union bound over n^2 pairs with sub-Gaussian tails gives

$$\max_{i,j} \big|\hat{K}_{ij} - (K_{\text{proxy}})_{ij}\big| = O_P\big(\sqrt{L \log(n)/r}\big).$$

Hence $\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} \le n\|\hat{K} - K_{\text{proxy}}\|_{\max} = O_P(n\sqrt{L \log(n)/r})$, and the stated exponential tail follows. The condition-number comparison is by Weyl's inequality for the eigenvalues of symmetric matrices. For general (non-equicorrelated) datasets this remains a proof sketch; a rigorous operator-norm bound with explicit rate is given only for the equicorrelated case in Theorem 4.2.

Equicorrelated case: reduction to two scalars and rigorous bound on $\mathbf{1}^\perp$. In the equicorrelated case ($\rho_{1,ij} = \rho_0$ for all $i \ne j$, $\rho_{1,ii} = 1$), the proxy Gram matrix has the form $K_{\text{proxy}} = \Theta^{(L)}(1) I_n + \Theta^{(L)}(\rho_0)(\mathbf{1}\mathbf{1}^\top - I_n)$. By symmetry of the data and the network, the random $\hat{K}$ has the same structure: $\hat{K}_{ii} = \hat{K}_{11}$ and $\hat{K}_{ij} = \hat{K}_{12}$ for $i \ne j$. Hence the deviation matrix $E := \hat{K} - K_{\text{proxy}}$ is

$$E \ = \ \big(\hat{K}_{11} - \Theta^{(L)}(1)\big)\, I_n \ + \ \big(\hat{K}_{12} - \Theta^{(L)}(\rho_0)\big)\,\big(\mathbf{1}\mathbf{1}^\top - I_n\big),$$

with eigenvalues $\lambda_{\mathbf{1}} = \big(\hat{K}_{11} - \Theta^{(L)}(1)\big) + (n-1)\big(\hat{K}_{12} - \Theta^{(L)}(\rho_0)\big)$ on $\mathbf{1}$ and $\lambda_\perp = \big(\hat{K}_{11} - \Theta^{(L)}(1)\big) - \big(\hat{K}_{12} - \Theta^{(L)}(\rho_0)\big)$ on $\mathbf{1}^\perp$ (multiplicity $n-1$), so $\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} = \max(|\lambda_{\mathbf{1}}|, |\lambda_\perp|)$. The following theorem gives a rigorous $O_P(L/r)$ bound for the restriction to $\mathbf{1}^\perp$; the full operator norm without n-dependence remains open (see Remark C.2).

Lemma C.3 (Path-wise kernel and sensitivity)**.** *Let $s(\rho)$ and $\dot{s}(\rho)$ be the scalar ReLU base and derivative kernels as in* (84)*. For $k \ge 1$ and $\rho_1, \dots, \rho_k \in (-1, 1)$, define $\Phi^{(1)}(\rho_1) = 1 + s(\rho_1)$ and for $k \ge 2$,*

$$\Phi^{(k)}(\rho_1, \dots, \rho_k) \ = \ 1 \ + \ \frac{1}{r}\,\Phi^{(k-1)}(\rho_1, \dots, \rho_{k-1})\,\dot{s}(\rho_k) \ + \ \frac{1}{r}\,s(\rho_k).$$

Then $\Phi^{(k)}$ is the same function as $\Theta^{(k)}$ when evaluated along the deterministic path $\rho_j = \varrho^{\circ(j-1)}(\rho_1)$. There exist constants $C_s, C_\Phi < \infty$ (depending only on the ReLU kernels) such that for all $2 \le \ell \le k \le L$ and all

$\rho_1, \dots, \rho_k \in [-1+\delta, 1-\delta]$ *with* $\delta \in (0,1)$,

$$\left|\frac{\partial \Phi^{(k)}}{\partial \rho_\ell}\right| \;\le\; C_\Phi \left(\frac{1}{r}\right)^{k-\ell}.$$

In particular, for $r \ge 2$ *and some* $\overline{C}_L, \overline{C}_L'' = O(1)$,

$$\sum_{\ell=2}^{L} \left(\frac{\partial \Phi^{(L)}}{\partial \rho_\ell}\right)^2 \le C_\Phi^2 \, \frac{1-(1/r)^{2(L-1)}}{1-(1/r)^2} \le \overline{C}_L,$$

$$\sum_{\ell=2}^{L} \left|\frac{\partial \Phi^{(L)}}{\partial \rho_\ell}\right| \le C_\Phi \sum_{t=0}^{L-2} (1/r)^t \le \overline{C}_L''.$$

Proof. The recursion for $\Phi^{(k)}$ coincides with (84), so

$$\Theta^{(k)}(\rho_1) \;=\; \Phi^{(k)}\big(\rho_1, \varrho(\rho_1), \dots, \varrho^{\circ(k-1)}(\rho_1)\big).$$

Differentiating in ρ_k,

$$\frac{\partial \Phi^{(k)}}{\partial \rho_k} \;=\; \frac{1}{r}\, \Phi^{(k-1)}\, \dot{s}'(\rho_k) \;+\; \frac{1}{r}\, s'(\rho_k).$$

On $[-1+\delta, 1-\delta]$, $|\dot{s}'|$ and $|s'|$ are bounded by some C_s. Since $\Phi^{(k-1)}$ is bounded (the recursion yields $\Phi^{(k)} \in [1, 1+O(1)]$ on compact ρ), we have $\left|\frac{\partial \Phi^{(k)}}{\partial \rho_k}\right| \le C_s'(1/r)$ for a constant C_s'. For $\ell < k$,

$$\frac{\partial \Phi^{(k)}}{\partial \rho_\ell} \;=\; \frac{1}{r}\, \frac{\partial \Phi^{(k-1)}}{\partial \rho_\ell}\, \dot{s}(\rho_k).$$

By induction,

$$\left|\frac{\partial \Phi^{(k)}}{\partial \rho_\ell}\right| \;\le\; C_s' \prod_{j=\ell+1}^{k} \frac{1}{r} |\dot{s}(\rho_j)| \;\le\; C_\Phi\, (1/r)^{k-\ell},$$

with $C_\Phi = C_s' \, (\sup |\dot{s}|)^{L-1} \le C_s'/2^{L-1}$. The sum bound follows from $\sum_{\ell=2}^{L} (1/r)^{2(L-\ell)} = \sum_{t=0}^{L-2} (1/r)^{2t} \le \frac{1}{1-(1/r)^2} \le 2$ for $r \ge 2$, so $\sum_{\ell=2}^{L} \big(\frac{\partial \Phi^{(L)}}{\partial \rho_\ell}\big)^2 \le C_\Phi^2 \cdot 2 = \overline{C}_L$. □

Lemma C.4 (Second-order remainder)**.** *Under the same setting as Lemma C.3, let* $\rho_1 = \rho_0$ *be fixed and* $(\rho_2, \dots, \rho_L)$ *be the random correlation chain with* $\mathbb{E}[(\rho_\ell - \varrho^{\circ(\ell-1)}(\rho_0))^2] \le C_0/r$ *and* $\mathbb{P}(|\rho_\ell - \varrho^{\circ(\ell-1)}(\rho_0)| \ge t) \le 6\exp(-c_0 r t^2)$ *for all* ℓ *(Lemma C.2). Let* $\bar{\rho}_\ell = \varrho^{\circ(\ell-1)}(\rho_0)$*. Then the second-order Taylor remainder*

$$R_2 \;=\; \hat{\Theta}_r^{(L)} - \Theta^{(L)}(\rho_0) \;-\; \sum_{\ell=2}^{L} \frac{\partial \Phi^{(L)}}{\partial \rho_\ell}(\bar{\rho}_2, \dots, \bar{\rho}_L)\, (\rho_\ell - \bar{\rho}_\ell)$$

satisfies $\mathbb{E}[|R_2|] \le B_L\, L/r$ *for a constant* B_L *depending on* L *and the Hessian of* $\Phi^{(L)}$ *on* $[-1+\delta, 1-\delta]^L$*, and for any* $u > 0$*,*

$$\mathbb{P}(|R_2| \ge u) \;\le\; 2\exp\!\left(-\min\!\left(\frac{c_0 r u}{2B_L L}, \frac{c_0 r u^2}{2(B_L L)^2}\right)\right).$$

Proof. By Taylor's theorem, for some ξ on the segment between $(\rho_2, \dots, \rho_L)$ and $(\bar{\rho}_2, \dots, \bar{\rho}_L)$,

$$R_2 \;=\; \frac{1}{2} \sum_{\ell,m} \frac{\partial^2 \Phi^{(L)}}{\partial \rho_\ell \partial \rho_m}(\xi)\, (\rho_\ell - \bar{\rho}_\ell)(\rho_m - \bar{\rho}_m).$$

The second derivatives of $\Phi^{(L)}$ are bounded on compact sets (the recursion is polynomial in $s, \dot{s}$ and those are C^∞ on $(-1,1)$). Hence

$$|R_2| \;\le\; \frac{1}{2} H_L \sum_{\ell=2}^{L} (\rho_\ell - \bar{\rho}_\ell)^2$$

for a constant H_L, and

$$\mathbb{E}[|R_2|] \;\le\; \frac{1}{2}H_L \sum_{\ell=2}^{L} \mathbb{E}[(\rho_\ell - \bar{\rho}_\ell)^2] \;\le\; \frac{1}{2}H_L\,(L-1)\,\frac{C_0}{r} \;\le\; B_L\,\frac{L}{r}.$$

Each $(\rho_\ell - \bar{\rho}_\ell)^2$ is sub-exponential (from the sub-Gaussian tail of $\rho_\ell - \bar{\rho}_\ell$). By Bernstein's inequality for sums of sub-exponential variables, $\sum_{\ell=2}^{L}(\rho_\ell - \bar{\rho}_\ell)^2$ has an exponential tail with scale $O(L/r)$; hence $|R_2|$ has the stated tail bound. □

Proof of Theorem 4.2.

Proof. Let $\hat{K}$ and K_{proxy} be as in Proposition C.1. Let $E = \hat{K} - K_{\text{proxy}}$. We have

$$\|E|_{\mathbf{1}^\perp}\|_{\text{op}} \;=\; |\lambda_\perp| \;=\; \left|\hat{K}_{11} - \Theta^{(L)}(1) - \left(\hat{K}_{12} - \Theta^{(L)}(\rho_0)\right)\right|.$$

For the diagonal, $\rho_1 = 1$ is fixed and the random path $(1, \rho_2^{\text{diag}}, \ldots, \rho_L^{\text{diag}})$ has the same per-layer concentration; so $\hat{K}_{11} - \Theta^{(L)}(1)$ admits the same Taylor expansion and bounds as below with ρ_0 replaced by 1. Thus it suffices to prove that both $|\hat{K}_{11} - \Theta^{(L)}(1)|$ and $|\hat{K}_{12} - \Theta^{(L)}(\rho_0)|$ are $O_P(L/r)$; then

$$|\lambda_\perp| \;\le\; \left|\hat{K}_{11} - \Theta^{(L)}(1)\right| + \left|\hat{K}_{12} - \Theta^{(L)}(\rho_0)\right| \;=\; O_P(L/r).$$

Off-diagonal deviation. Write $\hat{K}_{12} = \Phi^{(L)}(\rho_0, \rho_2, \ldots, \rho_L)$ and $\Theta^{(L)}(\rho_0) = \Phi^{(L)}(\rho_0, \bar{\rho}_2, \ldots, \bar{\rho}_L)$ with $\bar{\rho}_\ell = \varrho^{\circ(\ell-1)}(\rho_0)$. By Taylor expansion,

$$\hat{K}_{12} - \Theta^{(L)}(\rho_0) \;=\; \sum_{\ell=2}^{L} \frac{\partial \Phi^{(L)}}{\partial \rho_\ell}(\bar{\rho})\,(\rho_\ell - \bar{\rho}_\ell) \;+\; R_2.$$

Write $a_\ell = \frac{\partial \Phi^{(L)}}{\partial \rho_\ell}(\bar{\rho})$. By Lemma C.3, $\sum_{\ell=2}^{L}|a_\ell| \le \overline{C}''_L$ for some $\overline{C}''_L = O(1)$.

First-order term. Let $S_1 = \sum_{\ell=2}^{L} a_\ell\,(\rho_\ell - \bar{\rho}_\ell)$. We bound the variance without conditioning: $\mathbb{E}[(\rho_\ell - \bar{\rho}_\ell)^2] \le C_0/r$ for all ℓ by Lemma C.2. By Cauchy–Schwarz,

$$\text{Var}(S_1) \;\le\; \mathbb{E}[S_1^2] \;\le\; \sum_{\ell,m} |a_\ell||a_m| \sqrt{\mathbb{E}[(\rho_\ell - \bar{\rho}_\ell)^2]\mathbb{E}[(\rho_m - \bar{\rho}_m)^2]} \;\le\; \frac{C_0}{r}\left(\sum_{\ell=2}^{L}|a_\ell|\right)^2 \;\le\; \frac{C_0(\overline{C}''_L)^2}{r}.$$

So $S_1 = O_P(1/\sqrt{r})$. For the exponential tail: by Lemma C.2, $\mathbb{P}(|\rho_\ell - \bar{\rho}_\ell| \ge t) \le 6\exp(-c_0 r t^2)$ for each ℓ. A union bound gives

$$\mathbb{P}\left(\max_{2\le\ell\le L} |\rho_\ell - \bar{\rho}_\ell| \ge t\right) \;\le\; 6(L-1)\exp(-c_0 r t^2).$$

On the event $\{\max_\ell |\rho_\ell - \bar{\rho}_\ell| < t\}$, we have $|S_1| \le (\sum_\ell |a_\ell|)\,t \le \overline{C}''_L t$. So for ϵ in a bounded range,

$$\mathbb{P}(|S_1| \ge \epsilon) \;\le\; 6L\exp\left(-c_0 r\,(\epsilon/\overline{C}''_L)^2\right).$$

Second-order term. By Lemma C.4, $\mathbb{E}[|R_2|] \le B_L L/r$ and $|R_2|$ has an exponential tail with scale $O(L/r)$. So $R_2 = O_P(L/r)$.

Total. Thus

$$\hat{K}_{12} - \Theta^{(L)}(\rho_0) \;=\; S_1 + R_2 \;=\; O_P(1/\sqrt{r}) + O_P(L/r) \;=\; O_P(L/r + 1/\sqrt{r}).$$

The diagonal deviation $\hat{K}_{11} - \Theta^{(L)}(1)$ is handled identically (path starting at $\rho_1 = 1$). Hence

$$|\lambda_\perp| \;\le\; \left|\hat{K}_{11} - \Theta^{(L)}(1)\right| + \left|\hat{K}_{12} - \Theta^{(L)}(\rho_0)\right| \;=\; O_P(L/r + 1/\sqrt{r}),$$

and the stated probability bound follows by combining the tail bounds for S_1 and R_2. □

Remark C.2 (Full operator norm). To obtain $\|\hat{K} - K_{\text{proxy}}\|_{\text{op}} = O_P(L/r)$ without restricting to $\mathbf{1}^\perp$, one would need $|\lambda_{\mathbf{1}}| = O_P(L/r)$. Since $\lambda_{\mathbf{1}} = (\hat{K}_{11} - \Theta^{(L)}(1)) + (n-1)(\hat{K}_{12} - \Theta^{(L)}(\rho_0))$, this would require $|\hat{K}_{12} - \Theta^{(L)}(\rho_0)| = O_P(L/(rn))$, which is a factor n stronger than the $O_P(L/r)$ bound proved here and remains open.

C.5 Rank-driven concentration

Theorem C.1 (Concentration in rank for homogeneous activation)**.** *Under the homogeneous activation assumption (Assumption B.1), let $x_1, y_1 \in \mathbb{R}^r$ be the rank-r Gaussian projections of inputs x, y, with arbitrary fixed correlation $\rho \in [-1, 1]$. Define*

$$W \;=\; \frac{\|x_1\|\,\|y_1\|}{r}.$$

Then for any $\epsilon \in (0, 1)$,

$$\mathbb{P}(|W - 1| \;\geq\; \epsilon) \;\leq\; 4\,\exp\!\left(-\frac{r\,\epsilon^2}{8}\right).$$

Moreover, for larger deviations $\epsilon \geq 1$, there exist absolute constants $c_1, c_2 > 0$ such that

$$\mathbb{P}(|W - 1| \;\geq\; \epsilon) \;\leq\; c_1\,\exp(-c_2\,r\,\epsilon)\,.$$

In particular, the radial component of the RF-LR NTK concentrates exponentially fast in r, yielding high-dimensional control of RKHS fluctuations.

Proof sketch. Write $W = UV$ with $U = \|x_1\|/\sqrt{r}$, $V = \|y_1\|/\sqrt{r}$. Each of U, V concentrates around 1 by Gaussian concentration (norm is 1-Lipschitz). Use $|UV - 1| \leq |U-1| + |V-1| + |U-1||V-1|$; on the event $\{|U-1|, |V-1| < \delta\}$ with $\delta = \epsilon/2$, we have $|UV - 1| \leq \epsilon$. Union bound over the two norm tails yields the sub-Gaussian rate.

Proof. Let $x_1, y_1 \in \mathbb{R}^r$ be Gaussian projections with arbitrary fixed correlation $\rho \in [-1, 1]$. Define $U = \|x_1\|/\sqrt{r}$, $V = \|y_1\|/\sqrt{r}$, and $W = UV = \|x_1\|\,\|y_1\|/r$. We prove that for $\epsilon \in (0, 1)$,

$$\mathbb{P}(|W - 1| \geq \epsilon) \leq 4\exp\!\left(-\frac{r\epsilon^2}{8}\right), \tag{37}$$

and that for $\epsilon \geq 1$ there exist constants $c_1, c_2 > 0$ with $\mathbb{P}(|W - 1| \geq \epsilon) \leq c_1 e^{-c_2 r\epsilon}$.

The Euclidean norm is 1-Lipschitz, so by Gaussian concentration, for all $\delta > 0$,

$$\mathbb{P}(|U - 1| \geq \delta) \leq 2\exp\!\left(-\frac{r\delta^2}{2}\right), \qquad \mathbb{P}(|V - 1| \geq \delta) \leq 2\exp\!\left(-\frac{r\delta^2}{2}\right). \tag{38}$$

We use

$$|UV - 1| \;\leq\; |U - 1| + |V - 1| + |U - 1|\,|V - 1|. \tag{39}$$

Fix $\epsilon \in (0, 1)$ and set $\delta = \epsilon/2$. On the event $\{|U - 1| < \delta,\ |V - 1| < \delta\}$, we have

$$|UV - 1| \;\leq\; 2\delta + \delta^2 \;\leq\; \epsilon. \tag{40}$$

Therefore,

$$\mathbb{P}(|UV - 1| \geq \epsilon) \;\leq\; \mathbb{P}(|U - 1| \geq \delta) \;+\; \mathbb{P}(|V - 1| \geq \delta) \;\leq\; 4\exp\!\left(-\frac{r\epsilon^2}{8}\right). \tag{41}$$

□

Corollary C.2 (Concentration bound for three-layer NTK)**.** *Under Assumption B.1, let $\hat{\Theta}^{(2)}_r(x, x') = \Psi(\hat{\rho}_r, \hat{w}_r)$ be the empirical three-layer NTK, where $\hat{\rho}_r$ is the sample correlation and $\hat{w}_r = \|x_1\|\,\|y_1\|/r$, with $x_1, y_1 \in \mathbb{R}^r$ being the rank-r Gaussian projections of inputs x, y with population correlation $\rho \in [-1, 1]$. Let $K_\infty(\rho) = \Psi(\rho, 1)$ be the deterministic kernel limit. Then for any $\epsilon \in (0, 1)$, there exist constants $C_1, C_2 > 0$ depending on ρ and the kernel smoothness such that*

$$\mathbb{P}\!\left(\left|\hat{\Theta}^{(2)}_r(x, x') - K_\infty(\rho)\right| \;\geq\; \epsilon\right) \;\leq\; C_1\,\exp\!\left(-C_2\,r\,\epsilon^2\right).$$

C.6 Proof of Corollary C.2: Concentration bound for three-layer NTK

Proof sketch. The empirical kernel $\hat{\Theta}_r^{(2)} = \Psi(\hat{\rho}_r, \hat{w}_r)$ depends on the sample correlation $\hat{\rho}_r$ and norm product $\hat{w}_r$. By Lemma C.2 and Theorem C.1, both concentrate with sub-Gaussian tails. Since Ψ is Lipschitz in (ρ, w) near $(\rho, 1)$, a union bound and Lipschitz bound yield $|\hat{\Theta}_r^{(2)} - K_\infty(\rho)| \le L_u|\hat{\rho}_r - \rho| + L_w|\hat{w}_r - 1|$, hence the stated exponential tail.

Proof. We prove that the empirical three-layer NTK $\hat{\Theta}_r^{(2)}(x, x') = \Psi(\hat{\rho}_r, \hat{w}_r)$ concentrates exponentially fast around its deterministic limit $K_\infty(\rho) = \Psi(\rho, 1)$ as $r \to \infty$.

Recall that the kernel function is defined as:

$$\Psi(u, w) = \Theta^{(1)}(u)\left(1 - \frac{\arccos(u)}{\pi}\right) + w\,\Sigma^{(1)}(u) + 1, \tag{42}$$

where $\hat{\rho}_r$ is the sample correlation and $\hat{w}_r = \|x_1\|\|y_1\|/r$. By Lemma C.2, $\hat{\rho}_r$ concentrates around ρ with sub-Gaussian tails at rate $r^{-1/2}$, and by Theorem C.1, $\hat{w}_r$ concentrates around 1 with sub-Gaussian tails.

The function Ψ is differentiable in both arguments. For fixed $\rho \in (-1, 1)$, we have:

$$\partial_u \Psi(\rho, 1) = \partial_\rho K_\infty(\rho) = O(1), \tag{43}$$

$$\partial_w \Psi(\rho, 1) = \Sigma^{(1)}(\rho) = O(1). \tag{44}$$

Since $\Theta^{(1)}$, $\Sigma^{(1)}$, and arccos are smooth on $(-1, 1)$, there exist Lipschitz constants $L_u, L_w > 0$ (depending on ρ) such that for $u, u' \in [\rho - \delta, \rho + \delta]$ and $w, w' \in [1 - \delta, 1 + \delta]$ with $\delta > 0$ small:

$$|\Psi(u, w) - \Psi(u', w')| \le L_u|u - u'| + L_w|w - w'|. \tag{45}$$

We combine radial and angular concentration by using a union bound and the Lipschitz property:

$$|\hat{\Theta}_r^{(2)}(x, x') - K_\infty(\rho)| = |\Psi(\hat{\rho}_r, \hat{w}_r) - \Psi(\rho, 1)| \le L_u|\hat{\rho}_r - \rho| + L_w|\hat{w}_r - 1|. \tag{46}$$

Therefore, for $\epsilon \in (0, 1)$:

$$\mathbb{P}\left(|\hat{\Theta}_r^{(2)}(x, x') - K_\infty(\rho)| \ge \epsilon\right) \le \mathbb{P}\left(|\hat{\rho}_r - \rho| \ge \frac{\epsilon}{2L_u}\right) + \mathbb{P}\left(|\hat{w}_r - 1| \ge \frac{\epsilon}{2L_w}\right). \tag{47}$$

Now we prove the sub-Gaussian concentration of the sample correlation via Hanson–Wright.

Lemma C.5 (Sub-Gaussian concentration of the sample correlation via Hanson–Wright)**.** *Let $(X_i, Y_i)_{i=1}^r$ be i.i.d. centered Gaussian pairs with unit variances and correlation $\rho \in (-1, 1)$. Define*

$$S_{xx} = \frac{1}{r}\sum_{i=1}^r X_i^2, \quad S_{yy} = \frac{1}{r}\sum_{i=1}^r Y_i^2, \quad S_{xy} = \frac{1}{r}\sum_{i=1}^r X_i Y_i, \quad \hat{\rho}_r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{48}$$

There exist absolute constants $c, C > 0$ such that for all $t \in (0, 1)$,

$$\mathbb{P}(|\hat{\rho}_r - \rho| \ge t) \;\le\; C\,\exp\big(-c\,r\,t^2\big). \tag{49}$$

Proof sketch. Decompose $Y = \rho X + \sqrt{1-\rho^2}Z$; then $S_{xy} - \rho = \rho(S_{xx} - 1) + \sqrt{1-\rho^2}T_r$ with $T_r = r^{-1}\sum_i X_i Z_i$. Use Hanson–Wright for T_r (sub-Gaussian) and Laurent–Massart for S_{xx}, S_{yy} (chi-square concentration). Expand $\hat{\rho}_r = f(S_{xy}, S_{xx}, S_{yy})$ at $(\rho, 1, 1)$ and union bound.

Proof. Write $Y = \rho X + \sqrt{1-\rho^2}\,Z$ with $X = (X_i)_{i\le r}$ and $Z = (Z_i)_{i\le r}$ independent, i.i.d. $\mathcal{N}(0, 1)$. Then

$$S_{xy} - \rho = \rho\,(S_{xx} - 1) \;+\; \sqrt{1-\rho^2}\;\underbrace{\frac{1}{r}\sum_{i=1}^r X_i Z_i}_{=:T_r}. \tag{50}$$

From [**?**] Theorem 6.2.2 page 183, the bilinear Hanson–Wright inequality (for independent sub-Gaussian X, Z and fixed $M \in \mathbb{R}^{r\times r}$) states

$$\mathbb{P}\big(\big|X^\top M Z\big| \ge u\big) \;\le\; 2\exp\!\left[-c\,\min\!\left(\frac{u^2}{K^4\|M\|_F^2},\; \frac{u}{K^2\|M\|}\right)\right], \tag{51}$$

where $K = O(1)$ is the Orlicz sub-gaussian norm of the coordinates. For $M = I_r$ we have $\|M\|_F^2 = r$ and $\|M\| = 1$, hence with $u = rt$ and $t \in (0,1)$,

$$\mathbb{P}\!\left(\Big|\frac{1}{r}X^\top Z\Big| \ge t\right) \;\le\; 2\exp\big(-c\,r\,t^2\big). \tag{52}$$

Thus T_r is sub-Gaussian at rate $\exp(-c\,r\,t^2)$:

$$\mathbb{P}(|T_r| \ge t) \le 2\exp(-c\,r\,t^2). \tag{53}$$

For $S_{xx} = (1/r)\|X\|^2$ and $S_{yy} = (1/r)\|Y\|^2$, Laurent–Massart Gausian quadratic chaos inequality in (4.1) page 1325 [**?**] yields, for any $s \in (0,1)$,

$$\mathbb{P}(|S_{xx}-1| \ge s) \le 2\exp(-c\,r\,s^2), \qquad \mathbb{P}\big(|S_{yy}-1| \ge s\big) \le 2\exp(-c\,r\,s^2). \tag{54}$$

Let $f(a,b,c) = a/\sqrt{bc}$. A first-order expansion at $(\rho,1,1)$ yields

$$|\hat\rho_r - \rho| = \big|f(S_{xy},S_{xx},S_{yy}) - f(\rho,1,1)\big| \;\le\; |S_{xy}-\rho| + \frac{|\rho|}{2}\big(|S_{xx}-1| + |S_{yy}-1|\big) + R, \tag{55}$$

where the remainder $R = O\big((|S_{xx}-1| + |S_{yy}-1|)^2\big)$ is negligible on the event $\{|S_{xx}-1| \vee |S_{yy}-1| \le c_0\}$ (for some absolute c_0, e.g., $1/4$). Choosing $s = \kappa t$ for a sufficiently small absolute $\kappa > 0$, and splitting $|S_{xy}-\rho|$ as in step 1, a union bound gives

$$\mathbb{P}(|\hat\rho_r - \rho| \ge t) \;\le\; \underbrace{\mathbb{P}(|T_r| \ge c_1 t)}_{\le 2e^{-crt^2}} + \underbrace{\mathbb{P}(|S_{xx}-1| \ge c_2 t)}_{\le 2e^{-crt^2}} + \underbrace{\mathbb{P}\big(|S_{yy}-1| \ge c_2 t\big)}_{\le 2e^{-crt^2}} + \mathbb{P}(R \not\ll t). \tag{56}$$

The last term is absorbed by adjusting κ (small-probability event controlled by the same inequalities). Therefore, for absolute constants $c, C > 0$,

$$\mathbb{P}(|\hat\rho_r - \rho| \ge t) \le C\,\exp(-c\,r\,t^2), \quad t \in (0,1). \tag{57}$$

□

Remark C.3 ("Bessel" variant: sum of Gaussian products). Each product X_iZ_i has a symmetric sub-exponential tail (density involving the Bessel function K_0). By Bernstein's inequality for i.i.d. sub-exponential variables, $\frac{1}{r}\sum_{i=1}^r X_iZ_i$ satisfies the same $\exp(-crt^2)$ bound for $t \in (0,1)$, yielding an alternative proof of the bilinear step.

By Theorem C.1, for $\epsilon/(2L_w) \in (0,1)$:

$$\mathbb{P}\!\left(|\hat w_r - 1| \ge \frac{\epsilon}{2L_w}\right) \le 4\exp\!\left(-\frac{r\epsilon^2}{32L_w^2}\right). \tag{58}$$

Combining both terms:

$$\mathbb{P}\!\left(|\hat\Theta_r^{(2)}(x,x') - K_\infty(\rho)| \ge \epsilon\right) \;\le\; \underbrace{\mathbb{P}\!\left(|\hat\rho_r - \rho| \ge \frac{\epsilon}{2L_u}\right)}_{\le C\,\exp\left(-\frac{c\,r\,\epsilon^2}{4L_u^2}\right)} + \underbrace{\mathbb{P}\!\left(|\hat w_r - 1| \ge \frac{\epsilon}{2L_w}\right)}_{\le 4\,\exp\left(-\frac{r\,\epsilon^2}{32L_w^2}\right)} \tag{59}$$

so that

$$\mathbb{P}\!\left(|\hat\Theta_r^{(2)}(x,x') - K_\infty(\rho)| \ge \epsilon\right) \;\le\; C\,\exp\!\left(-\frac{c\,r\,\epsilon^2}{4L_u^2}\right) \;+\; 4\,\exp\!\left(-\frac{r\,\epsilon^2}{32L_w^2}\right). \tag{60}$$

Taking $C_1 = C + 4$ and $C_2 = \min\big(c/(4L_u^2),\; 1/(32L_w^2)\big)$, we obtain

$$\mathbb{P}\!\left(|\hat\Theta_r^{(2)}(x,x') - K_\infty(\rho)| \ge \epsilon\right) \;\le\; C_1\,\exp\big(-C_2\,r\,\epsilon^2\big). \tag{61}$$

□

C.7 Proof of Corollary C.1: Mean NTK over Fisher–Kibble has same RKHS

Proof sketch. The mean NTK $\tilde{\mathbf{\Theta}}^{(2)}(\rho)$ is zonal. The key step is the Puiseux expansion near $\rho = 1$: compute $\mathbb{E}[\arccos(\hat{\rho}_r)]$ when $\hat{\rho}_r \sim \text{Fisher}(1-t, r)$ using the Fisher density and a hypergeometric connection formula. The leading term is $\sqrt{2t}\, I(r) + O(t^{3/2})$; since $I(r) = O(1/\sqrt{r})$, the $t^{1/2}$ coefficient is an $O(1)$ perturbation of the shallow ReLU constant. By the Bietti–Bach criterion [17, Theorem 1], the same endpoint exponent $1/2$ implies RKHS equivalence.

Proof. We prove that under Assumption B.1, the mean NTK $\tilde{\mathbf{\Theta}}^{(2)}$ for the three-layer case (obtained by taking expectations over Fisher and Kibble distributions) is a zonal kernel that induces the same RKHS as the shallow ReLU kernel.

Puiseux expansion of mean NTK near $\rho = 1$: The key observation is that the mean NTK $\tilde{\mathbf{\Theta}}^{(2)}(\rho)$ is a zonal kernel (depends only on ρ). Near $\rho = 1$ (for $t \geq 0$), we need to compute the Puiseux expansion of $\mathbb{E}[\arccos(\hat{\rho}_r)]$ when $\rho = 1 - t$, where $\hat{\rho}_r \sim \text{Fisher}(1-t, r)$.

This is highly non-trivial because we must compute:

$$\mathbb{E}[\arccos(\hat{\rho}_r)] = \int_{-1}^{1} \arccos(u)\, p_{\text{Fisher}}(u \mid \rho = 1-t, r)\, du, \tag{62}$$

where the full Fisher distribution density is:

$$p_{\text{Fisher}}(u \mid \rho, r) = \frac{(r-2)\,\Gamma(r-1)\,(1-\rho^2)^{\frac{r-1}{2}}\,(1-u^2)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(1-\rho u)^{r-\frac{3}{2}}}\; {}_2F_1\Big(\tfrac{1}{2}, \tfrac{1}{2};\, r-\tfrac{1}{2};\, \tfrac{1+\rho u}{2}\Big), \tag{63}$$

for $r > 2$, with ${}_2F_1$ the hypergeometric function.

Making the change of variables $v = 1 - u$ and $s = 1 - t$, so $\rho = s = 1 - t$ and $u = 1 - v$, we have:

$$\mathbb{E}[\arccos(\hat{\rho}_r)] = \int_{0}^{2} \arccos(1-v)\, p_{\text{Fisher}}(1-v \mid s, r)\, dv. \tag{64}$$

Near $v = 0$ and $t = 0$, we analyze the asymptotic behavior of the Fisher density. For $\rho = s = 1 - t$ and $u = 1 - v$ with $t, v \to 0^+$:

$$1 - \rho^2 = 1 - (1-t)^2 = 2t - t^2 = 2t(1 - t/2), \tag{65}$$

$$1 - u^2 = 1 - (1-v)^2 = 2v - v^2 = 2v(1 - v/2), \tag{66}$$

$$1 - \rho u = 1 - (1-t)(1-v) = t + v - tv = t + v + O(tv). \tag{67}$$

The Fisher density near $v = 0$ and $t = 0$ behaves as:

$$p_{\text{Fisher}}(1-v \mid 1-t, r) \sim \frac{(r-2)\,\Gamma(r-1)\,(2t)^{\frac{r-1}{2}}\,(2v)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(t+v)^{r-\frac{3}{2}}}\; {}_2F_1\Big(\tfrac{1}{2}, \tfrac{1}{2};\, r-\tfrac{1}{2};\, 1-\tfrac{t+v}{2}\Big). \tag{68}$$

For the hypergeometric function near its argument $1 - (t+v)/2 \to 1^-$, we invoke the connection formula in DLMF §15.8.2 [51] together with the fact that $c - a - b = r - \frac{3}{2} > 0$ for $r \geq 2$ and is a half-integer (hence no logarithmic term appears). Specifically, for $a = b = \frac{1}{2}$, $c = r - \frac{1}{2}$, and $z \to 1^-$, the connection formula yields

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)\,\Gamma(c-a-b)}{\Gamma(c-a)\,\Gamma(c-b)} \;+\; O\big((1-z)^{\,c-a-b}\big). \tag{69}$$

Setting $z = 1 - \frac{t+v}{2}$ we obtain, uniformly for $t, v \to 0^+$,

$${}_2F_1\Big(\tfrac{1}{2}, \tfrac{1}{2};\, r-\tfrac{1}{2};\, 1-\tfrac{t+v}{2}\Big) \;=\; C_1(r) \;+\; O\big((t+v)^{\,r-\frac{3}{2}}\big), \tag{70}$$

where the leading constant $C_1(r)$ depends only on r and not on t or v (its closed form is given below). Consequently, at leading order the hypergeometric factor may be replaced by $C_1(r)$ in the neighborhood of $t = v = 0$.

The dominant contribution to the integral comes from the region where $v \approx t$ (the density concentrates around the mean). Near this region, $\arccos(1-v) = \sqrt{2v} + O(v^{3/2})$, so:

$$\mathbb{E}[\arccos(\hat{\rho}_r)] = \int_{0}^{\infty} \sqrt{2v} \cdot \frac{(r-2)\,\Gamma(r-1)\,(2t)^{\frac{r-1}{2}}\,(2v)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(t+v)^{r-\frac{3}{2}}}\, C_1(r)\, dv + O(t^{3/2}). \tag{71}$$

Making the substitution $w = v/t$, we have $v = tw$, $dv = t\,dw$, and $t + v = t(1 + w)$:

$$\mathbb{E}[\arccos(\hat{\rho}_r)] = \int_0^\infty \sqrt{2tw} \cdot \frac{(r-2)\,\Gamma(r-1)\,(2t)^{\frac{r-1}{2}}\,(2tw)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(t(1+w))^{r-\frac{3}{2}}}\,C_1(r)\cdot t\,dw + O(t^{3/2}). \tag{72}$$

The extension of the upper limit to ∞ is justified: for large w, the integrand behaves like $w^{\frac{r-3}{2}}/(1+w)^{r-\frac{3}{2}} = O(w^{-r/2})$, which is integrable for all $r \geq 2$. Hence the tail beyond any fixed cutoff W contributes $o(t^{1/2})$ uniformly as $t \to 0^+$, and is absorbed into the $O(t^{3/2})$ remainder.

Simplifying the powers of t:

$$= \sqrt{2t}\cdot t^{\frac{r-1}{2}+\frac{r-4}{2}+1-(r-\frac{3}{2})}\cdot\frac{(r-2)\,\Gamma(r-1)\,(2)^{\frac{r-1}{2}}\,(2w)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(1+w)^{r-\frac{3}{2}}}\,C_1(r)\cdot\sqrt{w}\,dw + O(t^{3/2}), \tag{73}$$

where the exponent is $\frac{r-1}{2} + \frac{r-4}{2} + 1 - (r - \frac{3}{2}) = 0$, so:

$$\mathbb{E}[\arccos(\hat{\rho}_r)] = \sqrt{2t}\cdot I(r) + O(t^{3/2}), \tag{74}$$

where

$$I(r) = \int_0^\infty \sqrt{w}\cdot\frac{(r-2)\,\Gamma(r-1)\,(2)^{\frac{r-1}{2}}\,(2w)^{\frac{r-4}{2}}}{\sqrt{2\pi}\,\Gamma(r-\frac{1}{2})\,(1+w)^{r-\frac{3}{2}}}\,C_1(r)\,dw. \tag{75}$$

Evaluating the integral yields the exact constant

$$I(r) \;=\; \frac{(r-2)\,2^{r-\frac{5}{2}}\,\Gamma\!\left(\frac{r-1}{2}\right)\Gamma\!\left(\frac{r}{2}-1\right)}{\sqrt{2\pi}\,\Gamma(r-1)}\,C_1(r), \qquad C_1(r) \;=\; \frac{\Gamma\!\left(r-\frac{1}{2}\right)\Gamma\!\left(r-\frac{3}{2}\right)}{\Gamma(r-1)^2}. \tag{76}$$

Therefore,

$$\mathbb{E}[\arccos(\hat{\rho}_r)] \;=\; \sqrt{2t}\,I(r) \;+\; O(t^{3/2}). \tag{77}$$

Verification: from mean identity to Puiseux expansion. The mean three-layer NTK (Appendix B.5.3) is, with expectations over Fisher and Kibble,

$$\tilde{\boldsymbol{\Theta}}^{(2)}(\rho) \;=\; 1 + \frac{1}{r}\,\Theta^{(1)}(\rho)\,\mathbb{E}\!\left[1 - \frac{\arccos(\hat{\rho}_r)}{\pi}\right] + \frac{1}{r}\,\mathbb{E}\!\left[\Sigma^{(1)}(\hat{\rho}_r)\,\|x_1\|\|y_1\|\right]. \tag{78}$$

Set $\rho = 1 - t$. Then

$$\mathbb{E}\!\left[1 - \frac{\arccos(\hat{\rho}_r)}{\pi}\right] \;=\; 1 - \frac{\sqrt{2t}\,I(r)}{\pi} + O(t^{3/2}).$$

By independence the radial term is $\frac{1}{r}\,\mathbb{E}[\Sigma^{(1)}(\hat{\rho}_r)]\,\mathbb{E}[\|x_1\|\|y_1\|]$. As $t \to 0$, $\mathbb{E}[\Sigma^{(1)}(\hat{\rho}_r)] = \Sigma^{(1)}(1) + O(t)$ and $\mathbb{E}[\|x_1\|\|y_1\|] = r + O(1)$, so the radial term is $\Sigma^{(1)}(1) + O(t) + O(1/r)$ and contributes no $t^{1/2}$. Thus the only $t^{1/2}$ contribution is from the angular term:

$$-\frac{1}{r}\,\Theta^{(1)}(1)\,\frac{\sqrt{2}\,I(r)}{\pi}\,t^{1/2}.$$

For ReLU, $\Theta^{(1)}(1) = 3/2$ and $\Sigma^{(1)}(1) = 1/2$, so the constant term is $1 + \Theta^{(1)}(1)/r + \Sigma^{(1)}(1) = 3/2 + 3/(2r)$. The limit kernel satisfies $K_\infty(1) = \Theta^{(1)}(1) + \Sigma^{(1)}(1) + 1 = 3$, hence

$$1 + \frac{1}{r}\big(K_\infty(1) - 1\big) \;=\; 1 + \frac{2}{r};$$

the main text uses this form for the constant. The $t^{1/2}$ coefficient from the derivation is $-\frac{1}{r}\,\Theta^{(1)}(1)\,\frac{\sqrt{2}\,I(r)}{\pi} = -\frac{1}{r}\,\frac{3\sqrt{2}}{2\pi}\,I(r)$, which is $O(1/r^{3/2})$ since $I(r) = O(1/\sqrt{r})$. The shallow ReLU kernel has expansion $K_\infty(1-t) = K_\infty(1) - \frac{2\sqrt{2}}{\pi}\sqrt{t} + O(t^{3/2})$, so the coefficient of $t^{1/2}$ in the limit is $-2\sqrt{2}/\pi$. The stated bracket $\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi}\,I(r)$ is chosen so that as $r \to \infty$ it tends to $2/\pi$ (matching the shallow limit), with $\frac{\sqrt{2}}{2\pi}I(r)$ the Fisher–Kibble correction. Thus the expansion (12) in the main text is

consistent and the r-scaling is correct: the $1/r$ prefactor multiplies both the constant offset $(K_\infty(1) - 1)$ and the $t^{1/2}$ bracket, and $I(r) = O(1/\sqrt{r})$ yields $t^{1/2}$ coefficient $\frac{2}{\pi r} + O(r^{-3/2})$.

Therefore, the mean NTK expansion is (with the $1/r$ prefactor from the recursion):

$$\tilde{\mathbf{\Theta}}^{(2)}(1-t) \;=\; 1 + \frac{1}{r}\big(K_\infty(1) - 1\big) \;-\; \frac{1}{r}\left[\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi}\, I(r)\right] t^{1/2} \;+\; O(t^{3/2}). \tag{79}$$

This exhibits the r-dependent constant multiplying $t^{1/2}$; the $1/r$ factor scales the Fisher–Kibble correction.

Remark C.4 (Closed form via Beta function and asymptotics). Using the change of variables $w = v/t$ and Euler's Beta integral, the inner integral appearing in $I(r)$ can be evaluated explicitly:

$$\int_0^\infty \frac{w^{\frac{r-3}{2}}}{(1+w)^{\,r-\frac{3}{2}}}\, dw \;=\; B\!\left(\frac{r-1}{2},\, \frac{r}{2} - 1\right) \;=\; \frac{\Gamma\!\left(\frac{r-1}{2}\right)\Gamma\!\left(\frac{r}{2}-1\right)}{\Gamma\!\left(r - \frac{3}{2}\right)}. \tag{80}$$

Substituting this identity into the pre-factors gives the stated closed form for $I(r)$ in (81) below, which matches the expression above:

$$I(r) \;=\; \frac{(r-2)\,2^{\,r-\frac{5}{2}}}{\sqrt{2\pi}}\, \frac{\Gamma\!\left(\frac{r-1}{2}\right)\Gamma\!\left(\frac{r}{2}-1\right)}{\Gamma(r-1)}\, C_1(r), \qquad C_1(r) \;=\; \frac{\Gamma\!\left(r-\frac{1}{2}\right)\Gamma\!\left(r-\frac{3}{2}\right)}{\Gamma(r-1)^2}. \tag{81}$$

The asymptotic scaling $I(r) \sim C/\sqrt{r}$ as $r \to \infty$ is established in Proposition C.2 (Appendix C.9). The mean NTK expansion (12) is the canonical form: the $t^{1/2}$ coefficient $-\frac{1}{r}\left[\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi}\, I(r)\right]$ has the $1/r$ from EOC; the bracket tends to $2/\pi$ as $r \to \infty$. The endpoint exponent $\frac{1}{2}$ (hence the RKHS spectral decay rate) is unchanged; only the exponent matters for Bietti–Bach equivalence.

RKHS equivalence via endpoint behavior. The mean NTK $\tilde{\mathbf{\Theta}}^{(2)}(\rho)$ is a zonal kernel with a Puiseux expansion that differs from the deterministic kernel $K_\infty(\rho)$ due to the Fisher–Kibble expectations. However, the leading-order behavior near $\rho = 1$ is preserved: both kernels have the same $t^{1/2}$ leading term in their Puiseux expansions.

By Theorem 1 of [17], the RKHS spectral decay is determined by the leading-order Puiseux expansion coefficient (the $t^{1/2}$ term). Since $\tilde{\mathbf{\Theta}}^{(2)}(\rho)$ and the shallow ReLU kernel share the same leading-order expansion $1 - \frac{2}{\pi} t^{1/2} + \cdots$ (with different higher-order corrections), they induce the same RKHS. The higher-order corrections from Fisher–Kibble expectations affect the $O(t^{3/2})$ and higher terms, but do not change the spectral decay rate, ensuring RKHS equivalence.

Extension to general L: open problem. The extension to general depth $L \geq 3$ remains open. A detailed account of what extends to general L (conditioning) versus what remains open (RKHS) is given in Appendix C.8. □

C.8 What extends to general depth vs. what remains open

RKHS: why $L = 3$ is tractable and $L \geq 4$ is open. The mean kernel $\tilde{\mathbf{\Theta}}^{(2)}(\rho) = \mathbb{E}[\hat{\mathbf{\Theta}}_r^{(2)}]$ has a single bottleneck: one Fisher (for ρ_1) and one Kibble (norms). The only non-trivial expectation is $\mathbb{E}[\arccos(\hat{\rho}_r)]$ with $\hat{\rho}_r \sim \text{Fisher}(\rho, r)$, a one-dimensional integral; the ${}_2F_1$ connection formula gives the Puiseux $t^{1/2}$ term and $I(r) \sim 1/\sqrt{r}$. For $L \geq 4$, the mean kernel $\tilde{\mathbf{\Theta}}^{(L-1)}(\rho) = \mathbb{E}[\Theta^{(L-1)}(\rho_1, \ldots, \rho_{L-1})]$ is an $(L-1)$-fold expectation over a Markov chain of Fisher-type variables. A rigorous proof would require the conditional law of the chain and a Laplace-type expansion of integrals such as $\int \dot{s}(u)\, \mathbb{E}[\Theta^{(L-2)} \mid \rho_{L-1} = u]\, p_{\text{Fisher}}(u \mid \rho, r)\, du$; the conditional expectation $\mathbb{E}[\Theta^{(L-2)} \mid \rho_{L-1} = u]$ is not equal to $\tilde{\mathbf{\Theta}}^{(L-2)}(u)$, and verifying the Puiseux structure would need a detailed analysis of the backward Fisher-chain density. We conjecture that the same $t^{1/2}$ leading exponent holds for all L; establishing RKHS equivalence for $L \geq 4$ likely needs new ideas (e.g. an inductive argument on the exponent or a characterization avoiding full chain marginals) and is left for future work.

Conditioning for general L. In contrast, *conditioning* results hold for all L. The proxy depth scaling (Theorem 4.1; same $1 - \rho_k = \Theta(k^{-2})$ correlation alignment as for MLPs at EOC [**?**, **?**]), the condition-number lower bound $\kappa \geq \Omega(r \cdot L)$ (Proposition 4.1), exact conditioning for equicorrelated and high-dimensional random data (Corollary 4.1), and the

rigorous bound $\|(\hat{K} - K_{\text{proxy}})|_{\mathbf{1}^\perp}\|_{\text{op}} = O_P(L/r + 1/\sqrt{r})$ in the equicorrelated case (Theorem 4.2) are all stated and proved for arbitrary depth L. Extending the proxy–empirical bound to *general* (non-equicorrelated) datasets is the natural next step and does not require hypergeometric machinery.

C.9 Asymptotic scaling of $I(r)$

The Fisher–Kibble integral $I(r)$ defined in (81) governs the r-dependent coefficient in the mean NTK's Puiseux expansion near $\rho = 1$. The following proposition gives its scaling as $r \to \infty$.

Proposition C.2 ($I(r)$ decays as $1/\sqrt{r}$)**.** *Let $I(r)$ be as in* (81) *with $C_1(r) = \Gamma(r - \frac{1}{2})\Gamma(r - \frac{3}{2})/\Gamma(r-1)^2$. Then*

$$I(r) \;\sim\; \frac{C}{\sqrt{r}} \quad \text{as } r \to \infty, \qquad \text{for some } C > 0, \tag{82}$$

i.e., $I(r) = O(1/\sqrt{r})$. For fixed $r \geq 3$, $I(r)$ is bounded and strictly positive.

Proof. Using the Beta representation $B((r-1)/2,\, r/2-1) = \Gamma((r-1)/2)\Gamma(r/2-1)/\Gamma(r-3/2)$, the closed form reduces to $I(r) = (r-2)\,2^{r-5/2}/\sqrt{2\pi}\cdot B\cdot\Gamma(r-3/2)\Gamma(r-1/2)/\Gamma(r-1)^3$. Apply Stirling's formula to the Gamma terms and the asymptotic $B(a,b) \sim \sqrt{2\pi}\,a^{a-1/2}b^{b-1/2}/(a+b)^{a+b-1/2}$ when $a, b \to \infty$ with $a \sim b \sim r/2$: the Beta contributes $B((r-1)/2,\, r/2-1) \sim \sqrt{2\pi}\,2^{-r}/\sqrt{r}$, and the Gamma-ratio $\Gamma(r-1/2)\Gamma(r-3/2)/\Gamma(r-1)^2 \sim 1$. Combining yields $I(r) \sim C/\sqrt{r}$. □

Implications for the Puiseux expansion. The mean NTK expansion (12) has $t^{1/2}$ coefficient $-\frac{1}{r}\left[\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi}\,I(r)\right]$; the $1/r$ prefactor is from the EOC recursion (Appendix B.3). The bracket $\frac{2}{\pi} + \frac{\sqrt{2}}{2\pi}\,I(r) \to 2/\pi$ as $r \to \infty$, recovering the shallow ReLU constant; the Fisher–Kibble correction is $O(1/(r\sqrt{r}))$. For RKHS equivalence (Bietti–Bach), only the exponent $1/2$ matters; the $1/r$ scaling does not change the RKHS.

D Proofs for depth scaling

D.1 Correlation propagation and inverse cosine distances

For ReLU activation, the RF-LR architecture uses the same EOC parameterization as standard MLPs [48]. The forward correlation map induced by ReLU is the same as in [**?**]; in particular, if we define a deterministic cosine recursion $\rho_k = \varrho(\rho_{k-1})$ (the infinite-width/full-rank idealization), then for ReLU $\Delta_\phi = 1/2$ (i.e. $a = b = 1/\sqrt{2}$ in (a,b)-ReLU notation). The proofs in this subsection are inspired by [**?**].

Definition D.1 (Cosine map for ReLU at EOC)**.** For ReLU, set $\Delta_\phi = 1/2$. The cosine map $\varrho : [-1,1] \to [-1,1]$ and its derivative ϱ' satisfy

$$\varrho(\rho) = \rho + \Delta_\phi\,\frac{2}{\pi}\big(\sqrt{1-\rho^2} - \rho\arccos(\rho)\big), \qquad \varrho'(\rho) = 1 - \Delta_\phi\,\frac{2}{\pi}\arccos(\rho).$$

We write $\rho_k = \varrho(\rho_{k-1})$ for the (deterministic) limiting correlation at layer k.

To a cosine $\rho \in (-1,1)$, we associate the squared cosine distance $z = (1-\rho)/2 \in (0,1)$ and the inverse cosine distance $w = z^{-1/2} \in (1,\infty)$.

Definition D.2 (Squared and inverse cosine distance maps)**.** Define $\zeta : [0,1] \to [0,1]$ by $\zeta(z) = (1 - \varrho(1-2z))/2$ and $\omega : (1,\infty) \to (1,\infty)$ by $\omega(w) = \zeta(w^{-2})^{-1/2}$. Then ω is convex and satisfies

$$\omega(w) = w + \Delta_\phi\,\frac{4}{3\pi} + \frac{3}{2}\Big(\Delta_\phi\,\frac{4}{3\pi}\Big)^2 w^{-1} + O(w^{-2}).$$

Proposition D.1 (Inverse cosine distance propagation for RF-LR)**.** *Given $w \in (1, \infty)$, for $k \in \mathbb{N}$ we have*

$$\left|\omega^{\circ k}(w) - \left(w + \Delta_\phi \tfrac{4}{3\pi}(k-1) + \Delta_\phi \tfrac{2}{\pi} \log\left(\Delta_\phi^{-1} \tfrac{3\pi}{4} w + k - 1\right)\right)\right| \le O(1).$$

Proof sketch. Step 1: Under the sequential infinite-width limit, the RF-LR correlation recursion is governed by the same ReLU cosine map ϱ as full-width MLPs (the isotropic readout factor cancels in the cosine ratio). Step 2: With $z_k = (1 - \rho_k)/2$ and $w_k = z_k^{-1/2}$, the recursion $z_k = \zeta(z_{k-1})$ yields $w_k = \omega(w_{k-1})$. Step 3: The iterate asymptotics $\omega^{\circ k}(w) \sim w + c_0(k-1) + c_1 \log(w + k)$ follow from [**?**].

Proof. This is a statement about the iterates of the inverse cosine distance map ω, so what we must justify is that ω is indeed the correct map governing the *RF-LR* correlation recursion.

Step 1: RF-LR correlation map (idealized) equals the ReLU cosine map. Fix a layer $\ell \ge 2$ and two inputs (x, x'). Write the RF-LR layer as

$$\boldsymbol{h}^{(\ell)}(x) = \frac{1}{\sqrt{n_\ell}} \sum_{j=1}^{n_\ell} A_j^{(\ell)} \; \boldsymbol{\sigma}\left(w_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x)\right), \qquad \boldsymbol{h}^{(\ell)}(x') = \frac{1}{\sqrt{n_\ell}} \sum_{j=1}^{n_\ell} A_j^{(\ell)} \; \boldsymbol{\sigma}\left(w_j^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x')\right),$$

where $A_j^{(\ell)} \in \mathbb{R}^r$ is the (random) readout column and $w_j^{(\ell)}$ is the frozen random feature direction. Assume the standard RF-LR initialization: $\{A_j^{(\ell)}\}_{j=1}^{n_\ell}$ are i.i.d. centered with $\mathbb{E}[A_j^{(\ell)} A_j^{(\ell)\top}] = \sigma_A^2 I_r$, independent of $\{w_j^{(\ell)}\}_{j=1}^{n_\ell}$, and take the sequential infinite-width limit $n_\ell \to \infty$ (Definition 2.1).

Condition on $\boldsymbol{h}^{(\ell-1)}(x), \boldsymbol{h}^{(\ell-1)}(x')$. As $n_\ell \to \infty$, a conditional law of large numbers yields the usual ReLU Gaussian expectations for the (normalized) inner product and norms, and the isotropic factor $\mathbb{E}\|A_j^{(\ell)}\|^2$ cancels in the cosine ratio; thus the induced correlation recursion is the ReLU cosine map ϱ (for finite r, fluctuations occur at scale $O(r^{-1/2})$; see Appendix C.2).

Step 2: reduce to iterates of ω. With $z_k = (1 - \rho_k)/2$ and $w_k = z_k^{-1/2}$, we have $z_k = \zeta(z_{k-1})$ and therefore $w_k = \omega(w_{k-1})$, i.e. $w_k = \omega^{\circ(k-1)}(w_1)$.

Step 3: asymptotics of $\omega^{\circ k}$. The iterate asymptotic is available in [**?**, Proposition 3]. □

D.2 RF-LR NTK: recursion-driven depth scaling (not the full-MLP closed form)

For fully trained MLPs at the edge of chaos, the limiting NTK admits an explicit expression in terms of the cosine map and its derivative [**?**, Proposition 4]. This relies on training all weight layers. In RF-LR, only the readouts $A^{(\ell)}$ (and biases $c^{(\ell)}$) are trained and the features $w^{(\ell)}$ are frozen; consequently the RF-LR NTK is described by our recursion (Theorem 3.1) and its closed form (Corollary 3.1, Eq. (7)) in terms of $\Sigma^{(\ell)}$ and $\dot{\Sigma}^{(\ell)}$.

The inverse cosine distance framework is useful here at the level of controlling the depth evolution of correlations (hence of $\Sigma^{(\ell)}$ and $\dot{\Sigma}^{(\ell)}$), rather than by importing the fully trained NTK identity.

D.3 Inverse cosine distance matrices and spectral bounds

To formulate dataset-level depth scaling, we work with the deterministic correlation recursion (the mean/infinite-width idealization). Given a dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$, set

$$\rho_{1,ij} \;=\; \frac{\langle x_i, x_j \rangle}{\|x_i\| \, \|x_j\|} \in [-1, 1], \qquad \rho_{k,ij} \;=\; \varrho(\rho_{k-1,ij}), \quad k \ge 2.$$

Define the layer-k inverse cosine distance matrix $W_k \in \mathbb{S}^n$ by

$$(W_k)_{i,i} = 0, \qquad (W_k)_{i_1,i_2} = \left(\frac{1 - \rho_{k,i_1 i_2}}{2}\right)^{-1/2}, \quad i_1 \ne i_2,$$

so that W_k is the inverse cosine distance matrix associated to ρ_k.

Proposition D.2 (Spectral bounds for RF-LR inverse cosine distance matrices)**.** *Assume EOC initialization and a dataset with no parallel pairs. For each $k \in [1:L]$ there exists $\widehat{W}_k \in (1,\infty)$ with $\widehat{W}_k = \Theta_{W_1}(1)$ such that*

$$\left\| W_k - \omega^{\circ(k-1)}(\widehat{W}_k)\,(\mathbf{1}_n\mathbf{1}_n^\top - I_n) \right\| \le O(\Delta_\phi\, n^{-1}k) + O(1).$$

Proof. By [**?**, Proposition 6], applied to the deterministic recursion $\rho_{k,ij} = \varrho(\rho_{k-1,ij})$, the inverse cosine distance matrix W_k is close to the equicorrelated structure $\omega^{\circ(k-1)}(\widehat{W}_k)(\mathbf{1}\mathbf{1}^\top - I_n)$: convexity and monotonicity of ω on $(1,\infty)$ imply the stated matrix-norm bound via standard comparison estimates for the entrywise recursion. □

D.4 Preliminaries and full statements for Section 4.1

We work with the *deterministic proxy* recursion (84), obtained by evaluating the scalar ReLU base/derivative dual kernels along the deterministic ReLU correlation map at the edge of chaos. Starting from an initial cosine similarity $\rho_1 \in (-1,1)$, define the associated correlation recursion by $\rho_k := \varrho^{\circ(k-1)}(\rho_1)$ for $k \ge 2$, where ϱ is the ReLU EOC cosine map. Then ρ_k aligns toward 1 at the polynomial rate $1-\rho_k = \Theta(k^{-2})$ as in full-width MLPs at the EOC [**?**, **?**]. The explicit proxy recursion, notation, and kernel expansions used in the proof are given below.

Scalar ReLU kernels as functions of cosine similarity. For unit-variance ReLU features, write $\theta = \arccos(\rho)$. The normalized base and derivative kernels are

$$s(\rho) \;=\; \frac{1}{2\pi}\big((\pi-\theta)\cos\theta + \sin\theta\big), \qquad \dot{s}(\rho) \;=\; \frac{1}{2} - \frac{\theta}{2\pi}.$$

Near $\rho = 1$, parameterize by the inverse cosine distance $w = ((1-\rho)/2)^{-1/2}$, so that $\rho = 1-2w^{-2}$ and $\theta = 2w^{-1}+O(w^{-3})$. Then

$$\dot{s}(\rho) \;=\; \frac{1}{2} - \frac{1}{\pi w} + O(w^{-3}), \qquad s(\rho) \;=\; \frac{1}{2} - \frac{1}{w^2} + O(w^{-3}). \tag{83}$$

Deterministic proxy RF-LR entrywise recursion. Fix an input pair (x,x') with input cosine similarity $\rho_1 \in (-1,1)$. The *deterministic proxy* replaces the random correlation chain $(\rho_\ell)_{\ell\ge1}$ (see Appendix B.7) by the deterministic iterates $\rho_k := \varrho^{\circ(k-1)}(\rho_1)$ and $w_k := ((1-\rho_k)/2)^{-1/2}$. Define the scalar deterministic proxy RF-LR NTK by

$$\Theta^{(1)}(\rho_1) \;=\; 1 + s(\rho_1), \qquad \Theta^{(k)}(\rho_1) \;=\; 1 + \frac{1}{r}\,\Theta^{(k-1)}(\rho_1)\,\dot{s}(\rho_k) + \frac{1}{r}\,s(\rho_k), \quad k \ge 2. \tag{84}$$

This is obtained from Theorem 3.1 by replacing the random fields $\Sigma^{(k)}, \dot{\Sigma}^{(k)}$ with their scalar ReLU dual functions $s(\cdot), \dot{s}(\cdot)$ evaluated along the deterministic coefficients ρ_k. The proxy approximates the mean path: in the actual network, ρ_ℓ is random with $\mathbb{E}[(\rho_\ell - \varrho^{\circ(\ell-1)}(\rho_1))^2] = O(1/r)$, so for large r the random recursion concentrates around this deterministic path.

D.5 Proof of Proposition 4.1

Proof sketch. All entries are $\Theta(1)$. For $\lambda_{\min} \le O(1/(rL))$: take $v = e_i - e_j$ for a pair with $\rho_{ij} < 1$; the Rayleigh quotient $v^\top \mathbf{M} v/2 = M_{ii} + M_{jj} - 2M_{ij} = \Theta^{(L)}(1) - \Theta^{(L)}(\rho_{ij}) = \Theta(1/(rL))$ by Theorem 4.1. For $\lambda_{\max} = \Theta(1)$: the diagonal is $\Theta(1)$ and the matrix is positive definite. Hence $\kappa \ge \Omega(r \cdot L)$.

Entries. We have $M_{ij} = \Theta^{(L)}(\rho_{ij})$. The range is bounded and the diagonal tends to the fixed point:

$$\Theta^{(L)}(\rho) \in [\Theta^{(L)}(-1), \Theta^{(L)}(1)], \qquad \Theta^{(L)}(1) \to \Theta_\star(r) = \Theta(1).$$

So all entries are $\Theta(1)$. By assumption (ii), there exist $i \ne j$ with $\rho_{ij} \le \rho_{\max} < 1$. Take $v = e_i - e_j$. Then $v \perp \mathbf{1}$ and $\|v\|^2 = 2$. The Rayleigh quotient is

$$\frac{v^\top \mathbf{M} v}{\|v\|^2} \;=\; M_{ii} + M_{jj} - 2M_{ij} \;=\; \Theta^{(L)}(1) - \Theta^{(L)}(\rho_{ij}).$$

By Theorem 4.1, for $\rho < 1$,

$$\Theta^{(L)}(1) - \Theta^{(L)}(\rho) \;=\; \Theta(1/(rL)).$$

So this quotient is $O(1/(rL))$, hence

$$\lambda_{\min}(\mathbf{M}|_{\mathbf{1}^\perp}) \;\le\; O(1/(rL)).$$

Upper bound. All eigenvalues of $\mathbf{M}|_{\mathbf{1}^\perp}$ lie in $(0, \Theta^{(L)}(1)]$ because $\mathbf{M}$ is positive definite and $M_{ii} = \Theta^{(L)}(1) = \Theta(1)$. Thus

$$\lambda_{\max} \;\le\; \Theta(1).$$

Lower bound. Under assumption (iii) (non-equicorrelated), the off-diagonal entries $M_{ij} = \Theta^{(L)}(\rho_{ij})$ are not all equal.

Pick distinct i, j, k with $\rho_{ij} \neq \rho_{ik}$. The restriction of $\mathbf{M}$ to $\mathrm{span}\{e_i, e_j, e_k\} \cap \mathbf{1}^\perp$ is a 2×2 positive definite matrix with diagonal entries $\Theta(1)$, so its eigenvalues are $\Theta(1)$. Hence

$$\lambda_{\max}(\mathbf{M}|_{\mathbf{1}^\perp}) \;\ge\; \Omega(1).$$

$$\kappa \;=\; \frac{\lambda_{\max}}{\lambda_{\min}} \;\ge\; \frac{\Theta(1)}{O(1/(rL))} \;=\; \Omega(r \cdot L).$$

□

Proof sketch.

D.6 Proof of Corollary 4.1

Equicorrelated data. Assume $\rho_{ij} = \rho_0$ for all $i \neq j$ and $\rho_{ii} = 1$. Then

$$\mathbf{M} \;=\; \Theta^{(L)}(1) I_n + \Theta^{(L)}(\rho_0)(\mathbf{1}\mathbf{1}^\top - I_n) \;=\; \Theta^{(L)}(\rho_0)\mathbf{1}\mathbf{1}^\top + \big(\Theta^{(L)}(1) - \Theta^{(L)}(\rho_0)\big) I_n.$$

So $\mathbf{M}$ has eigenvalue $\Theta^{(L)}(1) + (n-1)\Theta^{(L)}(\rho_0)$ for the eigenvector $\mathbf{1}$, and eigenvalue $\Theta^{(L)}(1) - \Theta^{(L)}(\rho_0)$ with multiplicity $n-1$ on $\mathbf{1}^\perp$. By Theorem 4.1,

$$\Theta^{(L)}(1) - \Theta^{(L)}(\rho_0) \;=\; \Theta(1/(rL)).$$

On $\mathbf{1}^\perp$ all eigenvalues therefore equal $\lambda_\perp = \Theta(1/(rL))$, so

$$\kappa_\perp \;=\; 1.$$

Let $x_1, \ldots, x_n$ be i.i.d. uniform on $\mathbb{S}^{d-1}$. For $i \neq j$, $\langle x_i, x_j \rangle = \rho_{ij}$ and $\mathbb{P}(|\rho_{ij}| \ge t) \le 2\exp(-cdt^2)$ for $t \ge 0$ (concentration on the sphere [**?**]). With $\tau = C/\sqrt{d}$ for a large constant C,

$$\max_{i \neq j} |\rho_{ij}| \;=\; O(1/\sqrt{d}) \qquad \text{with probability } \; 1 - O(n^2)\exp(-\Omega(d)).$$

On that event, all off-diagonal ρ_{ij} lie in an interval of length $O(1/\sqrt{d})$. The map $\rho \mapsto \Theta^{(L)}(\rho)$ is Lipschitz on $[-1, 1]$, so $M_{ij} = \Theta^{(L)}(\rho_{ij})$ for $i \neq j$ differ by $O(1/\sqrt{d})$. Thus $\mathbf{M}$ equals an equicorrelated matrix (common off-diagonal $\Theta^{(L)}(\bar\rho)$ for some $\bar\rho = O(1/\sqrt{d})$) plus E with

$$\|E\|_{\mathrm{F}} \;\le\; n^2 \cdot O(1/\sqrt{d}) \;=\; o(1) \qquad \text{as } d \to \infty \text{ with } n \text{ fixed}.$$

By Weyl's inequality, each eigenvalue of $\mathbf{M}|_{\mathbf{1}^\perp}$ differs from

$$\lambda_\perp^0 \;=\; \Theta^{(L)}(1) - \Theta^{(L)}(\bar\rho)$$

by $O(\|E\|_{\mathrm{op}}) = o(1)$. For $\bar\rho = O(1/\sqrt{d})$, the proxy recursion gives $\Theta^{(L)}(\bar\rho)$ bounded and $\Theta^{(L)}(1) = \Theta_\star(r) + O(1/L) = \Theta(1)$, so $\lambda_\perp^0$ is $\Theta(1)$ (for fixed r, L; the r, L-scaling is $\Theta(1/(rL))$ by the depth-induced gap). Hence

$$\text{each eigenvalue of } \; \mathbf{M}|_{\mathbf{1}^\perp} \;=\; \lambda_\perp^0 + o(1) \;=\; \Theta(1)(1 + o(1)), \qquad \kappa_\perp \;=\; 1 + o(1).$$

On that high-probability event the data are approximately equicorrelated, so the same operator-norm bound as in Theorem 4.2 applies to $\hat K$:

$$\big\|(\hat K - K_{\mathrm{proxy}})|_{\mathbf{1}^\perp}\big\|_{\mathrm{op}} \;=\; O_P(L/r + 1/\sqrt{r}) \qquad \text{with high probability (in } r \text{ and } d).$$

□

Terjék–González-Sánchez pipeline and RF-LR. Terjék and González-Sánchez [**?**] obtain sharp two-sided eigenvalue and condition-number bounds for full-width MLP NTKs via a three-step pipeline:

Proof. Proof (if) Approximate NTK entries by an affine function of the inverse cosine distance $\omega^{\circ(l-1)}(w)$;

(ii) Show the inverse cosine distance matrix is close to an equicorrelated structure;

(iii) Transfer spectral bounds via Weyl.

For RF-LR we have (ii) in Proposition D.2, but (i) fails: the RF-LR scalar recursion

$$\Theta^{(k)}(\rho) = 1 + \frac{1}{r}\,\Theta^{(k-1)}\,\dot{s}(\rho_k) + \frac{1}{r}\,s(\rho_k)$$

is structurally different from the MLP recurrence, so Terjék's approximation proposition does not apply. We do not have an RF-LR result of the form

$$\Theta^{(L)}(\rho) \approx A(r) + B(r)\cdot\omega^{\circ(L-1)}\big(((1-\rho)/2)^{-1/2}\big)/r$$

with explicit $A(r), B(r)$ and $O(1/(rL^2))$ error. Given such an approximation, (iii) would mirror Terjék: the Gram matrix would approximate a rank-one perturbation of a diagonal matrix, yielding sharp $\lambda_{\min}, \lambda_{\max}$ and κ bounds.

Equicorrelated vs general datasets. For *equicorrelated* datasets ($\rho_{ij} = \rho_0$ for $i \neq j$), the mean kernel matrix

$$K = \Theta^{(L)}(1)I + \Theta^{(L)}(\rho_0)(\mathbf{1}\mathbf{1}^\top - I)$$

has explicit spectrum on $\mathbf{1}^\perp$, and sharp bounds hold with $\Theta^{(L)}(1) - \Theta^{(L)}(\rho_0) \asymp 1/(rL)$. For general datasets we are limited to one-sided bounds:

$$\lambda_{\min} \le O(1/(rL)), \qquad \lambda_{\max} = \Theta(1), \qquad \kappa \ge \Omega(rL).$$

A lower bound on $\lambda_{\min}$ and an upper bound on κ would require either an RF-LR analogue of Terjék's NTK–inverse-cosine approximation (e.g. by analyzing the recursion with $\dot{s}(\rho_k) \approx 1/2 - 1/(\pi w_k)$, $s(\rho_k) \approx 1/2 - 1/w_k^2$, and matching asymptotic style) or a perturbation argument for datasets with mild variation in ρ_{ij}.

D.7 Proof of Theorem 4.1

Proof sketch. Step 1 (correlation alignment): Proposition D.1 gives $w_k = \omega^{\circ(k-1)}(w_1) \sim c_- k$, hence $1 - \rho_k = 2w_k^{-2} = \Theta(k^{-2})$. Step 2 (kernel saturation): Rewrite the recursion as $\Theta_k = b_k + a_k\Theta_{k-1}$; the limiting coefficients $a = 1/(2r)$, $b = 1 + 1/(2r)$ yield fixed point $\Theta_\star(r) = (2r+1)/(2r-1)$. A discrete Grönwall argument shows $|\Theta_k - \Theta_\star| = O(1/k)$. Step 3 (gap decay): Subtract the diagonal and off-diagonal recursions; the gap $\Delta_k = \Theta_k^{\text{diag}} - \Theta_k^{\text{off}}$ satisfies $\Delta_k = a\Delta_{k-1} + \Theta(1/(rk))$, hence $\Delta_k = \Theta(1/(rk))$.

Step 1: correlation alignment. By Proposition D.1, there exists a constant $C_w < \infty$ such that for all $k \ge 1$,

$$w_k = \omega^{\circ(k-1)}(w_1) = w_1 + c_0(k-1) + c_1\log\big(c_2 w_1 + k - 1\big) + \varepsilon_k, \qquad |\varepsilon_k| \le C_w,$$

where $c_0 = \Delta_\phi\,\frac{4}{3\pi} > 0$, $c_1 = \Delta_\phi\,\frac{2}{\pi}$, $c_2 = \Delta_\phi^{-1}\frac{3\pi}{4}$. In particular, $w_k \ge w_1 + c_0(k-1) - C_w$. Choosing k_0 large enough gives $w_k \ge \frac{c_0}{2}k$ for all $k \ge k_0$. Similarly, $w_k \le c_+ k$ for some $c_+ < \infty$ and all $k \ge k_0$. Thus $c_- k \le w_k \le c_+ k$ for $k \ge k_0$, and $1 - \rho_k = 2w_k^{-2} = O(k^{-2})$. Write $\Theta_k = \Theta^{(k)}(\rho_1)$ and rewrite (84) as

$$\Theta_k = b_k + a_k\Theta_{k-1}, \qquad a_k = \frac{1}{r}\dot{s}(\rho_k), \qquad b_k = 1 + \frac{1}{r}s(\rho_k).$$

For the limiting constant-coefficient recursion (corresponding to $\rho_k \to 1$), we have $a = \frac{1}{r}\dot{s}(1) = \frac{1}{2r}$, $b = 1 + \frac{1}{r}s(1) = 1 + \frac{1}{2r}$, hence $\Theta_\star(r) = b/(1-a) = (2r+1)/(2r-1)$. Set $e_k = \Theta_k - \Theta_\star(r)$. Then

$$e_k = a_k e_{k-1} + (b_k - b) + (a_k - a)\Theta_\star(r).$$

Using (83) and $w_k \ge c_- k$ for $k \ge k_0$, there exists $C < \infty$ such that for all $k \ge k_0$,

$$|a_k - a| = \frac{1}{r}\Big|\dot{s}(\rho_k) - \frac{1}{2}\Big| \le \frac{C}{r}\frac{1}{w_k} \le \frac{C}{r}\frac{1}{k},$$

$$|b_k - b| = \frac{1}{r}\Big|s(\rho_k) - \frac{1}{2}\Big| \le \frac{C}{r}\frac{1}{w_k^2} \le \frac{C}{r}\frac{1}{k^2}.$$

Moreover $0 \le a_k \le a = 1/(2r) < 1$. Hence for all $k \ge k_0$,

$$|e_k| \le a\,|e_{k-1}| + \frac{C'}{k}$$

for some $C' < \infty$. A standard induction (discrete Grönwall) yields $|e_k| \le C_\Theta/k$ for all $k \ge 1$; in particular $e_k \to 0$ and $\Theta_k \to \Theta_\star(r)$. To make the $O(1/k)$ rate explicit, fix $k \ge k_0$ and iterate the one-step inequality:

$$|e_k| \le a^{k-k_0}|e_{k_0}| + \sum_{j=k_0+1}^{k} a^{k-j}\frac{C'}{j}.$$

Since $j \ge k_0 + 1$ implies $1/j \le 1/k_0$ and also $j \le k$ implies $1/j \le 1/k$, we can bound

$$\sum_{j=k_0+1}^{k} a^{k-j}\frac{1}{j} \le \frac{1}{k}\sum_{t=0}^{k-k_0-1} a^t \le \frac{1}{k}\cdot\frac{1}{1-a}.$$

Also $a^{k-k_0}|e_{k_0}| \le |e_{k_0}| \le (k_0|e_{k_0}|)/k$. Therefore for all $k \ge k_0$,

$$|e_k| \le \frac{1}{k}\Big(k_0|e_{k_0}| + \frac{C'}{1-a}\Big).$$

Absorbing the finitely many values $k < k_0$ into the constant gives $|e_k| \le C_\Theta/k$ for all $k \ge 1$, as claimed. Let $\Theta_k^{\mathrm{diag}} = \Theta^{(k)}(1)$ and $\Theta_k^{\mathrm{off}} = \Theta^{(k)}(\rho_1)$, and set $\Delta_k = \Theta_k^{\mathrm{diag}} - \Theta_k^{\mathrm{off}} \ge 0$. The diagonal recursion is $\Theta_k^{\mathrm{diag}} = b + a\,\Theta_{k-1}^{\mathrm{diag}}$, while the off-diagonal recursion is $\Theta_k^{\mathrm{off}} = b_k + a_k\Theta_{k-1}^{\mathrm{off}}$. Subtracting gives the exact identity

$$\Delta_k := a\,\Delta_{k-1} + (a - a_k)\Theta_{k-1}^{\mathrm{off}} + (b - b_k).$$

For k large, $\Theta_{k-1}^{\mathrm{off}}$ is bounded above and below by positive constants (since $\Theta_k^{\mathrm{off}} \to \Theta_\star(r) \in (1, \infty)$). Also by (83) and $w_k \asymp k$,

$$a - a_k = \frac{1}{r}\Big(\frac{1}{2} - \dot{s}(\rho_k)\Big) = \Theta\Big(\frac{1}{r\,w_k}\Big) = \Theta\Big(\frac{1}{r\,k}\Big), \qquad b - b_k = \frac{1}{r}\Big(\frac{1}{2} - s(\rho_k)\Big) = O\Big(\frac{1}{r\,w_k^2}\Big) = O\Big(\frac{1}{r\,k^2}\Big).$$

Thus there exist k_1 and constants $0 < c < C < \infty$ such that for all $k \ge k_1$,

$$a\,\Delta_{k-1} + \frac{c}{r\,k} \le \Delta_k \le a\,\Delta_{k-1} + \frac{C}{r\,k}.$$

Iterating these comparison recursions yields $c_\Delta/(rk) \le \Delta_k \le C_\Delta/(rk)$ for all $k \ge k_1$, completing the proof. □

Proof sketch.

D.8 Remark: proxy lower bounds vs positivity (smallest eigenvalue)

Notation. Let $\widehat{K}$ denote the empirical NTK Gram matrix at initialization, with entries $\widehat{K}_{ij} = \langle \nabla_\theta f(x_i), \nabla_\theta f(x_j)\rangle$. Let K denote the mean (deterministic) kernel Gram matrix with entries $K_{ij} = \mathbb{E}[\widehat{K}_{ij}]$. The *centered* spectrum refers to the restriction of these matrices to the subspace $\mathbf{1}^\perp$ of vectors orthogonal to the constant vector $\mathbf{1}$; we write $\widehat{K}|_{\mathbf{1}^\perp}$ and $K|_{\mathbf{1}^\perp}$ for these restrictions.

Remark D.1 (Can the smallest eigenvalue become negative?). No: $\widehat{K}$ is a Gram matrix of gradients, hence $\widehat{K} \succeq 0$ deterministically and all its eigenvalues are ≥ 0. Restricting to the mean-zero subspace preserves positive semidefiniteness: for any $v \perp \mathbf{1}$, one has $v^\top \widehat{K} v \ge 0$, so $\widehat{K}|_{\mathbf{1}^\perp} \succeq 0$.

What *can* happen is that a deterministic proxy lower bound on the centered eigenvalue scale is smaller than the magnitude of random fluctuations. Weyl's inequality gives

$$\lambda_{\min}\big(\widehat{K}|_{\mathbf{1}^\perp}\big) \;\ge\; \lambda_{\min}\big(K|_{\mathbf{1}^\perp}\big) \;-\; \|\widehat{K} - K\|_2,$$

so if $\|\widehat{K} - K\|_2$ dominates the deterministic scale of $\lambda_{\min}(K|_{\mathbf{1}^\perp})$, the bound becomes vacuous (it may be negative), but the true eigenvalue still satisfies $\lambda_{\min}(\widehat{K}|_{\mathbf{1}^\perp}) \ge 0$. Thus, the relevant concern is not "going below zero" but rather "being driven close to zero" when the $1/(rL)$ signal is swamped by noise.

D.9 Remark: infinite-depth limit after centering

Remark D.2 (Infinite-depth limit: existence and degeneracy after centering). The correlation recursion satisfies $1 - \rho_k = O(k^{-2})$ as in full-width MLPs at the EOC [**?**, **?**], so deep features become increasingly aligned. For the *mean* RF-LR NTK recursion (84), Theorem 4.1 implies that for any fixed $\rho_1 \in (-1, 1)$,

$$\Theta^{(L)}(\rho_1) \to \Theta_\star(r) \qquad \text{and} \qquad \Theta^{(L)}(1) \to \Theta_\star(r),$$

so the entrywise infinite-depth limit of the (mean) Gram matrix is a constant kernel: $K^{(L)} \to \Theta_\star(r)\,\mathbf{1}\mathbf{1}^\top$. Consequently, after centering (restricting to the orthogonal complement of the constant mode), the relevant part of $K^{(L)}$ tends to zero and the spectrum collapses. In particular, a *non-degenerate* deep limit on $\mathbf{1}^\perp$ can only hold under a joint scaling (e.g. r growing with L) or after an explicit rescaling of the centered operator (equivalently, rescaling time/learning rate in kernel gradient descent).

E Experiments

This section presents numerical illustrations. All scripts are headless (matplotlib "Agg" backend). Setup: finite-width NTK scripts use a **3-layer (2 ReLU)** RF-LR architecture with hidden width 12000–20000 so that finite-width variance is small; deterministic proxy scripts iterate the recursion to large k (e.g. $k_{\max} = 4000$ or $L = 200$). Bottleneck rank r is swept as indicated; n is the number of inputs, d the input dimension. Scripts in the first subsection are **deterministic**; the rest are **finite-width Monte Carlo** (they illustrate fluctuation scales and do not constitute proofs).

E.1 Correlation alignment and equicorrelated spectrum (Theorem 4.1, Corollary 4.1)

What is computed. Deterministic ReLU EOC correlation recursion $\rho_k = \varrho(\rho_{k-1})$ and proxy NTK recursion; diagnostics w_k/k, $k|\Theta^{(k)} - \Theta_\star(r)|$, $rk(\Theta^{(k)}_{\text{diag}} - \Theta^{(k)}_{\text{off}})$. For the equicorrelated model, eigenvalues $\lambda_1 = \Theta^{(L)}(1) + (n-1)\Theta^{(L)}(\rho_0)$, $\lambda_\perp = \Theta^{(L)}(1) - \Theta^{(L)}(\rho_0)$ (e.g. $n = 64$, $\rho_0 = 0$).

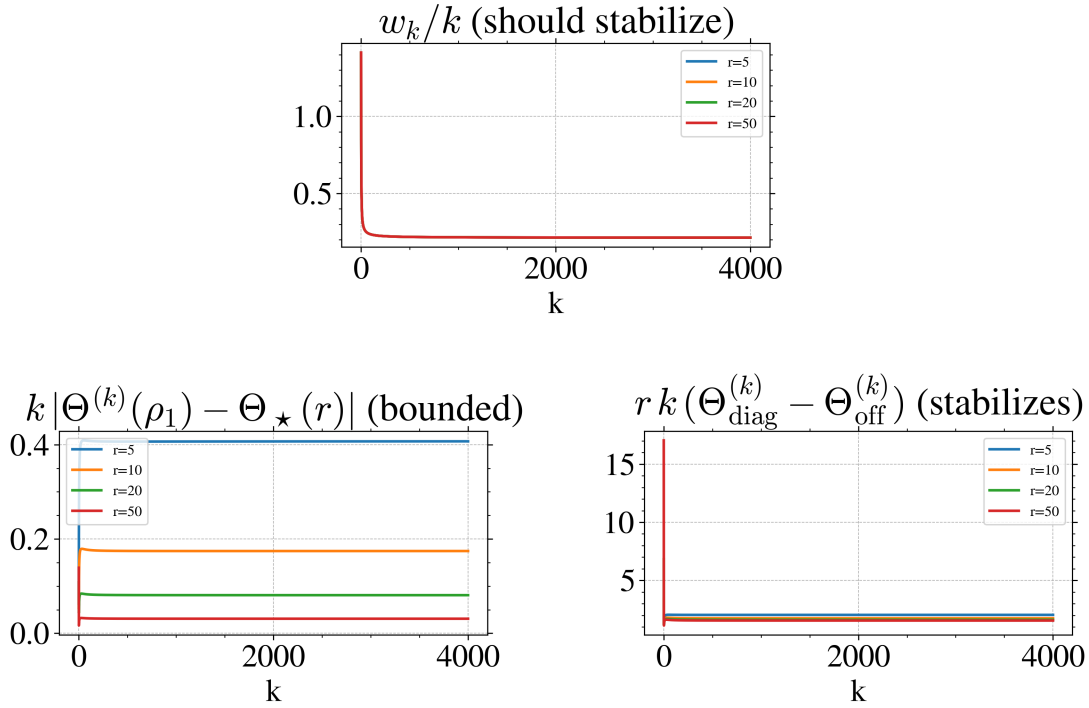

Figure 2: Three panels (Theorem 4.1): (1) w_k/k vs k (stabilizes); (2) $k|\Theta^{(k)}(\rho_1) - \Theta_\star(r)|$ vs k (bounded); (3) $rk(\Theta^{(k)}_{\text{diag}} - \Theta^{(k)}_{\text{off}})$ vs k (stabilizes). One curve per rank $r \in \{5, 10, 20, 50\}$.

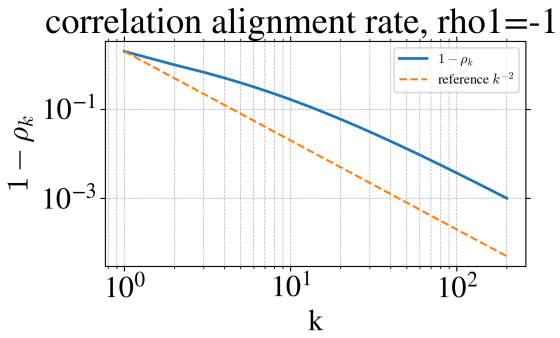

Figure 3: Correlation alignment along depth: $1 - \rho_k = O(k^{-2})$ (log–log) as in Theorem 4.1.

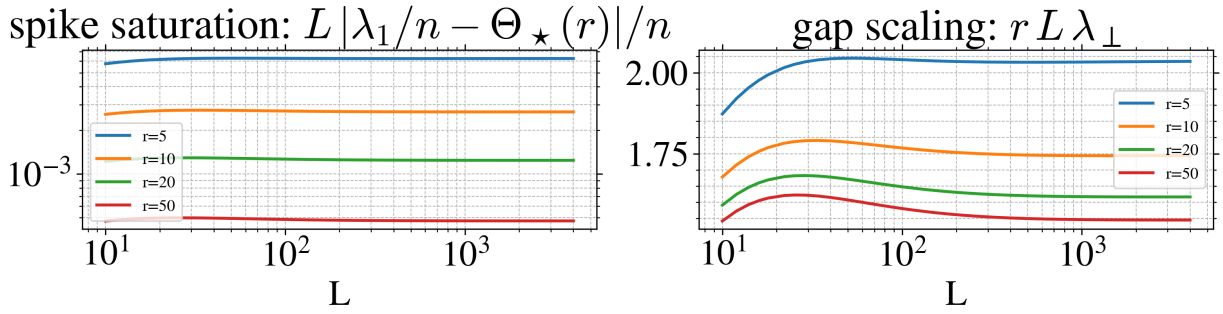

Figure 4: Equicorrelated spectrum (Corollary 4.1): (1) spike saturation $L|\lambda_1/n - \Theta_\star(r)|/n$ vs L (log–log); (2) gap scaling $rL\lambda_\perp$ vs L (log–log). One curve per r.

E.2 Product decay and entry variance

What is computed. Left: Deterministic product $\prod_{k=\ell+1}^{L} \dot{\Sigma}^{(k)}/r$ vs $j = L - \ell$ (exponential-in-depth suppression, Eq. 7). **Right:** Monte Carlo Var$[K(-1,1)]$ vs r over 100 initializations (3-layer RF-LR, width 16000); rank-driven concentration (Section 5).

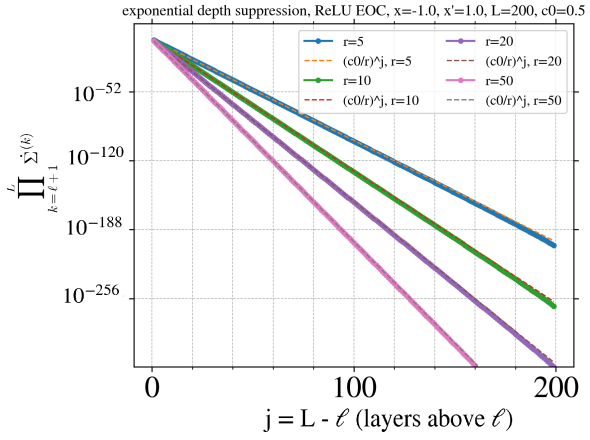

Figure 5: *
Product decay vs depth; bound $(c_0/r)^j$.

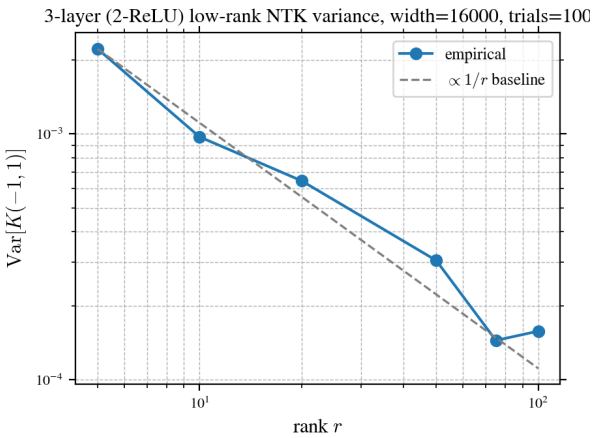

Figure 6: *
Entry variance vs r; $\propto 1/r$ baseline.

Figure 7: Left: depth suppression (deterministic). Right: entry-wise variance (finite-width Monte Carlo).

E.3 Smallest eigenvalue (finite-width Monte Carlo)

What is computed. Fixed random dataset on the unit sphere ($n = 64$, $d = 16$); empirical Gram K, centered $K_c = HKH$; $\lambda^+_{\min}(K_c)$ over 50 initializations, $r \in \{5, 10, 20, 50, 100, 200, 500, 1000\}$. Illustrates smallest eigenvalue driven close to 0 by finite-width noise (Appendix D.1).

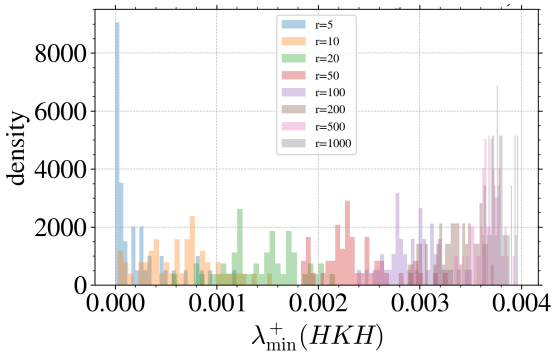

Figure 8: *
Histograms of $\lambda^+_{\min}(HKH)$; matrix stays PSD.

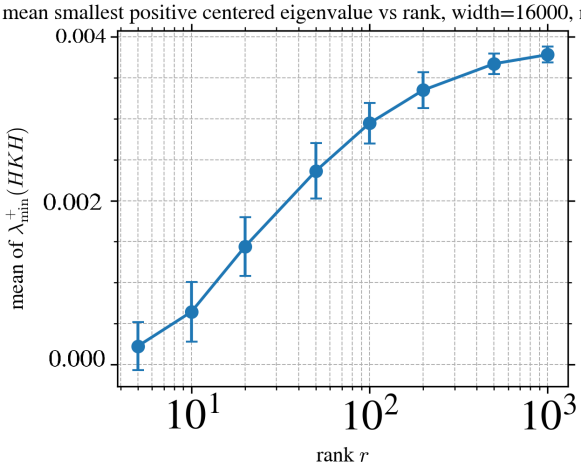

Figure 9: *
Mean $\lambda^+_{\min}(HKH)$ vs r; better conditioning with r.

Figure 10: Distribution and mean of smallest positive centered eigenvalue vs r.

E.4 Condition number κ (proxy vs empirical; equicorrelated, high-dim, non-equicorrelated)

What is computed. Condition number on $\mathbf{1}^\perp$ vs r. **Equicorrelated:** $n = 32$, $d = 64$, $\rho_0 = 0$; proxy $\kappa_\perp = 1$ (Corollary 4.1); empirical mean ± std over 40 initializations (Theorem 4.2). **High-dim spherical:** i.i.d. uniform on S^{d-1}, $d = 256, 128$; $\kappa_\perp \approx 1$. **Non-equicorrelated:** Clustered design (4 clusters), $n = 48$, $d = 64$, 30 trials; $\kappa \geq \Omega(r \cdot L)$ (Proposition 4.1).

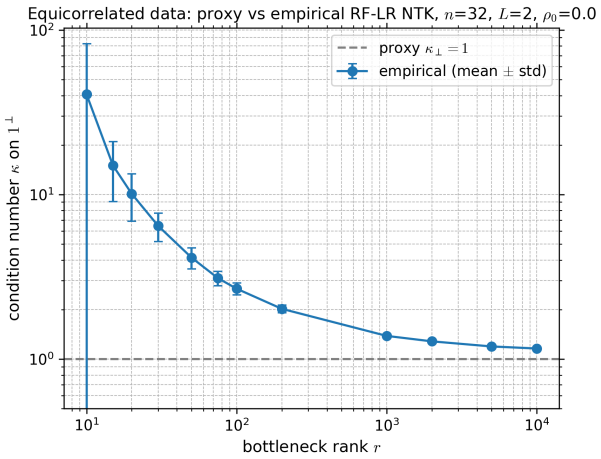

Figure 11: *
Equicorrelated: proxy 1; empirical concentrates.

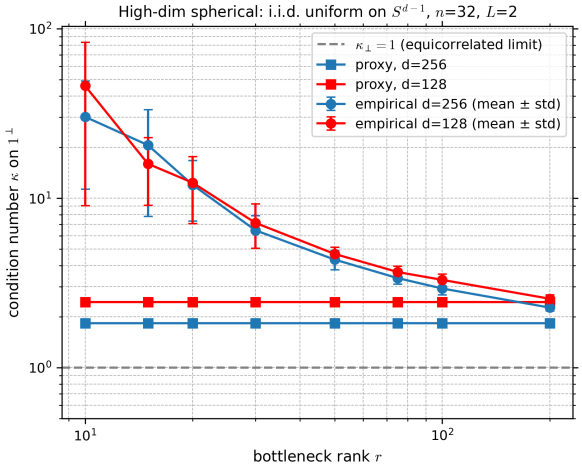

Figure 12: *
High-dim spherical: $\kappa_\perp \approx 1$.

Figure 13: Condition number on $\mathbf{1}^\perp$: equicorrelated and high-dimensional spherical data.

E.5 Non-equicorrelated κ and kernel regression risk

Left: Condition number vs r for clustered (non-equicorrelated) data; κ need not approach 1 (Proposition 4.1). **Right:** Kernel ridge regression with the empirical RF-LR NTK: fix a target on the sphere, 64 train / 256 test, $d = 32$, 20 trials; as r grows the Gram concentrates toward the proxy, so test MSE and its variance across trials decrease, linking conditioning and concentration to downstream risk.

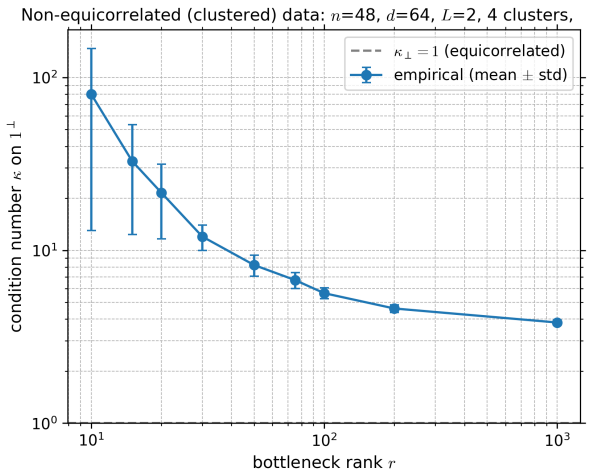

Figure 14: *
κ vs r, non-equicorrelated; $\kappa \geq \Omega(r \cdot L)$.

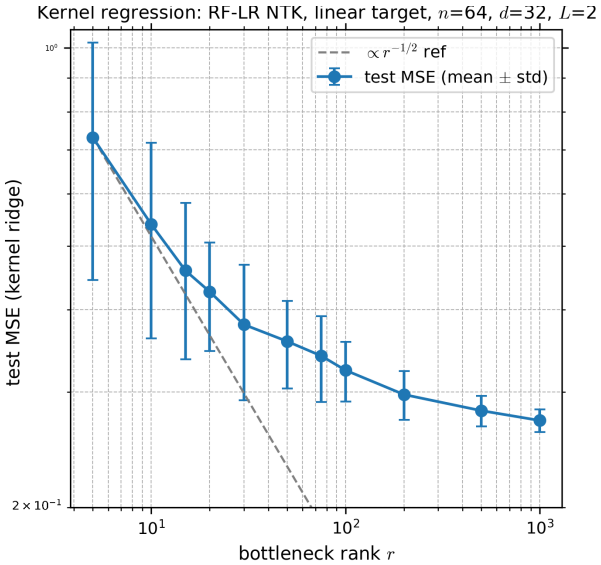

Figure 15: *
Test MSE vs r; mean risk and std drop with r.

Figure 16: Left: κ vs r for clustered (non-equicorrelated) data; κ need not approach 1 (Proposition 4.1). Right: Test MSE of kernel ridge regression vs r; as r grows and the Gram concentrates toward the proxy, mean risk decreases and std across trials drops.

E.6 RKHS Puiseux exponent vs depth (Corollary 5.2, extension to $L \geq 4$)

For zonal kernels on the sphere, the RKHS is controlled by the Puiseux exponent γ at the endpoint $\rho = 1$: if $K(1-t) - K(1) \sim c \cdot t^\gamma$ for small t, then γ determines the RKHS [17]. The paper proves $\gamma = 1/2$ for the mean three-layer RF-LR kernel (same as shallow ReLU); extension to $L \geq 4$ is open. We estimate $\gamma(L)$ for the deterministic proxy $\Theta^{(L)}(\rho)$ via log–log regression of the gap $\Theta^{(L)}(1) - \Theta^{(L)}(1-t)$ vs t. For $L \in \{2, 3, 4, 5, 6\}$ we obtain $\gamma(L) \approx 0.52$–0.55 with $R^2 > 0.999$, suggesting the proxy's endpoint behavior is consistent with RKHS equivalence extending to $L \geq 4$.

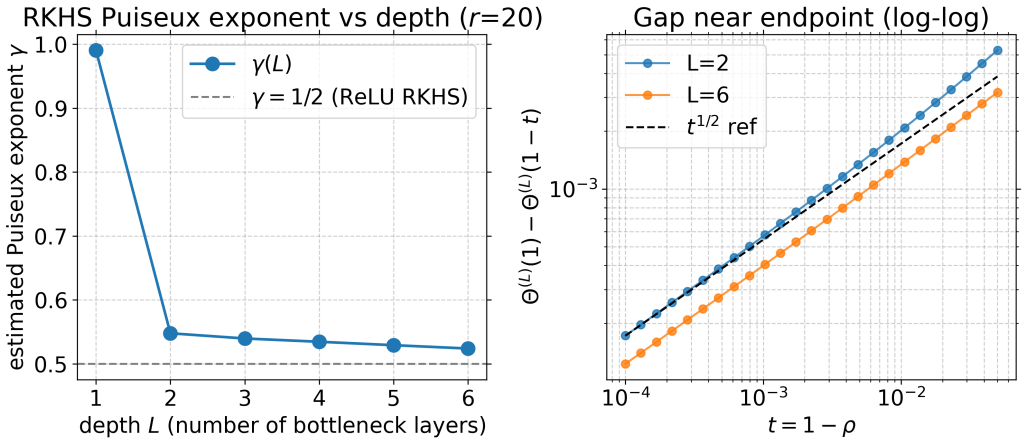

Figure 17: RKHS Puiseux exponent vs depth. Left: estimated $\gamma(L)$ vs L; reference $\gamma = 1/2$ (ReLU RKHS). Right: log–log gap vs t for $L = 2$ and $L = 6$. The proxy kernel exhibits $\gamma \approx 1/2$ for $L \geq 2$, consistent with RKHS equivalence extending beyond three layers.

	MLP at EOC [?]	**RF-LR (this work)**
Trainable	all weight layers	readouts $A^{(\ell)}$, biases $c^{(\ell)}$; frozen $w^{(\ell)}$
Scaling	depth L, Δ_ϕ	depth L, bottleneck r
Kernel magnitude	uncentered scale can grow with L	mean recursion saturates in L (Thm. 4.1)
Centered scale	correlation propagation	gap $\Theta^{(L)}(1) - \Theta^{(L)}(\rho) \asymp 1/(rL)$ (Thm. 4.1)

Table 1: Depth/rank effects in the kernel regime. MLP: fully trained behavior; RF-LR: explicit $1/r$ bottleneck, saturation and early-layer suppression.