Low-Rank Neural Networks and Finite-Width NTK at the Edge of Convexity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Low-rank neural networks are usually sold as smaller, compressed and pruned models. This paper asks a sharper question: when does low rank still preserve the optimization geometry that makes wide networks trainable? We answer this through the finite-width neural tangent kernel. Low-rank NTK training still carries an interpretable bottleneck-feature geometry, because rank controls which feature maps enter the tangent kernel. Our main result is that low-rank training is governed by a genuine three-way compromise between width, depth, and rank. To the best of our knowledge, this is the first NTK analysis showing that an optimization certificate can impose a depth-rank tradeoff of this form. In particular, for low-rank bottleneck dynamics, preserving the NTK spectral margin has a proven conservative cubic-depth sufficient rule, $r \gtrsim L^3$, up to dataset-size and logarithmic factors. From the finite-network experiments, we conjecture an effective $L^{3/2}$ law for scalar-output contracted cumulants. We also give a full-rank-compatible parametrization, separating true low-rank effects from ordinary finite-width effects. True finite-network NTK experiments confirm the exact full-rank matching, the predicted operator scaling, the rank-depth tradeoff, and the finite-network cumulant growth. The result is a criterion for when low-rank networks remain trainable for structural reasons, not merely parameter-efficient ones.

1 Introduction

Low-rank neural networks are usually motivated by efficiency: fewer parameters, lower memory traffic, and cheaper adaptation. This motivation is incomplete. Rank also changes the geometry of training. Low-rank bottlenecks do not merely shrink a dense model: they restrict the network to a small set of bottleneck feature maps, so feature selection already enters through the architecture. Thus saying that an NTK analysis has "no features" is misleading for low-rank networks; the question is whether the fixed or linearized low-rank feature geometry remains stable enough to explain training. In the NTK regime, gradient descent is close to kernel regression when the empirical NTK Gram matrix is positive and stable. Thus a low-rank network is useful not only when it is small, but when its finite NTK stays close enough to a positive limiting kernel. We call this the edge of convexity: the regime where the network is finite and structured, but the NTK Gram matrix still provides a convex local training geometry.

Empirically, low-rank bottlenecks can give a much better parameter-accuracy tradeoff than dense baselines. This raises the next theoretical question at very small rank: if only r bottleneck feature maps are available, which features are recovered, and is the recovery explained by a kernel/NTK mechanism or by genuine feature-learning dynamics such as mean-field or μP scaling? The present paper answers the controlled NTK side of that question. It shows that even in the nominally lazy

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Table 1: MNIST parameter-efficiency snapshot. Low-rank bottlenecks recover most of the dense baseline accuracy with far fewer trainable parameters.

Model	Trainable parameters	Test accuracy (%)
Dense MLP	669,706	98.39
Low-rank, $r = 5$	7,695	93.72
Low-rank, $r = 10$	10,260	96.24
Low-rank, $r = 15$	12,825	96.97
Low-rank, $r = 25$	17,955	97.00
Low-rank, $r = 50$	30,780	97.01
Low-rank trainable factors, $r = 32$	440,362	98.30

regime, low-rank bottlenecks introduce nontrivial rank-channel geometry, memory suppression, and spectral constraints that any feature-learning theory must build on.

This question is difficult because three limits interact. Width controls ordinary finite-width fluctuations, depth amplifies NTK tensor corrections, and rank changes the channel contractions themselves. Treating low rank as merely replacing $1/n$ by $1/r$ misses the full-rank endpoint: if a rank-n factorization is just an orthogonal change of variables, the low-rank correction must vanish exactly. Conversely, bottleneck random-feature low-rank networks do not reduce to dense MLPs and can suppress old NTK paths geometrically. A useful theory must separate these two cases.

We study random-feature low-rank (RF-LR) scalar-output ReLU networks with layer weights of the form

$$W^{(\ell)} = \frac{\sigma_w}{\sqrt{n_{\ell-1}}}\sqrt{\frac{n_\ell}{r_\ell}}\,U^{(\ell)}B^{(\ell)}, \qquad (U^{(\ell)})^\top U^{(\ell)} = I_{r_\ell}, \tag{1}$$

where $U^{(\ell)}$ is frozen and sampled uniformly from the Stiefel manifold $\mathrm{St}(n_\ell, r_\ell) = \{U \in \mathbb{R}^{n_\ell \times r_\ell} : U^\top U = I_{r_\ell}\}$, while $B^{(\ell)}$ is trainable. Thus the Stiefel object is the left factor $U^{(\ell)}$ itself: its columns form an orthonormal r_ℓ-frame in $\mathbb{R}^{n_\ell}$, chosen uniformly among all such frames. Equivalently, $U^{(\ell)}$ selects a random r_ℓ-dimensional subspace with no preferred direction. When $r_\ell = n_\ell$, the map $B^{(\ell)} \mapsto U^{(\ell)}B^{(\ell)}$ is an orthogonal reparametrization of a dense MLP. Thus the full-rank endpoint is exact, not asymptotic; Section 4 identifies the corresponding rank-defect scale.

2 Related Work

The NTK formalizes lazy infinite-width training as kernel regression [1, 2]. Lee et al. clarified that wide networks of any depth evolve approximately as linear models around initialization, making the empirical NTK spectrum the relevant stability object [2]. Many global-convergence results rely on keeping this empirical NTK close to a positive limiting kernel [3, 4, 5]. Spectral refinements show that strong guarantees can hold even when only part of a deep network is very wide [6, 7]; recent EOC analyses further study concentration and spectra of MLP NTKs at the edge of chaos [8, 9]. These works provide the dense full-rank baseline; they do not identify the finite-rank correction created by factorized layers or the cancellation that occurs at the full-rank endpoint.

A complementary line studies how finite networks depart from their infinite-width limits. Yang's Tensor Programs give a general language for infinite-width limits, parametrization choices, and feature-learning scalings across architectures [10, 11, 12, 13], while the neural tangent hierarchy describes finite-width departures from frozen-kernel dynamics [18]. Feynman-diagram methods organize these corrections by connected cumulants and tensor contractions [19, 20]. Our analysis follows this perturbative viewpoint, but changes the stochastic source: dense channel contractions are replaced by Stiefel-projector and rank-channel cumulants.

NTK theory is also not meant to replace feature-learning theories. Mean-field limits describe distributional feature evolution beyond the fixed-kernel regime [15, 16, 17], and μP, depth-μP, tensor-program, and DMFT analyses identify scalings where features continue to move at infinite width [12, 13, 14]. Our results are complementary: they quantify the finite-rank and finite-width stability of the lazy/proxy geometry before one studies full feature-learning dynamics.

Low-rank adaptation and low-rank training have also been analyzed in NTK-like settings, including guarantees that sufficiently large low-rank adapters avoid bad local minima in certain regimes [22]. Empirical NTK tools make direct finite-network checks feasible [21], and recent quantitative GP approximation during training gives another complementary control problem [23]. In contrast, our object is the operator deviation of a deep low-rank empirical NTK and the resulting Weyl spectral-preservation criterion.

Contributions. First, full-rank factorization is only a coordinate change: when $r = n$, the empirical NTK matches the dense MLP pathwise. Away from full rank, the correction is a projector defect, not the naive replacement $n \mapsto r$.

The second message is that depth turns small rank defects into a stability constraint. Low-rank randomness enters locally at each layer, but NTK legs are transported through the whole network. One transported leg creates a larger accumulated correction than a pure NNGP fluctuation, and two transported NTK legs create the worst depth growth. This is the mechanism behind the conservative cubic rank-depth rule for bottleneck dynamics: deep low-rank networks are hard to keep in the same convex NTK regime unless rank grows with depth.

The third message is practical. The theory gives an operator-level stability certificate under an explicit well-spread data assumption, and it remains closed under the NTK descendants that control short-time training drift. The experiments then check the pieces separately: exact full-rank matching, disappearance of the rank defect at $r = n$, the predicted depth normalization, and direct finite-network cumulants. The observed slopes are effective exponents of mixed observable quantities, so they also point to a sharper future theory beyond the worst-case cubic envelope.

3 Models and Full-Rank Matching

This section defines the factorized network, proves exact matching at full rank, and explains why the left factor is frozen.

For a fixed dataset $x_1, \dots, x_m$, the scalar-output empirical NTK is

$$\widehat{\Theta}_{ab} = \nabla_\theta f(x_a) \cdot \nabla_\theta f(x_b). \tag{2}$$

In (1), the trainable parameters are the right factors $B^{(\ell)}$ and the readout. We keep $U^{(\ell)}$ frozen so the tangent space is a fixed subspace and all rank-dependent randomness enters through $G = (n/r)UU^\top$. Training U would move the subspace and add U-B dNTK interaction terms.

Theorem 3.1 (Pathwise full-rank matching)**.** *If $r_\ell = n_\ell$ for every hidden layer, then for every realization, dataset, and depth, the empirical NTK in B-coordinates equals the dense MLP NTK in coordinates $\widetilde{W}^{(\ell)} = U^{(\ell)}B^{(\ell)}$. The same equality holds for the NNGP, dNTK, ddNTK, and their tensor cumulants.*

Proof. When $r = n$, U is orthogonal, so $\Delta B \mapsto U\Delta B$ preserves Frobenius inner products. Hence $J_B J_B^\top = J_W J_W^\top$, and higher derivatives and cumulants are invariant under the same coordinate change. □

Proposition 3.2 (Why the left factor is frozen)**.** *For one layer $W = \alpha UB$, with $U^\top U = I_r$ and $H = \partial\mathcal{L}/\partial W$, training only B gives*

$$\nabla_B \mathcal{L} = \alpha U^\top H, \qquad \dot{W}_B = -\eta\alpha^2 UU^\top H. \tag{3}$$

If U is also trained on the Stiefel manifold, the normal projected component contributes

$$\dot{W}_U = -\eta\alpha^2 (I - UU^\top) H B^\top B \quad \textit{up to rotations inside } \operatorname{span}(U) \textit{ that can be absorbed into } B. \tag{4}$$

Thus moving U adds U-B interaction blocks proportional to $B^\top B$, absent from the frozen-left model and not controlled solely by cumulants of $G = (n/r)UU^\top$.

Proof idea. Project the Euclidean gradient onto the B-directions and the Stiefel tangent space, then map both variations back to $W = \alpha UB$; this gives (3) and (4). The full differential calculation is in Appendix D. □

Freezing U therefore isolates the rank-width effect and keeps the diagrammatics closed under the Stiefel projector vertices of Lemma 4.1.

The RF-LR bottleneck has a different recursion,

$$\Theta^{(\ell)} = 1 + \frac{1}{r}\Theta^{(\ell-1)}\dot{\Sigma}^{(\ell)} + \frac{1}{r}\Sigma^{(\ell)}. \tag{5}$$

Here we suppress the sample pair (x, x'). If $(g_x^{(\ell)}, g_{x'}^{(\ell)})$ is the Gaussian preactivation pair induced by the previous-layer covariance, then

$$\Sigma^{(\ell)}(x, x') := \mathbb{E}\left[\sigma(g_x^{(\ell)})\sigma(g_{x'}^{(\ell)})\right], \qquad \dot{\Sigma}^{(\ell)}(x, x') := \mathbb{E}\left[\sigma'(g_x^{(\ell)})\sigma'(g_{x'}^{(\ell)})\right]. \tag{6}$$

Thus $\Sigma^{(\ell)}$ is the NNGP covariance term injected at layer ℓ, while $\dot{\Sigma}^{(\ell)}$ is the derivative covariance multiplying an old NTK line. At the ReLU fixed point $\dot{\Sigma}^{(\ell)} \to 1/2$, so an old NTK contribution is multiplied by approximately $(2r)^{-(L-k)}$. This is a different model, not a perturbation of the dense MLP.

4 Stiefel Projector Vertices

This section explains the random projector $G = (n/r)UU^\top$, which is the only way the frozen random left factor enters the NTK corrections. Let $U \in \mathbb{R}^{n\times r}$ be sampled uniformly from the Stiefel manifold and

$$G = \frac{n}{r}UU^\top. \tag{7}$$

We study this Stiefel projector because the trainable gradients with respect to B are always projected through the random subspace selected by U; therefore cumulants of G are the low-rank replacement for dense neuron-index contractions. Orthogonal invariance gives the full covariance tensor.

Lemma 4.1 (Stiefel projector covariance)**.** *For all indices,*

$$\begin{aligned} \mathrm{Cov}(G_{ij}, G_{kl}) &= \gamma_{n,r}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} - \frac{2}{n}\delta_{ij}\delta_{kl}\right), \\ \gamma_{n,r} &= \frac{n(n-r)}{r(n-1)(n+2)} = \frac{1}{r} - \frac{1}{n} + O(n^{-2}). \end{aligned} \tag{8}$$

In particular $\gamma_{n,n} = 0$. The mixed perturbation parameter is therefore $1/n + \gamma_{n,r}$: the $1/n$ term is the ordinary dense finite-width correction, while $\gamma_{n,r}$ is the rank defect. The rank-defect part vanishes at $r = n$, exactly as required by pathwise full-rank matching.

Proof. The covariance of a Stiefel projector has the invariant form $a(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + b\delta_{ij}\delta_{kl}$. The identity $\mathrm{tr}\, G = n$ gives $2a + nb = 0$. The standard second moment of a Stiefel projector gives $a = \gamma_{n,r}$, hence (8). □

Higher cumulants are given by orthogonal Weingarten contractions. Connected s-point cumulants scale as $O_s(r^{1-s})$ and vanish in the full-rank endpoint. Thus the low-rank Feynman rule is simple: dense Gaussian-ReLU vertices remain, and each low-rank correction inserts connected cumulants of G.

The rank-channel power counting is the usual connected-cumulant counting for iid empirical averages: a connected s-point rank vertex scales as r^{1-s}. Lemma E.1 in Appendix E gives the precise statement and defines the channel atom $Z_a^{(\ell)}$. This is why finite-rank bias terms are typically $O(r^{-1})$, while the full-rank endpoint must be treated through the projector defect $G - I$, not by the informal replacement $n \mapsto r$.

4.1 Triangular tensor recursion

The full pairing-basis dictionary for V, D, F, A, B is in Appendix G.7. In the main text we use the compressed block notation $\mathcal{V} = (V_{1234}, V_{1324}, V_{1423})$, $\mathcal{C} = (D_{1234}, F_{1324}, F_{1423})$, and $\mathcal{A} = (A_{1234}, B_{1324}, B_{1423})$.

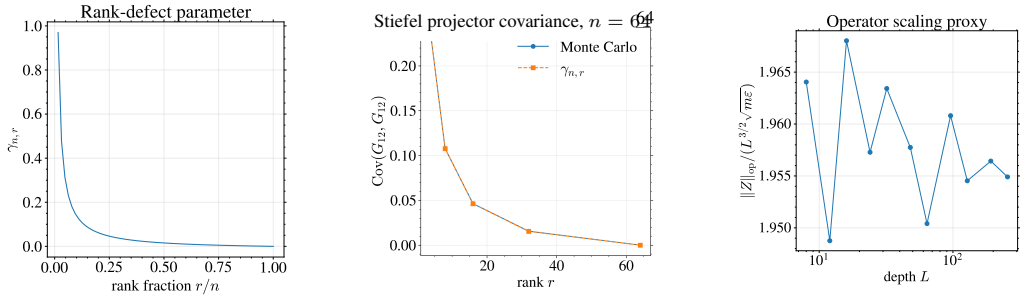


Figure 1: Rank-projector and operator checks. Left: the rank-defect scale $\gamma_{n,r}$ vanishes at full rank. Middle: empirical Stiefel-projector covariance matches the closed form. Right: the normalized operator proxy $\|Z\|_{\mathrm{op}}/(L^{3/2}\sqrt{m\varepsilon})$ is approximately stable across depth.

The layer recursion is triangular at first perturbative order:

$$\mathcal{V}_{\ell+1} = \mathcal{L}_V^{(\ell)}\mathcal{V}_\ell + \mathcal{S}_V^{(\ell)}, \tag{9}$$

$$\mathcal{C}_{\ell+1} = \mathcal{L}_C^{(\ell)}\mathcal{C}_\ell + \mathcal{B}_C^{(\ell)}\mathcal{V}_\ell + \mathcal{S}_C^{(\ell)}, \tag{10}$$

$$\mathcal{A}_{\ell+1} = \mathcal{L}_A^{(\ell)}\mathcal{A}_\ell + \mathcal{B}_A^{(\ell)}\mathcal{C}_\ell + \mathcal{Q}_A^{(\ell)}[\mathcal{V}_\ell, \mathcal{V}_\ell] + \mathcal{S}_A^{(\ell)}. \tag{11}$$

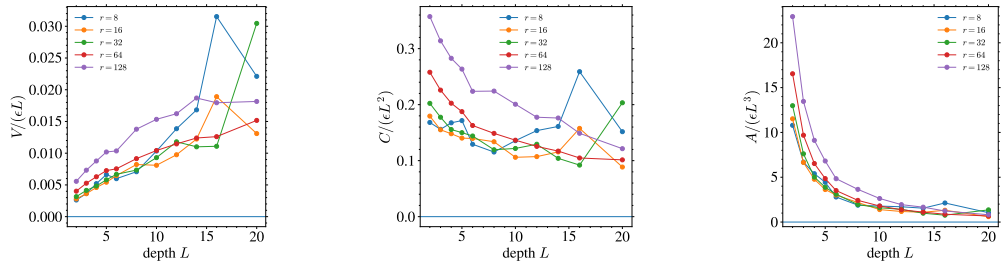


Figure 2: True finite-network cumulants from autodiff NTKs. These scalar-output cumulants are not isolated Feynman coefficients; they test the predicted polynomial control and rank-width scaling on actual finite networks.

The source bounds used in Proposition 5.2 are

$$\|\mathcal{S}_V^{(\ell)}\| \le C, \qquad \|\mathcal{S}_C^{(\ell)}\| \le C(1+\ell+\|\mathcal{V}_\ell\|), \qquad \|\mathcal{S}_A^{(\ell)}\| \le C(1+\ell^2+\ell\|\mathcal{V}_\ell\|+\|\mathcal{C}_\ell\|+\|\mathcal{V}_\ell\|^2). \tag{12}$$

These formulas are the visible mechanism behind the hierarchy: V feeds D, F, and D, F feed A, B.

5 Tensor Scaling at ReLU Edge of Chaos

This section turns the projector cumulants into depth scalings for the tensor blocks V, D, F, A, B. For normalized ReLU, the edge-of-chaos correlation asymptotics used in the RF-LR analysis give $\theta_\ell = 3\pi/\ell + O(\log \ell/\ell^2)$, and

$$\dot{\mathcal{C}}(\rho_\ell) = 1 - \frac{3}{\ell} + O\left(\frac{\log \ell}{\ell^2}\right). \tag{13}$$

Transporting an external NTK leg from layer k to L gives the factor $(k/L)^\lambda$ for an exponent $\lambda \in \{0, 3\}$, depending on whether the Gram entry is diagonal or off-diagonal.

Lemma 5.1 (Depth-transport accumulation). *Let $T_\ell \in \mathbb{R}^d$, $\ell \ge 2$, satisfy $T_2 = 0$ and*

$$T_{\ell+1} = \left(I - \frac{M}{\ell} + E_\ell\right) T_\ell + \ell^\beta (s + e_\ell), \tag{14}$$

where $\|E_\ell\| \le C \log^p(\ell)/\ell^2$, $\|e_\ell\| \le C \log^q(\ell)/\ell$, *every eigenvalue of* M *has real part greater than* $-(\beta+1)$, *and* $M + (\beta+1)I$ *is invertible. Then*

$$T_L = L^{\beta+1} \left(M + (\beta+1)I\right)^{-1} s + O\left(L^\beta \log^{p+q+2} L\right). \tag{15}$$

Proof idea. Iterating (14) gives a variation-of-constants formula; Appendix A gives the full discrete proof with the product estimates. The product of propagators from k to L is $(k/L)^M$ up to a summable perturbative error. Thus, at leading order,

$$T_L = L^{-M} \sum_{k \le L} k^{M+\beta I} s + O(L^\beta \log^{p+q+2} L).$$

Euler-Maclaurin yields

$$L^{-M} \sum_{k \le L} k^{M+\beta I} = L^{\beta+1} \int_0^1 u^{M+\beta I} \, \mathrm{d}u + O(L^\beta),$$

and the matrix integral is $(M + (\beta+1)I)^{-1}$. □

Let $\bar{\varepsilon} = \max_{\ell \le L}(1/n_\ell + \gamma_{n_\ell, r_\ell})$. The finite-width tensors obey

$$V_L = O(\bar{\varepsilon} L), \qquad C_L = O(\bar{\varepsilon} L^2), \qquad A_L = O(\bar{\varepsilon} L^3), \tag{16}$$

where C denotes the mixed NNGP-NTK block (D, F), and A denotes the NTK-NTK block (A, B).

Proposition 5.2 (Conditional tensor hierarchy)**.** *Assume the low-rank cumulant recursion closes on the same tensor sectors as the dense finite-width Feynman hierarchy and that the ReLU edge-of-chaos transport is uniform away from the diagonal singularity. Then the isolated tensor coordinates satisfy*

$$V_L = O(\bar{\varepsilon} L), \qquad D_L, F_L = O(\bar{\varepsilon} L^2), \qquad A_L, B_L = O(\bar{\varepsilon} L^3). \tag{17}$$

Consequently, the perturbative mean-NTK correction is controlled in the regime $L^3 \bar{\varepsilon} \ll 1$.

Proof idea. Appendix A gives the complete block-recursion proof. The fresh NNGP four-point source contributes $O(\bar{\varepsilon})$ per layer and has no transported NTK leg in the dominant radial mode, so summation gives $V_L = O(\bar{\varepsilon} L)$. The mixed blocks D, F have one transported NTK leg and therefore one depth-transport accumulation with $\beta = 1$, yielding $O(\bar{\varepsilon} L^2)$ by Lemma 5.1. The NTK-NTK blocks A, B have two transported NTK legs, equivalently source exponent $\beta = 2$, giving $O(\bar{\varepsilon} L^3)$. These are isolated tensor coordinates; scalar-output cumulants can mix these sectors. □

Endpoint constants. At the ReLU endpoint, set $X_0 = 2(g_+)^2$ and $Y_0 = 2\mathbf{1}_{\{g>0\}}$. Then

$$\mathrm{Var}(X_0) = 5, \qquad \mathrm{Cov}(X_0, Y_0) = 1, \qquad \mathrm{Var}(Y_0) = 1. \tag{18}$$

The table should be read together with the plots. On the standard long run, the visible slopes are indeed close across ranks, with effective exponents around 1.5–2.1 for the same-off-diagonal observables. This supports the interpretation that the proven L^3 tensor bound is a safe upper envelope, not a claim that every scalar-output cumulant must scale exactly as L^3. A natural conjecture is that some contracted observables obey a sharper mode-dependent law between $L^{3/2}$ and L^2, while the cubic tensor hierarchy remains the conservative certificate. These are effective slopes over the plotted depth range, not asymptotic exponents.

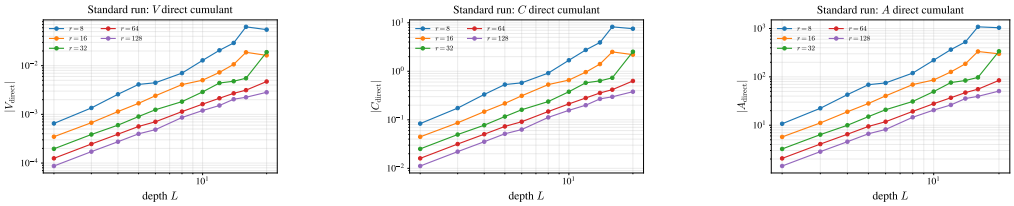


Figure 3: Direct finite-network cumulant magnitudes from the standard long run. From left to right: V_{direct}, C_{direct}, and A_{direct}. These scalar-output observables mix several diagrammatic sectors, so their slopes are effective diagnostics rather than isolated tensor exponents.

Table 2: Effective log-log depth slopes from the same standard long run as Figure 3, using $L \in \{2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 20\}$. These slopes are diagnostics of scalar-output contracted observables, not isolated diagrammatic tensor exponents.

Rank r	V_{direct}	C_{direct}	A_{direct}
8	2.01	2.03	2.06
16	1.76	1.79	1.81
32	1.79	1.81	1.83
64	1.56	1.58	1.59
128	1.55	1.57	1.58

For a transported off-diagonal leg, $\lambda = 3$. Indeed, the leg is multiplied from layer k to layer L by

$$\prod_{j=k+1}^{L} \dot{\mathcal{C}}(\rho_j) = \prod_{j=k+1}^{L} \left(1 - \frac{3}{j} + O\left(\frac{\log j}{j^2}\right)\right) = \left(\frac{k}{L}\right)^3 (1 + o(1)),$$

because the logarithm of the product is $-3\sum_{j=k+1}^{L} j^{-1} + O(1/k)$. This is the source of the exponent 3. Appendix K gives the ReLU moment and Riemann-sum calculation behind the mixed and NTK-NTK endpoint integrals:

$$B_\lambda = \frac{1}{\lambda+1}\left(5 - \frac{4}{\lambda+2}\right), \qquad A_{\lambda,\mu} = \frac{1}{(\lambda+1)(\mu+1)}\left(5 - \frac{4}{\lambda+2} - \frac{4}{\mu+2} + \frac{4}{\lambda+\mu+3}\right). \tag{19}$$

Thus for off-diagonal modes,

$$V : 5, \qquad C : B_3 = \frac{21}{20}, \qquad A : A_{3,3} = \frac{173}{720}. \tag{20}$$

The diagonal/off-diagonal endpoint table is

Tensor coordinate	leading coefficient
$V(e, f)$	5
$D(e, f), F(e, f)$, transported leg diagonal	$B_0 = 3$
$D(e, f), F(e, f)$, transported leg off-diagonal	$B_3 = 21/20$
$A(e, f), B(e, f)$, both legs diagonal	$A_{0,0} = 7/3$
$A(e, f), B(e, f)$, one diagonal and one off-diagonal leg	$A_{0,3} = 43/60$
$A(e, f), B(e, f)$, both legs off-diagonal	$A_{3,3} = 173/720$

Equivalently, for four distinct generic inputs and the three pairing vectors in (51)–(53),

$$\mathcal{V}_L = 5\,\bar{\varepsilon} L\,\mathbf{1}_3 + O(\bar{\varepsilon}\log^2 L), \tag{21}$$

$$\mathcal{C}_L = \frac{21}{20}\,\bar{\varepsilon} L^2\,\mathbf{1}_3 + O(\bar{\varepsilon} L\log^2 L), \tag{22}$$

$$\mathcal{A}_L = \frac{173}{720}\,\bar{\varepsilon} L^3\,\mathbf{1}_3 + O(\bar{\varepsilon} L^2\log^2 L), \tag{23}$$

up to the finite pairing-basis projection factors already absorbed in the $O(\bar{\varepsilon})$ normalization. This is the calculation from which the main L, L^2, L^3 hierarchy is extracted.

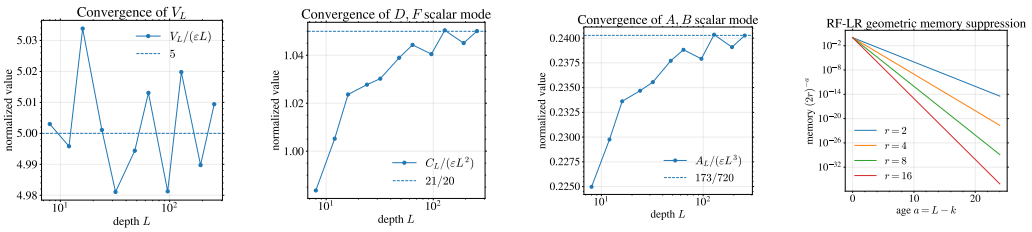


Figure 4: Tensor constants and RF-LR memory. The first three panels show convergence to the isolated constants 5, $21/20$, and $173/720$. The last panel shows geometric suppression of old NTK memory in the RF-LR bottleneck.

These constants are tensor coefficients after pairing-basis extraction, not arbitrary scalar-output cumulants. Direct finite-network cumulants mix Wick/Wishart terms, radial modes, readout contractions, and several Feynman sectors; Appendix H reports the corresponding log-log diagnostics.

A concrete three-layer delta-method calculation is given in Appendix B. It shows explicitly that right-factor-only low rank produces an r^{-1} mean correction and $r^{-1/2}$ single-seed fluctuations, while the diagonal ReLU correction cancels by scale invariance.

6 Operator Concentration and Spectral Preservation

The previous tensor estimates control individual NTK entries. For training stability, the relevant question is stronger: whether the whole empirical NTK matrix stays close enough to its limiting kernel that its smallest useful eigenvalue remains positive. Let $\Theta_{\infty,L}$ be the deterministic limiting NTK and $\widehat{\Theta}_L$ the empirical finite NTK. Entrywise tensor bounds give

$$\mathrm{Var}((\widehat{\Theta}_L)_{ab}) \lesssim L^3\bar{\varepsilon}, \qquad |\mathbb{E}(\widehat{\Theta}_L)_{ab} - (\Theta_{\infty,L})_{ab}| \lesssim L^3\bar{\varepsilon}. \tag{24}$$

A layer martingale and Matrix Freedman yield the operator form. Matrix Freedman is the matrix-valued analogue of Freedman's martingale concentration inequality: it controls the operator norm of a sum of conditionally mean-zero random matrices using a bound on the largest increment and on the predictable quadratic variation. The well-spread data assumption below means that no small group of examples dominates the random feature directions; equivalently, the matrix variance of the empirical NTK fluctuation grows like m, not m^2. This is not true for every deterministic dataset: duplicated, nearly duplicated, or highly aligned examples can violate it. It is a standard incoherence-type assumption, and it holds with high probability for many random generic datasets, for example iid sub-Gaussian or spherical inputs with bounded aspect ratio and no near-collisions, but we keep it explicit rather than hiding it inside the theorem.

Theorem 6.1 (Finite-network operator concentration)**.** *Assume $L^3\bar{\varepsilon} \le c$. Assume further that the data are well-spread in the layerwise feature bases, so the entrywise tensor variance bounds upgrade to a matrix variance proxy of order $mL^3\bar{\varepsilon}$. With probability at least $1-\delta$,*

$$\|\widehat{\Theta}_L - \Theta_{\infty,L}\|_{\mathrm{op}} \le CL^{3/2}\sqrt{m\bar{\varepsilon}}\,\mathrm{polylog}(m,L,1/\delta) + CmL^3\bar{\varepsilon}\,\mathrm{polylog}(m,L,1/\delta). \tag{25}$$

Without the well-spread data assumption, the same argument still gives a robust bound, but the variance term is weaker:

$$CL^{3/2}\sqrt{m\bar{\varepsilon}} \quad \textit{is replaced by} \quad CmL^{3/2}\sqrt{\bar{\varepsilon}}.$$

The deterministic bias term $CmL^3\bar{\varepsilon}$ is unchanged.

Proof sketch. Expose the random weights and projectors layer by layer and decompose $\widehat{\Theta}_L - \mathbb{E}\widehat{\Theta}_L$ as a matrix martingale. Proposition 5.2 gives conditional entrywise variance $O(L^3\bar{\varepsilon})$ and conditional bias $O(L^3\bar{\varepsilon})$. Under the well-spread data assumption, the predictable quadratic variation is $O(mL^3\bar{\varepsilon})$ in operator norm rather than $O(m^2L^3\bar{\varepsilon})$. Matrix Freedman gives the first term in (25), with logarithmic losses from truncation and union over tensor coordinates. The deterministic bias contributes the second term. Without this assumption, the same argument uses the Frobenius-to-operator bound on the variance proxy and loses a factor $\sqrt{m}$. □

If $\mu_\infty(L) = \lambda_{\min}(P_m\Theta_{\infty,L}P_m)$ is the limiting centered spectral margin, Weyl's inequality gives preservation whenever the right-hand side of (25) is at most $\mu_\infty(L)/2$. In the RF-LR bottleneck, the baseline is not the dense MLP margin; the centered proxy margin can scale as $1/(rL)$. Combining this with the bottleneck deviation gives the conservative design rule $r \gtrsim mL^3$, up to logarithmic factors, when the RF-LR perturbation scale is $1/r$.

6.1 RF-LR memory and rank-depth rules

The RF-LR recursion (5) is not a full-rank-compatible perturbation. Its old-kernel memory obeys an exact product rule:

$$\Theta^{(L)}_{\mathrm{old}\leftarrow k} = \Theta^{(k)} \prod_{j=k+1}^{L} \frac{\dot{\Sigma}^{(j)}}{r}. \tag{26}$$

At the normalized ReLU edge of chaos, $\dot{\Sigma}^{(j)} = 1/2 + O(j^{-1})$, hence

$$\left|\Theta^{(L)}_{\text{old}\leftarrow k}\right| \le C\left(\frac{1}{2r}\right)^{L-k}\left(\frac{L}{k}\right)^{C}. \tag{27}$$

Low rank therefore shortens old NTK memory geometrically. The cost is that the centered deterministic margin can shrink with r and L.

Proposition 6.2 (Unprojected and centered RF-LR stability rules)**.** *Consider an RF-LR bottleneck whose centered limiting margin satisfies $\mu_{\mathrm{RF}}(L,r) \gtrsim (rL)^{-1}$. If the empirical operator deviation is controlled by the conservative scale $L^{3/2}\sqrt{m/r} + mL^3/r$, then spectral preservation follows from*

$$r \gtrsim mL^3 \ \mathrm{polylog}(m, L, 1/\delta). \tag{28}$$

If the radial mode is projected out and an angularized centered decomposition improves the variance scale by one power of L, then the sufficient rule improves to

$$r \gtrsim mL^2 \ \mathrm{polylog}(m, L, 1/\delta). \tag{29}$$

Proof. Weyl's inequality requires the operator deviation to be at most a fixed fraction of $\mu_{\mathrm{RF}}(L,r)$. In the unprojected RF-LR model, the radial contribution remains present in the A-sector, whose tensor scale is cubic in depth. The sufficient inequality is therefore dominated by the mL^3/r contribution after absorbing logarithms and constants, giving (28). In the centered/angularized model, the radial sector is removed before concentration and the dominant tensor variance is quadratic in depth. Repeating the same Weyl argument gives (29). The second statement is conditional on this projection and should not be read as the default RF-LR rule. □

Training-time drift. dNTK and ddNTK descendants use the same Gaussian-ReLU atoms and Stiefel-projector cumulants as the NTK itself; differentiation changes deterministic leg labels, not the low-rank random vertex set. The full statement is in Appendix C.

Experiments. The main figures above give the projector, tensor-constant, memory, and slope diagnostics next to the claims they support. Appendix H reports the remaining true finite-network empirical NTKs, the deep cumulant stress test, and exact full-rank matching.

The true finite-network NTK plots, log-log cumulant panels, effective slope table, and depth-200 full-rank matching table are reported in Appendix H. They support the same claims: operator deviations follow the predicted normalization, finite-network cumulants exhibit polynomial effective slopes, and the full-rank endpoint matches the dense NTK to numerical precision.

7 Conclusion

Low rank changes finite-width NTK theory not only by reducing parameters, but by changing the stability threshold for training. The MLP-compatible model isolates full-rank-compatible rank defects, whereas RF-LR bottlenecks suppress old NTK memory by approximately $(2r)^{-(L-k)}$ while also shrinking the centered deterministic gap. Combining these effects gives the main design message: preserving the NTK edge of convexity has a conservative cubic-depth sufficient rule, $r \gtrsim mL^3$, up to logarithmic factors. The experiments do not contradict this theorem; they show that scalar-output contracted observables can have smaller effective exponents, around 1.5–2.1, suggesting a sharper mode-dependent law between $L^{3/2}$ and L^2. The rank-width diagrammatics developed here connect these criteria to conditional operator concentration and to finite-network experiments.

The consequence is that rank should be treated as a stability budget, not only as a compression knob. If the rank is too small, the model may still have fewer parameters, but its empirical NTK can drift far enough from the dense geometry to lose the spectral margin that makes lazy training predictable. The results also explain why full-rank factorization is harmless, why RF-LR bottlenecks behave differently, and why finite-network experiments can show milder effective depth laws than the conservative theorem. In practice, the theory gives a cautious certificate: it identifies when low rank is expected to preserve the optimization geometry, and where sharper mode-dependent predictions are still needed.

To our knowledge, this is the first framework that reconciles low-rank feature geometry with an NTK convexity certificate: low rank can select interpretable bottleneck features while still preserving the spectral conditions that make lazy training predictable.

References

[1] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. *NeurIPS*, 2018.

[2] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *NeurIPS*, 2019.

[3] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *ICML*, 2019.

[4] Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. *ICML*, 2019.

[5] S. Arora, S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *NeurIPS*, 2019.

[6] Q. Nguyen and M. Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *NeurIPS*, 2020.

[7] Q. Nguyen, M. Mondelli, and G. Montúfar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks. *ICML*, 2021.

[8] D. Terjék and D. González-Sánchez. MLPs at the EOC: Concentration of the NTK. arXiv:2501.14724, 2025.

[9] D. Terjék and D. González-Sánchez. MLPs at the EOC: Spectrum of the NTK. arXiv:2501.13225, 2025.

[10] G. Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. *NeurIPS*, 2019.

[11] G. Yang and E. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. *ICML*, 2021.

[12] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. *NeurIPS*, 2021.

[13] G. Yang, D. Yu, C. Zhu, and S. Hayou. Tensor programs VI: Feature learning in infinite-depth neural networks. *ICLR*, 2024.

[14] B. Bordelon and C. Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *NeurIPS*, 2022.

[15] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *PNAS*, 2018.

[16] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *NeurIPS*, 2018.

[17] G. M. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. arXiv:1805.00915, 2018.

[18] J. Huang and H.-T. Yau. Dynamics of deep neural networks and neural tangent hierarchy. *ICML*, 2020.

[19] E. Dyer and G. Gur-Ari. Asymptotics of wide networks from Feynman diagrams. arXiv:1909.11304, 2019.

[20] M. Guillen, P. Misof, and J. E. Gerken. Finite-width neural tangent kernels from Feynman diagrams. arXiv:2508.11522, 2025.

[21] R. Novak, J. Sohl-Dickstein, and S. Schoenholz. Fast finite width neural tangent kernel. *ICML*, 2022.

[22] U. Jang, J. D. Lee, and E. K. Ryu. LoRA training in the NTK regime has no spurious local minima. *ICML*, 2024.

[23] E. Mosig García, A. Agazzi, and D. Trevisan. Quantitative convergence of trained single layer neural networks to Gaussian processes. arXiv:2509.24544, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as a finite-width and finite-rank NTK analysis, and the assumptions are made explicit in the theorem statements. The empirical MNIST table is presented only as motivation, not as evidence for the main theoretical claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion states the narrow scope of the results, including the restriction to signed/isometric low-rank factors and RF-LR bottlenecks, the status of the endpoint constants, the logarithmic losses, and the open dNTK/ddNTK coefficient problem.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions.
   - The authors should reflect on the scope of the claims made.
   - The authors should reflect on the factors that influence the performance of the approach.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [Yes]

Justification: The theorem, proposition, and lemma statements specify the finite-rank, projector, ReLU transport, well-spread-data, and RF-LR assumptions. Proof sketches are in the main text, with the tensor hierarchy, projector cumulants, descendant arguments, and endpoint calculations expanded in the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material.
- Any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper?

Answer: [Yes]

Justification: Appendix H gives widths, ranks, depths, sample sizes, dimensions, repetitions, normalizations, and metrics for the true finite-network and proxy experiments. Appendix I gives the MNIST architecture, optimizer, learning rate, batch size, epochs, seed, and clipping rule.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Reproducibility can be accomplished through code, detailed instructions, model access, or other means appropriate to the research performed.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: An anonymized supplemental code bundle contains the Feynman/NTK experiment scripts and the MNIST scripts, together with a README describing the contents. Synthetic data are generated by the scripts, and MNIST is downloaded through the standard dataset loader.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines for more details.
- While code release is encouraged, a justified [No] is acceptable.
- The instructions should contain the exact command and environment needed to reproduce the results.
- The authors should provide instructions on data access and preparation.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines.
- At submission time, anonymized versions should be released if applicable.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details necessary to understand the results?

Answer: [Yes]

Justification: The appendices specify the finite-network initialization protocol, uniform Stiefel sampling, autograd NTK computation, ranks, widths, depths, sample sizes, seeds or repetition counts, and MNIST optimizer/training settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The true finite-network and proxy experiments report repeated-initialization summaries, including medians and means, with repetition counts stated in Appendix H. The MNIST snapshot is explicitly described as a compact motivating run rather than a statistical claim.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims.
- The factors of variability that the error bars are capturing should be clearly stated.
- The method for calculating the error bars should be explained.
- The assumptions made should be given.
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- For asymmetric distributions, authors should avoid misleading symmetric error bars.
- If error bars are reported, the text should explain how they were calculated and reference the corresponding figures or tables.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

Answer: [No]

Justification: The paper reports the model sizes and repetition counts but does not yet give precise hardware, memory, wall-clock time, or total compute estimates. The experiments are small PyTorch/autograd studies and MNIST runs, but the exact compute resources should be added before submission if required.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: The work is a theoretical and diagnostic study using synthetic data and MNIST, with no human-subject study, private data, deployment, or high-risk released model.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is foundational NTK theory and finite-network diagnostics, with no deployed system, user-facing model, human-subject data, or direct application domain. Its plausible impact is indirect, through better understanding of efficient low-rank training.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include malicious or unintended uses, fairness considerations, privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications.
    - The authors should consider possible harms from intended use, incorrect results, and misuse.
    - If there are negative societal impacts, authors may discuss mitigation strategies.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The paper does not release a high-risk dataset or model. The released material is code for synthetic NTK diagnostics and a standard MNIST benchmark.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards.
    - Datasets scraped from the Internet could pose safety risks.
    - Many papers do not require safeguards, but authors should make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper uses MNIST through standard loaders and does not redistribute the dataset, but the current manuscript does not explicitly list MNIST license terms. This should be added to the final supplemental documentation if strict asset-license disclosure is required.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license should be included for each asset.
    - For scraped data, copyright and terms of service should be provided.
    - If assets are released, license, copyright information, and terms of use should be provided.
    - For existing datasets that are re-packaged, both the original and derived licenses should be provided.
    - If this information is not available online, authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new supplemental asset is the experiment-code bundle, documented by a README that lists the Feynman/NTK scripts, MNIST scripts, and compact result files.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets if applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board approvals or equivalent approvals were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval may be required for human-subject research.
    - Procedures may vary significantly between institutions and locations.
    - For initial submissions, do not include any information that would break anonymity.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the mathematical methods, experiments, or scientific conclusions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to the NeurIPS LLM policy for what should or should not be described.

A Proof Details for Transport and Tensor Hierarchy

This appendix gives the full proof behind Lemma 5.1 and Proposition 5.2. The main text keeps only the proof idea.

Proof of Lemma 5.1. Write

$$A_\ell := I - \frac{M}{\ell} + E_\ell, \qquad \Phi_{L,k} := A_{L-1}A_{L-2}\cdots A_k \quad (L > k),$$

and $\Phi_{k,k} := I$. Since $T_2 = 0$, iteration of (14) gives the exact variation-of-constants identity

$$T_L = \sum_{k=2}^{L-1} \Phi_{L,k+1}\, k^\beta (s + e_k). \tag{30}$$

Let $\Phi^0_{L,k}$ denote the unperturbed product with $E_\ell = 0$. Standard finite-dimensional product estimates for regularly varying matrix products give

$$\Phi^0_{L,k+1} = \exp\left(-M \sum_{j=k+1}^{L-1} \log\left(1 + \frac{1}{j}\right)\right)\left(I + O(k^{-1})\right) = L^{-M}k^M + O\big(L^{-M}k^{M-1}\big), \tag{31}$$

uniformly for $2 \le k < L$, with the usual harmless logarithmic enlargement when M is not diagonalizable. The perturbation is summable:

$$\sum_{\ell\ge 2} \|E_\ell\| \le C \sum_{\ell\ge 2} \frac{\log^p \ell}{\ell^2} < \infty.$$

Discrete Duhamel expansion of the product,

$$\Phi_{L,k+1} - \Phi^0_{L,k+1} = \sum_{j=k+1}^{L-1} \Phi_{L,j+1} E_j \Phi^0_{j,k+1},$$

together with (31) and the summability of E_j, yields

$$\Phi_{L,k+1} = L^{-M}k^M + R_{L,k}, \qquad \sum_{k=2}^{L-1} k^\beta \|R_{L,k}\| = O\big(L^\beta \log^{p+2} L\big). \tag{32}$$

Substituting (32) into (30) gives

$$T_L = L^{-M} \sum_{k=2}^{L-1} k^{M+\beta I} s + O\big(L^\beta \log^{p+2} L\big) + L^{-M} \sum_{k=2}^{L-1} k^{M+\beta I} e_k + O\big(L^\beta \log^{p+q+2} L\big). \tag{33}$$

The error from e_k is $O(L^\beta \log^{q+1} L)$, hence is absorbed by the displayed remainder. The spectral assumption $\Re\lambda(M) > -(\beta+1)$ makes the matrix Riemann integral convergent. Euler-Maclaurin summation for matrix powers gives

$$L^{-M} \sum_{k=2}^{L-1} k^{M+\beta I} = L^{\beta+1} \int_0^1 u^{M+\beta I}\, \mathrm{d}u + O(L^\beta \log L). \tag{34}$$

Since

$$\int_0^1 u^{M+\beta I}\, \mathrm{d}u = (M + (\beta+1)I)^{-1},$$

we obtain (15). This proves the lemma.

Proof of Proposition 5.2. The diagrammatic recursion closes on the three blocks

$$\mathcal{V}_\ell, \qquad \mathcal{C}_\ell = (D_\ell, F_\ell), \qquad \mathcal{A}_\ell = (A_\ell, B_\ell).$$

The ReLU transport assumptions imply deterministic block recursions of the form

$$\mathcal{V}_{\ell+1} = \left(I - \frac{M_V}{\ell} + E_\ell^V\right)\mathcal{V}_\ell + \Xi_\ell^V, \tag{35}$$

$$\mathcal{C}_{\ell+1} = \left(I - \frac{M_C}{\ell} + E_\ell^C\right)\mathcal{C}_\ell + B_\ell^C\mathcal{V}_\ell + \Xi_\ell^C, \tag{36}$$

$$\mathcal{A}_{\ell+1} = \left(I - \frac{M_A}{\ell} + E_\ell^A\right)\mathcal{A}_\ell + B_\ell^A\mathcal{C}_\ell + Q_\ell^A[\mathcal{V}_\ell, \mathcal{V}_\ell] + \Xi_\ell^A, \tag{37}$$

where each E_l^* satisfies the summable bound in Lemma 5.1. The source tensors Ξ^V, Ξ^C, Ξ^A are connected rank-state or projector vertices. By Lemma E.1 and Lemma 4.1, their second-order size is bounded by $C\bar{\varepsilon}$ uniformly in ℓ:

$$\|\Xi_\ell^V\| + \|\Xi_\ell^C\| + \|\Xi_\ell^A\| \le C\bar{\varepsilon}. \tag{38}$$

The deterministic feed-through maps are uniformly bounded in the pairing basis, and the off-diagonal NTK transport contributes one power of ℓ per transported NTK leg after summation. Equivalently, the source exponents entering Lemma 5.1 are $\beta = 0$ for $\mathcal{V}$, $\beta = 1$ for $\mathcal{C}$, and $\beta = 2$ for $\mathcal{A}$.

For $\mathcal{V}$, (35) and (38) give Lemma 5.1 with $\beta = 0$, hence

$$\|\mathcal{V}_L\| \le C\bar{\varepsilon}L.$$

For $\mathcal{C}$, the effective source in (36) is $B_\ell^C\mathcal{V}_\ell + \Xi_\ell^C$. The bound on $\mathcal{V}_\ell$ gives

$$\|B_\ell^C\mathcal{V}_\ell + \Xi_\ell^C\| \le C\bar{\varepsilon}\ell,$$

so Lemma 5.1 with $\beta = 1$ yields

$$\|\mathcal{C}_L\| \le C\bar{\varepsilon}L^2.$$

For $\mathcal{A}$, the effective source in (37) is

$$B_\ell^A\mathcal{C}_\ell + Q_\ell^A[\mathcal{V}_\ell, \mathcal{V}_\ell] + \Xi_\ell^A.$$

Using the previous two bounds,

$$\|B_\ell^A\mathcal{C}_\ell\| \le C\bar{\varepsilon}\ell^2, \qquad \|Q_\ell^A[\mathcal{V}_\ell, \mathcal{V}_\ell]\| \le C\bar{\varepsilon}^2\ell^2.$$

In the perturbative regime $\bar{\varepsilon} \le 1$, the quadratic term is bounded by $C\bar{\varepsilon}\ell^2$, and $\Xi_\ell^A \le C\bar{\varepsilon}$ is smaller. Lemma 5.1 with $\beta = 2$ gives

$$\|\mathcal{A}_L\| \le C\bar{\varepsilon}L^3.$$

Projecting these vector-block estimates onto the coordinates V, D, F, A, B gives (17). Since the perturbative NTK mean correction is a finite linear combination of these blocks, its largest contribution is controlled by $C\bar{\varepsilon}L^3$, which is small when $L^3\bar{\varepsilon} \ll 1$.

B Three-Layer Finite-Rank Calculation

The simplest place where the low-rank simplification is visible is a three-layer network, i.e. two ReLU feature maps. Let

$$Q_x = \frac{1}{r}\sum_{a=1}^{r} X_a^2, \qquad Q_y = \frac{1}{r}\sum_{a=1}^{r} Y_a^2, \qquad S = \frac{1}{r}\sum_{a=1}^{r} X_a Y_a, \qquad C_r = \frac{S}{\sqrt{Q_x Q_y}}.$$

For the right-factor-only low-rank model, the empirical second hidden-layer kernel has the schematic form

$$K_{2,r} = \sqrt{Q_x Q_y}\,\kappa(C_r), \qquad \dot{K}_{2,r} = \dot{\kappa}(C_r), \qquad \widehat{\Theta}_r^{(3)} = K_{2,r} + S\,\dot{K}_{2,r},$$

where κ is the normalized ReLU covariance map.

Proposition B.1 (Delta-method correction for three layers)**.** *For fixed correlation $\rho \in (-1, 1)$, there is an explicit smooth function $\Delta^{(3)}(\rho)$, depending only on $\kappa, \dot\kappa$ and the covariance matrix of (X^2, Y^2, XY), such that*

$$\mathbb{E}\,\widehat{\Theta}_r^{(3)}(\rho) = \Theta_\infty^{(3)}(\rho) + \frac{1}{r}\Delta^{(3)}(\rho) + O(r^{-2}), \qquad \widehat{\Theta}_r^{(3)} - \mathbb{E}\widehat{\Theta}_r^{(3)} = O_{\mathbb{P}}(r^{-1/2}). \tag{39}$$

For scale-invariant ReLU on the diagonal, $\Delta^{(3)}(1) = 0$.

Proof. Write $\widehat{\Theta}_r^{(3)} = F(Q_x, Q_y, S)$. By the multivariate central limit theorem and Lemma E.1,

$$(Q_x, Q_y, S) = (1, 1, \rho) + r^{-1/2}\xi + O_{\mathbb{P}}(r^{-1})$$

with ξ asymptotically Gaussian and covariance determined by the moments of (X^2, Y^2, XY). Taylor expansion gives

$$\mathbb{E}F(Q_x, Q_y, S) = F(1, 1, \rho) + \frac{1}{2r}\operatorname{tr}\left(\nabla^2 F(1, 1, \rho)\,\Sigma_\rho\right) + O(r^{-2}),$$

which is (39). The $O_{\mathbb{P}}(r^{-1/2})$ fluctuation is the first-order delta-method term. On the diagonal $X = Y$, ReLU scale invariance makes the normalized correlation and derivative atoms deterministic along the radial direction, cancelling the first mean correction. □

C NTK Descendants and Training-Time Drift

This appendix records the short argument behind the training-time drift statement. Let $J_c = \nabla_\theta f(x_c)$, $\Xi_{ab;c} = D\Theta_{ab}[J_c]$, and $\Psi_{ab;cd} = D^2\Theta_{ab}[J_c, J_d]$. These descendants add derivative legs to the same layer maps; they do not introduce new independent random variables.

Claim. The joint cumulants of (K, Θ, Ξ, Ψ) are generated by the same Gaussian-ReLU atoms and Stiefel-projector cumulants of $G = (n/r)UU^\top$. Indeed, each derivative with respect to a trainable right factor $B^{(\ell)}$ inserts a deterministic backpropagated leg and the same frozen projector $UU^\top$. Higher derivatives only differentiate ReLU gates and covariance maps, replacing $\kappa, \dot\kappa$ by tensors such as $\ddot\kappa$. Since U is fixed, there is no moving-projector derivative. Thus the connected diagrams for K, Θ, Ξ, Ψ have the same random vertices as the NTK diagrams and differ only in deterministic leg labels.

Under square-loss gradient flow, $\dot\Theta_{ab,t} = -\sum_c \Xi_{ab;c,t} e_{c,t}$. The descendant concentration bound therefore yields, in a lazy window,

$$\|\Theta_t - \Theta_0\|_{\mathrm{op}} \le Ct\|e_0\|_2 \left(L^3\sqrt{m\bar\varepsilon} + mL^3\bar\varepsilon\right) \operatorname{polylog}(m, L, 1/\delta).$$

D Full Proof of the Frozen-Left Decomposition

We prove Proposition 3.2. Consider one layer

$$W = \alpha UB, \qquad U^\top U = I_r, \qquad B \in \mathbb{R}^{r\times d},$$

and let $H = \partial\mathcal{L}/\partial W$ be the Euclidean loss gradient in dense-weight coordinates. The Frobenius inner product is used throughout.

First freeze U. For a variation dB,

$$dW = \alpha U\,dB, \qquad d\mathcal{L} = \langle H, \alpha U\,dB\rangle = \langle \alpha U^\top H, dB\rangle.$$

Hence $\nabla_B\mathcal{L} = \alpha U^\top H$. Gradient descent in B gives

$$\dot B = -\eta\alpha U^\top H, \qquad \dot W_B = \alpha U\dot B = -\eta\alpha^2 UU^\top H,$$

which is (3). Therefore the B-only tangent directions are exactly the fixed linear subspace

$$\mathcal{T}_B W = \{\alpha U\Delta B : \Delta B \in \mathbb{R}^{r\times d}\}.$$

The corresponding contribution to the empirical NTK is a fixed-projector block, because every dense gradient is projected through $UU^\top$.

Now allow U to vary on the Stiefel manifold $\mathrm{St}(n,r) = \{U : U^\top U = I_r\}$. Its tangent space is

$$T_U\mathrm{St}(n,r) = \{\Delta U : U^\top \Delta U + \Delta U^\top U = 0\}.$$

Every tangent vector decomposes as

$$\Delta U = U\Omega + U_\perp K, \qquad \Omega^\top = -\Omega, \qquad U^\top U_\perp = 0,$$

where $U\Omega$ rotates the basis inside $\mathrm{span}(U)$ and $U_\perp K = (I - UU^\top)\Delta U$ moves the subspace itself. For a variation dU,

$$dW = \alpha\, dU\, B, \qquad d\mathcal{L} = \langle H, \alpha dU B\rangle = \langle \alpha H B^\top, dU\rangle.$$

Thus the Euclidean gradient in U-coordinates is $\alpha HB^\top$. Its normal subspace-moving component is

$$(I - UU^\top)\alpha HB^\top.$$

The skew component $U\Omega$ only rotates coordinates inside the same represented subspace: by replacing U with UR and B with $R^\top B$, the product UB is unchanged. This is why the main text states (4) up to rotations inside $\mathrm{span}(U)$ that can be absorbed into B.

Keeping only the subspace-moving component, Stiefel gradient descent contributes

$$\dot U_\perp = -\eta\alpha(I - UU^\top)HB^\top.$$

Mapping this velocity back to dense-weight coordinates gives

$$\dot W_U = \alpha \dot U_\perp B = -\eta\alpha^2(I - UU^\top)HB^\top B,$$

which is (4). This term is absent when U is frozen. It depends on the current right factor through $B^\top B$, so it cannot be represented by cumulants of the fixed Stiefel projector $G = (n/r)UU^\top$ alone.

Finally, if both U and B are trained, the tangent space contains the sum of B-directions $\alpha U\Delta B$, Stiefel subspace directions $\alpha\Delta U B$, and their cross terms in the NTK Gram matrix. Along gradient flow these directions also vary in time, producing dNTK terms involving $\dot U$, $\dot B$, and products such as $B^\top B$. The frozen-left model deliberately removes these moving-subspace interactions, leaving a closed fixed-projector diagrammatic calculus.

E Rank Cumulant Power Counting

Lemma E.1 (Rank cumulant power counting)**.** *Fix a layer ℓ and condition on all previous layers. For rank channel a, let*

$$Z_a^{(\ell)} := \left\{ \left(g_a^{(\ell)}(x_u), \sigma(g_a^{(\ell)}(x_u)), \sigma'(g_a^{(\ell)}(x_u))\right)_{u=1}^m, \text{ and the corresponding derivative atoms} \right\}.$$

Here $g_a^{(\ell)}(x_u)$ is the scalar preactivation of rank channel a on sample x_u, after conditioning on layer $\ell - 1$. Thus $Z_a^{(\ell)}$ is the local random object from which one-channel quantities such as $\sigma\sigma$, $\sigma\sigma'$, $\sigma'\sigma'$, and derivative insertions are evaluated. The variables $Z_1^{(\ell)}, \dots, Z_r^{(\ell)}$ are conditionally iid. Write them as Z_a, and assume the local functions below have uniformly bounded moments of all orders. For

$$X_i^{(r)} = r^{-a_i} \sum_{a=1}^r f_i(Z_a), \qquad i = 1, \dots, s,$$

the connected cumulant satisfies

$$\mathrm{cum}\left(X_1^{(r)}, \dots, X_s^{(r)}\right) = r^{1-\sum_i a_i}\, \mathrm{cum}\left(f_1(Z), \dots, f_s(Z)\right). \tag{40}$$

In particular, normalized empirical averages have s-point cumulants of order r^{1-s}.

Proof. By multilinearity,

$$\mathrm{cum}(X_1^{(r)}, \dots, X_s^{(r)}) = r^{-\sum_i a_i} \sum_{a_1,\dots,a_s=1}^r \mathrm{cum}(f_1(Z_{a_1}), \dots, f_s(Z_{a_s})).$$

Independence kills every term whose indices are not all in the same connected block. Only $a_1 = \cdots = a_s$ contributes, giving r identical terms and hence (40). □

F Projector Cumulants Beyond Second Order

This appendix records the higher-order cumulant scaling of the same projector G used in the main proof. For $G = (n/r)UU^\top$, the mean is $\mathbb{E}G = I_n$, the trace is deterministic, and every connected cumulant containing the trace direction vanishes. Orthogonal Weingarten calculus gives, for fixed cumulant order s,

$$\mathrm{cum}(G_{i_1 j_1}, \ldots, G_{i_s j_s}) = O_s(r^{1-s}) \tag{41}$$

uniformly when $r \le n$ and $r \to \infty$, with the full-rank endpoint cancelling after replacing r^{-1} by the defect scale $\gamma_{n,r}$ at second order. Diagrammatically, this is why a connected rank vertex is counted by the number of connected projector insertions, not simply by the number of factors U in the algebraic expression.

G Low-Rank Diagrammatic Rules

This appendix gives the graphical rules behind the main tensor recursions. This appendix records the graphical calculus underlying the perturbative statements in the main text. The external objects are the same as in finite-width NTK Feynman diagrams: NNGP/preactivation legs, NTK fluctuation legs, and derivative-NTK descendants. The difference is internal. In the low-rank model, stochastic vertices are empirical rank cumulants and frozen-projector cumulants rather than unconstrained neural-index contractions.

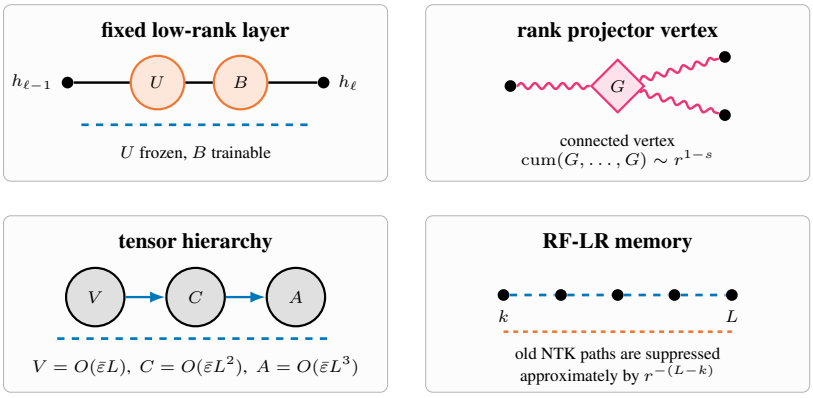


Figure 5: Main diagrammatic objects. Solid black lines denote NNGP/preactivation transport, dashed blue lines denote NTK legs, orange vertices denote low-rank factors, pink diamonds denote Stiefel-projector cumulants, and gray blobs denote transported tensor cumulants.

G.1 Dense NTK rules and low-rank replacement

This subsection makes explicit how the finite-width NTK Feynman rules are imported and modified [20]. The dense calculus has the following ingredients.

1. External filled dots are the observed sample labels. Solid external lines denote preactivation or NNGP objects $z_\alpha, K_{\alpha\beta}$. Colored dashed external lines denote NTK fluctuations $\widehat{\Delta\Theta}_{\alpha\beta}$. Double or multi-dashed descendants denote dNTK and ddNTK insertions.
2. Cubic vertices attach two external legs to one internal line. The decoration of the internal line is the local Gaussian-ReLU atom generated by Taylor expansion of the layer map, for example $\widehat{\Delta\Omega}_{i,\alpha\beta}$, $\sigma_{i,\alpha}\sigma'_{i,\beta}$, or $\sigma'_{i,\alpha}\sigma'_{i,\beta}$. In the dense network each such vertex carries the usual n_ℓ^{-1} counting from the neural index.
3. A white propagator is a Gaussian expectation $\langle\cdot\rangle_{K^{(\ell)}}$. It contracts the decorations of internal lines under the infinite-width NNGP covariance. The selection rules say that propagators connect allowed internal lines, equal neural indices are identified, undecorated dashed NTK lines do not enter the Gaussian expectation, and paired tangent colors contribute the corresponding $\Theta^{(\ell)}_{\alpha\beta}$ factor.

4. Quartic blobs encode the rank-four tensors that close the order-$1/n$ dense hierarchy. The V blob has four K-legs, D, F have two K-legs and two Θ-legs, and A, B have four Θ-legs. The same external grammar also defines the higher-derivative tensors P, Q, R, S, T, U for dNTK and ddNTK.

5. To compute a cumulant or mean correction, one draws all connected diagrams with the prescribed external legs and the desired order in $1/n$, applies the propagator selection rules, translates each diagram into a Gaussian expectation of derivatives of local atoms, and sums the admissible diagrams.

Our low-rank rules keep this external grammar exactly. The replacement happens only inside the stochastic source:

$$\begin{aligned}&\text{dense neural-index source} \longrightarrow \quad \text{rank-state cumulants } r_\ell^{1-s}\kappa_s^{(\ell)}\\ &\qquad\qquad\qquad\qquad \text{and projector cumulants } \operatorname{cum}(G^{(\ell)},\dots,G^{(\ell)})\\ &\frac{1}{n_\ell}\text{ dense second cumulant} \longrightarrow \frac{1}{n_\ell}\text{ dense part } + \gamma_{n_\ell,r_\ell}\text{ projector part } + \frac{1}{r_\ell}\text{ rank-state part.}\end{aligned}$$

At second order the projector contribution is exactly Lemma 4.1. It is proportional to

$$\gamma_{n,r} = \frac{n(n-r)}{r(n-1)(n+2)},$$

so it vanishes at $r = n$. This is the main structural difference from a naive dense replacement $n \mapsto r$: dense Feynman rules have no reason to cancel at full rank, whereas the isometric low-rank projector rules do.

G.2 Worked example: one NTK-mean correction

We now compute one representative diagram in algebraic form. Let $M_A^{(\ell)} = r_\ell^{-1}\sum_{a=1}^{r_\ell}\phi_A^{(\ell)}(Z_a)$ be a rank-state observable, where A labels a local atom such as $\sigma\sigma$, $\sigma\sigma'$, or $\sigma'\sigma'$. Suppose the one-layer NTK update can be written locally as a smooth map

$$\widehat{\Theta}_{12}^{(\ell+1)} = \Phi_{12}\Big(M^{(\ell)}, G^{(\ell)}\Big) \text{ transported lower-layer terms.}$$

Taylor expand Φ_{12} around the deterministic state $(\bar{M}, I)$. The connected rank-state part of the mean correction is

$$\mathbb{E}\widehat{\Theta}_{12}^{(\ell+1)} - \Theta_{12}^{(\ell+1)} = \frac{1}{2r_\ell}\sum_{A,B}\partial_A\partial_B\Phi_{12}(\bar{M}, I)\,\kappa_{AB}^{(\ell)} + \cdots, \tag{42}$$

where

$$\kappa_{AB}^{(\ell)} = \operatorname{cum}\left(\phi_A^{(\ell)}(Z), \phi_B^{(\ell)}(Z)\right).$$

This is the low-rank analogue of a Misof propagator diagram: their white Gaussian propagator evaluates a dense neural-index contraction; here the connected empirical-rank vertex evaluates the cumulant of one rank channel and contributes the explicit factor r_ℓ^{-1}.

The frozen projector gives a second, full-rank-compatible correction. If $G = I + \Delta G$, then the leading projector diagram is

$$\begin{aligned}&\frac{1}{2}\sum_{ij,kl}\partial_{G_{ij}}\partial_{G_{kl}}\Phi_{12}(\bar{M}, I)\ \operatorname{Cov}(G_{ij}, G_{kl})\\ &\qquad = \frac{\gamma_{n_\ell,r_\ell}}{2}\sum_{ij,kl}\partial_{G_{ij}}\partial_{G_{kl}}\Phi_{12}(\bar{M}, I)\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} - \frac{2}{n_\ell}\delta_{ij}\delta_{kl}\right).\end{aligned} \tag{43}$$

Equations (42) and (43) are the same diagrammatic operation as in the dense NTK calculus: choose the external Θ_{12} legs, attach the allowed local vertices, evaluate the connected stochastic source, and sum. The only change is the source value. Dense Misof diagrams count neural-index loops by $1/n$; our diagrams count rank-channel cumulants by $1/r$ and Stiefel-projector cumulants by $\gamma_{n,r}$.

For comparison, the dense next-to-leading NTK mean calculation has five admissible diagrams [20]: transported $\Theta^{\{1\}}$, transported $K^{\{1\}}$, and source diagrams from V, D, F. In our notation this becomes the schematic low-rank expansion

$$\Theta^{\{1\}(\ell+1)}_{\mathrm{LR},12} = \mathsf{T}_\Theta[\Theta^{\{1\}(\ell)}_{\mathrm{LR}}]_{12}\mathsf{T}_K[K^{\{1\}(\ell)}_{\mathrm{LR}}]_{12}\mathsf{S}_V[V_\ell]_{12}\mathsf{S}_D[D_\ell]_{12}\mathsf{S}_F[F_\ell]_{12}, \tag{44}$$

but the tensors V, D, F now decompose into dense finite-width, rank-state, and projector-defect sources. Thus the external five-diagram grammar is imported unchanged, while the internal weights are changed so that exact full-rank matching is preserved.

G.3 Summary of the low-rank results encoded by the rules

The diagrammatic rules above encode the paper's main results as follows.

1. Exact full-rank matching: when $r_\ell = n_\ell$, $U^{(\ell)}$ is orthogonal, $G = I$, every centered projector vertex vanishes, and the factorized NTK equals the dense NTK pathwise.
2. Stiefel projector covariance: the second projector vertex is $\mathrm{Cov}(G_{ij}, G_{kl}) = \gamma_{n,r}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} - 2\delta_{ij}\delta_{kl}/n)$, with $\gamma_{n,n} = 0$.
3. Rank cumulant power counting: a connected s-point rank-state vertex contributes r^{1-s}. Normalized rank averages therefore have s-point cumulants of order r^{1-s}.
4. Tensor hierarchy: the imported V, C, A external grammar remains triangular, with $V = O(\bar{\varepsilon}L)$, $C = O(\bar{\varepsilon}L^2)$, and $A = O(\bar{\varepsilon}L^3)$ under the stated ReLU transport assumptions.
5. Endpoint constants: after pairing-basis extraction, the isolated off-diagonal constants are 5 for V, $21/20$ for the mixed C sector, and $173/720$ for the A sector.
6. Operator criterion: the tensor bounds imply $\|\widehat{\Theta}_L - \Theta_{\infty,L}\|_{\mathrm{op}} \lesssim L^{3/2}\sqrt{m\bar{\varepsilon}} + mL^3\bar{\varepsilon}$, up to logarithmic factors and the well-spread data condition.
7. RF-LR memory and rank-depth rule: in the bottleneck model, old NTK lines are multiplied by $\prod_{j=k+1}^{L} \dot{\Sigma}^{(j)}/r$, giving geometric memory suppression and the conservative stability rule $r \gtrsim mL^3$, with the centered/angularized mL^2 rule only under the additional projection hypothesis.

G.4 Feynman diagram building blocks and tensor assembly

The diagrams below should be read as Feynman diagrams, not as decorative notation. A filled dot is an observed sample label. A solid external leg represents a preactivation or NNGP entry K_{ab}. A dashed colored leg represents an NTK fluctuation $\Delta\Theta_{ab}$. Double or multi-dashed legs represent dNTK and ddNTK descendants. The tensor is determined by the external legs:

$$\begin{array}{rcl} K, K & \rightsquigarrow & V, \\ K, \Theta & \rightsquigarrow & D, F, \\ \Theta, \Theta & \rightsquigarrow & A, B, \\ \mathrm{d}\Theta, K & \rightsquigarrow & P, Q, \\ \mathrm{dd}\Theta \text{ external legs} & \rightsquigarrow & R, S, T, U. \end{array}$$

The ordering of sample labels chooses the pairing coordinate. For example, the two K-leg cumulants $\mathrm{cum}(\delta K_{12}, \delta K_{34})$, $\mathrm{cum}(\delta K_{13}, \delta K_{24})$, and $\mathrm{cum}(\delta K_{14}, \delta K_{23})$ give V_{1234}, V_{1324}, and V_{1423}. Similarly, one solid pair and one dashed pair give D_{1234}, F_{1324}, F_{1423}, while two dashed pairs give A_{1234}, B_{1324}, B_{1423}.

Once the external legs are fixed, every connected diagram is built from the same three operations: transport an older tensor through deterministic K- or Θ-leg derivatives, feed a lower block into a higher block, or attach a fresh low-rank source. The fresh source is where our rules differ from the dense Misof rules: dense neural-index contractions are replaced by rank-state cumulants $r^{1-s}\kappa_s$ and Stiefel-projector cumulants, whose second-order scale is $\gamma_{n,r}$.

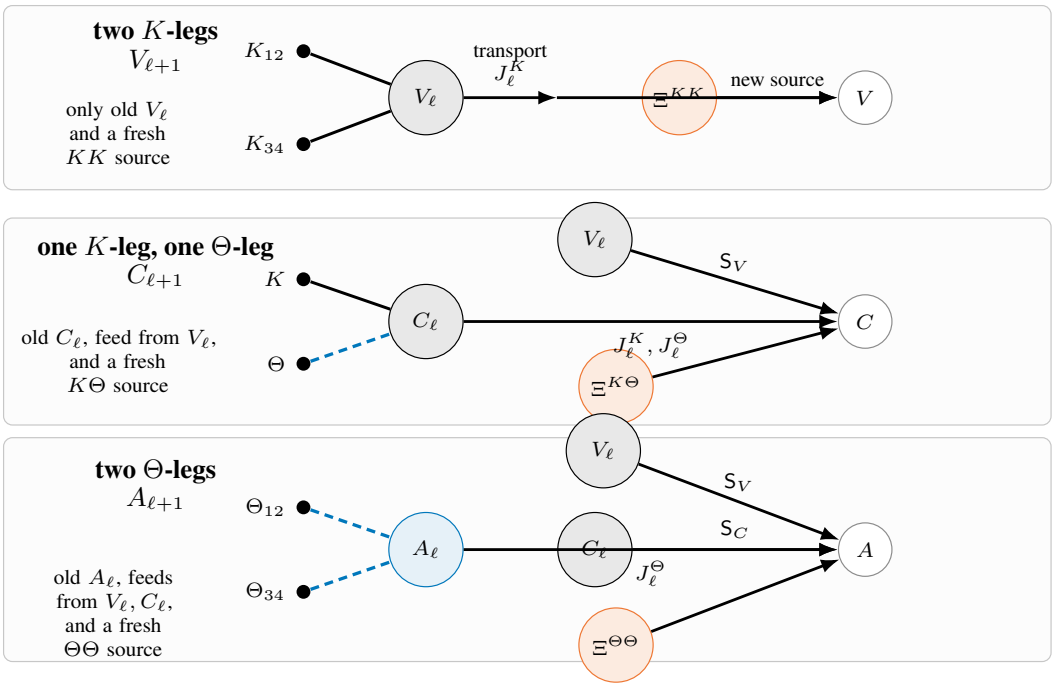


Figure 6: Diagrammatic reduction of the tensor hierarchy. External K- and Θ-legs choose the block being computed. Connected terms either transport an older tensor, feed a lower block into a higher one, or create a fresh low-rank source Ξ.

For the NTK-NTK block this gives the schematic reduction

$$\operatorname{cum}(\delta\Theta_{12}^{(\ell+1)}, \delta\Theta_{34}^{(\ell+1)}) = \underbrace{J_\ell^\Theta \operatorname{cum}(\delta\Theta_{12}^{(\ell)}, \delta\Theta_{34}^{(\ell)})(J_\ell^\Theta)^\top}_{\text{transported old NTK-NTK diagram}} + \underbrace{\mathsf{S}_C[C_\ell] + \mathsf{S}_V[V_\ell]}_{\text{mixed diagrams feeding into NTK-NTK}} + \underbrace{\Xi_\ell^{\Theta\Theta}}_{\text{new connected rank/projector source}}. \tag{45}$$

The same operation with two K-legs gives $V_{\ell+1}$, and with one K-leg and one Θ-leg gives $C_{\ell+1}$. Thus the apparently large expansion collapses because every connected diagram has only two possible roles: it either transports an older tensor through J_ℓ^K or J_ℓ^Θ, or it creates a fresh source Ξ. This is the mechanism behind the triangular recursion in (11).

G.5 Basic tensor recursions

The diagrammatic rules produce the same external tensor hierarchy as the dense finite-width expansion, but with low-rank sources. Let J_ℓ^K and J_ℓ^Θ denote the deterministic linear transports for NNGP and NTK legs. The NNGP covariance block satisfies schematically

$$V_{\ell+1} = J_\ell^K V_\ell (J_\ell^K)^\top + \Xi_\ell^{KK}, \qquad \Xi_\ell^{KK} = r_\ell^{-1}\kappa_2(S^K, S^K) + \gamma_{n_\ell, r_\ell}\mathfrak{h}_\ell^{KK}. \tag{46}$$

For the mixed NNGP-NTK block $C = (D, F)$,

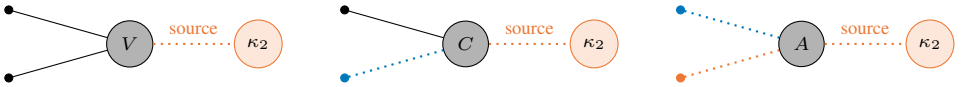
$$C_{\ell+1} = J_\ell^K C_\ell (J_\ell^\Theta)^\top + \mathsf{S}_V[V_\ell] + \Xi_\ell^{K\Theta}. \tag{47}$$

For the NTK-NTK block $A = (A, B)$,

$$A_{\ell+1} = J_\ell^\Theta A_\ell (J_\ell^\Theta)^\top + \mathsf{S}_C[C_\ell] + \mathsf{S}_V[V_\ell] + \Xi_\ell^{\Theta\Theta}. \tag{48}$$

The sources Ξ^{KK}, $\Xi^{K\Theta}$, and $\Xi^{\Theta\Theta}$ are precisely the rank and projector cumulant vertices with the corresponding external legs. The next display should be read only as a source dictionary. It is not a separate calculation: it says which connected second cumulant is injected into each tensor block.

The rank/projector blob labelled κ_2 contributes the small factor r^{-1}, $1/n$, or $\gamma_{n,r}$. The number of external Θ-legs then determines how many NTK transports are accumulated over depth, and therefore whether the block grows like L, L^2, or L^3.



(49)

Equivalently, the diagrammatic message is

External legs	Tensor block	Fresh source size	Depth accumulation
K, K	$V = \mathrm{cum}(K, K)$	$O(\bar{\varepsilon})$	no transported NTK leg, hence $O(\bar{\varepsilon}L)$
K, Θ	$C = (D, F)$	$O(\bar{\varepsilon})$	one transported NTK leg, hence $O(\bar{\varepsilon}L^2)$
Θ, Θ	$A = (A, B)$	$O(\bar{\varepsilon})$	two transported NTK legs, hence $O(\bar{\varepsilon}L^3)$

Thus the diagrams separate two effects that are easy to confuse. The orange/rank blob gives the perturbative smallness; it does not decide the power of L. The external K- and Θ-legs decide the depth power through deterministic transport.

G.6 First NTK mean correction

The familiar finite-width correction to the NTK mean has the same external five-term form:

$$\Theta_{\ell+1}^{\{1\}} = \mathsf{T}_\Theta[\Theta_\ell^{\{1\}}] + \mathsf{T}_K[K_\ell^{\{1\}}] + \mathsf{S}_V[V_\ell] + \mathsf{S}_D[D_\ell] + \mathsf{S}_F[F_\ell]. \tag{50}$$

The low-rank difference is that the source tensors V, D, F are further resolved into the rank-state and projector vertices above. Thus the low-rank calculus does not change the external NTK grammar; it changes the internal expansion parameter from neural-index contractions to empirical-rank cumulants and frozen-projector cumulants.

G.7 Full tensor dictionary

For four sample labels $1, 2, 3, 4$, write

$$\delta K_{ab}^{(\ell)} := K_{ab}^{(\ell)} - \mathbb{E}K_{ab}^{(\ell)}, \qquad \delta\Theta_{ab}^{(\ell)} := \Theta_{ab}^{(\ell)} - \mathbb{E}\Theta_{ab}^{(\ell)}.$$

All coordinates below are centered two-point cumulants of these pair observables. Since the variables are centered, $\mathrm{cum}(X, Y) = \mathbb{E}[XY]$. In particular, $V_{1234}^{(\ell)} = \mathbb{E}[\delta K_{12}^{(\ell)} \delta K_{34}^{(\ell)}]$, $D_{1234}^{(\ell)} = \mathbb{E}[\delta K_{12}^{(\ell)} \delta\Theta_{34}^{(\ell)}]$, $F_{1324}^{(\ell)} = \mathbb{E}[\delta K_{13}^{(\ell)} \delta\Theta_{24}^{(\ell)}]$, $A_{1234}^{(\ell)} = \mathbb{E}[\delta\Theta_{12}^{(\ell)} \delta\Theta_{34}^{(\ell)}]$, and $B_{1324}^{(\ell)} = \mathbb{E}[\delta\Theta_{13}^{(\ell)} \delta\Theta_{24}^{(\ell)}]$. The rank-four tensor basis is organized by the three pairings $(12|34)$, $(13|24)$, and $(14|23)$. The NNGP-NNGP block is

$$\mathcal{V}_\ell = \begin{pmatrix} V_{1234}^{(\ell)} \\ V_{1324}^{(\ell)} \\ V_{1423}^{(\ell)} \end{pmatrix} := \begin{pmatrix} \mathrm{cum}(\delta K_{12}^{(\ell)}, \delta K_{34}^{(\ell)}) \\ \mathrm{cum}(\delta K_{13}^{(\ell)}, \delta K_{24}^{(\ell)}) \\ \mathrm{cum}(\delta K_{14}^{(\ell)}, \delta K_{23}^{(\ell)}) \end{pmatrix}. \tag{51}$$

The mixed NNGP-NTK block is

$$\mathcal{C}_\ell = \begin{pmatrix} D_{1234}^{(\ell)} \\ F_{1324}^{(\ell)} \\ F_{1423}^{(\ell)} \end{pmatrix} := \begin{pmatrix} \mathrm{cum}(\delta K_{12}^{(\ell)}, \delta\Theta_{34}^{(\ell)}) \\ \mathrm{cum}(\delta K_{13}^{(\ell)}, \delta\Theta_{24}^{(\ell)}) \\ \mathrm{cum}(\delta K_{14}^{(\ell)}, \delta\Theta_{23}^{(\ell)}) \end{pmatrix}. \tag{52}$$

The NTK-NTK block is

$$\mathcal{A}_\ell = \begin{pmatrix} A_{1234}^{(\ell)} \\ B_{1324}^{(\ell)} \\ B_{1423}^{(\ell)} \end{pmatrix} := \begin{pmatrix} \mathrm{cum}(\delta\Theta_{12}^{(\ell)}, \delta\Theta_{34}^{(\ell)}) \\ \mathrm{cum}(\delta\Theta_{13}^{(\ell)}, \delta\Theta_{24}^{(\ell)}) \\ \mathrm{cum}(\delta\Theta_{14}^{(\ell)}, \delta\Theta_{23}^{(\ell)}) \end{pmatrix}. \tag{53}$$

This is the tensor dictionary behind the compact V, C, A notation in the main text: $C = (D, F)$ and $A = (A, B)$. The dense finite-width Feynman sectors remain the same, but their low-rank stochastic vertices are cumulants of the frozen projector $G = (n/r)UU^\top$.

G.8 Pairing extraction

A rank-four orthogonally invariant tensor decomposes as

$$C_{i_1i_2i_3i_4} = c_1\delta_{i_1i_2}\delta_{i_3i_4} + c_2\delta_{i_1i_3}\delta_{i_2i_4} + c_3\delta_{i_1i_4}\delta_{i_2i_3}. \tag{54}$$

Let B_1, B_2, B_3 be the three pairings. The normalized pairing Gram matrix is

$$\overline{G}_N = \begin{pmatrix} 1 & 1/N & 1/N \\ 1/N & 1 & 1/N \\ 1/N & 1/N & 1 \end{pmatrix}, \qquad \overline{G}_N^{-1} = \frac{N}{N-1} I_3 - \frac{N}{(N-1)(N+2)} \mathbf{1}\mathbf{1}^\top. \tag{55}$$

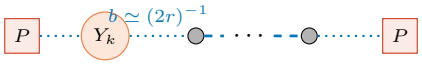
If $h = (\langle C, B_1\rangle, \langle C, B_2\rangle, \langle C, B_3\rangle)$ are measured contractions, the isolated tensor coefficients are $(c_1, c_2, c_3)^\top = \overline{G}_N^{-1} h$. This is why direct scalar-output cumulants should not be identified with the endpoint coefficients 5, $21/20$, and $173/720$ without a pairing-basis extraction.

G.9 Centered RF-LR memory lemma

For RF-LR, an old NTK line transported from layer k to L carries

$$B_{L:k} = \prod_{s=k+1}^{L} \frac{1}{r}\dot{\Sigma}^{(s)}. \tag{56}$$

At ReLU edge of chaos, $\dot{\Sigma}^{(s)} \to 1/2$, so $\|B_{L:k}\| \lesssim (2r)^{-(L-k)}$, up to polynomial endpoint factors. The corresponding centered transport diagram is



(57)

Let $P = I - \mathbf{1}\mathbf{1}^\top/m$ and $Z_L = P(\widehat{K}_L - K_L)P$. The centered fluctuation admits the linearized expansion

$$Z_L = \sum_{k=1}^{L} B_{L:k+1} Y_k + \sum_{k=1}^{L} B_{L:k+1} R_k, \tag{58}$$

where Y_k is a centered fresh rank source and R_k is quadratic in earlier fluctuations.

Lemma G.1 (Centered RF-LR memory)**.** *Assume, conditionally on the past,*

$$\left\|\mathbb{E}[Y_k^2 \mid \mathcal{F}_{k-1}]\right\|_{\mathrm{op}} \le C\frac{m\varepsilon_{\mathrm{RF}}}{r^2}, \qquad \|Y_k\|_{\psi_1} \le C\frac{\sqrt{\varepsilon_{\mathrm{RF}}}}{r}, \tag{59}$$

and $\|B_{L:k}\| \le q^{L-k}$ for some $q < 1$. If $\|R_k\|_{\mathrm{op}} \le Cm\varepsilon_{\mathrm{RF}}/r^2$ with high probability, then, up to logarithmic factors,

$$\|Z_L\|_{\mathrm{op}} \lesssim \frac{\sqrt{m\varepsilon_{\mathrm{RF}}\log(m/\delta)}}{r} + \frac{m\varepsilon_{\mathrm{RF}}\log(m/\delta)}{r^2} \tag{60}$$

with probability at least $1-\delta$.

Proof. Set $X_k = B_{L:k+1}Y_k$. The predictable matrix variance satisfies

$$\left\|\sum_{k=1}^{L} \mathbb{E}[X_k^2 \mid \mathcal{F}_{k-1}]\right\|_{\mathrm{op}} \le C\frac{m\varepsilon_{\mathrm{RF}}}{r^2}\sum_{k=1}^{L} q^{2(L-k)} \le C'\frac{m\varepsilon_{\mathrm{RF}}}{r^2}. \tag{61}$$

Matrix Freedman or Bernstein gives the first term in (60). For the quadratic remainder,

$$\left\|\sum_{k=1}^{L} B_{L:k+1}R_k\right\|_{\mathrm{op}} \le C\frac{m\varepsilon_{\mathrm{RF}}}{r^2}\sum_{k=1}^{L} q^{L-k} \le C'\frac{m\varepsilon_{\mathrm{RF}}}{r^2}. \tag{62}$$

□

Combining Lemma G.1 with Weyl's inequality gives the centered/angularized criterion. If the centered deterministic margin satisfies $\lambda_{\min}(PK_LP) \gtrsim (rL)^{-1}$, and if $\varepsilon_{\mathrm{RF}} \simeq 1/r$, then the fluctuation bound is below the margin when

$$r \gtrsim mL^2 \tag{63}$$

up to logarithms. Without removing the radial constant mode, a diffusive $\sqrt{L}$ factor can reappear, giving the conservative cubic-depth criterion used for the unprojected RF-LR statement.

H Extended True Finite-Network Experiments

This appendix keeps the detailed diagnostics out of the main ten-page narrative while preserving the evidence used to interpret the finite-network checks.

H.1 Experimental methodology

All true finite-network experiments use scalar-output ReLU networks with the signed/isometric factorization $W^{(\ell)} = \sigma_w n_{\ell-1}^{-1/2} \sqrt{n_\ell / r_\ell}\, U^{(\ell)} B^{(\ell)}$. The left factors $U^{(\ell)}$ are sampled uniformly from the Stiefel manifold and held fixed; the empirical NTK is computed by PyTorch autograd with respect to the trainable right factors $B^{(\ell)}$ and the readout. Inputs are synthetic, fixed across the compared initializations, and normalized to unit Euclidean norm. The reported quantities are not training accuracies: they are initialization-time NTK and NNGP diagnostics designed to test the rank-width perturbation theory.

For the clean true-NTK operator and rank-defect sweep, we use width $n = 128$, sample size $m = 8$, input dimension $d = 16$, ranks $r \in \{8, 16, 32, 64, 128\}$, and depths $L \in \{2, 3, 4, 5, 6, 8, 10\}$. For each pair (r, L), 200 independent initializations are drawn. For a fixed (r, L), we compute the empirical NTK Gram matrix $\widehat{\Theta}$, subtract its initialization mean across seeds, and report the median and mean operator norm of the fluctuation. The normalization in Figure 7 is $L^{3/2}\sqrt{m\varepsilon}$, with $\varepsilon = 1/n + \gamma_{n,r}$. The rank-defect panel fixes the largest depth in this sweep and varies r, checking that the excess low-rank contribution follows the defect scale and disappears at $r = n$. Exact full-rank matching is tested separately at $n = r = 128$, $m = 8$, $d = 16$, and depths up to $L = 200$; because the theorem is pathwise, this test measures numerical precision rather than a statistical average.

For direct finite-network cumulants, we again use $n = 128$, $m = 8$, $d = 16$, and ranks $r \in \{8, 16, 32, 64, 128\}$. For every (r, L), 2000 independent initializations are used. The standard long sweep uses depths $L \in \{2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 20\}$. The deep stress test uses $L \in \{20, 50, 100, 200\}$. For each initialization we compute the final hidden-feature Gram matrix $K = h_L h_L^\top / n$ and the empirical NTK $\Theta = JJ^\top$. Across seeds we estimate the centered second cumulants $V = \mathrm{cum}(K, K)$, $C = \mathrm{cum}(K, \Theta)$, and $A = \mathrm{cum}(\Theta, \Theta)$ on fixed same-off-diagonal and disjoint-off-diagonal sample pairings. The log-log slopes in Table 2 are ordinary least-squares fits of $\log|\text{cumulant}|$ against $\log L$ over the listed depth set for each rank. These slopes are diagnostic effective exponents of contracted scalar-output observables; they are not estimates of isolated pairing-basis tensor constants.

The proxy experiments in the main text are separate Monte Carlo checks of closed-form quantities. The Stiefel-projector covariance experiment uses $n = 64$, ranks $r \in \{4, 8, 16, 32, 64\}$, and 50,000 Stiefel samples to compare $\mathrm{Cov}(G_{12}, G_{12})$ with $\gamma_{n,r}$. The scalar endpoint-constant experiment uses $\varepsilon = 10^{-3}$, depths $L \in \{8, 12, 16, 24, 32, 48, 64, 96, 128, 192, 256\}$, and 200,000 Monte Carlo samples to check convergence of the normalized V, C, and A proxies to 5, $21/20$, and $173/720$. The operator proxy uses $m = 64$ and 300 repetitions per depth to check stability of the normalization $L^{3/2}\sqrt{m\varepsilon}$. The RF-LR memory panel uses ranks $2, 4, 8, 16$ and tracks the product of old-NTK transport factors over ages up to 24 layers.

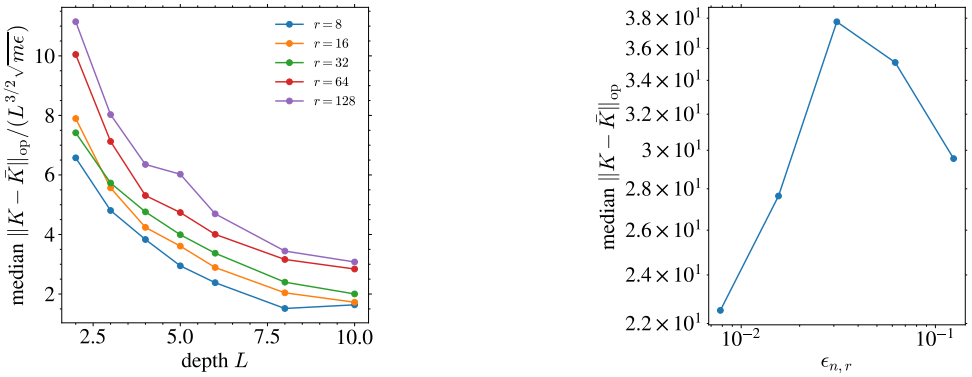


Figure 7: True finite-network NTK checks computed by autodiff. Left: empirical operator deviations follow the predicted normalization. Right: the low-rank excess deviation tracks the rank-defect scale; at the full-rank endpoint the rank-defect contribution disappears, while ordinary finite-width and numerical residuals can remain.

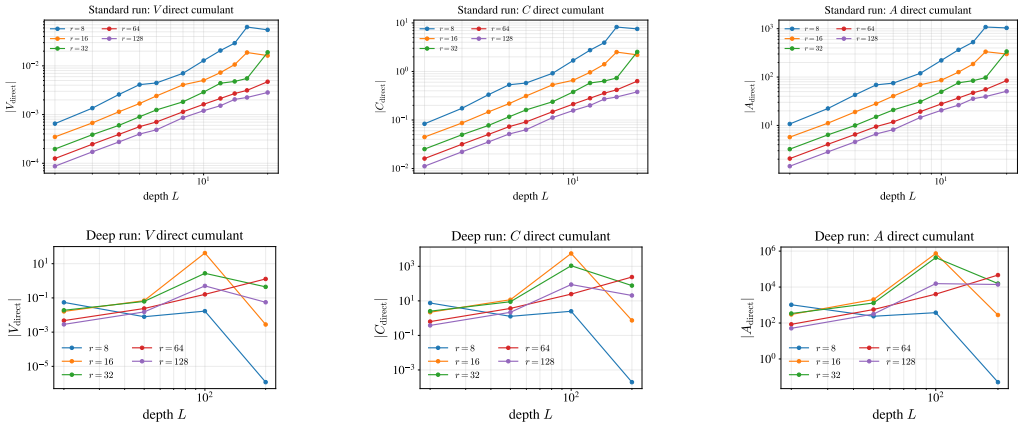


Figure 8: Log-log finite-network cumulant magnitudes. Top: the standard long run. Bottom: the $L \in \{20, 50, 100, 200\}$ deep stress test. These plots show effective depth slopes directly, before normalizing by the proposed L, L^2, and L^3 scales.

Table 3: Exact full-rank matching for true empirical NTKs at $n = r = 128$, $m = 8$, $d = 16$. Differences remain at numerical precision up to depth $L = 200$, confirming Theorem 3.1.

Depth	max absolute NTK difference	relative difference
2	$8.88 \cdot 10^{-16}$	$8.09 \cdot 10^{-17}$
4	$3.55 \cdot 10^{-15}$	$1.31 \cdot 10^{-16}$
6	$5.33 \cdot 10^{-15}$	$4.05 \cdot 10^{-16}$
8	$3.55 \cdot 10^{-15}$	$4.20 \cdot 10^{-16}$
10	$6.22 \cdot 10^{-15}$	$1.14 \cdot 10^{-15}$
20	$1.95 \cdot 10^{-14}$	$3.77 \cdot 10^{-15}$
50	$5.68 \cdot 10^{-14}$	$6.68 \cdot 10^{-16}$
100	$2.14 \cdot 10^{-15}$	$2.35 \cdot 10^{-14}$
200	$8.08 \cdot 10^{-14}$	$1.27 \cdot 10^{-14}$

I MNIST Parameter-Efficiency Snapshot

This auxiliary experiment is included only as empirical motivation for the rank bottleneck question; it is not used in the NTK proofs. The numerical summary is reported in Table 1. MNIST images are flattened to 784 dimensions and normalized with the standard MNIST mean and variance. All models are trained with cross-entropy using Adam with learning rate 10^{-3}, batch size 128, 30 epochs, seed 42, and gradient clipping at norm 1. The dense baseline is a ReLU MLP with hidden width 512 and two hidden layers. The low-rank models use the same width and a rank bottleneck; in the small-rank rows, one random feature side of the bottleneck is held fixed and the low-rank trainable maps and classifier are optimized.

J Why Strict Positive NMF Is Not Covered

This appendix explains why signed isometric low rank is not the same mathematical model as nonnegative matrix factorization. The model in (1) is signed and isometric. The right factor B is unconstrained and the frozen left factor U has orthonormal columns. Strict nonnegative matrix factorization imposes a cone constraint and changes the tangent space even at initialization. Its NTK is a projected or conic kernel, not the fixed linear-subspace kernel analyzed here. Therefore the results should not be cited as a theorem for positive NMF without a separate analysis of active constraints, boundary faces, and cone-projected gradients.

K Proof Details for Endpoint Constants

This appendix gives the elementary ReLU moment calculation that produces the endpoint constants used in the main text. For $g \sim \mathcal{N}(0,1)$, let

$$X_0 = 2(g_+)^2, \qquad Y_0 = 2\mathbf{1}_{\{g>0\}}.$$

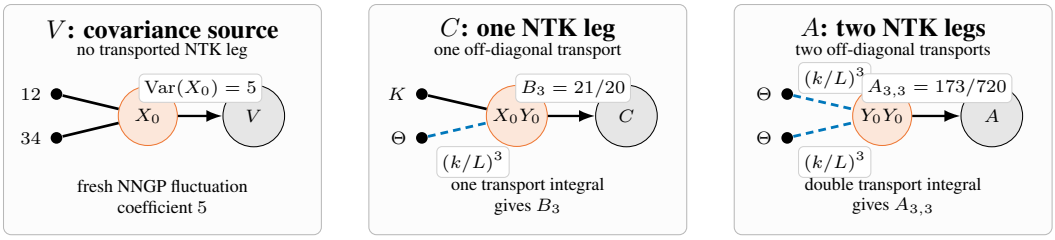


Figure 9: Endpoint diagrams for the constants in (20). Each panel separates the local ReLU atom from the external transport: V has no transported NTK leg, C has one off-diagonal NTK leg, and A has two.

Here X_0 is the scalar ReLU covariance atom and Y_0 is the scalar derivative atom appearing on an NTK leg. Since $\mathbb{E}[g_+^2] = 1/2$ and $\mathbb{E}[g_+^4] = 3/2$,

$$\mathbb{E}X_0 = 1, \qquad \mathbb{E}X_0^2 = 6, \qquad \mathrm{Var}(X_0) = 5.$$

Also $\mathbb{E}Y_0 = 1$, $\mathbb{E}Y_0^2 = 2$, and therefore $\mathrm{Var}(Y_0) = 1$. Finally,

$$\mathbb{E}X_0Y_0 = \mathbb{E}\big[4(g_+)^2\mathbf{1}_{\{g>0\}}\big] = 4\mathbb{E}[g_+^2] = 2, \qquad \mathrm{Cov}(X_0, Y_0) = 2 - 1 = 1.$$

Thus a fresh K-K source has endpoint coefficient 5.

It remains to explain the transported NTK legs. If a leg born at layer k is transported to depth L, the ReLU edge-of-chaos off-diagonal derivative satisfies

$$\prod_{j=k+1}^{L} \dot{\mathcal{C}}(\rho_j) = \prod_{j=k+1}^{L} \left(1 - \frac{3}{j} + O\left(\frac{\log j}{j^2}\right)\right) = \left(\frac{k}{L}\right)^3 (1 + o(1)).$$

More generally, write the transport exponent as λ. The finite depth sum for one transported NTK leg is a Riemann/Beta-sum limit,

$$B_\lambda = \lim_{L\to\infty} \frac{1}{L} \sum_{k=1}^{L} \left(\frac{k}{L}\right)^\lambda \left(1 + 4\frac{k}{L}\right) = \frac{1}{\lambda+1} + \frac{4}{\lambda+2} = \frac{1}{\lambda+1}\left(5 - \frac{4}{\lambda+2}\right).$$

The two-leg calculation is the same pairing-basis projection with two transported legs. Its endpoint sum is

$$A_{\lambda,\mu} = \frac{1}{(\lambda+1)(\mu+1)}\left(5 - \frac{4}{\lambda+2} - \frac{4}{\mu+2} + \frac{4}{\lambda+\mu+3}\right).$$

Consequently, for the isolated off-diagonal ReLU modes used in the main text, $\lambda = \mu = 3$, and

$$V = 5, \qquad C = B_3 = \frac{21}{20}, \qquad A = A_{3,3} = \frac{173}{720}.$$