How Does Orthogonalization Adapt to the Neural-Network Hessian Structure? A Gradient Self Outer-Product Analysis at Initialization

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Muon orthogonalizes a weight matrix's momentum before each step, and on neural networks this simple preconditioner beats entry-wise optimizers by a wide margin. Most existing analyses, however, work in a very abstract problem class, from which it is hard to see why orthogonalization should be particularly suited to neural networks. This work analyzes Muon's preconditioner in three concrete neural-network settings. The layer-wise Hessian of a neural network is known to be diagonally dominant within its row blocks. While Muon's implicit preconditioner has the matching Kronecker form $(\boldsymbol{V}\boldsymbol{V}^\top)^{1/2} \otimes \boldsymbol{I}$. The two align exactly when $\boldsymbol{V}\boldsymbol{V}^\top$ is itself diagonal, which raises a concrete question: when is $\boldsymbol{V}\boldsymbol{V}^\top$ (approximately) diagonal? To answer it, we compute $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ (equivalently $\mathbb{E}[\boldsymbol{V}\boldsymbol{V}^\top]$ at initialization) in closed form under Gaussian init, for three standard settings: symmetric matrix factorization, deep linear networks, and two-layer ReLU networks. In each case, the diagonal entries dominate the off-diagonal ones as the width grows. Hence $\boldsymbol{V}\boldsymbol{V}^\top$ is asymptotically diagonal: Muon's preconditioner aligns with the Hessian's row-block structure.

1. Introduction

Muon [14] optimizes a matrix-shaped weight $\boldsymbol{W}_t \in \mathbb{R}^{m\times n}$ by orthogonalizing the momentum buffer instead of using it directly. From a stochastic gradient $\boldsymbol{G}_t$,

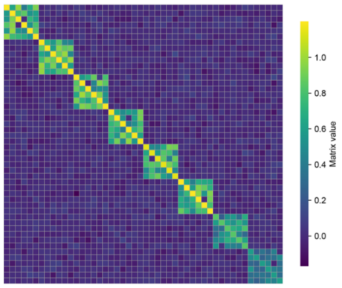
$$\boldsymbol{V}_t = \beta\boldsymbol{V}_{t-1} + (1-\beta)\boldsymbol{G}_t, \qquad \boldsymbol{D}_t = \mathrm{NS}_5(\boldsymbol{V}_t) \approx (\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2}\boldsymbol{V}_t, \qquad \boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t\boldsymbol{D}_t, \quad (1)$$

with $\boldsymbol{V}_0 = \boldsymbol{0}$ and NS_5 a fixed-point iteration for the orthogonal polar factor: if $\boldsymbol{V}_t = \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{Q}^\top$ is a thin SVD then $\boldsymbol{D}_t = \boldsymbol{P}\boldsymbol{Q}^\top = (\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2}\boldsymbol{V}_t$. A growing body of work studies Muon as a general optimization algorithm: momentum steepest descent, or a non-Euclidean trust-region step, in a spectral-norm geometry, with convergence rates and an implicit spectral-norm constraint on the iterates [1–4, 6, 15–17, 21, 24].

All of this works in an abstract problem class, and that is exactly its limitation: it cannot explain why Muon, a simple matrix preconditioner, dramatically outperforms entry-wise optimizers *on neural networks specifically*. The few network-specific results [9, 19, 26–28, 32] sharpen the picture in particular regimes but leave the basic structural question open: what makes the orthogonalization preconditioner suited to a neural network's curvature?

Preconditioner shape versus Hessian shape. Take a single step from initialization, where $\boldsymbol{V}_0 = \boldsymbol{G}_0$ and hence $\boldsymbol{V}_t\boldsymbol{V}_t^\top = \boldsymbol{G}_t\boldsymbol{G}_t^\top$. Vectorizing (1) row by row gives $\mathrm{vec}(\boldsymbol{D}_t) = \big((\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2} \otimes \boldsymbol{I}_n\big)\mathrm{vec}(\boldsymbol{V}_t)$, so Muon descends in the metric

$$H_{\mathrm{Muon}} = \boldsymbol{P}_t^{1/2} \otimes \boldsymbol{I}_n, \qquad \boldsymbol{P}_t := \boldsymbol{V}_t\boldsymbol{V}_t^\top \in \mathbb{R}^{m\times m}, \quad (2)$$



Hessian for Transformer Weight Matrix

$$\begin{pmatrix} \mathrm{p}_{11}\mathrm{I}_n & \mathrm{p}_{12}\mathrm{I}_n & \dots & \mathrm{p}_{1m}\mathrm{I}_n \\ \mathrm{p}_{21}\mathrm{I}_n & \mathrm{p}_{22}\mathrm{I}_n & \dots & \mathrm{p}_{2m}\mathrm{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{p}_{m1}\mathrm{I}_n & \mathrm{p}_{m2}\mathrm{I}_n & \dots & \mathrm{p}_{mm}\mathrm{I}_n \end{pmatrix}$$

$$\mathrm{H}_{\mathrm{MUON}} = \mathrm{P}^{\frac{1}{2}} \otimes \mathrm{I}_n$$

Preconditioner for Muon

Figure 1: Muon's implicit preconditioner $H_{\text{Muon}} = \boldsymbol{P}_t^{1/2} \otimes \boldsymbol{I}_n$ next to the layer-wise Hessian. When $\boldsymbol{P}_t = \boldsymbol{V}_t\boldsymbol{V}_t^\top$ is close to diagonal, H_{Muon} matches the dominant row-block pattern of the neural-network Hessian.

a Kronecker shape determined entirely by $\boldsymbol{P}_t$. On the other side, the layer-wise Hessian of a weight $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, unfolded into $m \times m$ blocks of size $n \times n$ along the m row directions, is *row-block diagonally dominant* [8, 33, 34]; since the Hessian is the ideal preconditioner under a quadratic model, that block pattern is the shape a good preconditioner should have. By (2), H_{Muon} takes exactly that block-diagonal form precisely when $\boldsymbol{P}_t$ is diagonal (Figure 1). So the question reduces to a clean property of the gradient.

> **Question.** Is $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ (equivalently, $\boldsymbol{G}_t\boldsymbol{G}_t^\top$ at initialization) diagonally dominant for neural-network weight matrices?

For neural networks the answer is yes, in the limit of large width. In practice this has been verified empirically along the entire training trajectory of GPT-2 and LLaMA [7]; here we give a theoretical answer at initialization in the three settings most used in neural-network theory, with the diagonal-versus-off-diagonal orders summarized at the end of Section 4. Two facts come out of the analysis.

1. The orthogonalization preconditioner aligns asymptotically with the neural-network Hessian: $\boldsymbol{P}_t = \boldsymbol{V}_t\boldsymbol{V}_t^\top$ becomes diagonal as the width grows, so H_{Muon} becomes block-diagonal in the Hessian's row-block structure.
2. Orthogonalization is asymptotically the row-wise normalization used by several recent stateless optimizers: when $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ is diagonal, $(\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2}\boldsymbol{V}_t$ is exactly each row of $\boldsymbol{V}_t$ divided by its own ℓ_2 norm. (Appendix D)

2. Related Work

Theory of Muon. One line analysis Muon in a general optimization setup (such as nonconvex smooth). Li and Hong [17], Shen et al. [24] give non-convex convergence rates for momentum

steepest descent in the spectral norm and its practical variants; Kovalev [15] reads gradient orthogonalization as a non-Euclidean trust-region step; Bernstein and Newhouse [2, 3], Chen et al. [4], Pethick et al. [21] describe Muon through norm-constrained linear minimization oracles and modular duality, with Chen et al. [4] extracting an implicit spectral-norm constraint on the weights; An et al. [1], Lau et al. [16] fit Muon into a wider family of matrix-gradient preconditioners; and Davis and Drusvyatskiy [6] asks when spectral updates help. None of this depends on the network, so none of it addresses why orthogonalization matches neural-network curvature. A smaller line does look at the network: Fan et al. [9], Vasudeva et al. [27] characterize the margin bias of spectral descent and its generalization benefit on separable or imbalanced data; Wang et al. [28] studies Muon on associative-memory learning; Zhang et al. [32] ties layer-wise preconditioning to provable feature learning; Su [26] asks whether orthogonalization is optimal under an isotropic curvature model; and Ma et al. [19] raises, in a matrix-factorization setup, the very question of how gradient orthogonalization meets the curvature of training. We answer it through the diagonal structure of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ across the three settings.

Row-wise normalization optimizers. A separate line drops optimizer state by normalizing the gradient row by row. Zhang et al. [34] groups learning rates by Hessian block, motivated directly by the row-block structure of the Transformer Hessian; Glentis et al. [10], Ma et al. [18], Scetbon et al. [23], Wen et al. [29] replace Adam's state with a row-wise ℓ_2 normalization (plus a second, cheap normalization); Gu and Xie [11], Xu et al. [31] look at the manifold-optimization and width-scaling sides of row/column normalization; Pethick et al. [21] reach a row-wise normalization from norm-constrained steepest descent; and Deng et al. [7] make the link to Muon explicit, showing row-momentum normalization matches Muon at scale. Our analysis explains the common thread: once $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ is diagonally dominant, row normalization and orthogonalization are the same operation.

3. Preliminaries

We study, at a Gaussian initialization, the $m \times m$ matrix $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ for a weight $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ with population loss $\mathcal{L}$ and gradient $\boldsymbol{G} = \partial\mathcal{L}/\partial\boldsymbol{W}$; this is the expected value of $\boldsymbol{P}_0 = \boldsymbol{V}_0\boldsymbol{V}_0^\top$ in (2). We use three settings, each stated with its own network and assumption. Throughout, $\boldsymbol{I}_d$ is the identity, tr the trace, $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm, $\sigma(z) = \max(z, 0)$ the ReLU with $\sigma'(z) = \mathbf{1}[z > 0]$, $\otimes$ the Kronecker product, and $a_n = \Theta(b_n)$ means $c\, b_n \le a_n \le C\, b_n$ for fixed $0 < c \le C$ as the relevant dimension grows.

Symmetric matrix factorization. For positive integers d, k, r with $k \ge r$, the over-parameterized loss is $\mathcal{L}(\boldsymbol{U}) := \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2$ with $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ and a fixed target $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$ that is symmetric positive semidefinite of rank r; the gradient is $\boldsymbol{G} = (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star)\boldsymbol{U} \in \mathbb{R}^{d\times k}$. The width is k.

Deep linear network. For $L \ge 1$ and layer dimensions $d_0, \ldots, d_L$, the network $f(\boldsymbol{x}) = \boldsymbol{W}_L \cdots \boldsymbol{W}_1\boldsymbol{x}$ with $\boldsymbol{W}_i \in \mathbb{R}^{d_i\times d_{i-1}}$ learns a fixed linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_L\times d_0}$ under $\mathcal{L} = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\,\|f(\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x}\|^2$. Write $\boldsymbol{B}_i := \boldsymbol{W}_{i-1}\cdots\boldsymbol{W}_1$ (with $\boldsymbol{B}_1 := \boldsymbol{I}$), $\boldsymbol{A}_i := \boldsymbol{W}_L\cdots\boldsymbol{W}_{i+1}$ (with $\boldsymbol{A}_L := \boldsymbol{I}$), and $\boldsymbol{R} := \boldsymbol{W}_L\cdots\boldsymbol{W}_1 - \boldsymbol{\Phi}$; the layer-$i$ gradient is $\boldsymbol{G}_i = \partial\mathcal{L}/\partial\boldsymbol{W}_i = \boldsymbol{A}_i^\top\boldsymbol{R}\,\boldsymbol{B}_i^\top$. The widths are the inner dimensions $d_1, \ldots, d_{L-1}$.

Two-layer ReLU network. For positive d_0, d_1, d_2, the network $f(x) = W_2\sigma(W_1x)$ with $W_1 \in \mathbb{R}^{d_1\times d_0}$, $W_2 \in \mathbb{R}^{d_2\times d_1}$ learns a fixed linear target $\Phi \in \mathbb{R}^{d_2\times d_0}$ under $\mathcal{L} = \frac{1}{2}\mathbb{E}_x \|f(x) - \Phi x\|^2$; write $h(x) := \sigma(W_1x)$, w_a for the a-th row of W_1, and $G_i := \partial\mathcal{L}/\partial W_i$. The hidden dimension is d_1.

Assumption 1 (Gaussian initialization) *All weight entries ($\boldsymbol{U}$; every $\boldsymbol{W}_i$; W_1 and W_2) are i.i.d. $\mathcal{N}(0,1)$; in the deep linear and ReLU settings the input is $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$ and is integrated out in $\mathcal{L}$; the target ($\boldsymbol{M}^\star$ or $\boldsymbol{\Phi}$) is deterministic and independent of the weights. Expectations are over the weights.*

4. Main Results

We state the closed form of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ in each setup, read off the dominance, and defer the proofs to Sections A to C.

Symmetric matrix factorization.

Theorem 1 *Under Assumption 1 (matrix factorization), with $\boldsymbol{W} := \boldsymbol{U}\boldsymbol{U}^\top \sim \mathcal{W}_d(k, \boldsymbol{I}_d)$ Wishart,*

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] = \alpha_3\, \boldsymbol{I}_d - 2\alpha_2\, \boldsymbol{M}^\star + k\,(\boldsymbol{M}^\star)^2, \tag{3}$$

where $\alpha_2 = k(k+d+1)$ and $\alpha_3 = k^3 + 3(d+1)k^2 + (d^2+3d+4)k$ are the (isotropic) diagonal entries of $\mathbb{E}[\boldsymbol{W}^2]$ and $\mathbb{E}[\boldsymbol{W}^3]$.

Reading off entries with $d, \boldsymbol{M}^\star$ fixed: the diagonal is $\alpha_3 - 2\alpha_2\boldsymbol{M}^\star_{ii} + k(\boldsymbol{M}^\star)^2_{ii} = \Theta(k^3)$ and the off-diagonal is $-2\alpha_2\boldsymbol{M}^\star_{ij} + k(\boldsymbol{M}^\star)^2_{ij} = \Theta(k^2)$ whenever $\boldsymbol{M}^\star_{ij} \neq 0$, so the ratio is $\Theta(k/\boldsymbol{M}^\star_{ij})$. As the width k grows, $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ is diagonal up to a factor k, and the large isotropic $\Theta(k^3)$ part is exactly what orthogonalization strips away, leaving the target-aligned $\boldsymbol{M}^\star$.

Deep linear network.

Theorem 2 *Under Assumption 1 (deep linear network), for every layer $i \in \{1, \dots, L\}$,*

$$\mathbb{E}[\boldsymbol{G}_i\boldsymbol{G}_i^\top] = V_i\, s_i\, \boldsymbol{I}_{d_i} + \boldsymbol{T}_i, \qquad \boldsymbol{T}_i = \begin{cases} \left(\prod_{j=1}^{L-1} d_j\right)\boldsymbol{\Phi}\boldsymbol{\Phi}^\top, & i = L, \\ \left(\prod_{j\neq i,\, 1\le j\le L-1} d_j\right)\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\ \boldsymbol{I}_{d_i}, & 1 \le i < L, \end{cases} \tag{4}$$

where V_i, s_i are positive dimension-only scalars given by two-step recursions in $(d_0, \dots, d_L)$ (Section B), with $V_L s_L \asymp d_{L-1}\left(\prod_{j=0}^{L-2} d_j\right)^2$.

For a hidden layer ($i < L$) the off-diagonal block $\boldsymbol{T}_i$ is a multiple of $\boldsymbol{I}_{d_i}$, so $\mathbb{E}[\boldsymbol{G}_i\boldsymbol{G}_i^\top]$ is exactly diagonal: linearity together with the sign symmetry $\boldsymbol{W}_i \mapsto \boldsymbol{D}\boldsymbol{W}_i$, $\boldsymbol{W}_{i+1} \mapsto \boldsymbol{W}_{i+1}\boldsymbol{D}^{-1}$ for diagonal $\boldsymbol{D}$ kills every off-diagonal expectation. At the output layer the diagonal noise $V_L s_L$ grows with the inner widths while the off-diagonal stays proportional to $\left(\prod_{j<L} d_j\right)(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ab}$, so the ratio still diverges.

Two-layer ReLU network. Let $\theta := \angle(u, v)$ for independent $u, v \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$, and set the dimension-free constants $\kappa_1 := \mathbb{E}\big[\frac{\sin^2\theta+(\pi-\theta)^2}{4\pi^2} + \frac{\sin\theta\cos\theta(\pi-\theta)}{2\pi^2}\big]$ and $\kappa_2 := \frac{\mathbb{E}[\sin\theta(\pi-\theta)]}{2\pi^2}$; as $d_0 \to \infty$, $\kappa_1 \to \frac{1}{4\pi^2} + \frac{1}{16}$ and $\kappa_2 \to \frac{1}{4\pi}$.

Theorem 3 *Under Assumption 1 (two-layer ReLU network), with* $H := \mathbb{E}_x[h(x)h(x)^\top]$ *and* $\alpha := \mathbb{E}_{W_1}[\mathrm{tr}(H^2)]$,

$$\mathbb{E}[G_2 G_2^\top] = \alpha\, \boldsymbol{I}_{d_2} + \tfrac{d_1}{4}\Phi\Phi^\top, \qquad \alpha = \Theta(d_0^2 d_1^2),$$

and $\mathbb{E}[G_1 G_1^\top]$ *is permutation-invariant in the hidden index with diagonal entries*

$$\mathbb{E}[(G_1 G_1^\top)_{aa}] = \tfrac{d_0 d_2^2}{4} + d_0 d_2\big[(d_1-1)\kappa_1 + \tfrac{1}{2}\big] + \tfrac{\|\Phi\|_{\mathrm{F}}^2}{4}$$

and off-diagonal entries $\mathbb{E}[(G_1 G_1^\top)_{ac}] = d_0 d_2\, \kappa_2 + O(d_2\sqrt{d_0})$ *for* $a \neq c$.

For G_1 the diagonal is $\Theta(d_0 d_2^2 + d_0 d_1 d_2)$ against an off-diagonal of $\Theta(d_0 d_2)$, a ratio of $\Theta(d_1)$, so $G_1 G_1^\top$ is asymptotically diagonal; for G_2 the target-free $\alpha = \Theta(d_0^2 d_1^2)$ beats the off-diagonal $\frac{d_1}{4}(\Phi\Phi^\top)_{ac}$ by $\Theta(d_0^2 d_1)$. The one nonzero off-diagonal, of order $\Theta(d_0 d_2)$, comes from a single term in the proof and traces back to ReLU's lack of odd symmetry, which is why the dominance here is asymptotic rather than exact. Simulations match every formula (Section C).

Summary and proof sketch. In all three settings the diagonal of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ outgrows the off-diagonal as the width grows (table below), so $\boldsymbol{P}_t = \boldsymbol{V}_t\boldsymbol{V}_t^\top$ is asymptotically diagonal, $H_{\mathrm{Muon}} = \boldsymbol{P}_t^{1/2} \otimes \boldsymbol{I}_n$ is asymptotically block-diagonal in the Hessian's row-block structure, and (Section D) orthogonalization is asymptotically a row-wise normalization.

Orders of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ at a Gaussian initialization

Model	Diagonal	Off-diagonal	Width
Matrix factorization	$\Theta(k^3)$	$\Theta(k^2)$	k
Deep linear, hidden layer i	$\Theta(V_i s_i)$	0	$d_1, \ldots, d_{L-1}$
Deep linear, output layer	$\Theta(V_L s_L)$	$\Theta\big(\prod_{j<L} d_j\big)$	$d_1, \ldots, d_{L-1}$
Two-layer ReLU, layer W_1	$\Theta(d_0 d_2^2 + d_0 d_1 d_2)$	$\Theta(d_0 d_2)$	d_1
Two-layer ReLU, layer W_2	$\Theta(d_0^2 d_1^2)$	$\Theta(d_1)$	d_1

The proofs all run the same way. We expand $\mathbb{E}[(\boldsymbol{G}\boldsymbol{G}^\top)_{ab}]$ by Gaussian (Wick/Isserlis) moments; a diagonal entry ($a = b$) and an off-diagonal one ($a \neq b$) differ in how many of the underlying i.i.d. Gaussian coordinates the surviving pairings tie together. On the diagonal the relevant quantities (gradient rows, columns of $\boldsymbol{U}$, hidden neurons) share an index and are strongly correlated, so many pairings survive and their contributions accumulate over the width; off the diagonal the correlations are weak, since independent Gaussian vectors are near-orthogonal in expectation, so the cross terms cancel or contribute only lower order. This is the same effect that makes a Gaussian vector's self-inner-product grow with its dimension while its cross inner-products do not. For the deep linear network the off-diagonal cancellation is exact at hidden layers because of the sign symmetry; ReLU breaks that symmetry and leaves an $O(1)$-per-pair residual, which still loses to the $\Theta(d_1)$ diagonal.

References

[1] Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

[2] Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024.

[3] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[4] Lizhang Chen, Jikai Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

[5] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 342–350, 2009.

[6] Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

[7] Shenyang Deng, Zhuoli Ouyang, Tianyu Pang, Zihang Liu, Ruochen Jin, Shuhua Yu, and Yaoqing Yang. RMNP: Row-momentum normalized preconditioning for scalable matrix-based optimization. *arXiv preprint arXiv:2603.20527*, 2026.

[8] Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the Hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

[9] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

[10] Athanasios Glentis, Jiaxiang Li, Andi Han, and Mingyi Hong. A minimalist optimizer design for LLM pretraining. *arXiv preprint arXiv:2506.16659*, 2025.

[11] Yufei Gu and Zeke Xie. Mano: Restriking manifold optimization for LLM training. *arXiv preprint arXiv:2601.23000*, 2026.

[12] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1–2):134–139, 1918.

[13] Svante Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, 1997.

[14] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. https://kellerjordan.github.io/posts/muon/.

[15] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

[16] Tim Tsz-Kit Lau, Qi Long, and Weijie Su. PolarGrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

[17] Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025.

[18] Chao Ma, Wenbo Gong, Meyer Scetbon, and Edward Meeds. SWAN: SGD with normalization and whitening enables stateless LLM training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[19] Jianhao Ma, Yiding Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026.

[20] Ivan Nourdin and Giovanni Peccati. *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*. Cambridge University Press, 2012.

[21] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

[22] Robert Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, IT-4(2):69–72, 1958.

[23] Meyer Scetbon, Chao Ma, Wenbo Gong, and Edward Meeds. Gradient multi-normalization for stateless and scalable LLM training. *arXiv preprint arXiv:2502.06742*, 2025.

[24] Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

[25] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

[26] Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.

[27] Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025.

[28] Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms Adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

[29] Ziqing Wen, Yanqi Shi, Jiahuan Wang, Ping Luo, Linbo Qiao, Dongsheng Li, and Tao Sun. SRON: State-free LLM training via row-wise gradient normalization. In *OpenReview preprint*, 2025. https://openreview.net/forum?id=BtQLBWr6zI.

[30] Gian-Carlo Wick. The evaluation of the collision matrix. *Physical Review*, 80(2):268–272, 1950.

[31] Ruihan Xu, Jiajin Li, and Yiping Lu. On the width scaling of neural optimizers under matrix operator norms I: Row/column normalization and hyperparameter transfer. *arXiv preprint arXiv:2603.09952*, 2026.

[32] Thomas T. Zhang, Behrad Moniri, Ansh Nagwekar, Faraz Rahman, Anton Xue, Hamed Hassani, and Nikolai Matni. On the concurrence of layer-wise preconditioning methods and provable feature learning. *arXiv preprint arXiv:2502.01763*, 2025.

[33] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need Adam: A Hessian perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[34] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

Contents

Appendix A. Symmetric Matrix Factorization

A.1. Setup

Let d, k, r be positive integers with $k \geq r$. The over-parameterized symmetric matrix-factorization loss is

$$\mathcal{L}(\boldsymbol{U}) \;:=\; \tfrac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2, \qquad \boldsymbol{U} \in \mathbb{R}^{d\times k}, \tag{5}$$

with target $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$ symmetric positive semidefinite of rank r. A direct computation gives the gradient

$$\boldsymbol{G} \;:=\; \nabla\mathcal{L}(\boldsymbol{U}) \;=\; (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star)\,\boldsymbol{U} \;\in\; \mathbb{R}^{d\times k}, \tag{6}$$

and the object of interest is the $d \times d$ matrix $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$.

A.2. Standing assumption

Assumption 2 *The entries U_{ij} are i.i.d. $\mathcal{N}(0,1)$, and $\boldsymbol{M}^\star$ is deterministic and independent of $\boldsymbol{U}$. Hence $\boldsymbol{W} := \boldsymbol{U}\boldsymbol{U}^\top$ is Wishart with k degrees of freedom and identity scale, $\boldsymbol{W} \sim \mathcal{W}_d(k, \boldsymbol{I}_d)$.*

A.3. Main results restated

Theorem (Theorem 1, restated). *Under Assumption 2,*

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] \;=\; \alpha_3(d,k)\,\boldsymbol{I}_d \;-\; 2\,\alpha_2(d,k)\,\boldsymbol{M}^\star \;+\; k\,(\boldsymbol{M}^\star)^2, \tag{7}$$

where the scalars are the diagonal entries of the first three Wishart moments,

$$\alpha_1(d,k) = k, \tag{8}$$
$$\alpha_2(d,k) = k(k+d+1), \tag{9}$$
$$\alpha_3(d,k) = k^3 + 3(d+1)k^2 + (d^2+3d+4)k. \tag{10}$$

Corollary 1 *Under Assumption 2, $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ is invariant under any orthogonal change of basis fixing $\boldsymbol{M}^\star$, with*

$$\mathbb{E}\big[(\boldsymbol{G}\boldsymbol{G}^\top)_{ii}\big] = \alpha_3 - 2\alpha_2\boldsymbol{M}^\star_{ii} + k(\boldsymbol{M}^\star)^2_{ii}, \tag{11}$$
$$\mathbb{E}\big[(\boldsymbol{G}\boldsymbol{G}^\top)_{ij}\big] = -2\alpha_2\boldsymbol{M}^\star_{ij} + k(\boldsymbol{M}^\star)^2_{ij} \qquad (i \neq j). \tag{12}$$

For fixed d, $\boldsymbol{M}^\star$ as $k \to \infty$, $\mathbb{E}[(\boldsymbol{G}\boldsymbol{G}^\top)_{ii}] = \Theta(k^3)$ and $\mathbb{E}[(\boldsymbol{G}\boldsymbol{G}^\top)_{ij}] = \Theta(k^2)$ whenever $\boldsymbol{M}^\star_{ij} \neq 0$, so the dominance ratio is $\Theta(k/\boldsymbol{M}^\star_{ij})$.

A.4. Preliminary lemmas

Lemma 1 *Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0,1)$ entries. Then for deterministic $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{N} \in \mathbb{R}^{m\times m}$,*

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top] = \operatorname{tr}(\boldsymbol{M})\,\boldsymbol{I}_m, \qquad \mathbb{E}[\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{X}] = \operatorname{tr}(\boldsymbol{N})\,\boldsymbol{I}_n. \tag{13}$$

Proof. Using $\mathbb{E}[\boldsymbol{X}_{ap}\boldsymbol{X}_{bq}] = \delta_{ab}\delta_{pq}$,

$$\mathbb{E}[(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)_{ab}] = \sum_{p,q} \boldsymbol{M}_{pq}\,\delta_{ab}\delta_{pq} = \delta_{ab}\operatorname{tr}(\boldsymbol{M}),$$

which is the first identity; the second follows by applying it to $\boldsymbol{X}^\top$ (entries i.i.d. $\mathcal{N}(0,1)$). ∎

Lemma 2 *Let $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top$ with $\boldsymbol{U}$ as in Assumption 2. For every $m \in \mathbb{N}$ there is a scalar $\alpha_m(d,k)$ with*

$$\mathbb{E}[\boldsymbol{W}^m] = \alpha_m(d,k)\,\boldsymbol{I}_d, \qquad \alpha_m(d,k) = \tfrac{1}{d}\,\mathbb{E}[\operatorname{tr}(\boldsymbol{W}^m)]. \tag{14}$$

Proof. For $\boldsymbol{Q} \in O(d)$, $\boldsymbol{Q}\boldsymbol{U} \overset{d}{=} \boldsymbol{U}$, hence $\boldsymbol{Q}\boldsymbol{W}^m\boldsymbol{Q}^\top \overset{d}{=} \boldsymbol{W}^m$ and

$$\mathbb{E}[\boldsymbol{W}^m] = \boldsymbol{Q}\,\mathbb{E}[\boldsymbol{W}^m]\,\boldsymbol{Q}^\top \qquad \text{for all } \boldsymbol{Q} \in O(d).$$

By Schur's lemma $\mathbb{E}[\boldsymbol{W}^m] = \alpha_m\boldsymbol{I}_d$; taking the trace gives $d\alpha_m = \mathbb{E}[\operatorname{tr}(\boldsymbol{W}^m)]$. ∎

Lemma 3 (Wick/Isserlis) *Let $X_1, \dots, X_{2n}$ be jointly centered Gaussian. Then*

$$\mathbb{E}[X_1 \cdots X_{2n}] = \sum_{P} \prod_{\{i,j\}\in P} \mathbb{E}[X_i X_j], \tag{15}$$

the sum over all pairings P of $\{1,\dots,2n\}$; odd moments vanish [12, 30].

Proof. Differentiate $\mathbb{E}[\exp(\boldsymbol{t}^\top\boldsymbol{X})] = \exp(\frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t})$ in $t_1, \dots, t_{2n}$ at $\boldsymbol{t} = \boldsymbol{0}$; cf. Janson [13, §1.4]. ∎

A.5. Proofs

Lemma 4 *With $\boldsymbol{W} := \boldsymbol{U}\boldsymbol{U}^\top$ and $\boldsymbol{B} := \boldsymbol{W} - \boldsymbol{M}^\star$,*

$$\boldsymbol{G}\boldsymbol{G}^\top = \boldsymbol{B}\boldsymbol{W}\boldsymbol{B} = \boldsymbol{W}^3 - \boldsymbol{W}^2\boldsymbol{M}^\star - \boldsymbol{M}^\star\boldsymbol{W}^2 + \boldsymbol{M}^\star\boldsymbol{W}\boldsymbol{M}^\star. \tag{16}$$

Proof. Since $\boldsymbol{B} = \boldsymbol{B}^\top$,

$$\boldsymbol{G}\boldsymbol{G}^\top = \boldsymbol{B}\boldsymbol{U}(\boldsymbol{B}\boldsymbol{U})^\top = \boldsymbol{B}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{B}^\top = \boldsymbol{B}\boldsymbol{W}\boldsymbol{B},$$

and substituting $\boldsymbol{B} = \boldsymbol{W} - \boldsymbol{M}^\star$ and expanding gives (16). ∎

Lemma 5 *Under Assumption 2, $\mathbb{E}[\boldsymbol{W}] = k\,\boldsymbol{I}_d$, hence $\alpha_1(d,k) = k$.*

Proof. Let $\boldsymbol{u}_a$ denote the a-th column of $\boldsymbol{U}$. Then

$$\mathbb{E}[\boldsymbol{W}] = \mathbb{E}\Big[\sum_{a=1}^{k} \boldsymbol{u}_a\boldsymbol{u}_a^\top\Big] = \sum_{a=1}^{k} \mathbb{E}[\boldsymbol{u}_a\boldsymbol{u}_a^\top] = k\,\boldsymbol{I}_d.$$

∎

Lemma 6 *Under Assumption 2, $\mathbb{E}[\boldsymbol{W}^2] = k(k+d+1)\,\boldsymbol{I}_d$, hence $\alpha_2(d,k) = k(k+d+1)$.*

Proof. By Lemma 2, $\mathbb{E}[\boldsymbol{W}^2] = \alpha_2 \boldsymbol{I}_d$. With columns $\boldsymbol{u}_a \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$,

$$\boldsymbol{W}^2 = \sum_{a,b=1}^{k} \langle \boldsymbol{u}_a, \boldsymbol{u}_b \rangle \, \boldsymbol{u}_a \boldsymbol{u}_b^\top.$$

Split into the k terms with $a = b$ and the $k(k-1)$ with $a \neq b$. By Lemma 3 and independence,

$$\mathbb{E}[\|\boldsymbol{u}_a\|^2 \, \boldsymbol{u}_a \boldsymbol{u}_a^\top] = (d+2)\,\boldsymbol{I}_d,$$
$$\mathbb{E}[\langle \boldsymbol{u}_a, \boldsymbol{u}_b \rangle \, \boldsymbol{u}_a \boldsymbol{u}_b^\top] = \sum_p \mathbb{E}[u_{a,p} \boldsymbol{u}_a] \, \mathbb{E}[u_{b,p} \boldsymbol{u}_b^\top] = \sum_p \boldsymbol{e}_p \boldsymbol{e}_p^\top = \boldsymbol{I}_d.$$

Summing, $\mathbb{E}[\boldsymbol{W}^2] = k(d+2)\boldsymbol{I}_d + k(k-1)\boldsymbol{I}_d = k(k+d+1)\boldsymbol{I}_d$. ∎

Lemma 7 *Under Assumption 2, $\mathbb{E}[\boldsymbol{W}^3] = \big[k^3 + 3(d+1)k^2 + (d^2+3d+4)k\big]\boldsymbol{I}_d$, hence $\alpha_3(d,k)$ as in (10).*

Proof. By Lemma 2, $\alpha_3 = \frac{1}{d}\mathbb{E}[\mathrm{tr}(\boldsymbol{W}^3)]$, and with $\boldsymbol{W} = \sum_a \boldsymbol{u}_a \boldsymbol{u}_a^\top$ and cyclicity of the trace,

$$\mathrm{tr}(\boldsymbol{W}^3) = \sum_{a,b,c} \langle \boldsymbol{u}_a, \boldsymbol{u}_b \rangle \, \langle \boldsymbol{u}_b, \boldsymbol{u}_c \rangle \, \langle \boldsymbol{u}_c, \boldsymbol{u}_a \rangle.$$

Partition $(a,b,c) \in \{1,\dots,k\}^3$ by multiplicity pattern:

$$\begin{aligned}
\#\{a=b=c\} = k: &\quad \mathbb{E}[\|\boldsymbol{u}_a\|^6] = d(d+2)(d+4),\\
\#\{|\{a,b,c\}| = 2\} = 3k(k-1): &\quad \mathbb{E}[\|\boldsymbol{u}_a\|^2 \langle \boldsymbol{u}_a, \boldsymbol{u}_c \rangle^2] = \mathbb{E}[\|\boldsymbol{u}_a\|^4] = d(d+2),\\
\#\{a,b,c \text{ distinct}\} = k(k-1)(k-2): &\quad \textstyle\sum_{p,q,r} \delta_{pr}\delta_{pq}\delta_{qr} = d,
\end{aligned}$$

where for the middle pattern, e.g. $a = b \neq c$ gives $\|\boldsymbol{u}_a\|^2 \langle \boldsymbol{u}_a, \boldsymbol{u}_c \rangle^2$ and, conditioning on $\boldsymbol{u}_a$, $\langle \boldsymbol{u}_a, \boldsymbol{u}_c \rangle \mid \boldsymbol{u}_a \sim \mathcal{N}(0, \|\boldsymbol{u}_a\|^2)$; the other two cases are symmetric. Hence

$$d\,\alpha_3 = k\,d(d+2)(d+4) + 3k(k-1)\,d(d+2) + k(k-1)(k-2)\,d,$$
$$\alpha_3 = k(d+2)(d+4) + 3k(k-1)(d+2) + k(k-1)(k-2) = k^3 + 3(d+1)k^2 + (d^2+3d+4)k.$$

∎

Proof of Theorem 1. Take expectations in (16). Since $\boldsymbol{M}^\star$ is deterministic, Lemma 2 gives

$$\mathbb{E}[\boldsymbol{W}^3] = \alpha_3 \boldsymbol{I}_d,$$
$$\mathbb{E}[\boldsymbol{W}^2 \boldsymbol{M}^\star] = \mathbb{E}[\boldsymbol{M}^\star \boldsymbol{W}^2] = \alpha_2 \boldsymbol{M}^\star,$$
$$\mathbb{E}[\boldsymbol{M}^\star \boldsymbol{W} \boldsymbol{M}^\star] = \boldsymbol{M}^\star \, \mathbb{E}[\boldsymbol{W}] \, \boldsymbol{M}^\star = k\,(\boldsymbol{M}^\star)^2,$$

so summing gives (7); the values of $\alpha_1, \alpha_2, \alpha_3$ are Lemmas 5 to 7. Corollary 1 follows by reading off entries: with $d, \boldsymbol{M}^\star$ fixed, $\alpha_3 = \Theta(k^3)$, $\alpha_2 = \Theta(k^2)$, $k(\boldsymbol{M}^\star)^2_\bullet = \Theta(k)$. ∎

Remark 1 *This computation answers the question raised by Ma et al. [19] of how gradient orthogonalization meets the curvature of training: the $\Theta(k)$ gap between the diagonal and off-diagonal scales is the regime in which orthogonalization, which suppresses the isotropic $\Theta(k^3)\boldsymbol{I}_d$ part, acts as a real preconditioner relative to the target-aligned signal $\boldsymbol{M}^\star$.*

Appendix B. Deep Linear Networks

B.1. Setup

Let $L \geq 1$ and $d_0, \ldots, d_L$ be positive integers. The depth-L linear network

$$f(\boldsymbol{x}) = \boldsymbol{W}_L \boldsymbol{W}_{L-1} \cdots \boldsymbol{W}_1 \boldsymbol{x}, \qquad \boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}, \tag{17}$$

learns a deterministic linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_L \times d_0}$ under the population squared loss

$$\mathcal{L}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L) \;:=\; \tfrac{1}{2}\, \mathbb{E}_{\boldsymbol{x}} \left\| \boldsymbol{W}_L \cdots \boldsymbol{W}_1 \boldsymbol{x} - \boldsymbol{\Phi} \boldsymbol{x} \right\|^2 . \tag{18}$$

Define the layer-i forward, backward and residual factors

$$\boldsymbol{B}_i := \boldsymbol{W}_{i-1} \cdots \boldsymbol{W}_1 \in \mathbb{R}^{d_{i-1} \times d_0}, \qquad \boldsymbol{A}_i := \boldsymbol{W}_L \cdots \boldsymbol{W}_{i+1} \in \mathbb{R}^{d_L \times d_i}, \qquad \boldsymbol{R} := \boldsymbol{W}_L \cdots \boldsymbol{W}_1 - \boldsymbol{\Phi}, \tag{19}$$

with $\boldsymbol{B}_1 := \boldsymbol{I}_{d_0}$ and $\boldsymbol{A}_L := \boldsymbol{I}_{d_L}$. The layer-$i$ gradient is

$$\boldsymbol{G}_i \;:=\; \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_i} \;=\; \boldsymbol{A}_i^\top \boldsymbol{R}\, \boldsymbol{B}_i^\top \;\in\; \mathbb{R}^{d_i \times d_{i-1}}. \tag{20}$$

The closed form below uses two pairs of dimension-only scalar sequences. For the backward factors $\{\boldsymbol{B}_i\}$,

$$r_1 = 1, \qquad v_1 = d_0^2, \tag{21}$$

$$r_i = v_{i-1} + (d_{i-1}+1) d_{i-2} r_{i-1}, \qquad v_i = d_{i-1}^2 v_{i-1} + 2 d_{i-1} d_{i-2} r_{i-1} \quad (i \geq 2); \tag{22}$$

for the forward factors $\{\boldsymbol{A}_i\}$,

$$s_L = 1, \qquad u_L = d_L^2, \tag{23}$$

$$s_{i-1} = u_i + (d_{i-1}+1) d_i s_i, \qquad u_{i-1} = d_{i-1}^2 u_i + 2 d_{i-1} d_i s_i \quad (i \leq L); \tag{24}$$

and the composite Frobenius coefficient is

$$V_i \;:=\; d_{i-1} r_i \;=\; \mathbb{E}\big[\big\| \boldsymbol{B}_i \boldsymbol{B}_i^\top \big\|_{\mathrm{F}}^2 \big]. \tag{25}$$

B.2. Standing assumption

Assumption 3 *The entries $(\boldsymbol{W}_i)_{ab}$ are i.i.d. $\mathcal{N}(0,1)$ across all i, a, b; $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$; and $\boldsymbol{\Phi}$ is deterministic and independent of $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$. Expectations $\mathbb{E}[\cdot]$ are over $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$ ($\boldsymbol{x}$ has been integrated out in $\mathcal{L}$).*

B.3. Main results restated

Theorem (Theorem 2, restated). *Under Assumption 3, for every layer $i \in \{1, \ldots, L\}$,*

$$\mathbb{E}[\boldsymbol{G}_i \boldsymbol{G}_i^\top] = V_i\, s_i\, \boldsymbol{I}_{d_i} + \boldsymbol{T}_i, \qquad \boldsymbol{T}_i = \begin{cases} \left(\prod_{j=1}^{L-1} d_j \right) \boldsymbol{\Phi} \boldsymbol{\Phi}^\top, & i = L, \\ \left(\prod_{j=1, j \neq i}^{L-1} d_j \right) \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\, \boldsymbol{I}_{d_i}, & 1 \leq i < L, \end{cases} \tag{26}$$

with V_i, s_i from (21)–(25).

Corollary 2 *Under Assumption 3, the diagonal entries of* $\mathbb{E}[\boldsymbol{G}_i\boldsymbol{G}_i^\top]$ *are*

$$\mathbb{E}\big[(\boldsymbol{G}_i\boldsymbol{G}_i^\top)_{aa}\big] = \begin{cases} V_i s_i + \big(\prod_{j\neq i,\,1\le j\le L-1} d_j\big)\,\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\,, & 1 \le i < L, \\ V_L s_L + \big(\prod_{j=1}^{L-1} d_j\big)(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{aa}, & i = L, \end{cases} \tag{27}$$

and, for $a \neq b$*, the off-diagonal entries are* $\mathbb{E}[(\boldsymbol{G}_i\boldsymbol{G}_i^\top)_{ab}] = 0$ *if* $1 \le i < L$ *and* $\big(\prod_{j=1}^{L-1} d_j\big)(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ab}$ *if* $i = L$*. In particular* $\mathbb{E}[\boldsymbol{G}_i\boldsymbol{G}_i^\top]$ *is a scalar multiple of* $\boldsymbol{I}_{d_i}$ *for every* $i < L$*, and unrolling (21)–(24) gives* $V_i s_i \asymp d_{i-1}\big(\prod_{j=0}^{i-2} d_j\big)^2\big(\prod_{j=i+1}^{L} d_j\big)^2$.

B.4. Preliminary lemmas

Lemma 8 *Same as Lemma 1.*

Lemma 9 (Wick/Isserlis) *Same as Lemma 3.*

Lemma 10 *Let* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ *have i.i.d.* $\mathcal{N}(0,1)$ *entries and let* $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, $\boldsymbol{N} \in \mathbb{R}^{m\times m}$ *be deterministic symmetric. Then*

$$\mathbb{E}\big[\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top\big] = \big(\operatorname{tr}(\boldsymbol{M})^2 + (m+1)\,\|\boldsymbol{M}\|_{\mathrm{F}}^2\big)\boldsymbol{I}_m, \tag{28}$$

$$\mathbb{E}\big[\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{X}\big] = \big(\operatorname{tr}(\boldsymbol{N})^2 + (n+1)\,\|\boldsymbol{N}\|_{\mathrm{F}}^2\big)\boldsymbol{I}_n. \tag{29}$$

Proof. We prove (28); (29) follows by applying it to $\boldsymbol{X}^\top$. For $a, b, c, d \in \{1, \dots, m\}$, Lemma 9 gives

$$\mathbb{E}[\boldsymbol{X}_{ap}\boldsymbol{X}_{bq}\boldsymbol{X}_{cr}\boldsymbol{X}_{ds}] = \delta_{ab}\delta_{pq}\delta_{cd}\delta_{rs} + \delta_{ac}\delta_{pr}\delta_{bd}\delta_{qs} + \delta_{ad}\delta_{ps}\delta_{bc}\delta_{qr},$$

and contracting against $\boldsymbol{M}_{pq}\boldsymbol{M}_{rs}$ with $\boldsymbol{M} = \boldsymbol{M}^\top$,

$$\mathbb{E}[(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)_{ab}(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)_{cd}] = \delta_{ab}\delta_{cd}\operatorname{tr}(\boldsymbol{M})^2 + (\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc})\,\|\boldsymbol{M}\|_{\mathrm{F}}^2\,.$$

Summing over $b = c$,

$$\mathbb{E}[(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)_{ac}] = \sum_b \mathbb{E}[(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)_{ab}(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)_{bc}] = \delta_{ac}\big(\operatorname{tr}(\boldsymbol{M})^2 + (m+1)\,\|\boldsymbol{M}\|_{\mathrm{F}}^2\big).$$

■

Lemma 11 *Let* $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ *have i.i.d.* $\mathcal{N}(0,1)$ *entries and* $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ *be deterministic symmetric. Then*

$$\mathbb{E}\big[\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)^2\big] = m^2\operatorname{tr}(\boldsymbol{M})^2 + 2m\,\|\boldsymbol{M}\|_{\mathrm{F}}^2\,. \tag{30}$$

Proof. Expanding,

$$\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)^2 = \sum_{a,a',p,q,p',q'} \boldsymbol{X}_{ap}\boldsymbol{M}_{pq}\boldsymbol{X}_{aq}\,\boldsymbol{X}_{a'p'}\boldsymbol{M}_{p'q'}\boldsymbol{X}_{a'q'};$$

by Lemma 9 the three pair contractions of $\mathbb{E}[\boldsymbol{X}_{ap}\boldsymbol{X}_{aq}\boldsymbol{X}_{a'p'}\boldsymbol{X}_{a'q'}]$ contribute (after summing over $\boldsymbol{M}\boldsymbol{M}$) $m^2\operatorname{tr}(\boldsymbol{M})^2$, $m\,\|\boldsymbol{M}\|_{\mathrm{F}}^2$ and $m\operatorname{tr}(\boldsymbol{M}^2) = m\,\|\boldsymbol{M}\|_{\mathrm{F}}^2$, which sum to (30). ■

B.5. Proofs

The proof of Theorem 2 has three steps: a $\boldsymbol{W}_i$-expansion; recursive moments of $\boldsymbol{B}_i\boldsymbol{B}_i^\top$ and $\boldsymbol{A}_i^\top\boldsymbol{A}_i$; and the target-dependent term.

Lemma 12 *Under Assumption 3, for every i,*

$$\mathbb{E}\big[\boldsymbol{G}_i\boldsymbol{G}_i^\top \;\big|\; \boldsymbol{A}_i, \boldsymbol{B}_i\big] = \operatorname{tr}\big((\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big)\,(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2 + \boldsymbol{A}_i^\top\boldsymbol{\Phi}\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{\Phi}^\top\boldsymbol{A}_i. \tag{31}$$

Proof. Substituting $\boldsymbol{R} = \boldsymbol{A}_i\boldsymbol{W}_i\boldsymbol{B}_i - \boldsymbol{\Phi}$ into (20),

$$\boldsymbol{G}_i\boldsymbol{G}_i^\top = (\boldsymbol{A}_i^\top\boldsymbol{A}_i)\boldsymbol{W}_i(\boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{B}_i^\top)\boldsymbol{W}_i^\top(\boldsymbol{A}_i^\top\boldsymbol{A}_i) + (\text{cross}) + \boldsymbol{A}_i^\top\boldsymbol{\Phi}\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{\Phi}^\top\boldsymbol{A}_i,$$

where (cross) collects the two terms linear in $\boldsymbol{W}_i$; since $\boldsymbol{W}_i \overset{d}{=} -\boldsymbol{W}_i$ is independent of $(\boldsymbol{A}_i, \boldsymbol{B}_i)$, $\mathbb{E}[(\text{cross}) \mid \boldsymbol{A}_i, \boldsymbol{B}_i] = \boldsymbol{0}$. By Lemma 8 with $\boldsymbol{K} = \boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{B}_i^\top$,

$$\mathbb{E}_{\boldsymbol{W}_i}\big[\boldsymbol{W}_i\boldsymbol{K}\boldsymbol{W}_i^\top \;\big|\; \boldsymbol{A}_i, \boldsymbol{B}_i\big] = \operatorname{tr}(\boldsymbol{K})\,\boldsymbol{I}_{d_i} = \operatorname{tr}\big((\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big)\boldsymbol{I}_{d_i},$$

which yields (31). ∎

Lemma 13 *Under Assumption 3, for every $i \in \{1, \dots, L\}$,*

$$\mathbb{E}\big[(\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big] = r_i\,\boldsymbol{I}_{d_{i-1}}, \qquad \mathbb{E}\big[\operatorname{tr}(\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big] = v_i, \tag{32}$$

with (r_i, v_i) from (21)–(22).

Proof. At $i = 1$, $\boldsymbol{B}_1 = \boldsymbol{I}_{d_0}$, so $(\boldsymbol{B}_1\boldsymbol{B}_1^\top)^2 = \boldsymbol{I}_{d_0}$ and $\operatorname{tr}(\boldsymbol{B}_1\boldsymbol{B}_1^\top)^2 = d_0^2$, matching $r_1 = 1$, $v_1 = d_0^2$. For $i \geq 2$, $\boldsymbol{B}_i = \boldsymbol{W}_{i-1}\boldsymbol{B}_{i-1}$ with $\boldsymbol{W}_{i-1} \in \mathbb{R}^{d_{i-1}\times d_{i-2}}$ independent of $\boldsymbol{B}_{i-1}$, so $\boldsymbol{B}_i\boldsymbol{B}_i^\top = \boldsymbol{W}_{i-1}\boldsymbol{M}\boldsymbol{W}_{i-1}^\top$ with $\boldsymbol{M} := \boldsymbol{B}_{i-1}\boldsymbol{B}_{i-1}^\top$. By Lemma 10,

$$\mathbb{E}\big[(\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2 \;\big|\; \boldsymbol{B}_{i-1}\big] = \big(\operatorname{tr}(\boldsymbol{M})^2 + (d_{i-1}+1)\,\|\boldsymbol{M}\|_{\mathrm{F}}^2\big)\boldsymbol{I}_{d_{i-1}},$$

and taking $\mathbb{E}$ over $\boldsymbol{B}_{i-1}$ with $\mathbb{E}[\operatorname{tr}(\boldsymbol{M}^2)] = d_{i-2}r_{i-1}$ gives $r_i = v_{i-1} + (d_{i-1}+1)d_{i-2}r_{i-1}$. Likewise, by Lemma 11,

$$\mathbb{E}\big[\operatorname{tr}(\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2 \;\big|\; \boldsymbol{B}_{i-1}\big] = d_{i-1}^2\operatorname{tr}(\boldsymbol{M})^2 + 2d_{i-1}\,\|\boldsymbol{M}\|_{\mathrm{F}}^2,$$

so $v_i = d_{i-1}^2 v_{i-1} + 2d_{i-1}d_{i-2}r_{i-1}$. ∎

Lemma 14 *Under Assumption 3, for every $i \in \{0, \dots, L\}$,*

$$\mathbb{E}\big[(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2\big] = s_i\,\boldsymbol{I}_{d_i}, \qquad \mathbb{E}\big[\operatorname{tr}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2\big] = u_i, \tag{33}$$

with (s_i, u_i) from (23)–(24).

Proof. At $i = L$, $\boldsymbol{A}_L = \boldsymbol{I}_{d_L}$, matching $s_L = 1$, $u_L = d_L^2$. For $i < L$, $\boldsymbol{A}_i = \boldsymbol{A}_{i+1}\boldsymbol{W}_{i+1}$ with $\boldsymbol{W}_{i+1} \in \mathbb{R}^{d_{i+1}\times d_i}$ independent of $\boldsymbol{A}_{i+1}$, so $\boldsymbol{A}_i^\top\boldsymbol{A}_i = \boldsymbol{W}_{i+1}^\top\boldsymbol{N}\boldsymbol{W}_{i+1}$ with $\boldsymbol{N} := \boldsymbol{A}_{i+1}^\top\boldsymbol{A}_{i+1}$. By Lemma 10 (right identity),

$$\mathbb{E}\big[(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2 \;\big|\; \boldsymbol{A}_{i+1}\big] = \big(\operatorname{tr}(\boldsymbol{N})^2 + (d_i+1)\,\|\boldsymbol{N}\|_{\mathrm{F}}^2\big)\boldsymbol{I}_{d_i},$$

so $s_i = u_{i+1} + (d_i+1)d_{i+1}s_{i+1}$; the recursion for u_i is that of v_i with $d_{i-1} \leftrightarrow d_i$ and $d_{i-2} \leftrightarrow d_{i+1}$, giving $u_{i-1} = d_{i-1}^2 u_i + 2d_{i-1}d_i s_i$. ∎

Lemma 15 *Under Assumption 3,* $\mathbb{E}[\boldsymbol{A}_i^\top \boldsymbol{\Phi} \boldsymbol{B}_i^\top \boldsymbol{B}_i \boldsymbol{\Phi}^\top \boldsymbol{A}_i] = \boldsymbol{T}_i$*, with* $\boldsymbol{T}_i$ *as in* (26)*.*

Proof. By independence of $\boldsymbol{A}_i, \boldsymbol{B}_i$ and Lemma 8 iterated through $\boldsymbol{B}_i = \boldsymbol{W}_{i-1} \cdots \boldsymbol{W}_1$,

$$\mathbb{E}[\boldsymbol{B}_i^\top \boldsymbol{B}_i] = \Big(\prod_{j=1}^{i-1} d_j\Big) \boldsymbol{I}_{d_0} \qquad \Longrightarrow \qquad \mathbb{E}[\boldsymbol{A}_i^\top \boldsymbol{\Phi} \boldsymbol{B}_i^\top \boldsymbol{B}_i \boldsymbol{\Phi}^\top \boldsymbol{A}_i] = \Big(\prod_{j=1}^{i-1} d_j\Big) \mathbb{E}[\boldsymbol{A}_i^\top \boldsymbol{\Phi} \boldsymbol{\Phi}^\top \boldsymbol{A}_i]$$

(the empty product is 1 when $i = 1$). Let $h_i(\boldsymbol{M}) := \mathbb{E}[\boldsymbol{A}_i^\top \boldsymbol{M} \boldsymbol{A}_i]$ for symmetric $\boldsymbol{M} \in \mathbb{R}^{d_L \times d_L}$. Then $h_L(\boldsymbol{M}) = \boldsymbol{M}$, and for $i < L$, by Lemma 8,

$$h_i(\boldsymbol{M}) = \mathbb{E}\big[\operatorname{tr}(\boldsymbol{A}_{i+1}^\top \boldsymbol{M} \boldsymbol{A}_{i+1})\big] \boldsymbol{I}_{d_i} = \operatorname{tr}\big(h_{i+1}(\boldsymbol{M})\big) \boldsymbol{I}_{d_i}.$$

Iterating,

$$h_i(\boldsymbol{M}) = \begin{cases} \boldsymbol{M}, & i = L, \\ \big(\prod_{j=i+1}^{L-1} d_j\big) \operatorname{tr}(\boldsymbol{M})\, \boldsymbol{I}_{d_i}, & i < L, \end{cases}$$

with $\prod_{j=L}^{L-1} d_j = 1$. Taking $\boldsymbol{M} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ (so $\operatorname{tr}(\boldsymbol{M}) = \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2$) and multiplying by $\prod_{j=1}^{i-1} d_j$ gives $\boldsymbol{T}_i$. ∎

Proof of Theorem 2. By Lemma 12,

$$\mathbb{E}[\boldsymbol{G}_i \boldsymbol{G}_i^\top] = \mathbb{E}\big[\operatorname{tr}((\boldsymbol{B}_i \boldsymbol{B}_i^\top)^2)\, (\boldsymbol{A}_i^\top \boldsymbol{A}_i)^2\big] + \mathbb{E}\big[\boldsymbol{A}_i^\top \boldsymbol{\Phi} \boldsymbol{B}_i^\top \boldsymbol{B}_i \boldsymbol{\Phi}^\top \boldsymbol{A}_i\big].$$

Since $\boldsymbol{A}_i, \boldsymbol{B}_i$ depend on disjoint sets of $\boldsymbol{W}_j$, they are independent, so by Lemmas 13 and 14,

$$\mathbb{E}\big[\operatorname{tr}((\boldsymbol{B}_i \boldsymbol{B}_i^\top)^2)\, (\boldsymbol{A}_i^\top \boldsymbol{A}_i)^2\big] = \mathbb{E}\big[\operatorname{tr}((\boldsymbol{B}_i \boldsymbol{B}_i^\top)^2)\big]\, \mathbb{E}\big[(\boldsymbol{A}_i^\top \boldsymbol{A}_i)^2\big] = d_{i-1} r_i \cdot s_i\, \boldsymbol{I}_{d_i} = V_i s_i\, \boldsymbol{I}_{d_i},$$

and the second term equals $\boldsymbol{T}_i$ by Lemma 15. Corollary 2 follows by reading off entries of (26): for $i < L$, $\boldsymbol{T}_i$ is a multiple of $\boldsymbol{I}_{d_i}$ so the off-diagonal vanishes; for $i = L$, $\boldsymbol{T}_L$ contributes $(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ab}$ entry-wise. The leading order of $V_i s_i$ follows by unrolling (21)–(24). ∎

Appendix C. Two-Layer ReLU Networks

C.1. Setup

Let d_0, d_1, d_2 be positive integers. The two-layer ReLU network

$$f(x) = W_2\, \sigma(W_1 x), \qquad W_1 \in \mathbb{R}^{d_1 \times d_0}, \quad W_2 \in \mathbb{R}^{d_2 \times d_1}, \tag{34}$$

acts on $x \in \mathbb{R}^{d_0}$ via $\sigma(z) = \max(z, 0)$, $\sigma'(z) = \mathbf{1}[z > 0]$. Write $h(x) := \sigma(W_1 x) \in \mathbb{R}^{d_1}$ and $w_a \in \mathbb{R}^{d_0}$ for the a-th row of W_1. The population loss against a deterministic linear target $\Phi \in \mathbb{R}^{d_2 \times d_0}$ is

$$\mathcal{L}(W_1, W_2) \;:=\; \tfrac{1}{2}\, \mathbb{E}_x \left\| W_2 \sigma(W_1 x) - \Phi x \right\|^2, \tag{35}$$

and $G_i := \partial \mathcal{L} / \partial W_i$, $i \in \{1, 2\}$. For independent $u, v \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$ let $\theta := \angle(u, v) = \arccos(\langle u, v \rangle / (\|u\|\, \|v\|)) \in [0, \pi]$, and define

$$\kappa_1 := \mathbb{E}\left[\frac{\sin^2\theta + (\pi - \theta)^2}{4\pi^2} + \frac{\sin\theta \cos\theta (\pi - \theta)}{2\pi^2} \right], \tag{36}$$

$$\kappa_2 := \frac{\mathbb{E}[\sin\theta (\pi - \theta)]}{2\pi^2}. \tag{37}$$

C.2. Standing assumption

Assumption 4 *$(W_1)_{ij}$ and $(W_2)_{ij}$ are i.i.d. $\mathcal{N}(0,1)$; $x \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$; Φ is deterministic and independent of W_1, W_2. Expectations $\mathbb{E}[\cdot]$ are over W_1, W_2 (x has been integrated out in $\mathcal{L}$).*

C.3. Main results restated

Theorem (Theorem 3, restated). *Under Assumption 4, with $H := \mathbb{E}_x[h(x)h(x)^\top]$ and $\alpha := \mathbb{E}_{W_1}[\mathrm{tr}(H^2)]$,*

$$\mathbb{E}[G_2 G_2^\top] = \alpha\, \boldsymbol{I}_{d_2} + \tfrac{d_1}{4}\, \Phi\Phi^\top, \qquad \alpha = \Theta(d_0^2 d_1^2) \text{ as } d_0, d_1 \to \infty, \tag{38}$$

and $\mathbb{E}[G_1 G_1^\top]$ is permutation-invariant in the hidden index with, as $d_0, d_2 \to \infty$,

$$\mathbb{E}\big[(G_1 G_1^\top)_{aa}\big] = \tfrac{d_0 d_2^2}{4} + d_0 d_2 \big[(d_1 - 1)\kappa_1 + \tfrac{1}{2}\big] + \tfrac{\|\Phi\|_{\mathrm{F}}^2}{4}, \tag{39}$$

$$\mathbb{E}\big[(G_1 G_1^\top)_{ac}\big] = d_0 d_2\, \kappa_2 + O(d_2 \sqrt{d_0}) \qquad (a \neq c). \tag{40}$$

As $d_0 \to \infty$, $\kappa_1 \to \frac{1}{4\pi^2} + \frac{1}{16}$ and $\kappa_2 \to \frac{1}{4\pi}$.

C.4. Preliminary lemmas

Lemma 16 *Same as Lemma 1.*

Lemma 17 (Gaussian integration by parts) *Let $x \sim \mathcal{N}(0, \boldsymbol{I}_n)$ and $f : \mathbb{R}^n \to \mathbb{R}$ be locally absolutely continuous along almost every line parallel to the j-th axis, with weak derivative $\partial_j f$, and $\mathbb{E}|x_j f(x)| < \infty$, $\mathbb{E}|\partial_j f(x)| < \infty$. Then*

$$\mathbb{E}[x_j f(x)] = \mathbb{E}[\partial_j f(x)]. \tag{41}$$

In particular this holds whenever f is Lipschitz.

Proof. A standard identity [20, 25]. The standard Gaussian density φ satisfies $\partial_j \varphi(x) = -x_j \varphi(x)$, so by Fubini and one-dimensional integration by parts in x_j (the boundary term vanishes by integrability),

$$\mathbb{E}[x_j f(x)] = -\int f\, \partial_j \varphi = \int \partial_j f\, \varphi = \mathbb{E}[\partial_j f(x)].$$

The Lipschitz case follows from Rademacher's theorem. ∎

Remark 2 *$z \mapsto \sigma(z)$ is 1-Lipschitz, so $f(x) = \sigma(w^\top x)$ falls under Lemma 17 with $\partial_j f(x) = \mathbf{1}[w^\top x > 0]\, w_j$ a.e. By contrast $g(x) = \mathbf{1}[w^\top x > 0] h(x)$ with h not vanishing on $\{w^\top x = 0\}$ is discontinuous across that hyperplane and is* not *covered by Lemma 17; for such integrands we use a half-space argument (Lemma 21) or differentiate a separately established Lipschitz identity (Lemma 19).*

Lemma 18 *For $Z \sim \mathcal{N}(0, s^2)$ with $s > 0$,*

$$\mathbb{E}[\sigma(Z)] = \frac{s}{\sqrt{2\pi}}, \qquad \mathbb{E}[\sigma(Z)^2] = \frac{s^2}{2}, \qquad \mathbb{E}[\sigma'(Z)] = \frac{1}{2}. \tag{42}$$

Proof. Direct computation: $\mathbb{E}[\sigma(Z)] = \int_0^\infty z\varphi_s(z)\,dz = s/\sqrt{2\pi}$; $\mathbb{E}[\sigma(Z)^2] = \int_0^\infty z^2\varphi_s(z)\,dz = s^2/2$; $\mathbb{E}[\sigma'(Z)] = \Pr(Z > 0) = 1/2$. ∎

Lemma 19 (Cho–Saul kernel) *Let $u, v \in \mathbb{R}^n \setminus \{0\}$, $x \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $\theta := \angle(u, v)$. Then [5]*

$$\mathbb{E}[\sigma(u^\top x)\sigma(v^\top x)] = \frac{\|u\|\,\|v\|}{2\pi}\big(\sin\theta + (\pi - \theta)\cos\theta\big), \tag{43}$$

$$\mathbb{E}[\sigma'(u^\top x)\sigma'(v^\top x)] = \frac{\pi - \theta}{2\pi}. \tag{44}$$

Proof. Both quantities depend on $(u^\top x, v^\top x)$, bivariate Gaussian with variances $\|u\|^2, \|v\|^2$ and correlation $\cos\theta$. Equation (44) is $\Pr(u^\top x > 0,\, v^\top x > 0)$, the angular measure of the cone $\{Z_1 > 0, Z_2 > 0\}$ for a standard bivariate Gaussian with correlation $\cos\theta$, namely $(\pi - \theta)/(2\pi)$. For (43), by Price's theorem [22],

$$\frac{\partial}{\partial\rho}\mathbb{E}[\sigma(u^\top x)\sigma(v^\top x)] = \|u\|\,\|v\|\;\mathbb{E}[\sigma'(u^\top x)\sigma'(v^\top x)] = \frac{\|u\|\,\|v\|\,(\pi - \theta)}{2\pi};$$

at $\rho = -1$ (i.e. $\theta = \pi$) the product vanishes a.s., and integrating ∂_ρ from -1 to $\cos\theta$ gives (43). ∎

Lemma 20 (Wick/Isserlis) *Same as Lemma 3.*

Lemma 21 *Let $x \sim \mathcal{N}(0, \boldsymbol{I}_n)$ and $u \in \mathbb{R}^n \setminus \{0\}$. Then*

$$\mathbb{E}\big[xx^\top\,\mathbf{1}[u^\top x > 0]\big] = \tfrac{1}{2}\,\boldsymbol{I}_n. \tag{45}$$

Proof. Write $A := \mathbb{E}[xx^\top\mathbf{1}[u^\top x > 0]]$ and $x = (u^\top x/\|u\|^2)u + x_\perp$ with $x_\perp$ the projection onto $u^\perp$. In an orthonormal basis containing $u/\|u\|$, the off-diagonal entries of A vanish by the reflection symmetry of $x_\perp$; the entry along $u/\|u\|$ is $\mathbb{E}[(u^\top x/\|u\|)^2\mathbf{1}[u^\top x > 0]] = \frac{1}{2}$ ($u^\top x$ symmetric), and each remaining diagonal entry is $\frac{1}{2}\mathbb{E}[x_j^2] = \frac{1}{2}$ by independence of $x_\perp$ from $u^\top x$ and $\Pr(u^\top x > 0) = \frac{1}{2}$. Hence $A = \frac{1}{2}\boldsymbol{I}_n$. ∎

C.5. Proof of the gradient self outer-product of W_2

Lemma 22 *$G_2 = W_2 H - \Phi C^\top$, where $H := \mathbb{E}_x[h(x)h(x)^\top] \in \mathbb{R}^{d_1\times d_1}$ and $C := \mathbb{E}_x[h(x)x^\top] \in \mathbb{R}^{d_1\times d_0}$.*

Proof. Differentiating (35) in W_2,

$$G_2 = \mathbb{E}_x[(W_2 h(x) - \Phi x)h(x)^\top] = W_2\,\mathbb{E}_x[h(x)h(x)^\top] - \Phi\,\mathbb{E}_x[xh(x)^\top] = W_2 H - \Phi C^\top.$$

∎

Lemma 23 $C = \frac{1}{2}W_1$.

Proof. $C_{aj} = \mathbb{E}_x[\sigma(w_a^\top x)x_j]$, and $f(x) := \sigma(w_a^\top x)$ is Lipschitz with $\partial_j f(x) = \mathbf{1}[w_a^\top x > 0](w_a)_j$ a.e. By Lemma 17 and Lemma 18,

$$C_{aj} = \mathbb{E}_x[\partial_j f(x)] = (w_a)_j\,\mathbb{E}_x[\sigma'(w_a^\top x)] = \tfrac{1}{2}(w_a)_j.$$

■

Proof of (38). By Lemma 22,

$$G_2 G_2^\top = W_2 H^2 W_2^\top - W_2 H C \Phi^\top - \Phi C^\top H W_2^\top + \Phi C^\top C \Phi^\top.$$

Since $\mathbb{E}[W_2] = 0$, the cross terms vanish under $\mathbb{E}_{W_2}[\cdot \mid W_1]$, and by Lemma 16 (left identity, $\boldsymbol{M} = H^2$),

$$\mathbb{E}[G_2 G_2^\top \mid W_1] = \operatorname{tr}(H^2)\, \boldsymbol{I}_{d_2} + \Phi C^\top C \Phi^\top. \tag{46}$$

By Lemma 23, $C^\top C = \frac{1}{4} W_1^\top W_1$, and Lemma 16 (right identity, $\boldsymbol{N} = \boldsymbol{I}_{d_1}$) gives $\mathbb{E}_{W_1}[W_1^\top W_1] = d_1 \boldsymbol{I}_{d_0}$, so $\mathbb{E}_{W_1}[C^\top C] = \frac{d_1}{4} \boldsymbol{I}_{d_0}$. Taking $\mathbb{E}_{W_1}$ of (46),

$$\mathbb{E}[G_2 G_2^\top] = \mathbb{E}_{W_1}[\operatorname{tr}(H^2)]\, \boldsymbol{I}_{d_2} + \tfrac{d_1}{4} \Phi \Phi^\top = \alpha\, \boldsymbol{I}_{d_2} + \tfrac{d_1}{4} \Phi \Phi^\top.$$

For the order of α, write $\operatorname{tr}(H^2) = \sum_a H_{aa}^2 + \sum_{a \neq b} H_{ab}^2$. By Lemma 18, $H_{aa} = \frac{1}{2} \|w_a\|^2$ with $\|w_a\|^2 \sim \chi^2_{d_0}$, so

$$\sum_a \mathbb{E}[H_{aa}^2] = \frac{d_1 d_0 (d_0 + 2)}{4} = \Theta(d_0^2 d_1).$$

By (43), $H_{ab} = \|w_a\| \, \|w_b\| \, g(\theta_{ab})$ with $g(\theta) = (\sin\theta + (\pi - \theta)\cos\theta)/(2\pi)$ and $\theta_{ab} = \angle(w_a, w_b)$ independent of $\|w_a\|, \|w_b\|$, so $\mathbb{E}[H_{ab}^2] = d_0^2\, \mathbb{E}[g(\theta)^2]$; since $\mathbb{E}[g(\theta)^2] \to g(\pi/2)^2 = 1/(4\pi^2) > 0$,

$$\sum_{a \neq b} \mathbb{E}[H_{ab}^2] = d_1 (d_1 - 1)\, d_0^2\, \mathbb{E}[g(\theta)^2] = \Theta(d_0^2 d_1^2).$$

The off-diagonal term dominates, so $\alpha = \Theta(d_0^2 d_1^2)$. ■

C.6. Proof of the gradient self outer-product of W_1

By the chain rule, with residual $r(x) := W_2 \sigma(W_1 x) - \Phi x$ and $D(x) := \operatorname{diag}(\sigma'(W_1 x))$,

$$G_1 = G_1^{(1)} - G_1^{(2)}, \qquad G_1^{(1)} := \mathbb{E}_x[D(x) W_2^\top W_2 \sigma(W_1 x) x^\top], \qquad G_1^{(2)} := \mathbb{E}_x[D(x) W_2^\top \Phi x x^\top]. \tag{47}$$

$G_1^{(1)}$ is quadratic in W_2 and $G_1^{(2)}$ is linear, so the cross terms in $G_1 G_1^\top$ are odd in W_2 and vanish under $\mathbb{E}_{W_2}[\cdot \mid W_1]$:

$$\mathbb{E}_{W_2}[G_1 G_1^\top \mid W_1] = \mathbb{E}_{W_2}[G_1^{(1)} (G_1^{(1)})^\top \mid W_1] + \mathbb{E}_{W_2}[G_1^{(2)} (G_1^{(2)})^\top \mid W_1]. \tag{48}$$

Lemma 24 *$G_1^{(2)} = \frac{1}{2} W_2^\top \Phi$, and $\mathbb{E}_{W_2}[G_1^{(2)} (G_1^{(2)})^\top \mid W_1] = \frac{\|\Phi\|_{\mathrm{F}}^2}{4} \boldsymbol{I}_{d_1}$.*

Proof. Let $v_a := \Phi^\top (W_2)_{:,a} \in \mathbb{R}^{d_0}$, so $(W_2^\top \Phi)_{ab} = (v_a)_b$. Then, by Lemma 21,

$$(G_1^{(2)})_{ab} = \mathbb{E}_x[\mathbf{1}[w_a^\top x > 0]\, x_b\, (x^\top v_a)] = \big(\mathbb{E}_x[\mathbf{1}[w_a^\top x > 0]\, x x^\top]\, v_a\big)_b = \tfrac{1}{2} (v_a)_b,$$

so $G_1^{(2)} = \frac{1}{2} W_2^\top \Phi$ and $G_1^{(2)} (G_1^{(2)})^\top = \frac{1}{4} W_2^\top \Phi \Phi^\top W_2$; Lemma 16 (right identity, $\boldsymbol{N} = \Phi \Phi^\top$) gives $\mathbb{E}_{W_2}[W_2^\top \Phi \Phi^\top W_2] = \|\Phi\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_1}$. ■

Lemma 25 *Define* $F_b(u,v) := \mathbb{E}_x[\sigma'(u^\top x)\sigma(v^\top x)x_b]$. *Then* $F_b(u,u) = \frac{1}{2}u_b$, *and for* $u \notin \operatorname{span}(v)$,

$$F_b(u,v) = u_b \cdot \frac{\|v\| \sin\theta}{2\pi \|u\|} + v_b \cdot \frac{\pi - \theta}{2\pi}, \qquad \theta := \angle(u,v). \tag{49}$$

Proof. For $u = v$, $\sigma'(u^\top x)\sigma(u^\top x) = \sigma(u^\top x)$ a.e., so by Lemma 17 (applied to the Lipschitz $x \mapsto \sigma(u^\top x)$) and Lemma 18, $F_b(u,u) = \mathbb{E}_x[\sigma(u^\top x)x_b] = \mathbb{E}_x[\sigma'(u^\top x)]u_b = \frac{1}{2}u_b$. For $u \neq v$, $f(x) = \sigma'(u^\top x)\sigma(v^\top x)$ is discontinuous across $\{u^\top x = 0\}$, so we differentiate the Cho–Saul kernel instead. With $K(u,v) := \mathbb{E}_x[\sigma(u^\top x)\sigma(v^\top x)]$, dominated convergence (since σ is Lipschitz with linear growth) gives

$$\frac{\partial K}{\partial u_b}(u,v) = \mathbb{E}_x[\sigma'(u^\top x)x_b\sigma(v^\top x)] = F_b(u,v).$$

By Lemma 19, $K(u,v) = \frac{\|u\|\|v\|}{2\pi}(\sin\theta + (\pi - \theta)\cos\theta)$, smooth for $u \notin \operatorname{span}(v)$; differentiating with $\partial \|u\| / \partial u_b = u_b / \|u\|$, $\partial \cos\theta / \partial u_b = \frac{1}{\|u\|}(v_b / \|v\| - \cos\theta\, u_b / \|u\|)$, $\partial\theta/\partial u_b = -\frac{1}{\sin\theta}\partial\cos\theta/\partial u_b$, and simplifying yields (49). ∎

Lemma 26 $\mathbb{E}_{W_2}[(W_2^\top W_2)_{ak}(W_2^\top W_2)_{cl}] = d_2^2\,\delta_{ak}\delta_{cl} + d_2\,\delta_{ac}\delta_{kl} + d_2\,\delta_{al}\delta_{ck}$.

Proof. $(W_2^\top W_2)_{ak} = \sum_{i=1}^{d_2}(W_2)_{ia}(W_2)_{ik}$, and applying Lemma 20 to the four Gaussian factors $(W_2)_{ia}, (W_2)_{ik}, (W_2)_{jc}, (W_2)_{jl}$ (summed over i, j) the three pairings give $\delta_{ak}\delta_{cl}d_2^2 + \delta_{ac}\delta_{kl}d_2 + \delta_{al}\delta_{ck}d_2$. ∎

Proof of (39)–(40). Permutation invariance follows from the exchangeability of the rows of W_1 and columns of W_2. The $G_1^{(2)}$ contribution to (48) is $\frac{\|\Phi\|_{\mathrm{F}}^2}{4}\boldsymbol{I}_{d_1}$ by Lemma 24. For $G_1^{(1)}$, write $(G_1^{(1)})_{ab} = \sum_{k=1}^{d_1}(W_2^\top W_2)_{ak}F_b(w_a, w_k)$, so

$$\mathbb{E}_{W_2}\left[(G_1^{(1)}(G_1^{(1)})^\top)_{ac} \;\middle|\; W_1\right] = \sum_{b,k,l} \mathbb{E}_{W_2}[(W_2^\top W_2)_{ak}(W_2^\top W_2)_{cl}]\, F_b(w_a, w_k)F_b(w_c, w_l),$$

and substituting Lemma 26 gives $T_1^{(ac)} + T_2^{(ac)} + T_3^{(ac)}$ with

$$T_1^{(ac)} = d_2^2 \textstyle\sum_b F_b(w_a,w_a)F_b(w_c,w_c), \tag{50}$$
$$T_2^{(ac)} = d_2\,\delta_{ac} \textstyle\sum_{k=1}^{d_1}\sum_b F_b(w_a,w_k)^2, \tag{51}$$
$$T_3^{(ac)} = d_2 \textstyle\sum_b F_b(w_a,w_c)F_b(w_c,w_a). \tag{52}$$

Term T_1. By Lemma 25, $F_b(w_a,w_a) = \frac{1}{2}(w_a)_b$, so $T_1^{(ac)} = \frac{d_2^2}{4}w_a^\top w_c$, giving $\mathbb{E}[T_1^{(aa)}] = \frac{d_0 d_2^2}{4}$ and $\mathbb{E}[T_1^{(ac)}] = 0$ for $a \neq c$.

Term T_2. For $u \neq v$, writing $A := \frac{\|v\|\sin\theta}{2\pi\|u\|}$, $B := \frac{\pi-\theta}{2\pi}$ in (49) and using $u^\top v = \|u\|\,\|v\|\cos\theta$,

$$\sum_b F_b(u,v)^2 = A^2\|u\|^2 + 2AB(u^\top v) + B^2\|v\|^2 = \|v\|^2\,\psi(\theta), \qquad \psi(\theta) := \frac{\sin^2\theta + (\pi-\theta)^2}{4\pi^2} + \frac{\sin\theta\cos\theta(\pi-\theta)}{2\pi^2}.$$

In $T_2^{(aa)}$ the term $k = a$ contributes $\sum_b F_b(w_a, w_a)^2 = \frac{1}{4}\|w_a\|^2$ (expectation $\frac{d_0}{4}$), and each $k \neq a$ contributes $\mathbb{E}[\|w_k\|^2]\mathbb{E}[\psi(\theta_{ak})] = d_0\kappa_1$, so

$$\mathbb{E}[T_2^{(aa)}] = d_2\Big[\tfrac{d_0}{4} + (d_1 - 1)d_0\kappa_1\Big] = d_0 d_2\big[(d_1 - 1)\kappa_1 + \tfrac{1}{4}\big],$$

and $T_2^{(ac)} = 0$ for $a \neq c$.

Term T_3. For $a = c$, $\sum_b F_b(w_a, w_a)^2 = \frac{1}{4}\|w_a\|^2$, so $\mathbb{E}[T_3^{(aa)}] = \frac{d_0 d_2}{4}$. For $a \neq c$, expanding $F_b(w_a, w_c)F_b(w_c, w_a)$ via (49) (with $\theta := \theta_{ac}$) and summing over b,

$$\sum_b F_b(w_a, w_c)F_b(w_c, w_a) = \frac{\|w_a\|\,\|w_c\|}{4\pi^2}\Big[\cos\theta\big(\sin^2\theta + (\pi - \theta)^2\big) + 2\sin\theta(\pi - \theta)\Big].$$

Here $\|w_a\|\,\|w_c\|$ is independent of θ_{ac} with $\mathbb{E}[\|w_a\|\,\|w_c\|] = d_0 + O(1)$; as $d_0 \to \infty$, θ_{ac} concentrates at $\pi/2$ with $\mathbb{E}[\cos\theta_{ac}] = 0$ and $\mathrm{Var}(\cos\theta_{ac}) = 1/d_0$, so $|\mathbb{E}[\cos\theta_{ac}\, q(\theta_{ac})]| = O(1/\sqrt{d_0})$ for bounded q, while $\mathbb{E}[\sin\theta_{ac}(\pi - \theta_{ac})] = \Theta(1)$. The $\cos\theta(\cdot)$ part contributes $O(d_2\sqrt{d_0})$ and the $2\sin\theta(\pi - \theta)$ part gives $d_0 d_2\kappa_2 + O(d_2)$, so $\mathbb{E}[T_3^{(ac)}] = d_0 d_2\kappa_2 + O(d_2\sqrt{d_0})$.

Combining. For $a = c$, adding the three terms and the $G_1^{(2)}$ contribution and merging the two $\frac{d_0 d_2}{4}$ pieces,

$$\mathbb{E}\big[(G_1 G_1^\top)_{aa}\big] = \tfrac{d_0 d_2^2}{4} + d_0 d_2\big[(d_1 - 1)\kappa_1 + \tfrac{1}{2}\big] + \tfrac{\|\Phi\|_{\mathrm{F}}^2}{4};$$

for $a \neq c$ only $T_3^{(ac)}$ survives, $\mathbb{E}[(G_1 G_1^\top)_{ac}] = d_0 d_2\kappa_2 + O(d_2\sqrt{d_0})$. Finally $\psi(\pi/2) = \frac{1 + (\pi/2)^2}{4\pi^2} = \frac{1}{4\pi^2} + \frac{1}{16}$ and $\sin(\pi/2)(\pi - \pi/2)/(2\pi^2) = \frac{1}{4\pi}$ give the high-dimensional limits of κ_1, κ_2. ∎

C.7. Numerical verification

End-to-end simulations match the formulas above without analytical shortcuts: sample W_1, W_2 with i.i.d. $\mathcal{N}(0, 1)$ entries and a large batch X with $x_i \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$, compute G_1, G_2 by automatic differentiation on the empirical loss, average $G_i G_i^\top$ over many initializations, and estimate $\alpha, \kappa_1, \kappa_2$ by Monte Carlo (never via $C = \frac{1}{2}W_1$, $G_1^{(2)} = \frac{1}{2}W_2^\top\Phi$, or the Cho–Saul kernel). At $d_0 = 25$, $d_1 = 20$, $d_2 = 12$, $\Phi_{ij} \sim \mathcal{N}(0, 9)$ (so $\|\Phi\|_{\mathrm{F}}^2 = 3211.10$), $n_{\text{init}} = 4000$, $n_x = 8000$:

Quantity	Simulation	Theory	Ratio
G_2 diagonal mean	11541.18	11549.25	0.999
G_2 off-diagonal projection coefficient $\hat{c}$	4.50	5.00	0.900
G_1 diagonal mean	2372.19	2364.78	1.003
G_1 off-diagonal mean	24.64	23.38	1.054

Figure 2 sweeps one of d_0, d_1, d_2 at a time; the empirical slopes track the predicted exponents (for instance G_1 off-diagonal vs d_1 has slope ≈ 0, vs d_0 and d_2 slope ≈ 1, while G_1 diagonal vs d_1 is sublinear at moderate d_1 because the $\frac{d_0 d_2^2}{4} + \frac{\|\Phi\|_{\mathrm{F}}^2}{4}$ constants still dominate the $d_0 d_1 d_2\kappa_1$ term).

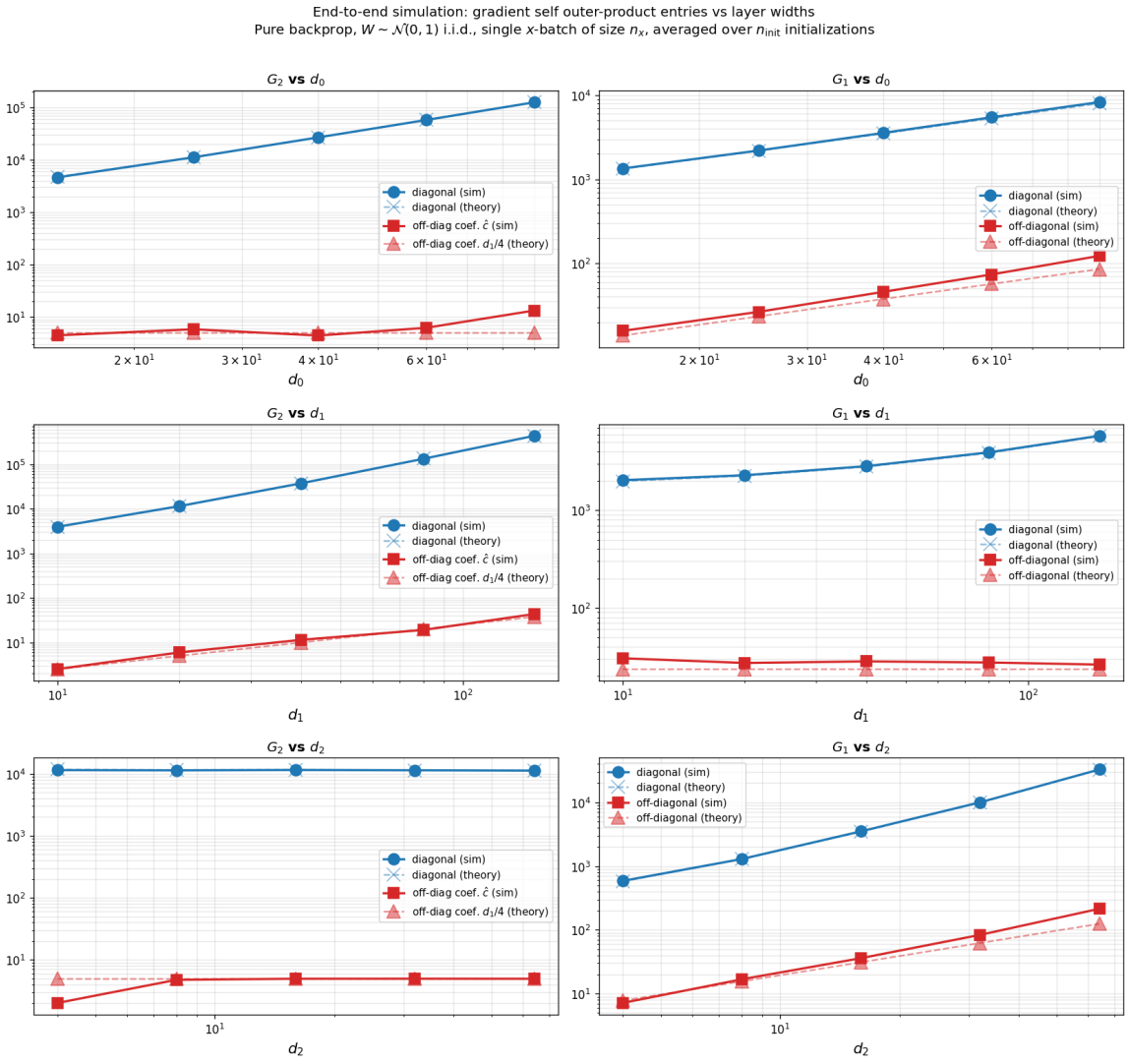


Figure 2: End-to-end simulation of the two-layer ReLU network. Rows: vary d_0 (d_1=20, d_2=12); vary d_1 (d_0=25, d_2=12); vary d_2 (d_0=25, d_1=20). Left: G_2; right: G_1. Diagonal entries (circles, simulation) match theory (crosses); off-diagonal entries (squares, simulation) match theory (triangles).

C.8. Discussion

At initialization the gradient of W_2 is mostly target-independent: the $\Theta(d_0^2 d_1^2)$ leading order of $\mathbb{E}[(G_2 G_2^\top)_{aa}]$ comes entirely from $\mathrm{tr}(H^2)$, which depends only on W_1, while the target term $\frac{d_1}{4}(\Phi\Phi^\top)_{aa}$ is dominated unless $\|\Phi\|_{\mathrm{F}}^2$ is exceptionally large. The nonzero $\Theta(d_0 d_2)$ off-diagonal of $\mathbb{E}[G_1 G_1^\top]$ comes from T_3 and reflects ReLU's lack of odd symmetry: in a deep linear network the symmetry $W_1 \mapsto DW_1$, $W_2 \mapsto W_2 D$ for $D = \mathrm{diag}(\pm 1)$ preserves the loss and forces off-diagonal

expectations to zero (cf. Section B), whereas $\sigma(-z) \neq -\sigma(z)$ breaks it. The dominance is therefore asymptotic in d_1 rather than exact.

Appendix D. Orthogonalization Is Asymptotically a Row Normalization

What does diagonal dominance say about the orthogonalization step itself? Suppose $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ is diagonal. Then $\boldsymbol{V}_t\boldsymbol{V}_t^\top = \operatorname{diag}(\boldsymbol{V}_t\boldsymbol{V}_t^\top)$, so $(\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2} = \operatorname{diag}(\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2}$, and since $\operatorname{diag}(\boldsymbol{V}_t\boldsymbol{V}_t^\top)_{ii} = \|(\boldsymbol{V}_t)_{i,:}\|_2^2$,

$$\left[(\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2}\boldsymbol{V}_t\right]_{i,:} = \frac{(\boldsymbol{V}_t)_{i,:}}{\|(\boldsymbol{V}_t)_{i,:}\|_2}. \tag{53}$$

That is, when $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ is diagonal the orthogonalization in (1) is just each row of the momentum divided by its own ℓ_2 norm, which is exactly the row normalization used by several recent stateless optimizers [7, 10, 18, 21, 23, 29]; equivalently, its implicit preconditioner is $\operatorname{diag}(\boldsymbol{P}_t)^{1/2} \otimes \boldsymbol{I}_n$, which equals H_{Muon} when $\boldsymbol{P}_t$ is diagonal. Putting (53) together with Theorems 1 to 3: at a Gaussian initialization, as the width grows, Muon's $\Theta(mn\min(m,n))$ orthogonalization collapses to a $\Theta(mn)$ row-wise ℓ_2 normalization with the same preconditioner shape. Deng et al. [7] check that this equivalence holds throughout the training of GPT-2 and LLaMA, and that the dominance ratio grows with model size, in line with the width dependence in Theorem 3. Carrying the analysis past initialization, and to deep nonlinear networks, is left for future work.