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

Anonymous authors
Paper under double-blind review

Abstract

Muon orthogonalizes a weight matrix's momentum before each step, and on neural networks this simple preconditioner beats AdamW in a wide range of tasks. Existing analyses explain Muon within general optimization frameworks or on a few specific models, and many of them view orthogonalization as a preconditioner. Since the Hessian is the optimal preconditioner under a quadratic approximation of the loss, these works leave a basic question open: what makes orthogonalization suited to the Hessian of a neural network? Prior work shows that the layer-wise Hessian of a neural network is row-block diagonally dominant, and Muon's implicit preconditioner, the Kronecker product $(\boldsymbol{V}\boldsymbol{V}^\top)^{1/2} \otimes \boldsymbol{I}$ with $\boldsymbol{V}$ the momentum, has the same row-block structure exactly when $\boldsymbol{V}\boldsymbol{V}^\top$ is diagonally dominant. Diagonal dominance of $\boldsymbol{V}\boldsymbol{V}^\top$ has been observed in language models, but why a neural network produces it has not been explained. This work provides a quantitative asymptotic theory for neural networks at initialization, and points out the intuition that the overparameterization of the weight matrices drives the diagonal dominance. We analyze $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$, which equals $\mathbb{E}[\boldsymbol{V}\boldsymbol{V}^\top]$ at initialization, in four canonical models: symmetric matrix factorization, deep linear networks, two-layer ReLU networks, and single-layer linear attention. In each case the diagonal entries outgrow the off-diagonal ones at explicit rates as the hidden dimensions grow; for the query, key and value weights of attention the off-diagonal vanishes altogether. Simulations of the four models and experiments on language models at initialization confirm the predicted scaling. The code for all our simulations and language-model experiments is available at this link.

1 Introduction

Muon (Jordan et al., 2024) 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 $\boldsymbol{V}_0 = \boldsymbol{G}_0$ 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 why Muon works within general optimization frameworks: as momentum steepest descent under a chosen operator norm (Bernstein & Newhouse, 2024b;a; Pethick et al., 2025; Lau et al., 2025), as a non-Euclidean trust-region step (Kovalev, 2025), under an isotropic curvature model (Su, 2025), with an implicit spectral-norm constraint on the iterates (Chen et al., 2025), and with convergence guarantees under smoothness or Hölder smoothness, heavy-tailed noise, inexact orthogonalization, variance reduction, and nonsmooth or convex Lipschitz objectives (Li & Hong, 2025; Shen et al., 2025; An et al., 2025; Kovalev & Borodich, 2026; Nagashima & Iiduka, 2026; Sato et al., 2026; Gratton & Toint, 2026; Hübler et al., 2026b; Iiduka, 2026; Choudhury et al., 2026; Shulgin et al., 2025; Kim & hwan Oh, 2026; Qian & Richtárik, 2026; Do et al., 2026; Chang et al., 2026; Qian et al., 2025; Li & Tsuchiya, 2026; Yang et al., 2026; Lai et al., 2026; Parshakova et al., 2026; Liu et al., 2026; Long et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

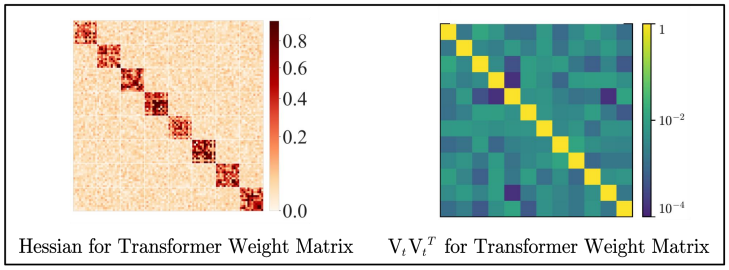


Figure 1: Left: Hessian of a Transformer weight matrix at initialization, whose mass concentrates on the diagonal blocks. Right: the momentum self outer-product $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ of a Transformer weight matrix at initialization (log scale), whose mass concentrates on the diagonal.

However, those works derive their analyses from abstract assumptions on the loss, such as smoothness or convexity, imposed uniformly over a problem class rather than from any specific model. The limitation of those works is that they cannot explain why Muon, a simple matrix preconditioner, works so well *for neural networks specifically*. The few theoretical results for specific models, such as matrix factorization and in-context learning with linear transformers (Ma et al., 2026), linear or bilinear classifiers on separable or imbalanced data (Fan et al., 2025; Vasudeva et al., 2025), and random-feature and transformer blocks at initialization (Davis & Drusvyatskiy, 2025), sharpen the picture in particular regimes. Nevertheless, since the Hessian is the optimal preconditioner under a quadratic approximation of the loss, those previous works leave a basic structural question for preconditioning methods open: **What makes the orthogonalization preconditioner suited to the Hessian of a neural network specifically? This paper gives a theoretical answer that Muon's implicit preconditioner shares a similar row-block diagonally dominant heterogeneous structure to the neural-network Hessian described by Zhang et al. (2024a;b).**

Preconditioner structure versus Hessian structure. Let us first take a view of the equivalent preconditioner form of orthogonalization. The orthogonalized update is $\boldsymbol{D}_t = (\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2}\boldsymbol{V}_t$, a left multiplication that mixes the rows of $\boldsymbol{V}_t$. Stacking the rows of $\boldsymbol{D}_t$ and of $\boldsymbol{V}_t$ into vectors turns it into $\mathrm{vec}(\boldsymbol{D}_t) = \big((\boldsymbol{V}_t\boldsymbol{V}_t^\top)^{-1/2} \otimes \boldsymbol{I}_n\big)\mathrm{vec}(\boldsymbol{V}_t)$, so Muon descends in the metric

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

a Kronecker structure determined entirely by $\boldsymbol{P}_t$. At initialization, $\boldsymbol{V}_0 = \boldsymbol{G}_0$ and hence $\boldsymbol{P}_0 = \boldsymbol{G}_0\boldsymbol{G}_0^\top$. Previous work has shown that the neural-network Hessian has a special heterogeneous structure: for a weight $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, the $m\times m$ blocks indexed by its rows (output neurons) are dominated by the diagonal ones, that is, the layer-wise Hessian is *row-block diagonally dominant* (Zhang et al., 2024a;b; Dong et al., 2025). Meanwhile, Deng et al. (2026) point out empirically that $\boldsymbol{G}_t\boldsymbol{G}_t^\top$ is diagonally dominant throughout training, and the larger the model, the stronger the dominance. When $\boldsymbol{P}_t$ is close to diagonal, H_{Muon} in (2) shares the row-block structure of the Hessian (Figure 1), so the question reduces to a clean property of the gradient.

> **Question.** How does Neural Networks produce such a diagonally dominant $\boldsymbol{V}_t\boldsymbol{V}_t^\top$ (equivalently, $\boldsymbol{G}_t\boldsymbol{G}_t^\top$ at initialization)?

This paper answers this question for neural networks in an asymptotic regime, going a step further than the empirical observation of Deng et al. (2026) by turning it into a quantitative asymptotic theoretical result at initialization. We prove that in four canonical neural-network models the diagonal of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ outgrows the off-diagonal as the hidden dimensions of the weight matrices grow, at rates summarized in Table 1. The overparameterization of the weight matrices is thus a key indicator of this diagonalization. Our contributions can be summarized as follows.

1. **Hidden-dimension asymptotics theory for Muon's preconditioner.** For symmetric matrix factorization, deep linear networks, two-layer ReLU networks and single-layer linear attention, we prove that Muon's implicit preconditioner becomes diagonally dominant as the hidden dimensions grow, with explicit rates (Theorems 1 to 4 and Table 1).
2. **Empirical evaluation of the predicted scaling, from simulations to language models.** Simulations of the four models reproduce the predicted rates (Section 5), and Qwen3.5, Llama-3.2 and Granite-4.0 at initialization show the dominance strengthening with model width (Section 6).

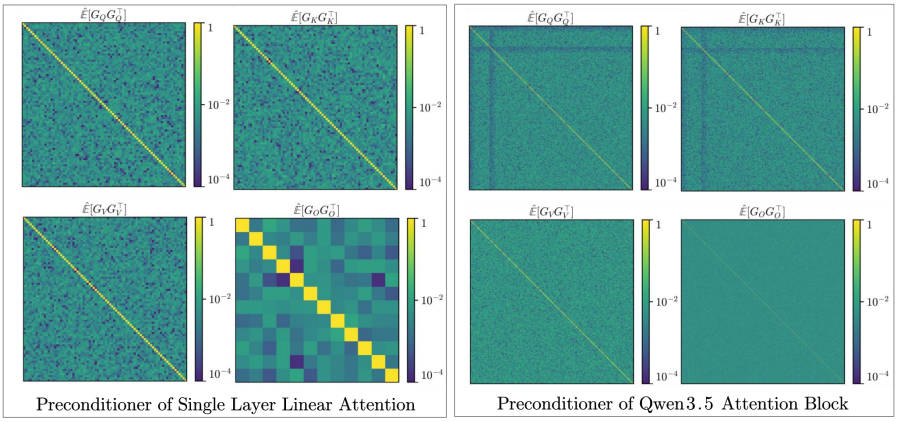


Figure 2: Empirical $\widehat{\mathbb{E}}[\boldsymbol{G}\boldsymbol{G}^\top]$ (log scale) of the query, key, value and output projections at initialization, for single-layer linear attention (left, details in Appendix G.4) and the first attention block of Qwen3.5 (right, details in Appendix H.1). In both cases the mass concentrates on the diagonal. Each heatmap is normalized by its empirical mean diagonal magnitude.

2 Related Work

Mechanistic analyses of Muon. Muon (Jordan et al., 2024) replaces the momentum of each weight matrix by its polar factor. This is the momentum form of steepest descent in the spectral norm, a view developed by Bernstein & Newhouse (2024b;a). Work on why this update suits neural networks follows two routes. The first route models the curvature of the loss. Su (2025) introduces an isotropic curvature model and shows that, under a growth condition on the curvature, the optimal update flattens the spectrum of the gradient, as orthogonalization does. Davis & Drusvyatskiy (2025) give a layerwise condition, comparing the nuclear-to-Frobenius ratio of the gradient with the stable rank of the incoming activations, under which one spectral step decreases the loss more than one Euclidean step. Ma et al. (2026) prove that a simplified Muon converges linearly on matrix factorization and on in-context learning with linear transformers, at a rate independent of the condition number and ahead of gradient descent and Adam. Nguyen et al. (2026) recast Muon as a preconditioned gradient method and show, under a Kronecker-factored curvature model, that spectral flattening improves the convergence factor. Wang et al. (2026) compare the one-step loss decrease of Muon and Adam under a second-order Taylor approximation and attribute Muon's advantage to a smaller curvature penalty, and Wu et al. (2026) probe the loss landscape along real training runs, estimating the loss-optimal step size along each singular direction of the momentum. Since the isotropic geometry behind Muon only approximates the true curvature, several curvature-aware variants have followed (Zhang et al., 2026c; Du & Su, 2026; Che, 2026; Zhang et al., 2026b). The second route studies the update itself. Beneventano et al. (2026) characterize the polar factor as the entropy-maximizing bounded update that keeps the singular directions of the gradient and derive its spectral dynamics. Shumaylov et al. (2026) find that random or inverted spectra train equally well, and Wang et al. (2025b) attribute Muon's advantage on heavy-tailed data to its more isotropic update spectrum. Zhang (2026) separate spectral control from gauge equivariance, Eschenhagen et al. (2026) relate Muon to Shampoo, and Zhang et al. (2026a) ask at which granularity attention weights should be orthogonalized. Both routes take a curvature model or a landscape as given. None of them compares the structure of Muon's implicit preconditioner with that of the network's Hessian, which is the question we study.

Block structure of the neural-network Hessian. The Hessian of a neural network is near block diagonal, with most of its mass in dense principal blocks. This observation goes back to Collobert (2004), and the same structure is assumed by the block-diagonal curvature approximations of Le Roux et al. (2008) and Martens & Grosse (2015). Dong et al. (2025) quantify the effect at random initialization: for linear and one-hidden-layer classifiers, the off-diagonal blocks vanish relative to the diagonal ones as the number of classes grows. The dense blocks within a layer are indexed by output neurons. This is the row-block structure we build on, and it is the partition by which

Adam-mini assigns one learning rate per block to the value, projection and MLP weights (Zhang et al., 2024b). Beyond being block diagonal, the Hessian is heterogeneous across parameter blocks. Zhang et al. (2024a) show that in Transformers, unlike CNNs, the Hessian spectrum varies strongly from one weight matrix to another, so that no single learning rate serves all of them. Ormaniec et al. (2024) derive the Hessian of a self-attention layer and trace this heterogeneity to the softmax and the query-key parameterization, Tomihari & Sato (2025) show that gradient heterogeneity across blocks complements it, and Wang et al. (2025a) exploit the sharpness disparity between blocks to speed up pre-training. Within a block, Zhang et al. (2025) prove that under anisotropic inputs a layer-wise Kronecker-factored preconditioner recovers the feature-learning rate that SGD and Adam miss. Our work builds on the block-diagonal part of this picture: a preconditioner should be block diagonal in the row-block basis, and Muon's $(\boldsymbol{V}\boldsymbol{V}^\top)^{1/2} \otimes \boldsymbol{I}$ has this form precisely when $\boldsymbol{V}\boldsymbol{V}^\top$ is diagonal. Further preconditioning methods are discussed in Appendix A.

3 Preliminaries

Throughout, $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ denotes a weight matrix, $\mathcal{L}$ the population loss and $\boldsymbol{G} = \partial\mathcal{L}/\partial\boldsymbol{W}$ the gradient of the loss with respect to that matrix. Our object of study is the $m \times m$ matrix $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ at a Gaussian initialization. This matrix is the expectation of $\boldsymbol{P}_0 = \boldsymbol{V}_0\boldsymbol{V}_0^\top$ in (2). We analyze it in four settings. Section 3.1 introduces the four models, and Section 3.2 states the initialization assumed for each. We write $\boldsymbol{I}_d$ for the $d \times d$ identity, tr for the trace and $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm. The ReLU is $\sigma(z) = \max(z, 0)$, with derivative $\sigma'(z) = \mathbf{1}[z > 0]$, and $\otimes$ is the Kronecker product. Finally, $a = \Theta(b)$ means that $c\,b \le a \le C\,b$ for some constants $0 < c \le C$.

3.1 Setup

Symmetric matrix factorization. For positive integers d, k, r with $k \ge r$, a weight $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ fits a fixed target $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$, symmetric positive semidefinite of rank r, under the over-parameterized loss

$$\mathcal{L}(\boldsymbol{W}) := \tfrac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2, \qquad \boldsymbol{G} := \frac{\partial\mathcal{L}}{\partial\boldsymbol{W}} = (\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}^\star)\,\boldsymbol{W} \in \mathbb{R}^{d\times k}. \tag{3}$$

The hidden dimension is k.

Deep linear network. For $L \ge 1$ and layer dimensions $d_0, \dots, d_L$, weights $\boldsymbol{W}_i \in \mathbb{R}^{d_i\times d_{i-1}}$ fit a fixed linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_L\times d_0}$ on inputs $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_0})$,

$$f(\boldsymbol{x}) = \boldsymbol{W}_L \cdots \boldsymbol{W}_1\boldsymbol{x}, \qquad \mathcal{L}(\boldsymbol{W}_1, \dots, \boldsymbol{W}_L) := \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{x}}\left\|f(\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x}\right\|^2. \tag{4}$$

With $\boldsymbol{B}_i := \boldsymbol{W}_{i-1}\cdots\boldsymbol{W}_1$ ($\boldsymbol{B}_1 := \boldsymbol{I}$), $\boldsymbol{A}_i := \boldsymbol{W}_L\cdots\boldsymbol{W}_{i+1}$ ($\boldsymbol{A}_L := \boldsymbol{I}$) and $\boldsymbol{R} := \boldsymbol{W}_L\cdots\boldsymbol{W}_1 - \boldsymbol{\Phi}$, the layer-$i$ gradient is

$$\boldsymbol{G}_i := \frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_i} = \boldsymbol{A}_i^\top\,\boldsymbol{R}\,\boldsymbol{B}_i^\top \in \mathbb{R}^{d_i\times d_{i-1}}. \tag{5}$$

The hidden dimensions are $d_1, \dots, d_{L-1}$.

Two-layer ReLU network. For positive d_0, d_1, d_2, weights $\boldsymbol{W}_1 \in \mathbb{R}^{d_1\times d_0}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_2\times d_1}$ fit a fixed linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_2\times d_0}$ on inputs $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_0})$,

$$f(\boldsymbol{x}) = \boldsymbol{W}_2\,\sigma(\boldsymbol{W}_1\boldsymbol{x}), \qquad \mathcal{L}(\boldsymbol{W}_1, \boldsymbol{W}_2) := \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{x}}\left\|f(\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x}\right\|^2. \tag{6}$$

With $\boldsymbol{h}(\boldsymbol{x}) := \sigma(\boldsymbol{W}_1\boldsymbol{x})$ the hidden features, $\boldsymbol{r}(\boldsymbol{x}) := f(\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x}$ the residual and $\boldsymbol{D}(\boldsymbol{x}) := \mathrm{diag}(\sigma'(\boldsymbol{W}_1\boldsymbol{x}))$, the gradients $\boldsymbol{G}_i := \partial\mathcal{L}/\partial\boldsymbol{W}_i$ are

$$\boldsymbol{G}_1 = \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{D}(\boldsymbol{x})\boldsymbol{W}_2^\top\boldsymbol{r}(\boldsymbol{x})\boldsymbol{x}^\top\right] \in \mathbb{R}^{d_1\times d_0}, \qquad \boldsymbol{G}_2 = \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{r}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top\right] \in \mathbb{R}^{d_2\times d_1}. \tag{7}$$

Let $\boldsymbol{w}_a$ denote the a-th row of $\boldsymbol{W}_1$. The hidden dimension is d_1.

Single-layer linear attention. For positive T, d_0, d_k, d_v, d_2, tokens $\boldsymbol{X} \in \mathbb{R}^{T\times d_0}$ (one per row) pass through weights $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_k\times d_0}$, $\boldsymbol{W}_V \in \mathbb{R}^{d_v\times d_0}$, $\boldsymbol{W}_O \in \mathbb{R}^{d_2\times d_v}$, which fit a fixed linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_2\times d_0}$,

$$\begin{aligned} \boldsymbol{A} &:= \tfrac{1}{\sqrt{d_k}}\,\boldsymbol{X}\boldsymbol{W}_Q^\top\boldsymbol{W}_K\boldsymbol{X}^\top, \qquad f(\boldsymbol{X}) = \boldsymbol{A}\,\boldsymbol{X}\boldsymbol{W}_V^\top\boldsymbol{W}_O^\top, \\ \mathcal{L}(\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V,\boldsymbol{W}_O) &:= \tfrac{1}{2}\left\|f(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{\Phi}^\top\right\|_{\mathrm{F}}^2. \end{aligned} \tag{8}$$

Let $\boldsymbol{C} := \boldsymbol{X}^\top\boldsymbol{X}$, $c_p := \mathrm{tr}(\boldsymbol{C}^p)$, $\boldsymbol{Z} := \boldsymbol{X}\boldsymbol{W}_V^\top$ and $\boldsymbol{R} := f(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{\Phi}^\top$. The gradients $\boldsymbol{G}_\bullet := \partial\mathcal{L}/\partial\boldsymbol{W}_\bullet$, $\bullet \in \{Q, K, V, O\}$, are

$$\begin{aligned} \boldsymbol{G}_Q &= \tfrac{1}{\sqrt{d_k}}\,\boldsymbol{W}_K\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{W}_O^\top\boldsymbol{R}^\top\boldsymbol{X}, & \boldsymbol{G}_K &= \tfrac{1}{\sqrt{d_k}}\,\boldsymbol{W}_Q\boldsymbol{X}^\top\boldsymbol{R}\boldsymbol{W}_O\boldsymbol{Z}^\top\boldsymbol{X}, \\ \boldsymbol{G}_V &= \boldsymbol{W}_O^\top\boldsymbol{R}^\top\boldsymbol{A}\boldsymbol{X}, & \boldsymbol{G}_O &= \boldsymbol{R}^\top\boldsymbol{A}\boldsymbol{Z}. \end{aligned} \tag{9}$$

The hidden dimensions are d_k and d_v. In most practical language models the input and output dimensions of an attention block both equal the model width, $d_0 = d_2$ (Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Grattafiori et al., 2024; Meta AI, 2024; Qwen Team, 2026; IBM Granite Team, 2025), and scaling a model up increases this width while d_k, d_v stay fixed, so for attention we also track d_0 and d_2.

3.2 Assumptions

Each setting is analyzed at a standard Gaussian initialization, stated separately below.

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

Assumption 2 (Deep linear network)**.** *The entries of $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$ are i.i.d. $\mathcal{N}(0,1)$; the input $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$ is integrated out in $\mathcal{L}$; and $\boldsymbol{\Phi}$ is deterministic and independent of the weights. Expectations are over $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$.*

Assumption 3 (Two-layer ReLU network)**.** *The entries of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are i.i.d. $\mathcal{N}(0,1)$; the input $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$ is integrated out in $\mathcal{L}$; and $\boldsymbol{\Phi}$ is deterministic and independent of $\boldsymbol{W}_1, \boldsymbol{W}_2$. Expectations are over $\boldsymbol{W}_1, \boldsymbol{W}_2$.*

Assumption 4 (Linear attention)**.** *$\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O$ are mutually independent with i.i.d. $\mathcal{N}(0,1)$ entries; the rows of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$ and independent of the weights, so that $\boldsymbol{C} \sim \mathcal{W}_{d_0}(T, \boldsymbol{I}_{d_0})$ is Wishart; and $\boldsymbol{\Phi}$ is deterministic. Expectations are over the weights and the tokens jointly.*

4 Main Results

We state the closed form of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ in each setup, discuss its diagonal dominance, and defer the proofs to Appendices B to E.

4.1 Symmetric matrix factorization

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

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] = \alpha_3\,\boldsymbol{I}_d - 2\alpha_2\,\boldsymbol{M}^\star + k\,(\boldsymbol{M}^\star)^2, \tag{10}$$

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 $\mathbb{E}[\boldsymbol{S}^2]$ and $\mathbb{E}[\boldsymbol{S}^3]$. Consequently, every diagonal entry $\alpha_3 - 2\alpha_2\boldsymbol{M}^\star_{ii} + k(\boldsymbol{M}^\star)^2_{ii}$ is $\Theta(k^3)$, and every off-diagonal entry $-2\alpha_2\boldsymbol{M}^\star_{ij} + k(\boldsymbol{M}^\star)^2_{ij}$ with $\boldsymbol{M}^\star_{ij} \neq 0$ is $\Theta(k^2)$.

As the hidden dimension k grows, the ratio of diagonal to off-diagonal entries is therefore $\Theta(k)$, and $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ approaches the isotropic matrix $\alpha_3\boldsymbol{I}_d$. At initialization the preconditioner thus reduces, to leading order, to a uniform rescaling of the momentum, and the target enters only through the lower-order terms $-2\alpha_2\boldsymbol{M}^\star + k(\boldsymbol{M}^\star)^2$.

4.2 DEEP LINEAR NETWORK

Theorem 2. *Under Assumption 2, for every layer $i \in \{1, \dots, L\}$,*

$$\mathbb{E}[\boldsymbol{G}_i \boldsymbol{G}_i^\top] = V_i\, s_i\, \boldsymbol{I}_{d_i} + \boldsymbol{T}_i, \qquad \boldsymbol{T}_i = \begin{cases} \big(\prod_{j=1}^{L-1} d_j\big)\, \boldsymbol{\Phi}\boldsymbol{\Phi}^\top, & i = L, \\ \big(\prod_{j\neq i,\, 1\le j\le L-1} d_j\big)\, \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\, \boldsymbol{I}_{d_i}, & 1 \le i < L, \end{cases} \tag{11}$$

where V_i, s_i are positive dimension-only scalars given by two-step recursions in $(d_0, \dots, d_L)$ (Appendix C), with

$$V_i s_i = \begin{cases} \Theta\big(d_i \prod_{j=1,\, j\neq i}^{L-1} d_j^2\big), & 1 \le i < L, \\ \Theta\big(\prod_{j=1}^{L-1} d_j^2\big), & i = L. \end{cases} \tag{12}$$

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. At the output layer the diagonal noise $V_L s_L = \Theta\big(\prod_{j=1}^{L-1} d_j^2\big)$ grows with the hidden dimensions while the off-diagonal stays proportional to $\big(\prod_{j=1}^{L-1} d_j\big)(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ab}$, so the ratio, of order $\prod_{j=1}^{L-1} d_j$, still diverges. For equal hidden dimensions $d_1 = \cdots = d_{L-1} = d$ the output-layer diagonal is $\Theta(d^{2(L-1)})$ against an off-diagonal of $\Theta(d^{L-1})$.

4.3 TWO-LAYER RELU NETWORK

Let $\theta := \angle(\boldsymbol{u}, \boldsymbol{v})$ for independent $\boldsymbol{u}, \boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$, and set the 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 $\kappa_2 := \frac{\mathbb{E}[\sin\theta(\pi-\theta)]}{2\pi^2}$.

Theorem 3. *Under Assumption 3, with $\boldsymbol{H} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top]$ and $\alpha := \mathbb{E}_{\boldsymbol{W}_1}[\mathrm{tr}(\boldsymbol{H}^2)]$,*

$$\mathbb{E}[\boldsymbol{G}_2 \boldsymbol{G}_2^\top] = \alpha\, \boldsymbol{I}_{d_2} + \tfrac{d_1}{4} \boldsymbol{\Phi}\boldsymbol{\Phi}^\top, \qquad \alpha = \Theta(d_1^2),$$

and $\mathbb{E}[\boldsymbol{G}_1 \boldsymbol{G}_1^\top]$ is permutation-invariant in the hidden index with diagonal entries

$$\mathbb{E}[(\boldsymbol{G}_1 \boldsymbol{G}_1^\top)_{aa}] = \tfrac{d_0 d_2^2}{4} + d_0 d_2 \big[(d_1 - 1)\kappa_1 + \tfrac{1}{2}\big] + \tfrac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}$$

and off-diagonal entries $\mathbb{E}[(\boldsymbol{G}_1 \boldsymbol{G}_1^\top)_{ac}] = d_0 d_2\, \kappa_2 + O(d_2\sqrt{d_0})$ for $a \neq c$. Consequently, for $\boldsymbol{G}_1$ the diagonal entries are $\Theta(d_1)$ and the off-diagonal entries $\Theta(1)$, and for $\boldsymbol{G}_2$ they are $\Theta(d_1^2)$ and $\Theta(d_1)$.

As the hidden dimension d_1 grows, the ratio of diagonal to off-diagonal entries is therefore $\Theta(d_1)$ for both layers. For $\boldsymbol{G}_2$ the off-diagonal $\frac{d_1}{4}(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ac}$ comes only from the target, as at the output layer of Theorem 2. For $\boldsymbol{G}_1$ the off-diagonal $d_0 d_2 \kappa_2$ is target-free: it 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 (Appendix G.3).

4.4 SINGLE-LAYER LINEAR ATTENTION

Theorem 4. *Under Assumption 4,*

$$\mathbb{E}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top] = c_Q\, \boldsymbol{I}_{d_k}, \qquad \mathbb{E}[\boldsymbol{G}_K \boldsymbol{G}_K^\top] = c_K\, \boldsymbol{I}_{d_k}, \qquad \mathbb{E}[\boldsymbol{G}_V \boldsymbol{G}_V^\top] = c_V\, \boldsymbol{I}_{d_v},$$
$$\mathbb{E}[\boldsymbol{G}_O \boldsymbol{G}_O^\top] = \bar{\alpha}_O\, \boldsymbol{I}_{d_2} + d_v\, \gamma\, \boldsymbol{\Phi}\boldsymbol{\Phi}^\top, \tag{13}$$
$$\gamma = \tfrac{1}{d_0}\mathbb{E}[c_2^2] = T\big[(Td_0{+}8)(T{+}d_0)^2 + 2(Td_0{+}10)(T{+}d_0) + 5(Td_0{+}4)\big],$$

where $c_Q, c_K, c_V, \bar{\alpha}_O$ are explicit positive polynomials in the dimensions and $\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2$ (Appendix E); for fixed $\boldsymbol{\Phi}$,

$$c_Q \asymp c_K = \Theta\big(d_2^2 d_v^2 d_0^6 / d_k\big), \qquad c_V = \Theta\big(d_2(d_2{+}d_v)\, d_0^6\big), \qquad \bar{\alpha}_O = \Theta\big(d_v^2 d_0^6\big).$$

For $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$ the off-diagonal is exactly zero (Lemma 34). For $\boldsymbol{W}_O$, as at the output layer of Theorem 2, the off-diagonal $d_v \gamma (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ab} = \Theta(d_v d_0^3)$ comes only from the target while the diagonal $\bar{\alpha}_O = \Theta(d_v^2 d_0^6)$ is target-free noise, so the ratio $\Theta(d_v d_0^3)$ diverges with d_v and, faster, with d_0; d_k does not affect it. With $d_0 = d_2$ equal to the model width (Section 3.1), the ratio grows like the cube of the model width (Section 6).

4.5 SUMMARY AND PROOF SKETCH

In all four settings the diagonal of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ outgrows the off-diagonal as the hidden dimension grows (Table 1), so $\boldsymbol{P}_t = \boldsymbol{V}_t\boldsymbol{V}_t^\top$ is asymptotically diagonal, $H_{\text{Muon}} = \boldsymbol{P}_t^{1/2} \otimes \boldsymbol{I}_n$ is asymptotically block-diagonal in the Hessian's row-block structure, and (Appendix F) orthogonalization is asymptotically a row-wise normalization.

Model	Diagonal	Off-diagonal	Asymptotic dims
Matrix factorization	$\Theta(k^3)$	$\Theta(k^2)$	k
Deep linear, hidden layer i	$\Theta\big(d_i \prod_{j=1,\, j\neq i}^{L-1} d_j^2\big)$	0	$d_1, \ldots, d_{L-1}$
Deep linear, output layer	$\Theta\big(\prod_{j=1}^{L-1} d_j^2\big)$	$\Theta\big(\prod_{j=1}^{L-1} d_j\big)$	$d_1, \ldots, d_{L-1}$
Two-layer ReLU, layer $\boldsymbol{W}_1$	$\Theta(d_1)$	$\Theta(1)$	d_1
Two-layer ReLU, layer $\boldsymbol{W}_2$	$\Theta(d_1^2)$	$\Theta(d_1)$	d_1
Linear attention, $\boldsymbol{W}_Q, \boldsymbol{W}_K$	$\Theta(d_2^2 d_v^2 d_0^6/d_k)$	0	d_v, d_0, d_2
Linear attention, $\boldsymbol{W}_V$	$\Theta(d_2(d_2{+}d_v)d_0^6)$	0	d_v, d_0, d_2
Linear attention, $\boldsymbol{W}_O$	$\Theta(d_v^2 d_0^6)$	$\Theta(d_v d_0^3)$	d_v, d_0

Table 1: Summary of the orders of $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ at a Gaussian initialization in the hidden dimensions of each model and, for attention, the token dimensions d_0, d_2.

The four proofs share a similar strategy. Each row of $\boldsymbol{G}$ is a random vector, and the (a, b) entry of $\boldsymbol{G}\boldsymbol{G}^\top$ is the inner product of rows a and b, a measure of their correlation. A diagonal entry is the correlation of a row with itself, its squared norm, which adds up over all hidden coordinates. An off-diagonal entry is the correlation between two different rows, which are built from independent Gaussian vectors at initialization, so most of its terms cancel in expectation and only the lower-order terms carried by the fixed target remain. The gap between the orders of the diagonal and off-diagonal entries in Table 1 comes from this difference. In the proofs we compute these statistics of $\boldsymbol{G}\boldsymbol{G}^\top$ exactly, using moments of the Wishart distribution for matrix factorization, deep linear networks and linear attention, and Stein's lemma together with the arc-cosine kernel of Cho & Saul (2009) for the ReLU network, whose nonlinearity is not a polynomial. Each closed form is reproduced by the simulations in Section 5.

5 SIMULATION OF THEORETICAL RESULTS

Simulation setup. We verify the dimension scalings of Theorems 1 to 4 under Assumptions 1 to 4 by varying the model dimensions. For each model, we vary the dimensions in the corresponding theoretical results and examine the resulting structure of the expected gradient self outer-product $\boldsymbol{P} := \mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] \in \mathbb{R}^{m\times m}$, that is, the expectation of $\boldsymbol{P}_0$ in (2). For this matrix, we summarize the mean diagonal magnitude, the mean off-diagonal magnitude, and their ratio,

$$D(\boldsymbol{P}) = \frac{1}{m}\sum_{i=1}^{m} |P_{ii}|, \qquad \mathrm{OD}(\boldsymbol{P}) = \frac{1}{m(m-1)}\sum_{i\neq j} |P_{ij}|, \qquad R(\boldsymbol{P}) = \frac{D(\boldsymbol{P})}{\mathrm{OD}(\boldsymbol{P})}, \tag{14}$$

We estimate $\boldsymbol{P}$ by $\widehat{\mathbb{E}}[\boldsymbol{G}\boldsymbol{G}^\top] = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{G}^{(s)}\boldsymbol{G}^{(s)\top}$ over S independent initializations; $\widehat{D}, \widehat{\mathrm{OD}}, \widehat{R}$ denote these estimates. Where the expected off-diagonal is exactly zero, the nonzero $\widehat{\mathrm{OD}}$ is Monte Carlo variance (details in Appendix G). We next apply this procedure to matrix factorization, deep linear networks, two-layer ReLU networks, and single-layer linear attention, varying in each case the dimensions governing the corresponding theoretical scaling.

Symmetric matrix factorization. We begin with symmetric matrix factorization, varying $k \in \{5, 10, 20, 40, 80, 160, 320, 640\}$ with $d = 25$ and a fixed rank-5 target, using $S = 10{,}000$ initializations for each k. In Figure 3 (left), $\widehat{R}$ increases with k, in line with $R = \Theta(k)$ from Theorem 1; $\widehat{D}$ and $\widehat{\mathrm{OD}}$ follow $\Theta(k^3)$ and $\Theta(k^2)$ (Appendix G.1).

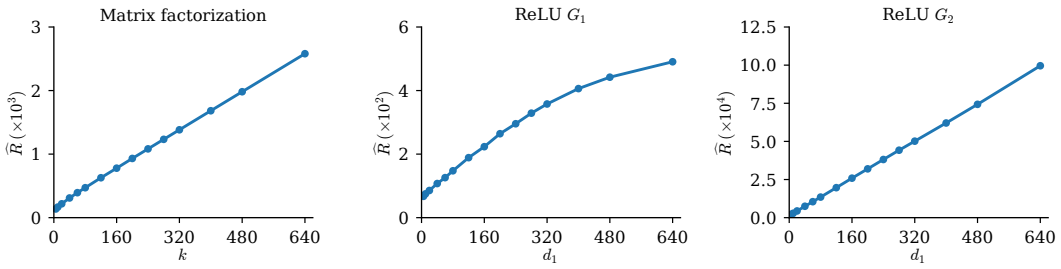


Figure 3: Diagonal-dominance ratio $\widehat{R}$ for symmetric matrix factorization against k (left; $d = 25$, rank-5 target) and for the two-layer ReLU network against d_1 (right; $d_0 = 512$, $d_2 = 12$).

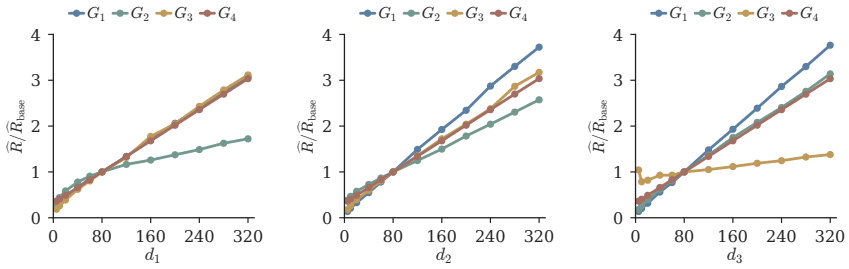


Figure 4: Diagonal-dominance ratio $\widehat{R}$ of $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_4$ for the depth-4 linear network when d_1, d_2 or d_3 is varied (left to right). Each curve is normalized by its value at the baseline ($d_1 = d_2 = d_3 = 80$) to facilitate comparison across layers with different absolute scales.

Deep linear network. We next consider a depth-4 linear network with baseline dimensions $[d_0, d_1, d_2, d_3, d_4] = [512, 80, 80, 80, 512]$. We vary each hidden dimension d_1, d_2, d_3 in turn over the same grid while holding the remaining dimensions fixed, using $S = 10{,}000$ initializations per configuration. In Figure 4, $\widehat{R}$ increases with every hidden dimension for all layers. For the output layer, Theorem 2 gives $D = \Theta((d_1 d_2 d_3)^2)$ and $\mathrm{OD} = \Theta(d_1 d_2 d_3)$, and hence $R = \Theta(d_1 d_2 d_3)$; accordingly, $\widehat{R}$ grows approximately linearly when any one hidden dimension is varied. For the hidden layers, the expected off-diagonal is zero (Appendix G.2).

Two-layer ReLU network. We then study the two-layer ReLU network, fixing $d_0 = 512$ and $d_2 = 12$ and varying the hidden width d_1 over the same grid. For each configuration, we use $S = 10{,}000$ weight initializations and estimate the population gradient using 128,000 Gaussian inputs. In Figure 3 (right), $\widehat{R}$ increases with d_1 for both layers, matching $R = \Theta(d_1)$ in Theorem 3 (Appendix G.3).

Single-layer linear attention. Finally, we consider single-layer linear attention with $T = 32$, $d_0 = 512$, and $d_2 = 12$. We examine four dimension sweeps: $d_k = d_v$ jointly, d_k with $d_v = 80$, d_v with $d_k = 80$, and $d_0 = d_2$ jointly. In Figure 5, $\widehat{R}$ of $\boldsymbol{G}_O$ grows approximately linearly with d_v, is nearly unchanged with d_k, and grows superlinearly with d_0, consistent with $R = \Theta(d_v d_0^3)$ from Theorem 4. For $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V$, the expected off-diagonal entries are zero. Their finite $\widehat{R}$ values nevertheless follow the corresponding diagonal trends: they increase with $d_k = d_v$, increase with d_v, and grow strongly with $d_0 = d_2$, while $\boldsymbol{G}_Q$ and $\boldsymbol{G}_K$ decrease with d_k, consistent with $D = \Theta(d_2^2 d_v^2 d_0^6 / d_k)$ (Appendix G.4).

6 Empirical Evaluation on Large Language Models

We next test whether the diagonal dominance of Theorem 4 persists in Transformer-based language models at random initialization: Qwen3.5-0.8B-Base (Qwen Team, 2026), Granite-4.0-1B-Base (IBM Granite Team, 2025) and Llama-3.2-1B (Meta AI, 2024). Figure 2 (right) shows the empirical $\widehat{\mathbb{E}}[\boldsymbol{G}\boldsymbol{G}^\top]$ of the four projections in Qwen3.5, whose mass concentrates on the diagonal. For

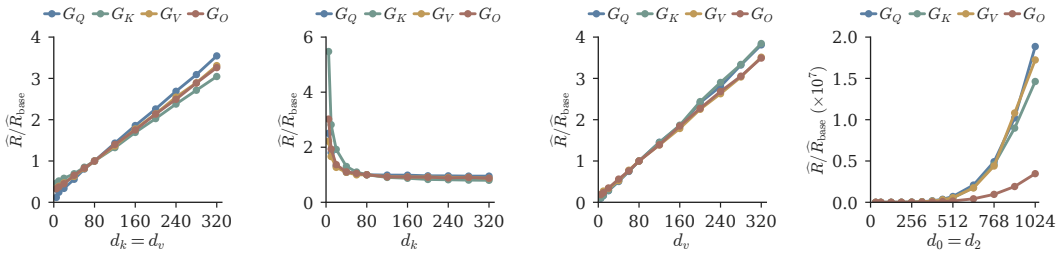


Figure 5: Diagonal-dominance ratio $\widehat{R}$ of the four attention projections when $d_k = d_v$, d_k, d_v, and $d_0 = d_2$ are varied (left to right). Each curve is normalized by its value at the baseline (first three: $T = 32$, $d_0 = 512$, $d_k = d_v = 80$, $d_2 = 12$; fourth: $T = 16, d_0 = d_2 = 32, d_k = d_v = 20$) to facilitate comparison across projections with different absolute scales.

the first head of the first eligible attention block of each model we vary the model width $d_0 = d_2$ at fixed head dimension $d_k = d_v$, using 25 random initializations and 64 fixed FineWeb-Edu (Penedo et al., 2024) sequences per configuration (details in Appendix H.2). Figure 6 shows $\widehat{R}$ increasing with model width for all four projections in all three models, as the $\Theta(d_v d_0^3)$ ratio of Theorem 4 predicts.

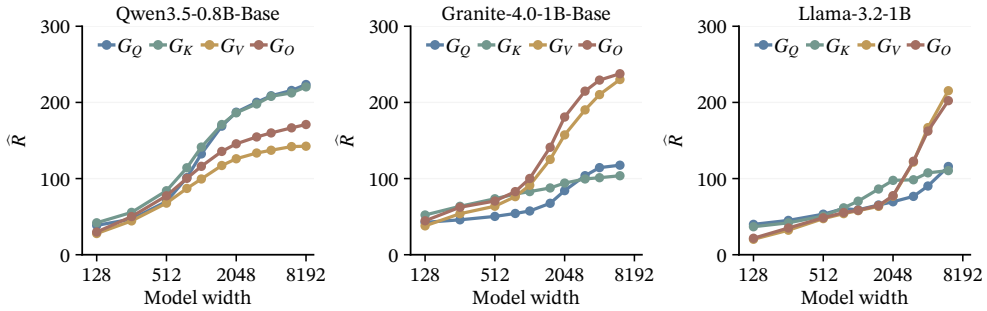


Figure 6: Model width scaling of $\widehat{R}$ for the query, key, value, and output projections in the first head of the first eligible attention block of the three language models at random initialization.

7 Conclusion and Limitations

Conclusion. This work provides an asymptotic theory that shows Muon's implicit preconditioner matches the row-block diagonally dominant structure of the neural-network Hessian, because $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ itself becomes diagonally dominant as the weight matrices are overparameterized, at the explicit rates of Table 1. Specifically, in matrix factorization the isotropic $\Theta(k^3)$ diagonal outgrows the target-driven $\Theta(k^2)$ off-diagonal. In deep linear networks the hidden layers are exactly diagonal, and the ratio at the output layer grows with the product of all hidden dimensions. In the two-layer ReLU network the nonlinearity leaves an off-diagonal of constant order, yet the ratio still grows linearly with the hidden dimension. In linear attention the query, key and value weights are exactly diagonal. For the output weight the ratio $\Theta(d_v d_0^3)$ grows with the head dimension and grows as the cube of the model width, so the dominance strengthens as models get wider. Our experiments on language models of increasing width match this intuition.

Limitations. Our theory is stated at initialization and in expectation: it does not characterize the training dynamics of $\boldsymbol{V}_t\boldsymbol{V}_t^\top$, although the empirical evidence of Deng et al. (2026) indicates that the dominance persists during training. It is also worth studying whether this dominance reflects a degenerate behavior of the model that emerges at large scale. We leave these questions to future work.

AI USE STATEMENT

In this work, we used generative AI tools to check the calculations behind our theoretical results and to assist in the writing of proofs by improving their presentation. We have not used generative AI tools to derive our theoretical results, to develop the theoretical model or conceptual framework, to formulate the mathematical claims, to propose or refine hypotheses, or to design the research methodology or experiments, and the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the manuscript, to create and edit software code, and to create or modify scientific figures. We have reviewed all AI-assisted work: all calculations and proofs were re-derived and checked by the authors, AI-generated code was tested for correctness by the authors, and all edits to the text were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical: it characterizes the expected gradient self outer-product of standard neural-network models at initialization. The experiments consist of synthetic simulations and training runs of publicly available language models on standard public corpora; they involve no human subjects, no personal or sensitive data, and no newly collected datasets. We do not foresee ethical concerns or direct negative societal impacts arising from this work.

REPRODUCIBILITY STATEMENT

All results are stated under explicit assumptions (Section 3.2) and proved in full in the appendix: Appendix B (symmetric matrix factorization), Appendix C (deep linear networks), Appendix D (two-layer ReLU networks) and Appendix E (single-layer linear attention). Each appendix restates the corresponding theorem together with its standing assumption and all supporting lemmas, and the closed-form Wishart trace moments used for the attention layer are listed explicitly in Lemma 33. The numerical checks (Appendix G) specify every dimension, sample size and initialization used, and the language-model experiments report the model sizes, datasets, optimizer settings and number of training steps; the code for both the simulations and the language-model runs is provided at the anonymous repository https://anonymous.4open.science/r/orthogonalization-hessian-CB2C/.

REFERENCES

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.

Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory-efficient adaptive optimization. In *Advances in Neural Information Processing Systems*, 2019.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of Muon. *arXiv preprint arXiv:2606.08388*, 2026.

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

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

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of Muon and beyond. *arXiv preprint arXiv:2509.15816*, 2026.

Tong Che. Second-order Muon done right: A principled marriage of spectral geometry and curvature. *arXiv preprint arXiv:2608.09763*, 2026.

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

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

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026.

Ronan Collobert. *Large Scale Machine Learning*. PhD thesis, Université de Paris VI, 2004.

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

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.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On muon optimization: From non-convergence to an error analysis with polar express and the Newton-Schulz polynomial from implementations. *arXiv preprint arXiv:2608.04607*, 2026.

Zhaorui Dong, Yushun Zhang, Zhi-Quan Luo, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks, 2025. URL https://arxiv.org/abs/2505.02809.

Zhehang Du and Weijie Su. The Newton-Muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and Hao-Jun Michael Shi. Clarifying shampoo: Adapting spectral descent to stochasticity and the parameter trajectory. *arXiv preprint arXiv:2602.09314*, 2026.

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

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

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

S. Gratton and Ph. L. Toint. A unified convergence theory for adaptive first-order methods in the nonconvex case, including adanorm, full and diagonal adagrad and Muon. *arXiv preprint arXiv:2604.17423*, 2026.

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

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. *arXiv preprint arXiv:1802.09568*, 2018.

Florian Hübler, Kai Lion, Antonio Orvieto, and Niao He. Muown implicitly performs angular stepsize decay. *arXiv preprint arXiv:2606.23637*, 2026a.

Florian Hübler, Thomas Pethick, and Suvrit Sra. Free heavy-tailed lunch for Muon: A theoretical justification of empirical success. *arXiv preprint arXiv:2606.14560*, 2026b.

IBM Granite Team. Granite-4.0-1b-base. https://huggingface.co/ibm-granite/granite-4.0-1b-base, 2025. Model card.

Hideaki Iiduka. Muon converges under heavy-tailed noise: Nonconvex hölder-smooth empirical risk minimization. *arXiv preprint arXiv:2603.15059*, 2026.

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.

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

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/.

Gyu Yeol Kim and Min hwan Oh. Convergence of Muon with Newton-Schulz. *arXiv preprint arXiv:2601.19156*, 2026.

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

Dmitry Kovalev and Ekaterina Borodich. Non-Euclidean SGD for structured optimization: Unified analysis and improved rates. *arXiv preprint arXiv:2511.11466*, 2026.

Lexiao Lai, Tianyi Lin, and Jiayu Zhang. Nonsmooth optimization via orthogonalized momentum. *arXiv preprint arXiv:2609.13677*, 2026.

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.

Nicolas Le Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In *Advances in Neural Information Processing Systems 20*, 2008.

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

Mingyi Li and Taira Tsuchiya. Muon with finite Newton-Schulz: The smoothing benefit in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2608.26288*, 2026.

Kai Lion, Florian Hübler, Bingcong Li, Antonio Orvieto, and Niao He. Muown: Row-norm control for Muon optimization. *arXiv preprint arXiv:2605.10797*, 2026.

Jinlin Liu, Song Chen, Jiaxu Liu, and Chao Xu. A continuous-time analysis of smoothed matrix-polar spectral gradient flows for Muon-type optimization. *arXiv preprint arXiv:2608.01911*, 2026.

Yanlin Long, Yufei Gu, and Zeke Xie. Dimension-free saddle-point escape in Muon. *arXiv preprint arXiv:2605.09331*, 2026.

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.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon, 2026. URL https://arxiv.org/abs/2601.13474.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. *arXiv preprint arXiv:1503.05671*, 2015.

Meta AI. Llama 3.2 1B. https://huggingface.co/meta-llama/Llama-3.2-1B, 2024. Model card.

Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham Kakade, and Lucas Janson. A new perspective on Shampoo's preconditioner. *arXiv preprint arXiv:2406.17748*, 2024.

Shuntaro Nagashima and Hideaki Iiduka. Improved convergence rates of Muon optimizer for nonconvex optimization. *arXiv preprint arXiv:2601.19400*, 2026.

Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral flattening is all Muon needs: How orthogonalization controls learning rate and convergence. *arXiv preprint arXiv:2605.13079*, 2026.

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

Weronika Ormaniec, Felix Dangel, and Sidak Pal Singh. What does it mean to be a Transformer? insights from a theoretical Hessian analysis. *arXiv preprint arXiv:2410.10986*, 2024.

Tianyu Pang, Yujie Fang, Zihang Liu, Shenyang Deng, Lei Hsiung, Shuhua Yu, and Yaoqing Yang. Htmuon: Improving muon via heavy-tailed spectral correction, 2026. URL https://arxiv.org/abs/2603.10067.

Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M. Gower. Muon does not converge on convex Lipschitz functions. *arXiv preprint arXiv:2605.08980*, 2026.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024.

Thomas Pethick. When to use what Schatten-p norm in deep learning? *arXiv preprint arXiv:2606.15268*, 2026.

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.

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

Xun Qian and Peter Richtárik. Convergence analysis of Muon-type methods with inexact lmo in the degenerate case. *arXiv preprint arXiv:2606.21581*, 2026.

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtárik. Muon is provably faster with momentum variance reduction. *arXiv preprint arXiv:2512.16598*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://huggingface.co/Qwen/Qwen3.5-0.8B-Base, February 2026. Qwen3.5-0.8B-Base model card.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer. *arXiv preprint arXiv:2507.01598*, 2026.

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

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

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact Muon update. *arXiv preprint arXiv:2510.19933*, 2025.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

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

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

Akiyoshi Tomihari and Issei Sato. Gradient heterogeneity complements Hessian heterogeneity in Transformer optimization. *arXiv preprint arXiv:2502.00213*, 2025.

Hugo Touvron et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. *arXiv preprint arXiv:2409.11321*, 2024.

Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Weinan E, and Lei Wu. The sharpness disparity principle in Transformers for accelerating language model pre-training. *arXiv preprint arXiv:2502.19002*, 2025a.

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*, 2025b.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026.

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.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Xiaodong Wu, Wenyi Yu, Chao Zhang, and Philip Woodland. Spectral allocation: Why Muon outperforms Adam, and how to improve Muon. *arXiv preprint arXiv:2608.25990*, 2026.

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.

Yixuan Yang, Yuqing He, and Song Li. Convergence of spectral descent for non-smooth optimization. *arXiv preprint arXiv:2605.26977*, 2026.

Haozhou Zhang. Polaradamw: Disentangling spectral control and schur gauge-equivariance in matrix optimisation. *arXiv preprint arXiv:2605.07067*, 2026.

Hongtao Zhang, Wenjie Zhou, Wei Chen, and Xueqi Cheng. When and why grouping attention heads accelerates Muon optimization. *arXiv preprint arXiv:2605.08933*, 2026a.

Qiaozhe Zhang, Jun Sun, and Yingzhuang Liu. Beyond the matrix sign: Quadratic spectral descent. *arXiv preprint arXiv:2609.07597*, 2026b.

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.

Yechen Zhang, Shuhao Xing, Junhao Huang, Kai Lv, Yunhua Zhou, Xipeng Qiu, Qipeng Guo, and Kai Chen. Mousse: Rectifying the geometry of Muon with curvature-aware preconditioning. *arXiv preprint arXiv:2603.09697*, 2026c.

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)*, 2024a.

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*, 2024b.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 Preliminaries **4**

3.1 Setup 4

3.2 Assumptions 5

4 Main Results **5**

4.1 Symmetric matrix factorization 5

4.2 Deep linear network 6

4.3 Two-layer ReLU network 6

4.4 Single-layer linear attention 6

4.5 Summary and proof sketch 7

5 Simulation of Theoretical Results **7**

6 Empirical Evaluation on Large Language Models **8**

7 Conclusion and Limitations **9**

A Extended Related Work **16**

B Symmetric Matrix Factorization **17**

B.1 Setup 17

B.2 Standing assumption 17

B.3 Main results restated 17

B.4 Preliminary lemmas 18

B.5 Proofs 18

C Deep Linear Networks **19**

C.1 Setup 19

C.2 Standing assumption . . . 20
C.3 Main results restated . . . 20
C.4 Preliminary lemmas . . . 20
C.5 Proofs . . . 21

D Two-Layer ReLU Networks **22**
D.1 Setup . . . 22
D.2 Standing assumption . . . 23
D.3 Main results restated . . . 23
D.4 Preliminary lemmas . . . 23
D.5 Proof of the gradient self outer-product of $\boldsymbol{W}_2$ . . . 24
D.6 Proof of the gradient self outer-product of $\boldsymbol{W}_1$ . . . 25
D.7 Discussion . . . 26

E Single-Layer Linear Attention **27**
E.1 Setup . . . 27
E.2 Standing assumption . . . 27
E.3 Main results restated . . . 27
E.4 Preliminary lemmas . . . 28
E.5 Proofs . . . 31
E.6 Discussion . . . 35

F Asymptotic Approximation between Orthogonalization and Row Normalization **35**

G Simulation of Theoretical Results **36**
G.1 Symmetric Matrix Factorization . . . 36
G.2 Deep Linear Network . . . 38
G.3 Two-Layer ReLU Network . . . 40
G.4 Single-Layer Attention . . . 41

H Empirical Evaluation on Large Language Models **44**
H.1 Gradient Gram Structure at Initialization . . . 45
H.2 Model Width Scaling . . . 45

A Extended Related Work

Mechanistic analyses of Muon and the heterogeneity of the neural-network Hessian are discussed in Section 2; this appendix covers the remaining line of work.

Preconditioning methods for neural-network optimization. Muon belongs to the family of optimizers defined by a linear minimization oracle (LMO) over a norm ball (Pethick et al., 2025; Chen et al., 2025; Kovalev, 2025), which also contains Scion and Gluon, unifies sign, spectral and Schatten-p updates (Lau et al., 2025; An et al., 2025; Pang et al., 2026; Pethick, 2026), and comes

with convergence theory for the idealized and the practical algorithm (Li & Hong, 2025; Shen et al., 2025; Kovalev & Borodich, 2026). Two-sided Kronecker preconditioners such as Shampoo and SOAP whiten both the row and the column side of a matrix gradient (Gupta et al., 2018; Morwani et al., 2024; Vyas et al., 2024); Anil et al. (2020) make this second-order preconditioning practical at the scale of modern deep learning, and Anil et al. (2019) go in the opposite direction, compressing the state of adaptive optimizers to a memory-efficient factored form. A parallel line keeps only the row side and drops optimizer state by preconditioning the gradient row by row: SWAN, gradient multi-normalization, a minimalist optimizer and SRON replace Adam's state with a row-wise ℓ_2 normalization (Ma et al., 2025; Scetbon et al., 2025; Glentis et al., 2025; Wen et al., 2025); Gu & Xie (2026); Xu et al. (2026) study its manifold-optimization and width-scaling sides; Pethick et al. (2025) reach it from norm-constrained steepest descent; Lion et al. (2026) make the row magnitudes an explicit optimizer variable, which Hübler et al. (2026a) show implicitly decays the angular step size; and Deng et al. (2026) show that row-momentum normalization matches Muon at scale. Our analysis explains the common thread: once $\boldsymbol{V}\boldsymbol{V}^\top$ is diagonally dominant, row normalization and orthogonalization are the same operation (Appendix F).

B Symmetric Matrix Factorization

B.1 Setup

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

$$\mathcal{L}(\boldsymbol{W}) \;:=\; \tfrac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2, \qquad \boldsymbol{W} \in \mathbb{R}^{d\times k}, \tag{15}$$

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

$$\boldsymbol{G} \;:=\; \nabla\mathcal{L}(\boldsymbol{W}) \;=\; (\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}^\star)\,\boldsymbol{W} \;\in\; \mathbb{R}^{d\times k}, \tag{16}$$

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

B.2 Standing assumption

Assumption 5. *The entries W_{ij} are i.i.d. $\mathcal{N}(0,1)$, and $\boldsymbol{M}^\star$ is deterministic and independent of $\boldsymbol{W}$. Hence $\boldsymbol{S} := \boldsymbol{W}\boldsymbol{W}^\top$ is Wishart with k degrees of freedom and identity scale, $\boldsymbol{S} \sim \mathcal{W}_d(k, \boldsymbol{I}_d)$.*

B.3 Main results restated

Theorem (Theorem 1, restated). *Under Assumption 5,*

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] \;=\; \alpha_3(d,k)\,\boldsymbol{I}_d \;-\; 2\,\alpha_2(d,k)\,\boldsymbol{M}^\star \;+\; k\,(\boldsymbol{M}^\star)^2, \tag{17}$$

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

$$\alpha_1(d,k) = k, \tag{18}$$
$$\alpha_2(d,k) = k(k+d+1), \tag{19}$$
$$\alpha_3(d,k) = k^3 + 3(d+1)k^2 + (d^2+3d+4)k. \tag{20}$$

Corollary 1. *Under Assumption 5, $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ is invariant under any orthogonal change of basis fixing $\boldsymbol{M}^\star$, with*

$$\mathbb{E}\big[(\boldsymbol{G}\boldsymbol{G}^\top)_{ii}\big] = \alpha_3 - 2\alpha_2\boldsymbol{M}^\star_{ii} + k(\boldsymbol{M}^\star)^2_{ii}, \tag{21}$$
$$\mathbb{E}\big[(\boldsymbol{G}\boldsymbol{G}^\top)_{ij}\big] = -2\alpha_2\boldsymbol{M}^\star_{ij} + k(\boldsymbol{M}^\star)^2_{ij} \qquad (i \neq j). \tag{22}$$

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 $\Theta(k/\boldsymbol{M}^\star_{ij})$.

B.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 $\boldsymbol{N} \in \mathbb{R}^{m\times m}$,*

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top] = \operatorname{tr}(\boldsymbol{M})\,\boldsymbol{I}_m, \qquad \mathbb{E}[\boldsymbol{X}^\top\boldsymbol{N}\boldsymbol{X}] = \operatorname{tr}(\boldsymbol{N})\,\boldsymbol{I}_n. \tag{23}$$

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 $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{W}^\top$ with $\boldsymbol{W}$ as in Assumption 5. For every $m \in \mathbb{N}$ there is a scalar $\alpha_m(d,k)$ with*

$$\mathbb{E}[\boldsymbol{S}^m] = \alpha_m(d,k)\,\boldsymbol{I}_d, \qquad \alpha_m(d,k) = \tfrac{1}{d}\,\mathbb{E}[\operatorname{tr}(\boldsymbol{S}^m)]. \tag{24}$$

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

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

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

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

$$\mathbb{E}[X_1\cdots X_{2n}] = \sum_{P}\prod_{\{i,j\}\in P}\mathbb{E}[X_iX_j], \tag{25}$$

the sum over all pairings P of $\{1,\dots,2n\}$; odd moments vanish (Isserlis, 1918; Wick, 1950).

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 (1997, §1.4). ∎

B.5 Proofs

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

$$\boldsymbol{G}\boldsymbol{G}^\top = \boldsymbol{B}\boldsymbol{S}\boldsymbol{B} = \boldsymbol{S}^3 - \boldsymbol{S}^2\boldsymbol{M}^\star - \boldsymbol{M}^\star\boldsymbol{S}^2 + \boldsymbol{M}^\star\boldsymbol{S}\boldsymbol{M}^\star. \tag{26}$$

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

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

and substituting $\boldsymbol{B} = \boldsymbol{S} - \boldsymbol{M}^\star$ and expanding gives (26). ∎

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

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

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

∎

Lemma 6. *Under Assumption 5, $\mathbb{E}[\boldsymbol{S}^2] = k(k+d+1)\,\boldsymbol{I}_d$, hence $\alpha_2(d,k) = k(k+d+1)$.*

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

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

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

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

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

Lemma 7. *Under Assumption 5,* $\mathbb{E}[\boldsymbol{S}^3] = \big[k^3 + 3(d+1)k^2 + (d^2+3d+4)k\big]\boldsymbol{I}_d$, *hence* $\alpha_3(d,k)$ *as in* (20).

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

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

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

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

where for the middle pattern, e.g. $a = b \neq c$ gives $\|\boldsymbol{w}_a\|^2 \langle \boldsymbol{w}_a, \boldsymbol{w}_c\rangle^2$ and, conditioning on $\boldsymbol{w}_a$, $\langle \boldsymbol{w}_a, \boldsymbol{w}_c\rangle \mid \boldsymbol{w}_a \sim \mathcal{N}(0, \|\boldsymbol{w}_a\|^2)$; 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 (26). Since $\boldsymbol{M}^\star$ is deterministic, Lemma 2 gives

$$\begin{aligned} \mathbb{E}[\boldsymbol{S}^3] &= \alpha_3 \boldsymbol{I}_d,\\ \mathbb{E}[\boldsymbol{S}^2\boldsymbol{M}^\star] = \mathbb{E}[\boldsymbol{M}^\star\boldsymbol{S}^2] &= \alpha_2 \boldsymbol{M}^\star,\\ \mathbb{E}[\boldsymbol{M}^\star\boldsymbol{S}\boldsymbol{M}^\star] = \boldsymbol{M}^\star\,\mathbb{E}[\boldsymbol{S}]\,\boldsymbol{M}^\star &= k\,(\boldsymbol{M}^\star)^2, \end{aligned}$$

so summing gives (17); the values of $\alpha_1, \alpha_2, \alpha_3$ are Lemmas 5 to 7. Corollary 1 follows from the entries of (17): 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. (2026) 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* $\boldsymbol{M}^\star$.

C Deep Linear Networks

C.1 Setup

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

$$f(\boldsymbol{x}) = \boldsymbol{W}_L\boldsymbol{W}_{L-1}\cdots\boldsymbol{W}_1\boldsymbol{x}, \qquad \boldsymbol{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}, \tag{27}$$

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

$$\mathcal{L}(\boldsymbol{W}_1, \dots, \boldsymbol{W}_L) \;:=\; \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{x}}\,\|\boldsymbol{W}_L\cdots\boldsymbol{W}_1\boldsymbol{x} - \boldsymbol{\Phi}\boldsymbol{x}\|^2\,. \tag{28}$$

Define the layer-i forward, backward and residual factors

$$\boldsymbol{B}_i := \boldsymbol{W}_{i-1}\cdots\boldsymbol{W}_1 \in \mathbb{R}^{d_{i-1}\times d_0}, \qquad \boldsymbol{A}_i := \boldsymbol{W}_L\cdots\boldsymbol{W}_{i+1} \in \mathbb{R}^{d_L \times d_i}, \qquad \boldsymbol{R} := \boldsymbol{W}_L\cdots\boldsymbol{W}_1 - \boldsymbol{\Phi}, \tag{29}$$

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

$$\boldsymbol{G}_i \;:=\; \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_i} \;=\; \boldsymbol{A}_i^\top \boldsymbol{R}\,\boldsymbol{B}_i^\top \;\in\; \mathbb{R}^{d_i \times d_{i-1}}. \tag{30}$$

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

$$r_1 = 1, \qquad v_1 = d_0^2, \tag{31}$$

$$r_i = v_{i-1} + (d_{i-1}+1)d_{i-2}r_{i-1}, \qquad v_i = d_{i-1}^2 v_{i-1} + 2d_{i-1}d_{i-2}r_{i-1} \quad (i \geq 2); \tag{32}$$

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

$$s_L = 1, \qquad u_L = d_L^2, \tag{33}$$

$$s_{i-1} = u_i + (d_{i-1}+1)d_i s_i, \qquad u_{i-1} = d_{i-1}^2 u_i + 2d_{i-1}d_i s_i \quad (i \leq L); \tag{34}$$

and the composite Frobenius coefficient is

$$V_i \;:=\; d_{i-1}r_i \;=\; \mathbb{E}\big[\big\|\boldsymbol{B}_i\boldsymbol{B}_i^\top\big\|_{\mathrm{F}}^2\big]. \tag{35}$$

C.2 Standing assumption

Assumption 6. *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 $\boldsymbol{W}_1, \dots, \boldsymbol{W}_L$. Expectations $\mathbb{E}[\cdot]$ are over $\boldsymbol{W}_1, \dots, \boldsymbol{W}_L$ ($\boldsymbol{x}$ has been integrated out in $\mathcal{L}$).*

C.3 Main results restated

Theorem (Theorem 2, restated). *Under Assumption 6, for every layer $i \in \{1, \dots, L\}$,*

$$\mathbb{E}[\boldsymbol{G}_i \boldsymbol{G}_i^\top] = V_i \, s_i \, \boldsymbol{I}_{d_i} + \boldsymbol{T}_i, \qquad \boldsymbol{T}_i = \begin{cases} \Big(\prod_{j=1}^{L-1} d_j\Big) \boldsymbol{\Phi}\boldsymbol{\Phi}^\top, & i = L, \\ \Big(\prod_{j=1,\, j\neq i}^{L-1} d_j\Big) \, \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 \, \boldsymbol{I}_{d_i}, & 1 \le i < L, \end{cases} \tag{36}$$

with V_i, s_i from (31)–(35).

Corollary 2. *Under Assumption 6, the diagonal entries of $\mathbb{E}[\boldsymbol{G}_i \boldsymbol{G}_i^\top]$ are*

$$\mathbb{E}\big[(\boldsymbol{G}_i \boldsymbol{G}_i^\top)_{aa}\big] = \begin{cases} V_i s_i + \big(\prod_{j\neq i,\, 1\le j\le L-1} d_j\big) \, \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2, & 1 \le i < L, \\ V_L s_L + \big(\prod_{j=1}^{L-1} d_j\big)(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{aa}, & i = L, \end{cases} \tag{37}$$

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 $i < L$. Moreover,

$$V_i s_i = \Theta\Big(d_i \prod_{j=1,\, j\neq i}^{L-1} d_j^2\Big) \quad (1 \le i < L), \qquad V_L s_L = \Theta\Big(\prod_{j=1}^{L-1} d_j^2\Big). \tag{38}$$

C.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*

$$\mathbb{E}\big[\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top\big] = \big(\mathrm{tr}(\boldsymbol{M})^2 + (m+1)\,\|\boldsymbol{M}\|_{\mathrm{F}}^2\big)\boldsymbol{I}_m, \tag{39}$$

$$\mathbb{E}\big[\boldsymbol{X}^\top \boldsymbol{N}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{N}\boldsymbol{X}\big] = \big(\mathrm{tr}(\boldsymbol{N})^2 + (n+1)\,\|\boldsymbol{N}\|_{\mathrm{F}}^2\big)\boldsymbol{I}_n. \tag{40}$$

Proof. We prove (39); (40) 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 $\boldsymbol{M} = \boldsymbol{M}^\top$,

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

Summing over $b = c$,

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

■

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*

$$\mathbb{E}\big[\mathrm{tr}(\boldsymbol{X}\boldsymbol{M}\boldsymbol{X}^\top)^2\big] = m^2\,\mathrm{tr}(\boldsymbol{M})^2 + 2m\,\|\boldsymbol{M}\|_{\mathrm{F}}^2. \tag{41}$$

Proof. Expanding,

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

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 $\boldsymbol{M}\boldsymbol{M}$) $m^2\,\mathrm{tr}(\boldsymbol{M})^2$, $m\,\|\boldsymbol{M}\|_{\mathrm{F}}^2$ and $m\,\mathrm{tr}(\boldsymbol{M}^2) = m\,\|\boldsymbol{M}\|_{\mathrm{F}}^2$, which sum to (41). ■

C.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 6, for every i,*

$$\mathbb{E}\big[\boldsymbol{G}_i\boldsymbol{G}_i^\top \mid \boldsymbol{A}_i, \boldsymbol{B}_i\big] = \mathrm{tr}\big((\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big)\,(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2 + \boldsymbol{A}_i^\top\boldsymbol{\Phi}\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{\Phi}^\top\boldsymbol{A}_i. \tag{42}$$

Proof. Substituting $\boldsymbol{R} = \boldsymbol{A}_i\boldsymbol{W}_i\boldsymbol{B}_i - \boldsymbol{\Phi}$ into (30),

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

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 $\boldsymbol{K} = \boldsymbol{B}_i\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{B}_i^\top$,

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

which yields (42). ∎

Lemma 13. *Under Assumption 6, for every $i \in \{1, \dots, L\}$,*

$$\mathbb{E}\big[(\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big] = r_i\,\boldsymbol{I}_{d_{i-1}}, \qquad \mathbb{E}\big[\mathrm{tr}(\boldsymbol{B}_i\boldsymbol{B}_i^\top)^2\big] = v_i, \tag{43}$$

with (r_i, v_i) from (31)–(32).

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 $\mathrm{tr}(\boldsymbol{B}_1\boldsymbol{B}_1^\top)^2 = d_0^2$, 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,

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

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

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

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

Lemma 14. *Under Assumption 6, for every $i \in \{0, \dots, L\}$,*

$$\mathbb{E}\big[(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2\big] = s_i\,\boldsymbol{I}_{d_i}, \qquad \mathbb{E}\big[\mathrm{tr}(\boldsymbol{A}_i^\top\boldsymbol{A}_i)^2\big] = u_i, \tag{44}$$

with (s_i, u_i) from (33)–(34).

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),

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

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 6, $\mathbb{E}[\boldsymbol{A}_i^\top\boldsymbol{\Phi}\boldsymbol{B}_i^\top\boldsymbol{B}_i\boldsymbol{\Phi}^\top\boldsymbol{A}_i] = \boldsymbol{T}_i$, with $\boldsymbol{T}_i$ as in (36).*

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,

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

Iterating,

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

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

Proof of Theorem 2. By Lemma 12,

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

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

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

and the second term equals $\boldsymbol{T}_i$ by Lemma 15. Corollary 2 follows from the entries of (36): 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.

It remains to prove (38). Let $\pi_i := \prod_{j=0}^{i-1} d_j$ and $\sigma_i := \sum_{m=0}^{i-1} d_m^{-1}$, so that $\pi_i = d_{i-1}\pi_{i-1}$ and $\sigma_i = \sigma_{i-1} + d_{i-1}^{-1}$. Since $V_i = d_{i-1}r_i$, the recursion (31)–(32) reads

$$\begin{aligned} V_1 &= d_0, & v_1 &= d_0^2, \\ V_i &= d_{i-1}v_{i-1} + d_{i-1}(d_{i-1}+1)V_{i-1}, & v_i &= d_{i-1}^2 v_{i-1} + 2d_{i-1}V_{i-1} \qquad (i \ge 2). \end{aligned}$$

Let $a_i := v_i/\pi_i^2$ and $b_i := V_i/(\pi_i^2\sigma_i)$, so $a_1 = b_1 = 1$. Dividing the two recursions by π_i^2 and by $\pi_i^2\sigma_i$ gives

$$a_i = a_{i-1} + \frac{2\sigma_{i-1}}{d_{i-1}}\,b_{i-1}, \qquad b_i = \frac{d_{i-1}^{-1}}{\sigma_i}\,a_{i-1} + \frac{\sigma_{i-1}}{\sigma_i}\,(1 + d_{i-1}^{-1})b_{i-1}.$$

The second identity expresses b_i as a convex combination of a_{i-1} and $(1 + d_{i-1}^{-1})b_{i-1}$, so $\min(a_{i-1}, b_{i-1}) \le b_i \le 2\max(a_{i-1}, b_{i-1})$; the first gives $a_{i-1} \le a_i \le a_{i-1} + 2(i-1)b_{i-1}$, using $\sigma_{i-1} \le i-1$ and $d_{i-1} \ge 1$. By induction, $1 \le a_i, b_i \le (2L)^L$ for every $i \le L$, that is,

$$v_i = \Theta(\pi_i^2), \qquad V_i = \Theta(\pi_i^2\sigma_i),$$

The recursion (33)–(34) has the same form read from the output side: with $S_i := d_i s_i$, $\rho_i := \prod_{j=i}^{L} d_j$ and $\tau_i := \sum_{m=i}^{L} d_m^{-1}$,

$$\begin{aligned} S_L &= d_L, & u_L &= d_L^2, \\ S_{i-1} &= d_{i-1}u_i + d_{i-1}(d_{i-1}+1)S_i, & u_{i-1} &= d_{i-1}^2 u_i + 2d_{i-1}S_i \qquad (i \le L), \end{aligned}$$

and the same argument gives $u_i = \Theta(\rho_i^2)$ and $S_i = \Theta(\rho_i^2\tau_i)$. Therefore

$$V_i s_i = \frac{V_i S_i}{d_i} = \Theta\Big(\frac{\pi_i^2\rho_i^2}{d_i}\,\sigma_i\tau_i\Big), \qquad \frac{\pi_i^2\rho_i^2}{d_i} = \frac{1}{d_i}\Big(\prod_{j=0}^{L} d_j\Big)^2.$$

Since $d_0^{-1} \le \sigma_i \le L$ and $d_L^{-1} \le \tau_i \le L$, the factor $\sigma_i\tau_i$ is bounded above and below, and $\big(\prod_{j=0}^{L} d_j\big)^2/d_i$ equals $d_0^2 d_L^2\, d_i \prod_{j=1,\,j\ne i}^{L-1} d_j^2$ for $i < L$ and $d_0^2 d_L \prod_{j=1}^{L-1} d_j^2$ for $i = L$, which gives (38). ∎

D Two-Layer ReLU Networks

D.1 Setup

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

$$f(\boldsymbol{x}) = \boldsymbol{W}_2\,\sigma(\boldsymbol{W}_1\boldsymbol{x}), \qquad \boldsymbol{W}_1 \in \mathbb{R}^{d_1\times d_0}, \quad \boldsymbol{W}_2 \in \mathbb{R}^{d_2\times d_1}, \tag{45}$$

acts on $\boldsymbol{x} \in \mathbb{R}^{d_0}$ via $\sigma(z) = \max(z, 0)$, $\sigma'(z) = \mathbf{1}[z > 0]$. Let $\boldsymbol{h}(\boldsymbol{x}) := \sigma(\boldsymbol{W}_1\boldsymbol{x}) \in \mathbb{R}^{d_1}$, and let $\boldsymbol{w}_a \in \mathbb{R}^{d_0}$ denote the a-th row of $\boldsymbol{W}_1$. The population loss against a deterministic linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_2 \times d_0}$ is

$$\mathcal{L}(\boldsymbol{W}_1, \boldsymbol{W}_2) := \tfrac{1}{2}\,\mathbb{E}_{\boldsymbol{x}} \left\| \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x} \right\|^2, \tag{46}$$

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

$$\kappa_1 := \mathbb{E}\left[\frac{\sin^2\theta + (\pi - \theta)^2}{4\pi^2} + \frac{\sin\theta\cos\theta(\pi - \theta)}{2\pi^2} \right], \tag{47}$$

$$\kappa_2 := \frac{\mathbb{E}[\sin\theta(\pi - \theta)]}{2\pi^2}. \tag{48}$$

D.2 Standing assumption

Assumption 7. *$(\boldsymbol{W}_1)_{ij}$ and $(\boldsymbol{W}_2)_{ij}$ are i.i.d. $\mathcal{N}(0, 1)$; $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_{d_0})$; $\boldsymbol{\Phi}$ is deterministic and independent of $\boldsymbol{W}_1, \boldsymbol{W}_2$. Expectations $\mathbb{E}[\cdot]$ are over $\boldsymbol{W}_1, \boldsymbol{W}_2$ ($\boldsymbol{x}$ has been integrated out in $\mathcal{L}$).*

D.3 Main results restated

Theorem (Theorem 3, restated). *Under Assumption 7, with $\boldsymbol{H} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top]$ and $\alpha := \mathbb{E}_{\boldsymbol{W}_1}[\mathrm{tr}(\boldsymbol{H}^2)]$,*

$$\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top] = \alpha\,\boldsymbol{I}_{d_2} + \tfrac{d_1}{4}\,\boldsymbol{\Phi}\boldsymbol{\Phi}^\top, \qquad \alpha = \Theta(d_1^2), \tag{49}$$

and $\mathbb{E}[\boldsymbol{G}_1\boldsymbol{G}_1^\top]$ is permutation-invariant in the hidden index with

$$\mathbb{E}\big[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{aa}\big] = \tfrac{d_0 d_2^2}{4} + d_0 d_2\big[(d_1 - 1)\kappa_1 + \tfrac{1}{2}\big] + \tfrac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}, \tag{50}$$

$$\mathbb{E}\big[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{ac}\big] = d_0 d_2\,\kappa_2 + O(d_2\sqrt{d_0}) \qquad (a \neq c). \tag{51}$$

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

D.4 Preliminary lemmas

Lemma 16. *Same as Lemma 1.*

Lemma 17 (Gaussian integration by parts)**.** *Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_n)$ 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 $\mathbb{E}|x_j f(\boldsymbol{x})| < \infty$, $\mathbb{E}|\partial_j f(\boldsymbol{x})| < \infty$. Then*

$$\mathbb{E}[x_j f(\boldsymbol{x})] = \mathbb{E}[\partial_j f(\boldsymbol{x})]. \tag{52}$$

In particular this holds whenever f is Lipschitz.

Proof. A standard identity (Stein, 1981; Nourdin & Peccati, 2012). The standard Gaussian density φ satisfies $\partial_j\varphi(\boldsymbol{x}) = -x_j\varphi(\boldsymbol{x})$, so by Fubini and one-dimensional integration by parts in x_j (the boundary term vanishes by integrability),

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

The Lipschitz case follows from Rademacher's theorem. ∎

Remark 2. *$z \mapsto \sigma(z)$ is 1-Lipschitz, so $f(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top\boldsymbol{x})$ falls under Lemma 17 with $\partial_j f(\boldsymbol{x}) = \mathbf{1}[\boldsymbol{w}^\top\boldsymbol{x} > 0]\,w_j$ a.e. By contrast $g(\boldsymbol{x}) = \mathbf{1}[\boldsymbol{w}^\top\boldsymbol{x} > 0]q(\boldsymbol{x})$ with q not vanishing on $\{\boldsymbol{w}^\top\boldsymbol{x} = 0\}$ 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 $s > 0$,*

$$\mathbb{E}[\sigma(Z)] = \frac{s}{\sqrt{2\pi}}, \qquad \mathbb{E}[\sigma(Z)^2] = \frac{s^2}{2}, \qquad \mathbb{E}[\sigma'(Z)] = \frac{1}{2}. \tag{53}$$

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 $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n \setminus \{0\}$, $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $\theta := \angle(\boldsymbol{u}, \boldsymbol{v})$. Then (Cho & Saul, 2009)*

$$\mathbb{E}[\sigma(\boldsymbol{u}^\top \boldsymbol{x})\sigma(\boldsymbol{v}^\top \boldsymbol{x})] = \frac{\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|}{2\pi}\big(\sin\theta + (\pi - \theta)\cos\theta\big), \tag{54}$$

$$\mathbb{E}[\sigma'(\boldsymbol{u}^\top \boldsymbol{x})\sigma'(\boldsymbol{v}^\top \boldsymbol{x})] = \frac{\pi - \theta}{2\pi}. \tag{55}$$

Proof. Both quantities depend on $(\boldsymbol{u}^\top \boldsymbol{x}, \boldsymbol{v}^\top \boldsymbol{x})$, bivariate Gaussian with variances $\|\boldsymbol{u}\|^2, \|\boldsymbol{v}\|^2$ and correlation $\cos\theta$. Equation (55) is $\Pr(\boldsymbol{u}^\top \boldsymbol{x} > 0,\, \boldsymbol{v}^\top \boldsymbol{x} > 0)$, 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 (54), by Price's theorem (Price, 1958),

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

at $\rho = -1$ (i.e. $\theta = \pi$) the product vanishes a.s., and integrating ∂_ρ from -1 to $\cos\theta$ gives (54). ∎

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

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

$$\mathbb{E}\big[\boldsymbol{x}\boldsymbol{x}^\top \mathbf{1}[\boldsymbol{u}^\top \boldsymbol{x} > 0]\big] = \tfrac{1}{2}\,\boldsymbol{I}_n. \tag{56}$$

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

D.5 Proof of the gradient self outer-product of $\boldsymbol{W}_2$

Lemma 22. $\boldsymbol{G}_2 = \boldsymbol{W}_2\boldsymbol{H} - \boldsymbol{\Phi}\boldsymbol{C}^\top$, *where* $\boldsymbol{H} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top] \in \mathbb{R}^{d_1 \times d_1}$ *and* $\boldsymbol{C} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{h}(\boldsymbol{x})\boldsymbol{x}^\top] \in \mathbb{R}^{d_1 \times d_0}$.

Proof. Differentiating (46) in $\boldsymbol{W}_2$,

$$\boldsymbol{G}_2 = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{W}_2\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top] = \boldsymbol{W}_2\,\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top] - \boldsymbol{\Phi}\,\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{h}(\boldsymbol{x})^\top] = \boldsymbol{W}_2\boldsymbol{H} - \boldsymbol{\Phi}\boldsymbol{C}^\top.$$

∎

Lemma 23. $\boldsymbol{C} = \frac{1}{2}\boldsymbol{W}_1$.

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

$$\boldsymbol{C}_{aj} = \mathbb{E}_{\boldsymbol{x}}[\partial_j f(\boldsymbol{x})] = (\boldsymbol{w}_a)_j\,\mathbb{E}_{\boldsymbol{x}}[\sigma'(\boldsymbol{w}_a^\top \boldsymbol{x})] = \tfrac{1}{2}(\boldsymbol{w}_a)_j.$$

∎

Proof of (49). By Lemma 22,

$$\boldsymbol{G}_2\boldsymbol{G}_2^\top = \boldsymbol{W}_2\boldsymbol{H}^2\boldsymbol{W}_2^\top - \boldsymbol{W}_2\boldsymbol{H}\boldsymbol{C}\boldsymbol{\Phi}^\top - \boldsymbol{\Phi}\boldsymbol{C}^\top\boldsymbol{H}\boldsymbol{W}_2^\top + \boldsymbol{\Phi}\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{\Phi}^\top.$$

Since $\mathbb{E}[\boldsymbol{W}_2] = 0$, the cross terms vanish under $\mathbb{E}_{\boldsymbol{W}_2}[\cdot \mid \boldsymbol{W}_1]$, and by Lemma 16 (left identity, $\boldsymbol{M} = \boldsymbol{H}^2$),

$$\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top \mid \boldsymbol{W}_1] = \operatorname{tr}(\boldsymbol{H}^2)\,\boldsymbol{I}_{d_2} + \boldsymbol{\Phi}\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{\Phi}^\top. \tag{57}$$

By Lemma 23, $\boldsymbol{C}^\top\boldsymbol{C} = \frac{1}{4}\boldsymbol{W}_1^\top\boldsymbol{W}_1$, and Lemma 16 (right identity, $\boldsymbol{N} = \boldsymbol{I}_{d_1}$) gives $\mathbb{E}_{\boldsymbol{W}_1}[\boldsymbol{W}_1^\top\boldsymbol{W}_1] = d_1\boldsymbol{I}_{d_0}$, so $\mathbb{E}_{\boldsymbol{W}_1}[\boldsymbol{C}^\top\boldsymbol{C}] = \frac{d_1}{4}\boldsymbol{I}_{d_0}$. Taking $\mathbb{E}_{\boldsymbol{W}_1}$ of (57),

$$\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top] = \mathbb{E}_{\boldsymbol{W}_1}[\operatorname{tr}(\boldsymbol{H}^2)]\,\boldsymbol{I}_{d_2} + \tfrac{d_1}{4}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \alpha\,\boldsymbol{I}_{d_2} + \tfrac{d_1}{4}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top.$$

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

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

By (54), $\boldsymbol{H}_{ab} = \|\boldsymbol{w}_a\|\,\|\boldsymbol{w}_b\|\; g(\theta_{ab})$ with $g(\theta) = (\sin\theta + (\pi-\theta)\cos\theta)/(2\pi)$ and $\theta_{ab} = \angle(\boldsymbol{w}_a, \boldsymbol{w}_b)$ independent of $\|\boldsymbol{w}_a\|, \|\boldsymbol{w}_b\|$, so $\mathbb{E}[\boldsymbol{H}_{ab}^2] = d_0^2\,\mathbb{E}[g(\theta)^2]$, where $\mathbb{E}[g(\theta)^2] > 0$, and

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

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

D.6 Proof of the gradient self outer-product of $\boldsymbol{W}_1$

By the chain rule, with residual $\boldsymbol{r}(\boldsymbol{x}) := \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}) - \boldsymbol{\Phi}\boldsymbol{x}$ and $\boldsymbol{D}(\boldsymbol{x}) := \operatorname{diag}(\sigma'(\boldsymbol{W}_1\boldsymbol{x}))$,

$$\begin{gathered}\boldsymbol{G}_1 = \boldsymbol{G}_1^{(1)} - \boldsymbol{G}_1^{(2)},\\ \boldsymbol{G}_1^{(1)} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{D}(\boldsymbol{x})\boldsymbol{W}_2^\top\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x})\boldsymbol{x}^\top], \qquad \boldsymbol{G}_1^{(2)} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{D}(\boldsymbol{x})\boldsymbol{W}_2^\top\boldsymbol{\Phi}\boldsymbol{x}\boldsymbol{x}^\top].\end{gathered} \tag{58}$$

$\boldsymbol{G}_1^{(1)}$ is quadratic in $\boldsymbol{W}_2$ and $\boldsymbol{G}_1^{(2)}$ is linear, so the cross terms in $\boldsymbol{G}_1\boldsymbol{G}_1^\top$ are odd in $\boldsymbol{W}_2$ and vanish under $\mathbb{E}_{\boldsymbol{W}_2}[\cdot \mid \boldsymbol{W}_1]$:

$$\mathbb{E}_{\boldsymbol{W}_2}[\boldsymbol{G}_1\boldsymbol{G}_1^\top \mid \boldsymbol{W}_1] = \mathbb{E}_{\boldsymbol{W}_2}[\boldsymbol{G}_1^{(1)}(\boldsymbol{G}_1^{(1)})^\top \mid \boldsymbol{W}_1] + \mathbb{E}_{\boldsymbol{W}_2}[\boldsymbol{G}_1^{(2)}(\boldsymbol{G}_1^{(2)})^\top \mid \boldsymbol{W}_1]. \tag{59}$$

Lemma 24. *$\boldsymbol{G}_1^{(2)} = \frac{1}{2}\boldsymbol{W}_2^\top\boldsymbol{\Phi}$, and $\mathbb{E}_{\boldsymbol{W}_2}[\boldsymbol{G}_1^{(2)}(\boldsymbol{G}_1^{(2)})^\top \mid \boldsymbol{W}_1] = \frac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}\boldsymbol{I}_{d_1}$.*

Proof. Let $\boldsymbol{v}_a := \boldsymbol{\Phi}^\top(\boldsymbol{W}_2)_{:,a} \in \mathbb{R}^{d_0}$, so $(\boldsymbol{W}_2^\top\boldsymbol{\Phi})_{ab} = (\boldsymbol{v}_a)_b$. Then, by Lemma 21,

$$(\boldsymbol{G}_1^{(2)})_{ab} = \mathbb{E}_{\boldsymbol{x}}[\mathbf{1}[\boldsymbol{w}_a^\top\boldsymbol{x} > 0]\,x_b\,(\boldsymbol{x}^\top\boldsymbol{v}_a)] = \big(\mathbb{E}_{\boldsymbol{x}}[\mathbf{1}[\boldsymbol{w}_a^\top\boldsymbol{x} > 0]\,\boldsymbol{x}\boldsymbol{x}^\top]\,\boldsymbol{v}_a\big)_b = \tfrac{1}{2}(\boldsymbol{v}_a)_b,$$

so $\boldsymbol{G}_1^{(2)} = \frac{1}{2}\boldsymbol{W}_2^\top\boldsymbol{\Phi}$ and $\boldsymbol{G}_1^{(2)}(\boldsymbol{G}_1^{(2)})^\top = \frac{1}{4}\boldsymbol{W}_2^\top\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{W}_2$; Lemma 16 (right identity, $\boldsymbol{N} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$) gives $\mathbb{E}_{\boldsymbol{W}_2}[\boldsymbol{W}_2^\top\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{W}_2] = \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\,\boldsymbol{I}_{d_1}$. ∎

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

$$F_b(\boldsymbol{u},\boldsymbol{v}) = u_b \cdot \frac{\|\boldsymbol{v}\|\sin\theta}{2\pi\,\|\boldsymbol{u}\|} + v_b \cdot \frac{\pi-\theta}{2\pi}, \qquad \theta := \angle(\boldsymbol{u},\boldsymbol{v}). \tag{60}$$

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

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

By Lemma 19, $K(\boldsymbol{u},\boldsymbol{v}) = \frac{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}{2\pi}(\sin\theta + (\pi-\theta)\cos\theta)$, smooth for $\boldsymbol{u} \notin \operatorname{span}(\boldsymbol{v})$; differentiating with $\partial\|\boldsymbol{u}\|/\partial u_b = u_b/\|\boldsymbol{u}\|$, $\partial\cos\theta/\partial u_b = \frac{1}{\|\boldsymbol{u}\|}(v_b/\|\boldsymbol{v}\| - \cos\theta\,u_b/\|\boldsymbol{u}\|)$, $\partial\theta/\partial u_b = -\frac{1}{\sin\theta}\partial\cos\theta/\partial u_b$, and simplifying yields (60). ∎

Lemma 26. $\mathbb{E}_{\boldsymbol{W}_2}[(\boldsymbol{W}_2^\top\boldsymbol{W}_2)_{ak}(\boldsymbol{W}_2^\top\boldsymbol{W}_2)_{cl}] = d_2^2\,\delta_{ak}\delta_{cl} + d_2\,\delta_{ac}\delta_{kl} + d_2\,\delta_{al}\delta_{ck}$.

Proof. $(\boldsymbol{W}_2^\top\boldsymbol{W}_2)_{ak} = \sum_{i=1}^{d_2}(\boldsymbol{W}_2)_{ia}(\boldsymbol{W}_2)_{ik}$, and applying Lemma 20 to the four Gaussian factors $(\boldsymbol{W}_2)_{ia}, (\boldsymbol{W}_2)_{ik}, (\boldsymbol{W}_2)_{jc}, (\boldsymbol{W}_2)_{jl}$ (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 (50)–(51). Permutation invariance follows from the exchangeability of the rows of $\boldsymbol{W}_1$ and columns of $\boldsymbol{W}_2$. The $\boldsymbol{G}_1^{(2)}$ contribution to (59) is $\frac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}\boldsymbol{I}_{d_1}$ by Lemma 24. For $\boldsymbol{G}_1^{(1)}$, we have $(\boldsymbol{G}_1^{(1)})_{ab} = \sum_{k=1}^{d_1}(\boldsymbol{W}_2^\top \boldsymbol{W}_2)_{ak} F_b(\boldsymbol{w}_a, \boldsymbol{w}_k)$, so

$$\mathbb{E}_{\boldsymbol{W}_2}\big[(\boldsymbol{G}_1^{(1)}(\boldsymbol{G}_1^{(1)})^\top)_{ac} \;\big|\; \boldsymbol{W}_1\big] = \sum_{b,k,l} \mathbb{E}_{\boldsymbol{W}_2}[(\boldsymbol{W}_2^\top \boldsymbol{W}_2)_{ak}(\boldsymbol{W}_2^\top \boldsymbol{W}_2)_{cl}]\, F_b(\boldsymbol{w}_a, \boldsymbol{w}_k) F_b(\boldsymbol{w}_c, \boldsymbol{w}_l),$$

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

$$T_1^{(ac)} = d_2^2 \textstyle\sum_b F_b(\boldsymbol{w}_a, \boldsymbol{w}_a) F_b(\boldsymbol{w}_c, \boldsymbol{w}_c), \tag{61}$$

$$T_2^{(ac)} = d_2\, \delta_{ac} \textstyle\sum_{k=1}^{d_1} \sum_b F_b(\boldsymbol{w}_a, \boldsymbol{w}_k)^2, \tag{62}$$

$$T_3^{(ac)} = d_2 \textstyle\sum_b F_b(\boldsymbol{w}_a, \boldsymbol{w}_c) F_b(\boldsymbol{w}_c, \boldsymbol{w}_a). \tag{63}$$

Term T_1. By Lemma 25, $F_b(\boldsymbol{w}_a, \boldsymbol{w}_a) = \frac{1}{2}(\boldsymbol{w}_a)_b$, so $T_1^{(ac)} = \frac{d_2^2}{4}\boldsymbol{w}_a^\top \boldsymbol{w}_c$, 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 $\boldsymbol{u} \neq \boldsymbol{v}$, let $A := \frac{\|\boldsymbol{v}\| \sin\theta}{2\pi\|\boldsymbol{u}\|}$ and $B := \frac{\pi-\theta}{2\pi}$ in (60); using $\boldsymbol{u}^\top \boldsymbol{v} = \|\boldsymbol{u}\|\,\|\boldsymbol{v}\| \cos\theta$,

$$\sum_b F_b(\boldsymbol{u}, \boldsymbol{v})^2 = A^2 \|\boldsymbol{u}\|^2 + 2AB(\boldsymbol{u}^\top \boldsymbol{v}) + B^2 \|\boldsymbol{v}\|^2 = \|\boldsymbol{v}\|^2\, \psi(\theta),$$

$$\psi(\theta) := \frac{\sin^2\theta + (\pi-\theta)^2}{4\pi^2} + \frac{\sin\theta\cos\theta(\pi-\theta)}{2\pi^2}.$$

In $T_2^{(aa)}$ the term $k = a$ contributes $\sum_b F_b(\boldsymbol{w}_a, \boldsymbol{w}_a)^2 = \frac{1}{4}\|\boldsymbol{w}_a\|^2$ (expectation $\frac{d_0}{4}$), and each $k \neq a$ contributes $\mathbb{E}[\|\boldsymbol{w}_k\|^2]\mathbb{E}[\psi(\theta_{ak})] = d_0\kappa_1$, 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 $a = c$, $\sum_b F_b(\boldsymbol{w}_a, \boldsymbol{w}_a)^2 = \frac{1}{4}\|\boldsymbol{w}_a\|^2$, so $\mathbb{E}[T_3^{(aa)}] = \frac{d_0 d_2}{4}$. For $a \neq c$, expanding $F_b(\boldsymbol{w}_a, \boldsymbol{w}_c)F_b(\boldsymbol{w}_c, \boldsymbol{w}_a)$ via (60) (with $\theta := \theta_{ac}$) and summing over b,

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

Here $\|\boldsymbol{w}_a\|\,\|\boldsymbol{w}_c\|$ is independent of θ_{ac} with $\mathbb{E}[\|\boldsymbol{w}_a\|\,\|\boldsymbol{w}_c\|] = d_0 + O(1)$; 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 $\boldsymbol{G}_1^{(2)}$ contribution and merging the two $\frac{d_0 d_2}{4}$ pieces,

$$\mathbb{E}\big[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{aa}\big] = \tfrac{d_0 d_2^2}{4} + d_0 d_2\big[(d_1-1)\kappa_1 + \tfrac{1}{2}\big] + \tfrac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4};$$

for $a \neq c$ only $T_3^{(ac)}$ survives, $\mathbb{E}[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{ac}] = d_0 d_2 \kappa_2 + O(d_2\sqrt{d_0})$. 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. ∎

D.7 Discussion

At initialization the gradient of $\boldsymbol{W}_2$ is mostly target-independent: the $\Theta(d_1^2)$ leading order of $\mathbb{E}[(\boldsymbol{G}_2\boldsymbol{G}_2^\top)_{aa}]$ comes entirely from $\mathrm{tr}(\boldsymbol{H}^2)$, which depends only on $\boldsymbol{W}_1$, while the target term $\frac{d_1}{4}(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{aa}$ is dominated unless $\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2$ is exceptionally large. The nonzero $\Theta(1)$ off-diagonal of $\mathbb{E}[\boldsymbol{G}_1\boldsymbol{G}_1^\top]$ comes from T_3 and reflects ReLU's lack of odd symmetry: in a deep linear network the symmetry $\boldsymbol{W}_1 \mapsto \boldsymbol{D}\boldsymbol{W}_1$, $\boldsymbol{W}_2 \mapsto \boldsymbol{W}_2\boldsymbol{D}$ for $\boldsymbol{D} = \mathrm{diag}(\pm 1)$ preserves the loss and forces off-diagonal expectations to zero (cf. Appendix C), whereas $\sigma(-z) \neq -\sigma(z)$ breaks it. The dominance is therefore asymptotic in d_1 rather than exact.

E SINGLE-LAYER LINEAR ATTENTION

E.1 SETUP

Let T, d_0, d_k, d_v, d_2 be positive integers. Single-layer linear attention with tokens $\boldsymbol{X} \in \mathbb{R}^{T\times d_0}$ (one per row) and weights $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d_k\times d_0}$, $\boldsymbol{W}_V \in \mathbb{R}^{d_v\times d_0}$, $\boldsymbol{W}_O \in \mathbb{R}^{d_2\times d_v}$, and its loss against a deterministic linear target $\boldsymbol{\Phi} \in \mathbb{R}^{d_2\times d_0}$, are

$$\boldsymbol{A} := \frac{1}{\sqrt{d_k}}\,\boldsymbol{X}\boldsymbol{W}_Q^\top\boldsymbol{W}_K\boldsymbol{X}^\top, \qquad f(\boldsymbol{X}) = \boldsymbol{A}\,\boldsymbol{X}\boldsymbol{W}_V^\top\boldsymbol{W}_O^\top, \qquad \mathcal{L} := \tfrac{1}{2}\,\big\|f(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{\Phi}^\top\big\|_{\mathrm{F}}^2. \tag{64}$$

Let

$$\boldsymbol{Z} := \boldsymbol{X}\boldsymbol{W}_V^\top, \qquad \boldsymbol{R} := f(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{\Phi}^\top, \qquad \boldsymbol{Y} := \boldsymbol{X}\boldsymbol{\Phi}^\top, \qquad \boldsymbol{C} := \boldsymbol{X}^\top\boldsymbol{X}, \qquad c_p := \mathrm{tr}(\boldsymbol{C}^p), \tag{65}$$

with $c_0 = \mathrm{tr}(\boldsymbol{C}^0) = d_0$, and define the three data moments

$$\begin{aligned} \mathsf{A}_1 &:= \mathbb{E}[c_1^2c_4 + c_2c_4 + c_2^3], \qquad \mathsf{A}_2 := \mathbb{E}[c_1^2c_2^2 + 2c_2c_4], \\ \mathsf{A}_3 &:= \mathbb{E}[c_1^2c_4 + c_1^2c_2^2 + 3c_2c_4 + c_2^3] = \mathsf{A}_1 + \mathsf{A}_2, \end{aligned} \tag{66}$$

which are explicit polynomials in (T, d_0) (Lemma 33). The gradients $\boldsymbol{G}_\bullet := \partial\mathcal{L}/\partial\boldsymbol{W}_\bullet$, $\bullet \in \{Q, K, V, O\}$, are

$$\begin{aligned} \boldsymbol{G}_Q &= \tfrac{1}{\sqrt{d_k}}\,\boldsymbol{W}_K\boldsymbol{X}^\top\boldsymbol{Z}\boldsymbol{W}_O^\top\boldsymbol{R}^\top\boldsymbol{X}, & \boldsymbol{G}_K &= \tfrac{1}{\sqrt{d_k}}\,\boldsymbol{W}_Q\boldsymbol{X}^\top\boldsymbol{R}\boldsymbol{W}_O\boldsymbol{Z}^\top\boldsymbol{X}, \\ \boldsymbol{G}_V &= \boldsymbol{W}_O^\top\boldsymbol{R}^\top\boldsymbol{A}\boldsymbol{X}, & \boldsymbol{G}_O &= \boldsymbol{R}^\top\boldsymbol{A}\boldsymbol{Z}. \end{aligned} \tag{67}$$

Muon acts per weight matrix, so the objects of study are the four gradient self outer-products $\boldsymbol{G}_Q\boldsymbol{G}_Q^\top$ and $\boldsymbol{G}_K\boldsymbol{G}_K^\top$ (both $d_k \times d_k$), $\boldsymbol{G}_V\boldsymbol{G}_V^\top$ ($d_v \times d_v$) and $\boldsymbol{G}_O\boldsymbol{G}_O^\top$ ($d_2 \times d_2$), as in Appendices B to D.

E.2 STANDING ASSUMPTION

Assumption 8. *$\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O$ are mutually independent with i.i.d. $\mathcal{N}(0,1)$ entries; the rows of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$, independent of the weights, so $\boldsymbol{C} \sim \mathcal{W}_{d_0}(T, \boldsymbol{I}_{d_0})$; $\boldsymbol{\Phi}$ is deterministic. Expectations $\mathbb{E}[\cdot]$ are over data and weights jointly, and $\mathbb{E}[\cdot \mid \boldsymbol{X}]$ over the weights only.*

This is Assumption 4 of the main text; fan-in scaling is treated in Remark 6.

E.3 MAIN RESULTS RESTATED

Every d_k dependence below enters through the single ratio $\mathsf{A}_2/\mathsf{A}_1$ of the moments (66).

Theorem (Theorem 4, restated). *Under Assumption 8, $\mathbb{E}[\boldsymbol{G}_Q\boldsymbol{G}_Q^\top] = c_Q\boldsymbol{I}_{d_k}$, $\mathbb{E}[\boldsymbol{G}_K\boldsymbol{G}_K^\top] = c_K\boldsymbol{I}_{d_k}$ and $\mathbb{E}[\boldsymbol{G}_V\boldsymbol{G}_V^\top] = c_V\boldsymbol{I}_{d_v}$ with*

$$\begin{aligned} c_Q &= \frac{1}{d_k}\Big(\mathbb{E}[c_2\beta_1] + d_v\gamma\,\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\Big) + \frac{1}{d_k^2}\Big(\mathbb{E}[c_2\beta_1] + \mathbb{E}[c_2\beta_2]\Big), \\ c_K &= \frac{1}{d_k}\Big(\mathbb{E}[c_2\beta_1] + d_v\gamma\,\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2\Big) + \frac{1}{d_k^2}\,\mathbb{E}[(c_1^2+c_2)\beta_1], \\ c_V &= d_2(d_2+d_v+1)\Big(\mathsf{A}_1 + \tfrac{1}{d_k}\mathsf{A}_3\Big) + \gamma\,\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2, \end{aligned} \tag{68}$$

where β_1, β_2 are the polynomials (75); and

$$\mathbb{E}\big[\boldsymbol{G}_O\boldsymbol{G}_O^\top\big] = \bar{\alpha}_O\,\boldsymbol{I}_{d_2} + d_v\,\gamma\,\boldsymbol{\Phi}\boldsymbol{\Phi}^\top, \tag{69}$$

with

$$\gamma = \tfrac{1}{d_0}\,\mathbb{E}[c_2^2] = T\Big[(Td_0+8)(T+d_0)^2 + 2(Td_0+10)(T+d_0) + 5(Td_0+4)\Big], \tag{70}$$

$$\bar{\alpha}_O = d_v(d_v+1)\,\mathsf{A}_1 + d_v\,\mathsf{A}_2 + \frac{d_v(d_v+1)}{d_k}\,\mathsf{A}_3 + \frac{2d_v}{d_k}\,\mathsf{A}_1 = \Theta(d_v^2d_0^6). \tag{71}$$

Corollary 3. *Under Assumption 8 and for fixed $\mathbf{\Phi}$,*

$$c_Q \asymp c_K = \Theta\big(d_2^2 d_v^2 d_0^6/d_k\big), \qquad c_V = \Theta\big(d_2(d_2{+}d_v)\, d_0^6\big), \qquad \bar{\alpha}_O = \Theta\big(d_v^2 d_0^6\big),$$

with $c_K - c_Q = \big(\mathbb{E}[c_1^2\beta_1] - \mathbb{E}[c_2\beta_2]\big)/d_k^2$ of relative size $O(1/d_k)$; and for $a \neq b$ with $(\mathbf{\Phi}\mathbf{\Phi}^\top)_{ab} \neq 0$,

$$\frac{\mathbb{E}[(\boldsymbol{G}_O \boldsymbol{G}_O^\top)_{aa}]}{\mathbb{E}[(\boldsymbol{G}_O \boldsymbol{G}_O^\top)_{ab}]} = \Theta\big(d_v d_0^3\big).$$

Thus the hidden dimension d_v and the token dimension d_0 drive the diagonal dominance of $\mathbb{E}[\boldsymbol{G}_O \boldsymbol{G}_O^\top]$, while d_k enters c_V, $\bar{\alpha}_O$ and the $\boldsymbol{W}_O$ ratio only through bounded factors.

E.4 Preliminary lemmas

Lemmas 29 to 32 are stated conditionally on $\boldsymbol{X}$, i.e. with $\boldsymbol{C}$ fixed and $c_p = \operatorname{tr}(\boldsymbol{C}^p)$; Lemma 33 supplies the moments of $\boldsymbol{C} \sim \mathcal{W}_{d_0}(T, \boldsymbol{I}_{d_0})$ needed to average over $\boldsymbol{X}$.

Lemma 27. *Same as Lemma 1.*

Lemma 28 (Wick/Isserlis)**.** *Same as Lemma 3.*

Lemma 29. *Let $\boldsymbol{W} \in \mathbb{R}^{n\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries, let $\boldsymbol{S} := \boldsymbol{W}^\top \boldsymbol{W} \sim \mathcal{W}_d(n, \boldsymbol{I}_d)$, and let $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ be deterministic symmetric. Then*

$$\mathbb{E}[S_{ab}S_{cd}] = n^2\delta_{ab}\delta_{cd} + n(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}), \qquad \mathbb{E}[\boldsymbol{S}\boldsymbol{M}\boldsymbol{S}] = n(n{+}1)\boldsymbol{M} + n\operatorname{tr}(\boldsymbol{M})\,\boldsymbol{I}_d, \quad (72)$$

hence $\mathbb{E}[\operatorname{tr}(\boldsymbol{M}\boldsymbol{S}\boldsymbol{M}\boldsymbol{S})] = n(n{+}1)\operatorname{tr}(\boldsymbol{M}^2) + n(\operatorname{tr}\boldsymbol{M})^2$ and $\mathbb{E}[\boldsymbol{S}^2] = n(n{+}d{+}1)\boldsymbol{I}_d$.

Proof. Since $S_{ab} = \sum_\alpha W_{\alpha a}W_{\alpha b}$, Lemma 28 gives

$$\mathbb{E}[S_{ab}S_{cd}] = \sum_{\alpha,\beta} \mathbb{E}[W_{\alpha a}W_{\alpha b}W_{\beta c}W_{\beta d}] = \sum_{\alpha,\beta}\big(\delta_{ab}\delta_{cd} + \delta_{\alpha\beta}\delta_{ac}\delta_{bd} + \delta_{\alpha\beta}\delta_{ad}\delta_{bc}\big),$$

which is the first identity. Contracting it against M_{bc} and using $\boldsymbol{M} = \boldsymbol{M}^\top$,

$$\mathbb{E}[(\boldsymbol{S}\boldsymbol{M}\boldsymbol{S})_{ad}] = \sum_{b,c} M_{bc}\,\mathbb{E}[S_{ab}S_{cd}] = n^2 M_{ad} + nM_{ad} + n\,\delta_{ad}\operatorname{tr}(\boldsymbol{M}),$$

which is the second identity; contracting once more against M_{da} gives the trace identity, and $\boldsymbol{M} = \boldsymbol{I}_d$ gives $\mathbb{E}[\boldsymbol{S}^2]$. ∎

Lemma 30. *Under Assumption 8, conditionally on $\boldsymbol{X}$, let $\boldsymbol{N} := \boldsymbol{W}_Q \boldsymbol{C} \boldsymbol{W}_Q^\top \in \mathbb{R}^{d_k\times d_k}$. Then*

$$\begin{aligned} \mathbb{E}[N_{ij}N_{kl} \mid \boldsymbol{X}] &= c_1^2\,\delta_{ij}\delta_{kl} + c_2(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \\ \mathbb{E}[(\operatorname{tr}\boldsymbol{N})^2 \mid \boldsymbol{X}] = d_k^2 c_1^2 + 2d_k c_2, &\qquad \mathbb{E}[\operatorname{tr}\boldsymbol{N}^2 \mid \boldsymbol{X}] = d_k^2 c_2 + d_k(c_1^2{+}c_2). \end{aligned} \quad (73)$$

The same identities hold with $\boldsymbol{C}$ replaced by any deterministic symmetric $\boldsymbol{E} \in \mathbb{R}^{d_0\times d_0}$ and (c_1, c_2) by $(\operatorname{tr}\boldsymbol{E}, \operatorname{tr}\boldsymbol{E}^2)$, and with $\boldsymbol{W}_Q$ replaced by $\boldsymbol{W}_K$.

Proof. Since $N_{ij} = \sum_{a,b}(\boldsymbol{W}_Q)_{ia}C_{ab}(\boldsymbol{W}_Q)_{jb}$, Lemma 28 gives

$$\mathbb{E}[(\boldsymbol{W}_Q)_{ia}(\boldsymbol{W}_Q)_{jb}(\boldsymbol{W}_Q)_{kc}(\boldsymbol{W}_Q)_{ld}] = \delta_{ij}\delta_{ab}\delta_{kl}\delta_{cd} + \delta_{ik}\delta_{ac}\delta_{jl}\delta_{bd} + \delta_{il}\delta_{ad}\delta_{jk}\delta_{bc},$$

and contracting against $C_{ab}C_{cd}$, with $\sum_{a,b} C_{ab}C_{ab} = \operatorname{tr}(\boldsymbol{C}^2) = c_2$, gives the first line of (73). Summing it over $i = j$ and $k = l$ gives $\mathbb{E}[(\operatorname{tr}\boldsymbol{N})^2 \mid \boldsymbol{X}]$; summing it over $j = k$ and $l = i$ gives $\mathbb{E}[\operatorname{tr}\boldsymbol{N}^2 \mid \boldsymbol{X}]$. Only the symmetry of $\boldsymbol{C}$ and the law of $\boldsymbol{W}_Q$ were used, which gives the last assertion. (These are Lemmas 10 and 11 with $m = d_k$ and $\boldsymbol{M} = \boldsymbol{C}$.) ∎

Lemma 31. *Under Assumption 8, conditionally on $\boldsymbol{X}$, let $\boldsymbol{N} \in \mathbb{R}^{d_k\times d_k}$ be symmetric and independent of $\boldsymbol{W}_K$, and let $\boldsymbol{S} := \boldsymbol{W}_K^\top \boldsymbol{N} \boldsymbol{W}_K$. Then $\mathbb{E}[S_{ab}S_{cd} \mid \boldsymbol{X}, \boldsymbol{N}] = (\operatorname{tr}\boldsymbol{N})^2\delta_{ab}\delta_{cd} + \operatorname{tr}(\boldsymbol{N}^2)(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc})$, and*

$$\begin{aligned} \mathbb{E}[\operatorname{tr}(\boldsymbol{S}\boldsymbol{C}^2\boldsymbol{S}\boldsymbol{C}^2) \mid \boldsymbol{X}, \boldsymbol{N}] &= (\operatorname{tr}\boldsymbol{N})^2 c_4 + \operatorname{tr}(\boldsymbol{N}^2)(c_4{+}c_2^2), \\ \mathbb{E}[(\operatorname{tr}(\boldsymbol{S}\boldsymbol{C}^2))^2 \mid \boldsymbol{X}, \boldsymbol{N}] &= (\operatorname{tr}\boldsymbol{N})^2 c_2^2 + 2\operatorname{tr}(\boldsymbol{N}^2)\, c_4. \end{aligned} \quad (74)$$

Proof. Since $S_{ab} = \sum_{i,j} (\boldsymbol{W}_K)_{ia} N_{ij} (\boldsymbol{W}_K)_{jb}$, Lemma 28 applied to $\mathbb{E}[(\boldsymbol{W}_K)_{ia}(\boldsymbol{W}_K)_{jb}(\boldsymbol{W}_K)_{kc}(\boldsymbol{W}_K)_{ld}]$ as in the proof of Lemma 30, followed by contraction against $N_{ij}N_{kl}$ with $\boldsymbol{N} = \boldsymbol{N}^\top$, gives the first identity. Contracting the first identity against $(\boldsymbol{C}^2)_{bc}(\boldsymbol{C}^2)_{da}$, the three terms give $(\operatorname{tr}\boldsymbol{N})^2 \operatorname{tr}(\boldsymbol{C}^4)$, $\operatorname{tr}(\boldsymbol{N}^2)\operatorname{tr}(\boldsymbol{C}^4)$ and $\operatorname{tr}(\boldsymbol{N}^2)\operatorname{tr}(\boldsymbol{C}^2)^2$, which is the first line of (74); contracting against $(\boldsymbol{C}^2)_{ba}(\boldsymbol{C}^2)_{dc}$ instead, they give $(\operatorname{tr}\boldsymbol{N})^2\operatorname{tr}(\boldsymbol{C}^2)^2$, $\operatorname{tr}(\boldsymbol{N}^2)\operatorname{tr}(\boldsymbol{C}^4)$ and $\operatorname{tr}(\boldsymbol{N}^2)\operatorname{tr}(\boldsymbol{C}^4)$, which is the second line. ∎

Lemma 32. *Under Assumption 8, conditionally on $\boldsymbol{X}$, let $\boldsymbol{H}_O := \boldsymbol{W}_O^\top \boldsymbol{W}_O \sim \mathcal{W}_{d_v}(d_2, \boldsymbol{I}_{d_v})$ and $\boldsymbol{E} := \boldsymbol{C}\boldsymbol{W}_V^\top \boldsymbol{H}_O \boldsymbol{W}_V \boldsymbol{C}$, and define*

$$\begin{aligned} \beta_1 &:= d_2 d_v \Big[(d_2 d_v + d_2 + d_v + 3)\, c_4 + (d_2 + d_v + 1)\, c_2^2 \Big], \\ \beta_2 &:= d_2 d_v \Big[(d_2 d_v + 2)\, c_2^2 + 2(d_2 + d_v + 1)\, c_4 \Big]. \end{aligned} \tag{75}$$

Then $\mathbb{E}[\operatorname{tr}(\boldsymbol{E}^2) \mid \boldsymbol{X}] = \beta_1$ and $\mathbb{E}[(\operatorname{tr}\boldsymbol{E})^2 \mid \boldsymbol{X}] = \beta_2$.

Proof. Let $\boldsymbol{S}_V := \boldsymbol{W}_V^\top \boldsymbol{H}_O \boldsymbol{W}_V$, so that $\operatorname{tr}\boldsymbol{E} = \operatorname{tr}(\boldsymbol{S}_V \boldsymbol{C}^2)$ and $\operatorname{tr}\boldsymbol{E}^2 = \operatorname{tr}(\boldsymbol{S}_V \boldsymbol{C}^2 \boldsymbol{S}_V \boldsymbol{C}^2)$. Since $\boldsymbol{H}_O$ is symmetric and independent of $\boldsymbol{W}_V$, Lemma 31 applies with $(\boldsymbol{W}_V, \boldsymbol{H}_O)$ in place of $(\boldsymbol{W}_K, \boldsymbol{N})$:

$$\begin{aligned} \mathbb{E}[\operatorname{tr}\boldsymbol{E}^2 \mid \boldsymbol{X}, \boldsymbol{H}_O] &= (\operatorname{tr}\boldsymbol{H}_O)^2 c_4 + \operatorname{tr}(\boldsymbol{H}_O^2)(c_4 + c_2^2), \\ \mathbb{E}[(\operatorname{tr}\boldsymbol{E})^2 \mid \boldsymbol{X}, \boldsymbol{H}_O] &= (\operatorname{tr}\boldsymbol{H}_O)^2 c_2^2 + 2\operatorname{tr}(\boldsymbol{H}_O^2)\, c_4. \end{aligned}$$

By Lemma 29 with $(n, d) = (d_2, d_v)$, $\mathbb{E}[(\operatorname{tr}\boldsymbol{H}_O)^2] = d_2 d_v (d_2 d_v + 2)$ and $\mathbb{E}[\operatorname{tr}\boldsymbol{H}_O^2] = d_2 d_v (d_2 + d_v + 1)$; substituting gives β_1 and β_2. ∎

Lemma 33 (Wishart trace moments)**.** *For $\boldsymbol{C} \sim \mathcal{W}_{d_0}(T, \boldsymbol{I}_{d_0})$: $\mathbb{E}[c_1] = T d_0$, $\mathbb{E}[c_2] = T d_0 (T + d_0 + 1)$, and*

$$\mathbb{E}[c_4] = T d_0 \Big[(T+d_0)^3 + 6(T+d_0)^2 + 3(T d_0 + 7)(T+d_0) + 5(T d_0 + 4) \Big], \tag{76}$$

$$\mathbb{E}[c_2^2] = T d_0 \Big[(T d_0 + 8)(T+d_0)^2 + 2(T d_0 + 10)(T+d_0) + 5(T d_0 + 4) \Big], \tag{77}$$

$$\begin{aligned} \mathbb{E}[c_2 c_4] = T d_0 \Big[&(T d_0 + 16)(T+d_0)^4 + (7 T d_0 + 136)(T+d_0)^3 \\ &+ \big(3T^2 d_0^2 + 99 T d_0 + 736\big)(T+d_0)^2 + \big(8 T^2 d_0^2 + 305 T d_0 + 1608\big)(T+d_0) \\ &+ 21 T^2 d_0^2 + 420 T d_0 + 1344 \Big], \end{aligned} \tag{78}$$

$$\begin{aligned} \mathbb{E}[c_2^3] = T d_0 \Big[&\big(T^2 d_0^2 + 24 T d_0 + 160\big)(T+d_0)^3 + \big(3T^2 d_0^2 + 84 T d_0 + 704\big)(T+d_0)^2 \\ &+ \big(15 T^2 d_0^2 + 344 T d_0 + 1664\big)(T+d_0) + 13 T^2 d_0^2 + 380 T d_0 + 1312 \Big], \end{aligned} \tag{79}$$

$$\mathbb{E}[c_1^2 c_4] = (T d_0 + 8)(T d_0 + 10)\, \mathbb{E}[c_4], \qquad \mathbb{E}[c_1^2 c_2^2] = (T d_0 + 8)(T d_0 + 10)\, \mathbb{E}[c_2^2]. \tag{80}$$

Consequently $\mathsf{A}_3 = \mathsf{A}_1 + \mathsf{A}_2$. As polynomials in d_0, all six moments have positive coefficients, of degree 4 for $\mathbb{E}[c_4], \mathbb{E}[c_2^2]$ and degree 6 for the other four; hence $\mathbb{E}[c_4], \mathbb{E}[c_2^2] = \Theta(d_0^4)$, $\mathbb{E}[c_2 c_4], \mathbb{E}[c_2^3], \mathbb{E}[c_1^2 c_4], \mathbb{E}[c_1^2 c_2^2] = \Theta(d_0^6)$, and therefore $\gamma = \frac{1}{d_0}\mathbb{E}[c_2^2] = \Theta(d_0^3)$ and $\mathsf{A}_1, \mathsf{A}_2 = \Theta(d_0^6)$.

Proof. Throughout the proof let $s := T + d_0$ and $q := T d_0$. Since $\boldsymbol{C} = \boldsymbol{X}^\top \boldsymbol{X}$ with $\boldsymbol{X} \in \mathbb{R}^{T \times d_0}$ standard Gaussian, every moment below is the expectation of a polynomial in the entries of $\boldsymbol{X}$, and Gaussian integration by parts, $\mathbb{E}[X_{ta}\Psi(\boldsymbol{X})] = \mathbb{E}[\partial\Psi/\partial X_{ta}]$ for polynomial Ψ, applies.

Step 1: a recursion in the total degree. For $k \geq 1$ and $g := \prod_{i=1}^n c_{m_i}$ with $m_i \geq 1$ ($g = 1$ if $n = 0$),

$$\mathbb{E}[c_k g] = \mathbb{E}\Big[\Big((T+k-1)\, c_{k-1} + \sum_{j=1}^{k-1} c_j\, c_{k-1-j} \Big) g \Big] + 2 \sum_{i=1}^{n} m_i\, \mathbb{E}\Big[c_{k+m_i-1} \prod_{l \neq i} c_{m_l} \Big]. \tag{81}$$

Indeed, $c_k = \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{C}^{k-1}) = \sum_{t,a} X_{ta} (\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta}$, so integration by parts in each X_{ta} gives

$$\mathbb{E}[c_k g] = \sum_{t,a} \mathbb{E}\Big[\frac{\partial (\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta}}{\partial X_{ta}}\, g \Big] + \sum_{t,a} \mathbb{E}\Big[(\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta}\, \frac{\partial g}{\partial X_{ta}} \Big]. \tag{82}$$

From $\partial C_{bc}/\partial X_{ta} = \delta_{ab}X_{tc} + \delta_{ac}X_{tb}$ and the product rule, for $m \geq 1$,

$$\frac{\partial (\boldsymbol{C}^m)_{bc}}{\partial X_{ta}} = \sum_{r=0}^{m-1} \Big[(\boldsymbol{C}^r)_{ba}\,(\boldsymbol{X}\boldsymbol{C}^{m-1-r})_{tc} + (\boldsymbol{C}^r\boldsymbol{X}^\top)_{bt}\,(\boldsymbol{C}^{m-1-r})_{ac} \Big], \qquad \frac{\partial c_m}{\partial X_{ta}} = \sum_b \frac{\partial (\boldsymbol{C}^m)_{bb}}{\partial X_{ta}} = 2m\,(\boldsymbol{X}\boldsymbol{C}^{m-1})_{ta}. \tag{83}$$

Since $(\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta} = \sum_b X_{tb}(\boldsymbol{C}^{k-1})_{ba}$, (83) with $m = k-1$ gives

$$\begin{aligned}\sum_{t,a} \frac{\partial (\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta}}{\partial X_{ta}} &= \sum_{t,a} (\boldsymbol{C}^{k-1})_{aa} + \sum_{r=0}^{k-2} \Big[\sum_{t,a} (\boldsymbol{X}\boldsymbol{C}^r)_{ta} (\boldsymbol{X}\boldsymbol{C}^{k-2-r})_{ta} \Big] \\ &\quad + \sum_{r=0}^{k-2} \operatorname{tr}(\boldsymbol{X}\boldsymbol{C}^r\boldsymbol{X}^\top)\,\operatorname{tr}(\boldsymbol{C}^{k-2-r}) \\ &= T\,c_{k-1} + (k-1)\,c_{k-1} + \sum_{r=0}^{k-2} c_{r+1}\,c_{k-2-r} = (T+k-1)\,c_{k-1} + \sum_{j=1}^{k-1} c_j\,c_{k-1-j},\end{aligned}$$

since $\sum_{t,a}(\boldsymbol{X}\boldsymbol{C}^r)_{ta}(\boldsymbol{X}\boldsymbol{C}^{k-2-r})_{ta} = \operatorname{tr}(\boldsymbol{C}^r\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{C}^{k-2-r}) = c_{k-1}$ and $\operatorname{tr}(\boldsymbol{X}\boldsymbol{C}^r\boldsymbol{X}^\top) = c_{r+1}$; this is the first term of (81). For the second, $\partial g/\partial X_{ta} = \sum_i (\partial c_{m_i}/\partial X_{ta}) \prod_{l\neq i} c_{m_l}$ and, by (83),

$$\begin{aligned}\sum_{t,a} (\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta}\,\frac{\partial c_m}{\partial X_{ta}} &= 2m \sum_{t,a} (\boldsymbol{X}\boldsymbol{C}^{k-1})_{ta}(\boldsymbol{X}\boldsymbol{C}^{m-1})_{ta} \\ &= 2m\,\operatorname{tr}(\boldsymbol{C}^{k-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{C}^{m-1}) = 2m\,c_{k+m-1}.\end{aligned}$$

Step 2: moments containing c_1, and (80). For $k = 1$ the first bracket of (81) is $Tc_0 = q$ and each c_{m_i} is reproduced, so

$$\mathbb{E}[c_1 g] = \big(q + 2\deg g\big)\mathbb{E}[g], \qquad \deg g := \textstyle\sum_i m_i, \tag{84}$$

and iterating, $\mathbb{E}[c_1^2 F] = (q+2d+2)(q+2d)\,\mathbb{E}[F]$ for F homogeneous of degree d in $\boldsymbol{C}$; $d = 4$ gives (80).

Step 3: degree at most four. By (81) with $g = 1$ and $k = 1, 2, 3$, and by (84),

$$\begin{aligned}&\mathbb{E}[c_1] = q, \qquad \mathbb{E}[c_1^2] = (q+2)\,q, \qquad \mathbb{E}[c_2] = (T+1+d_0)\mathbb{E}[c_1] = q(s+1), \\ &\mathbb{E}[c_1c_2] = (q+4)\mathbb{E}[c_2] = q(q+4)(s+1), \\ &\mathbb{E}[c_3] = (T+2+d_0)\mathbb{E}[c_2] + \mathbb{E}[c_1^2] = q\big[(s+1)(s+2) + q + 2\big] = q\big[s^2 + 3s + q + 4\big].\end{aligned}$$

Then (81) with $(k, g) = (4, 1)$ and $(2, c_2)$ gives

$$\begin{aligned}\mathbb{E}[c_4] &= (T+3+d_0)\mathbb{E}[c_3] + 2\mathbb{E}[c_1c_2] \\ &= q\big[(s+3)(s^2+3s+q+4) + 2(s+1)(q+4)\big] = q\big[s^3 + 6s^2 + 3(q+7)s + 5(q+4)\big], \\ \mathbb{E}[c_2^2] &= (T+1+d_0)\mathbb{E}[c_1c_2] + 4\mathbb{E}[c_3] \\ &= q\big[(s+1)^2(q+4) + 4(s^2+3s+q+4)\big] = q\big[(q+8)s^2 + 2(q+10)s + 5(q+4)\big],\end{aligned}$$

which are (76) and (77).

Step 4: degree five. By (81) with $(k, g) = (5, 1)$ and $(3, c_2)$, and by (84) for $\mathbb{E}[c_1c_3] = (q+6)\mathbb{E}[c_3]$ and $\mathbb{E}[c_1^2c_2] = (q+4)(q+6)\mathbb{E}[c_2]$,

$$\begin{aligned}\mathbb{E}[c_5] &= (T+4+d_0)\mathbb{E}[c_4] + 2\mathbb{E}[c_1c_3] + \mathbb{E}[c_2^2] \\ &= q\big[s^4 + 10s^3 + (6q+65)s^2 + 5(5q+32)s + (q+4)(2q+37)\big], \\ \mathbb{E}[c_2c_3] &= (T+2+d_0)\mathbb{E}[c_2^2] + \mathbb{E}[c_1^2c_2] + 4\mathbb{E}[c_4] \\ &= q\big[(q+12)s^3 + 4(q+15)s^2 + (q+7)(q+24)s + (q+4)(q+36)\big].\end{aligned}$$

Step 5: degree six. By (81) with $(k, g) = (4, c_2)$ and $(2, c_2^2)$, and by (84) for $\mathbb{E}[c_1 c_2^2] = (q{+}8)\mathbb{E}[c_2^2]$,

$$\mathbb{E}[c_2 c_4] = (T{+}3{+}d_0)\mathbb{E}[c_2 c_3] + 2\mathbb{E}[c_1 c_2^2] + 4\mathbb{E}[c_5] = (s{+}3)\mathbb{E}[c_2 c_3] + 2(q{+}8)\mathbb{E}[c_2^2] + 4\mathbb{E}[c_5],$$
$$\mathbb{E}[c_2^3] = (T{+}1{+}d_0)\mathbb{E}[c_1 c_2^2] + 8\mathbb{E}[c_2 c_3] = (s{+}1)(q{+}8)\mathbb{E}[c_2^2] + 8\mathbb{E}[c_2 c_3];$$

substituting Steps 3 and 4 and collecting powers of s gives (78) and (79).

Step 6: orders in d_0. All coefficients above are positive, so each moment is Θ of its highest power of d_0, read off from the displayed formulas: $d_0 \cdot d_0^3$ for $\mathbb{E}[c_4]$, $d_0 \cdot d_0 \cdot d_0^2$ for $\mathbb{E}[c_2^2]$, $d_0 \cdot d_0 \cdot d_0^4$ for $\mathbb{E}[c_2 c_4]$, $d_0 \cdot d_0^2 \cdot d_0^3$ for $\mathbb{E}[c_2^3]$, and d_0^2 times $\mathbb{E}[c_4]$ and $\mathbb{E}[c_2^2]$ for (80). Since A_1 and A_2 are positive combinations of degree-six moments by (66), they are $\Theta(d_0^6)$; $\mathsf{A}_3 = \mathsf{A}_1 + \mathsf{A}_2$ holds by definition. ■

E.5 Proofs

The proof of Theorem 4 has three steps: the isotropy of $\mathbb{E}[\boldsymbol{G}_\bullet \boldsymbol{G}_\bullet^\top \mid \boldsymbol{X}]$ for $\bullet \in \{Q, K, V\}$ (Lemma 34); the conditional closed forms of $\mathbb{E}[\boldsymbol{G}_O \boldsymbol{G}_O^\top \mid \boldsymbol{X}]$ (Lemma 35) and of the isotropic constants (Lemma 36); and the average over $\boldsymbol{X}$.

Lemma 34 (Isotropy)**.** *Under Assumption 8, conditionally on $\boldsymbol{X}$,*

$$\mathbb{E}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top \mid \boldsymbol{X}] = c_Q(\boldsymbol{C})\, \boldsymbol{I}_{d_k}, \qquad \mathbb{E}[\boldsymbol{G}_K \boldsymbol{G}_K^\top \mid \boldsymbol{X}] = c_K(\boldsymbol{C})\, \boldsymbol{I}_{d_k},$$
$$\mathbb{E}[\boldsymbol{G}_V \boldsymbol{G}_V^\top \mid \boldsymbol{X}] = c_V(\boldsymbol{C})\, \boldsymbol{I}_{d_v},$$

with $c_Q(\boldsymbol{C}) := \frac{1}{d_k}\mathbb{E}[\|\boldsymbol{G}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}]$, $c_K(\boldsymbol{C}) := \frac{1}{d_k}\mathbb{E}[\|\boldsymbol{G}_K\|_{\mathrm{F}}^2 \mid \boldsymbol{X}]$ and $c_V(\boldsymbol{C}) := \frac{1}{d_v}\mathbb{E}[\|\boldsymbol{G}_V\|_{\mathrm{F}}^2 \mid \boldsymbol{X}]$.

Proof. Fix $\boldsymbol{X}$ and let $W := (\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O)$. Let $\boldsymbol{D} \in \mathbb{R}^{d_k \times d_k}$ be diagonal with entries in $\{\pm 1\}$, and let $\tau_{\boldsymbol{D}}$ denote the map $(\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O) \mapsto (\boldsymbol{D}\boldsymbol{W}_Q, \boldsymbol{D}\boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O)$. Since $(\boldsymbol{D}\boldsymbol{W}_Q)^\top(\boldsymbol{D}\boldsymbol{W}_K) = \boldsymbol{W}_Q^\top \boldsymbol{W}_K$, the matrix $\boldsymbol{A}$ in (64), and hence $\mathcal{L}$, is invariant: $\mathcal{L} \circ \tau_{\boldsymbol{D}} = \mathcal{L}$. Differentiating this identity with respect to $\boldsymbol{W}_Q$ by the chain rule gives $\boldsymbol{D}^\top \boldsymbol{G}_Q(\tau_{\boldsymbol{D}} W) = \boldsymbol{G}_Q(W)$, i.e.

$$\boldsymbol{G}_Q(\tau_{\boldsymbol{D}} W) = \boldsymbol{D}\, \boldsymbol{G}_Q(W). \tag{85}$$

Under Assumption 8, $\tau_{\boldsymbol{D}} W \overset{d}{=} W$, so

$$\mathbb{E}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top \mid \boldsymbol{X}] = \mathbb{E}\big[\boldsymbol{G}_Q(\tau_{\boldsymbol{D}} W)\boldsymbol{G}_Q(\tau_{\boldsymbol{D}} W)^\top \,\big|\, \boldsymbol{X}\big] = \boldsymbol{D}\, \mathbb{E}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top \mid \boldsymbol{X}]\, \boldsymbol{D}.$$

For $a \neq b$, choosing $D_{aa} = 1$ and $D_{bb} = -1$ gives $\mathbb{E}[(\boldsymbol{G}_Q \boldsymbol{G}_Q^\top)_{ab} \mid \boldsymbol{X}] = -\mathbb{E}[(\boldsymbol{G}_Q \boldsymbol{G}_Q^\top)_{ab} \mid \boldsymbol{X}]$, hence 0. Likewise, for a permutation matrix $\boldsymbol{P} \in \mathbb{R}^{d_k \times d_k}$ the map $(\boldsymbol{W}_Q, \boldsymbol{W}_K) \mapsto (\boldsymbol{P}\boldsymbol{W}_Q, \boldsymbol{P}\boldsymbol{W}_K)$ preserves $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$ and the law of W, and the same argument gives $\mathbb{E}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top \mid \boldsymbol{X}] = \boldsymbol{P}\, \mathbb{E}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top \mid \boldsymbol{X}]\, \boldsymbol{P}^\top$; hence all diagonal entries are equal, and their common value is $\frac{1}{d_k}\mathbb{E}[\mathrm{tr}(\boldsymbol{G}_Q \boldsymbol{G}_Q^\top) \mid \boldsymbol{X}] = c_Q(\boldsymbol{C})$. The same argument with $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ exchanged gives the claim for $\boldsymbol{G}_K$.

For $\boldsymbol{G}_V$, let $\boldsymbol{D} \in \mathbb{R}^{d_v \times d_v}$ be diagonal with entries in $\{\pm 1\}$ and let $\tau_{\boldsymbol{D}}$ map $(\boldsymbol{W}_V, \boldsymbol{W}_O) \mapsto (\boldsymbol{D}\boldsymbol{W}_V, \boldsymbol{W}_O \boldsymbol{D})$, fixing $\boldsymbol{W}_Q, \boldsymbol{W}_K$. Since $f(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{X}(\boldsymbol{W}_O \boldsymbol{W}_V)^\top$ depends on $(\boldsymbol{W}_V, \boldsymbol{W}_O)$ only through $\boldsymbol{W}_O \boldsymbol{W}_V = (\boldsymbol{W}_O \boldsymbol{D})(\boldsymbol{D}\boldsymbol{W}_V)$, $\mathcal{L} \circ \tau_{\boldsymbol{D}} = \mathcal{L}$; differentiating with respect to $\boldsymbol{W}_V$ gives $\boldsymbol{G}_V(\tau_{\boldsymbol{D}} W) = \boldsymbol{D}\, \boldsymbol{G}_V(W)$, and $(\boldsymbol{D}\boldsymbol{W}_V, \boldsymbol{W}_O \boldsymbol{D}) \overset{d}{=} (\boldsymbol{W}_V, \boldsymbol{W}_O)$ under Assumption 8. The sign and permutation arguments above apply verbatim. ■

Lemma 35 (Conditional closed form for $\boldsymbol{W}_O$)**.** *Under Assumption 8, conditionally on $\boldsymbol{X}$,*

$$\mathbb{E}\big[\boldsymbol{G}_O \boldsymbol{G}_O^\top \,\big|\, \boldsymbol{X}\big] = \alpha_O(\boldsymbol{C})\, \boldsymbol{I}_{d_2} + d_v\, c_2\, \boldsymbol{\Phi}\, \boldsymbol{C}^2\, \boldsymbol{\Phi}^\top, \tag{86}$$

$$\begin{aligned}\alpha_O(\boldsymbol{C}) = {} & d_v(d_v{+}1)\big(c_1^2 c_4{+}c_2 c_4{+}c_2^3\big) + d_v\big(c_1^2 c_2^2{+}2c_2 c_4\big) \\ & + \tfrac{d_v(d_v+1)}{d_k}\big(c_1^2 c_4{+}c_1^2 c_2^2{+}3c_2 c_4{+}c_2^3\big) + \tfrac{2d_v}{d_k}\big(c_1^2 c_4{+}c_2 c_4{+}c_2^3\big).\end{aligned} \tag{87}$$

Proof. Let $\boldsymbol{\Omega} := \boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{A}^\top$. Substituting $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{W}_O^\top - \boldsymbol{Y}$ into $\boldsymbol{G}_O \boldsymbol{G}_O^\top = \boldsymbol{R}^\top \boldsymbol{\Omega} \boldsymbol{R}$,

$$\boldsymbol{G}_O \boldsymbol{G}_O^\top = \underbrace{\boldsymbol{W}_O \boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{\Omega} \boldsymbol{A}\boldsymbol{Z}\boldsymbol{W}_O^\top}_{\mathrm{T}_1} - \underbrace{\boldsymbol{W}_O \boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{\Omega}\, \boldsymbol{Y}}_{\mathrm{T}_2} - \underbrace{\boldsymbol{Y}^\top \boldsymbol{\Omega} \boldsymbol{A}\boldsymbol{Z}\boldsymbol{W}_O^\top}_{\mathrm{T}_3} + \underbrace{\boldsymbol{Y}^\top \boldsymbol{\Omega}\, \boldsymbol{Y}}_{\mathrm{T}_4}. \tag{88}$$

The terms $\mathrm{T}_2, \mathrm{T}_3$ are linear in $\boldsymbol{W}_O$, which is centered and independent of $(\boldsymbol{X}, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V)$, so $\mathbb{E}[\mathrm{T}_2 \mid \boldsymbol{X}] = \mathbb{E}[\mathrm{T}_3 \mid \boldsymbol{X}] = \boldsymbol{0}$.

The term T_4. By Lemma 27, $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top \mid \boldsymbol{X}, \boldsymbol{W}_Q, \boldsymbol{W}_K] = \boldsymbol{X}\,\mathbb{E}[\boldsymbol{W}_V^\top \boldsymbol{W}_V]\,\boldsymbol{X}^\top = d_v \boldsymbol{X}\boldsymbol{X}^\top$, hence

$$\mathbb{E}[\mathrm{T}_4 \mid \boldsymbol{X}, \boldsymbol{W}_Q, \boldsymbol{W}_K] = d_v\, \boldsymbol{\Phi}\big(\boldsymbol{X}^\top \boldsymbol{A}\, \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{X}\big)\boldsymbol{\Phi}^\top,$$
$$\boldsymbol{X}^\top \boldsymbol{A}\, \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{X} = \tfrac{1}{d_k} \boldsymbol{C}\boldsymbol{W}_Q^\top \boldsymbol{W}_K \boldsymbol{C}^2 \boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{C}.$$

By Lemma 27, $\mathbb{E}[\boldsymbol{W}_K \boldsymbol{C}^2 \boldsymbol{W}_K^\top \mid \boldsymbol{X}] = c_2 \boldsymbol{I}_{d_k}$ and then $\mathbb{E}[\boldsymbol{W}_Q^\top \boldsymbol{W}_Q] = d_k \boldsymbol{I}_{d_0}$, so

$$\mathbb{E}[\mathrm{T}_4 \mid \boldsymbol{X}] = d_v\, c_2\, \boldsymbol{\Phi}\boldsymbol{C}^2\boldsymbol{\Phi}^\top. \tag{89}$$

The term T_1. By Lemma 27 applied to $\boldsymbol{W}_O$,

$$\mathbb{E}[\mathrm{T}_1 \mid \boldsymbol{X}, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V] = \mathrm{tr}\big(\boldsymbol{Z}^\top \boldsymbol{A}^\top \boldsymbol{\Omega} \boldsymbol{A} \boldsymbol{Z}\big)\boldsymbol{I}_{d_2} = \mathrm{tr}\big((\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^\top)^2\big)\boldsymbol{I}_{d_2}.$$

Let $\boldsymbol{N} := \boldsymbol{W}_Q \boldsymbol{C} \boldsymbol{W}_Q^\top$, $\boldsymbol{S} := \boldsymbol{W}_K^\top \boldsymbol{N} \boldsymbol{W}_K$, $\boldsymbol{H}_V := \boldsymbol{W}_V^\top \boldsymbol{W}_V \sim \mathcal{W}_{d_0}(d_v, \boldsymbol{I}_{d_0})$ and $\boldsymbol{M} := \boldsymbol{C}\boldsymbol{S}\boldsymbol{C}$, which is symmetric. Then $\boldsymbol{A}^\top \boldsymbol{A} = \frac{1}{d_k} \boldsymbol{X}\boldsymbol{S}\boldsymbol{X}^\top$ and $\boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{X}\boldsymbol{H}_V\boldsymbol{X}^\top$, so

$$\mathrm{tr}\big((\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^\top)^2\big) = \tfrac{1}{d_k^2}\,\mathrm{tr}(\boldsymbol{M}\boldsymbol{H}_V\boldsymbol{M}\boldsymbol{H}_V).$$

Since $\boldsymbol{H}_V$ is independent of $\boldsymbol{M}$, Lemma 29 with $(n, d) = (d_v, d_0)$ gives

$$\mathbb{E}\big[\mathrm{tr}(\boldsymbol{M}\boldsymbol{H}_V\boldsymbol{M}\boldsymbol{H}_V) \,\big|\, \boldsymbol{X}, \boldsymbol{W}_Q, \boldsymbol{W}_K\big] = d_v(d_v{+}1)\,\mathrm{tr}(\boldsymbol{M}^2) + d_v(\mathrm{tr}\,\boldsymbol{M})^2.$$

Now $\mathrm{tr}\,\boldsymbol{M} = \mathrm{tr}(\boldsymbol{S}\boldsymbol{C}^2)$ and $\mathrm{tr}\,\boldsymbol{M}^2 = \mathrm{tr}(\boldsymbol{S}\boldsymbol{C}^2\boldsymbol{S}\boldsymbol{C}^2)$, so Lemma 31 (conditionally on $\boldsymbol{N}$) followed by Lemma 30 give

$$\begin{aligned}\mathbb{E}[\mathrm{tr}\,\boldsymbol{M}^2 \mid \boldsymbol{X}] &= d_k^2\big(c_1^2c_4{+}c_2c_4{+}c_2^3\big) + d_k\big(c_1^2c_4{+}c_1^2c_2^2{+}3c_2c_4{+}c_2^3\big),\\ \mathbb{E}[(\mathrm{tr}\,\boldsymbol{M})^2 \mid \boldsymbol{X}] &= d_k^2\big(c_1^2c_2^2{+}2c_2c_4\big) + 2d_k\big(c_1^2c_4{+}c_2c_4{+}c_2^3\big).\end{aligned} \tag{90}$$

Therefore $\mathbb{E}[\mathrm{T}_1 \mid \boldsymbol{X}] = \alpha_O(\boldsymbol{C})\boldsymbol{I}_{d_2}$ with $\alpha_O(\boldsymbol{C}) = \frac{1}{d_k^2}\big(d_v(d_v{+}1)\mathbb{E}[\mathrm{tr}\,\boldsymbol{M}^2 \mid \boldsymbol{X}] + d_v\mathbb{E}[(\mathrm{tr}\,\boldsymbol{M})^2 \mid \boldsymbol{X}]\big)$, which is (87); adding (89) gives (86). ∎

Lemma 36 (Conditional constants)**.** *Under Assumption 8, conditionally on* $\boldsymbol{X}$*, with* $t_\Phi := \mathrm{tr}(\boldsymbol{\Phi}\boldsymbol{C}^2\boldsymbol{\Phi}^\top)$*, the constants of Lemma 34 are*

$$c_Q(\boldsymbol{C}) = \frac{c_2}{d_k}\big(\beta_1 + d_v\, t_\Phi\big) + \frac{c_2}{d_k^2}\big(\beta_1 + \beta_2\big), \tag{91}$$

$$c_K(\boldsymbol{C}) = \frac{c_2}{d_k}\big(\beta_1 + d_v\, t_\Phi\big) + \frac{c_1^2 + c_2}{d_k^2}\,\beta_1, \tag{92}$$

$$c_V(\boldsymbol{C}) = d_2(d_2{+}d_v{+}1)\Big[\big(c_1^2c_4{+}c_2c_4{+}c_2^3\big) + \tfrac{1}{d_k}\big(c_1^2c_4{+}c_1^2c_2^2{+}3c_2c_4{+}c_2^3\big)\Big] + c_2\, t_\Phi. \tag{93}$$

Proof. Substituting $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{Z}\boldsymbol{W}_O^\top - \boldsymbol{Y}$ into (67) splits each gradient as $\boldsymbol{G}_\bullet = \boldsymbol{a}_\bullet - \boldsymbol{b}_\bullet$, where $\boldsymbol{a}_\bullet$ collects the term containing $\boldsymbol{A}\boldsymbol{Z}\boldsymbol{W}_O^\top$ and $\boldsymbol{b}_\bullet$ the term containing $\boldsymbol{Y}$. For $\bullet \in \{Q, K\}$, $\boldsymbol{a}_\bullet$ is quadratic and $\boldsymbol{b}_\bullet$ linear in $\boldsymbol{W}_O$; for $\bullet = V$, $\boldsymbol{a}_V$ is quadratic and $\boldsymbol{b}_V$ linear in $\boldsymbol{W}_O$ as well. Since $\boldsymbol{W}_O$ is centered and independent of everything else, $\mathbb{E}[\mathrm{tr}(\boldsymbol{a}_\bullet \boldsymbol{b}_\bullet^\top) \mid \boldsymbol{X}] = 0$ and

$$\mathbb{E}[\|\boldsymbol{G}_\bullet\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = \mathbb{E}[\|\boldsymbol{a}_\bullet\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] + \mathbb{E}[\|\boldsymbol{b}_\bullet\|_{\mathrm{F}}^2 \mid \boldsymbol{X}].$$

Throughout, $\boldsymbol{H}_O := \boldsymbol{W}_O^\top \boldsymbol{W}_O$, $\boldsymbol{E} := \boldsymbol{C}\boldsymbol{W}_V^\top \boldsymbol{H}_O \boldsymbol{W}_V \boldsymbol{C}$ (as in Lemma 32), $\boldsymbol{N} := \boldsymbol{W}_Q \boldsymbol{C} \boldsymbol{W}_Q^\top$ (as in Lemma 30), and we use the identities $\boldsymbol{X}^\top \boldsymbol{Z} = \boldsymbol{C}\boldsymbol{W}_V^\top$ and $\boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{X} = \frac{1}{\sqrt{d_k}} \boldsymbol{C}\boldsymbol{W}_K^\top \boldsymbol{W}_Q \boldsymbol{C}$.

Target parts. From (67), $\boldsymbol{b}_Q = \frac{1}{\sqrt{d_k}} \boldsymbol{W}_K \boldsymbol{C} \boldsymbol{W}_V^\top \boldsymbol{W}_O^\top \boldsymbol{\Phi}\boldsymbol{C}$. By Lemma 27 applied to $\boldsymbol{W}_K$ and cyclicity of the trace,

$$\mathbb{E}[\|\boldsymbol{b}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}, \boldsymbol{W}_V, \boldsymbol{W}_O] = \mathrm{tr}\big(\boldsymbol{W}_V \boldsymbol{C}^2 \boldsymbol{W}_V^\top \cdot \boldsymbol{W}_O^\top \boldsymbol{\Phi}\boldsymbol{C}^2\boldsymbol{\Phi}^\top \boldsymbol{W}_O\big),$$

and Lemma 27 applied to $\boldsymbol{W}_V$ ($\mathbb{E}[\boldsymbol{W}_V \boldsymbol{C}^2 \boldsymbol{W}_V^\top] = c_2 \boldsymbol{I}_{d_v}$) and to $\boldsymbol{W}_O$ ($\mathbb{E}[\boldsymbol{W}_O^\top \boldsymbol{\Phi}\boldsymbol{C}^2\boldsymbol{\Phi}^\top \boldsymbol{W}_O] = t_\Phi \boldsymbol{I}_{d_v}$) gives $\mathbb{E}[\|\boldsymbol{b}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = d_v c_2 t_\Phi$. The same computation applies to $\boldsymbol{b}_K =$

$\frac{1}{\sqrt{d_k}}\boldsymbol{W}_Q\boldsymbol{C}\boldsymbol{\Phi}^\top\boldsymbol{W}_O\boldsymbol{W}_V\boldsymbol{C}$ (Lemma 27 applied to $\boldsymbol{W}_Q$) and to $\boldsymbol{b}_V = \boldsymbol{W}_O^\top\boldsymbol{Y}^\top\boldsymbol{A}\boldsymbol{X}$ (Lemma 27 applied to $\boldsymbol{W}_O$, then (89) for the average of $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}^\top\boldsymbol{X}$ over $\boldsymbol{W}_K, \boldsymbol{W}_Q$), so

$$\mathbb{E}[\|\boldsymbol{b}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = \mathbb{E}[\|\boldsymbol{b}_K\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = \mathbb{E}[\|\boldsymbol{b}_V\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = d_v\, c_2\, t_\Phi.$$

Model part of $\boldsymbol{G}_Q$. Using $\boldsymbol{X}^\top\boldsymbol{Z} = \boldsymbol{C}\boldsymbol{W}_V^\top$ and $\boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{X} = \boldsymbol{W}_V\boldsymbol{X}^\top\boldsymbol{A}^\top\boldsymbol{X} = \frac{1}{\sqrt{d_k}}\boldsymbol{W}_V\boldsymbol{C}\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{C}$,

$$\boldsymbol{a}_Q = \tfrac{1}{d_k}\boldsymbol{W}_K\boldsymbol{E}\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{C} = \tfrac{1}{d_k}\boldsymbol{J}\boldsymbol{W}_Q\boldsymbol{C}, \qquad \|\boldsymbol{a}_Q\|_{\mathrm{F}}^2 = \tfrac{1}{d_k^2}\operatorname{tr}\big(\boldsymbol{J}^2\widetilde{\boldsymbol{N}}\big),$$

where $\boldsymbol{J} := \boldsymbol{W}_K\boldsymbol{E}\boldsymbol{W}_K^\top$ and $\widetilde{\boldsymbol{N}} := \boldsymbol{W}_Q\boldsymbol{C}^2\boldsymbol{W}_Q^\top$. Given $\boldsymbol{X}$, $\boldsymbol{J}$ and $\widetilde{\boldsymbol{N}}$ are independent, and $\mathbb{E}[\widetilde{\boldsymbol{N}} \mid \boldsymbol{X}] = c_2\boldsymbol{I}_{d_k}$ by Lemma 27, so $\mathbb{E}[\|\boldsymbol{a}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = \frac{c_2}{d_k^2}\mathbb{E}[\operatorname{tr}\boldsymbol{J}^2 \mid \boldsymbol{X}]$. By Lemma 30 with $\boldsymbol{E}$ in place of $\boldsymbol{C}$ and $\boldsymbol{W}_K$ in place of $\boldsymbol{W}_Q$, $\mathbb{E}[\operatorname{tr}\boldsymbol{J}^2 \mid \boldsymbol{X}, \boldsymbol{E}] = d_k^2\operatorname{tr}(\boldsymbol{E}^2) + d_k\big((\operatorname{tr}\boldsymbol{E})^2 + \operatorname{tr}(\boldsymbol{E}^2)\big)$, and by Lemma 32, $\mathbb{E}[\operatorname{tr}\boldsymbol{J}^2 \mid \boldsymbol{X}] = d_k^2\beta_1 + d_k(\beta_1{+}\beta_2)$. Hence $\mathbb{E}[\|\boldsymbol{a}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = c_2\beta_1 + \frac{c_2}{d_k}(\beta_1{+}\beta_2)$, and $c_Q(\boldsymbol{C}) = \frac{1}{d_k}\big(\mathbb{E}[\|\boldsymbol{a}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] + \mathbb{E}[\|\boldsymbol{b}_Q\|_{\mathrm{F}}^2 \mid \boldsymbol{X}]\big)$ is (91).

Model part of $\boldsymbol{G}_K$. Using $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{Z} = \frac{1}{\sqrt{d_k}}\boldsymbol{C}\boldsymbol{W}_Q^\top\boldsymbol{W}_K\boldsymbol{C}\boldsymbol{W}_V^\top$ and $\boldsymbol{Z}^\top\boldsymbol{X} = \boldsymbol{W}_V\boldsymbol{C}$,

$$\boldsymbol{a}_K = \tfrac{1}{d_k}\boldsymbol{N}\boldsymbol{W}_K\boldsymbol{E}, \qquad \|\boldsymbol{a}_K\|_{\mathrm{F}}^2 = \tfrac{1}{d_k^2}\operatorname{tr}\big(\boldsymbol{N}^2\,\boldsymbol{W}_K\boldsymbol{E}^2\boldsymbol{W}_K^\top\big).$$

Given $\boldsymbol{X}$, the matrices $\boldsymbol{N}$, $\boldsymbol{W}_K$ and $\boldsymbol{E}$ are independent. By Lemma 27, $\mathbb{E}[\boldsymbol{W}_K\boldsymbol{E}^2\boldsymbol{W}_K^\top \mid \boldsymbol{X}, \boldsymbol{E}] = \operatorname{tr}(\boldsymbol{E}^2)\boldsymbol{I}_{d_k}$; by Lemma 30, $\mathbb{E}[\operatorname{tr}\boldsymbol{N}^2 \mid \boldsymbol{X}] = d_k^2c_2 + d_k(c_1^2{+}c_2)$; and by Lemma 32, $\mathbb{E}[\operatorname{tr}\boldsymbol{E}^2 \mid \boldsymbol{X}] = \beta_1$. Hence $\mathbb{E}[\|\boldsymbol{a}_K\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = \big(c_2 + \frac{c_1^2+c_2}{d_k}\big)\beta_1$, which together with the target part gives (92).

Model part of $\boldsymbol{G}_V$. Here $\boldsymbol{a}_V = \boldsymbol{W}_O^\top\boldsymbol{W}_O\boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{H}_O\boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{X}$, so

$$\|\boldsymbol{a}_V\|_{\mathrm{F}}^2 = \operatorname{tr}(\boldsymbol{H}_O^2\boldsymbol{M}_V), \qquad \boldsymbol{M}_V := \boldsymbol{Z}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{Z} = \tfrac{1}{d_k^2}\boldsymbol{W}_V\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}^2\boldsymbol{S}\boldsymbol{C}\boldsymbol{W}_V^\top,$$

with $\boldsymbol{S}$ as in the proof of Lemma 35. Given $\boldsymbol{X}$, $\boldsymbol{H}_O$ is independent of $\boldsymbol{M}_V$ and $\mathbb{E}[\boldsymbol{H}_O^2] = d_2(d_2{+}d_v{+}1)\boldsymbol{I}_{d_v}$ by Lemma 29, so $\mathbb{E}[\|\boldsymbol{a}_V\|_{\mathrm{F}}^2 \mid \boldsymbol{X}] = d_2(d_2{+}d_v{+}1)\mathbb{E}[\operatorname{tr}\boldsymbol{M}_V \mid \boldsymbol{X}]$. Taking the expectation over $\boldsymbol{W}_V$ ($\mathbb{E}[\boldsymbol{W}_V^\top\boldsymbol{W}_V] = d_v\boldsymbol{I}_{d_0}$) and then over $\boldsymbol{W}_K, \boldsymbol{W}_Q$ by Lemmas 30 and 31,

$$\mathbb{E}[\operatorname{tr}\boldsymbol{M}_V \mid \boldsymbol{X}] = \tfrac{d_v}{d_k^2}\,\mathbb{E}[\operatorname{tr}(\boldsymbol{S}\boldsymbol{C}^2\boldsymbol{S}\boldsymbol{C}^2) \mid \boldsymbol{X}] = d_v\Big[\big(c_1^2c_4{+}c_2c_4{+}c_2^3\big) + \tfrac{1}{d_k}\big(c_1^2c_4{+}c_1^2c_2^2{+}3c_2c_4{+}c_2^3\big)\Big],$$

which is the first line of (90) divided by d_k^2. Adding the target part and dividing by d_v gives (93). ■

Proof of Theorem 4. Take expectations over $\boldsymbol{X}$ in Lemmas 34 to 36; by the tower property and Assumption 8,

$$\mathbb{E}[\boldsymbol{G}_O\boldsymbol{G}_O^\top] = \mathbb{E}[\alpha_O(\boldsymbol{C})]\,\boldsymbol{I}_{d_2} + d_v\,\boldsymbol{\Phi}\,\mathbb{E}[c_2\boldsymbol{C}^2]\,\boldsymbol{\Phi}^\top, \qquad \mathbb{E}[\boldsymbol{G}_\bullet\boldsymbol{G}_\bullet^\top] = \mathbb{E}[c_\bullet(\boldsymbol{C})]\,\boldsymbol{I} \quad (\bullet \in \{Q, K, V\}).$$

For $\boldsymbol{Q} \in O(d_0)$, $\boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q}^\top \overset{d}{=} \boldsymbol{C}$ and c_2 is invariant, so $\mathbb{E}[c_2\boldsymbol{C}^2] = \boldsymbol{Q}\,\mathbb{E}[c_2\boldsymbol{C}^2]\,\boldsymbol{Q}^\top$ for all $\boldsymbol{Q}$, and by Schur's lemma $\mathbb{E}[c_2\boldsymbol{C}^2] = \gamma\,\boldsymbol{I}_{d_0}$ with $\gamma = \frac{1}{d_0}\mathbb{E}[c_2\operatorname{tr}\boldsymbol{C}^2] = \frac{1}{d_0}\mathbb{E}[c_2^2]$; (77) gives the polynomial and the order in (70). This proves (69), and averaging (87) termwise with (66) gives the closed form of $\bar{\alpha}_O = \mathbb{E}[\alpha_O(\boldsymbol{C})]$ in (71). The same isotropy argument gives $\mathbb{E}[c_2t_\Phi] = \operatorname{tr}(\boldsymbol{\Phi}\,\mathbb{E}[c_2\boldsymbol{C}^2]\,\boldsymbol{\Phi}^\top) = \gamma\,\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2$, so averaging (91)–(93) gives (68), where by (75)

$$\begin{aligned}
\mathbb{E}[c_2\beta_1] &= d_2d_v\Big[(d_2d_v{+}d_2{+}d_v{+}3)\,\mathbb{E}[c_2c_4] + (d_2{+}d_v{+}1)\,\mathbb{E}[c_2^3]\Big],\\
\mathbb{E}[c_2\beta_2] &= d_2d_v\Big[(d_2d_v{+}2)\,\mathbb{E}[c_2^3] + 2(d_2{+}d_v{+}1)\,\mathbb{E}[c_2c_4]\Big], \qquad (94)\\
\mathbb{E}[(c_1^2{+}c_2)\beta_1] &= d_2d_v\Big[(d_2d_v{+}d_2{+}d_v{+}3)\big(\mathbb{E}[c_1^2c_4]{+}\mathbb{E}[c_2c_4]\big) + (d_2{+}d_v{+}1)\big(\mathbb{E}[c_1^2c_2^2]{+}\mathbb{E}[c_2^3]\big)\Big],
\end{aligned}$$

all explicit by Lemma 33. Finally, substituting $\mathsf{A}_3 = \mathsf{A}_1 + \mathsf{A}_2$ into the closed form of $\bar{\alpha}_O$,

$$\bar{\alpha}_O = d_v\Big[(d_v{+}1)\Big(1 + \tfrac{1}{d_k}\Big) + \tfrac{2}{d_k}\Big]\mathsf{A}_1 + d_v\Big[1 + \tfrac{d_v+1}{d_k}\Big]\mathsf{A}_2 = \Theta(d_v^2d_0^6), \qquad (95)$$

all terms being nonnegative and $\mathsf{A}_1, \mathsf{A}_2 = \Theta(d_0^6)$ by Lemma 33, which is the order stated in (71). ■

Proof of Corollary 3. All Wishart moments of Lemma 33 are positive, and $d_v, d_k \geq 1$ throughout.

Step 1: the β-moments. By (94), each of

$$\mathbb{E}[c_2\beta_1] = d_2 d_v \Big[(d_2 d_v{+}d_2{+}d_v{+}3)\,\mathbb{E}[c_2 c_4] + (d_2{+}d_v{+}1)\,\mathbb{E}[c_2^3]\Big],$$
$$\mathbb{E}[c_2\beta_2] = d_2 d_v \Big[(d_2 d_v{+}2)\,\mathbb{E}[c_2^3] + 2(d_2{+}d_v{+}1)\,\mathbb{E}[c_2 c_4]\Big],$$
$$\mathbb{E}[(c_1^2{+}c_2)\beta_1] = d_2 d_v \Big[(d_2 d_v{+}d_2{+}d_v{+}3)\big(\mathbb{E}[c_1^2 c_4]{+}\mathbb{E}[c_2 c_4]\big) + (d_2{+}d_v{+}1)\big(\mathbb{E}[c_1^2 c_2^2]{+}\mathbb{E}[c_2^3]\big)\Big]$$

is $d_2 d_v$ times a polynomial of degree one in d_v whose coefficients are positive combinations of the degree-six moments $\mathbb{E}[c_2 c_4], \mathbb{E}[c_2^3], \mathbb{E}[c_1^2 c_4], \mathbb{E}[c_1^2 c_2^2] = \Theta(d_0^6)$ of Lemma 33; since $d_2 d_v \geq \frac{1}{2}(d_2{+}d_v)$, each is

$$\mathbb{E}[c_2\beta_1] = \Theta(d_2^2 d_v^2 d_0^6), \qquad \mathbb{E}[c_2\beta_2] = \Theta(d_2^2 d_v^2 d_0^6), \qquad \mathbb{E}[(c_1^2{+}c_2)\beta_1] = \Theta(d_2^2 d_v^2 d_0^6). \tag{96}$$

Step 2: c_Q and c_K. Substituting (96) into (68),

$$c_Q = \frac{1}{d_k}\Big(\underbrace{\mathbb{E}[c_2\beta_1]}_{\Theta(d_2^2 d_v^2 d_0^6)} + \underbrace{d_v\gamma\,\|\mathbf{\Phi}\|_{\mathrm{F}}^2}_{\Theta(d_v d_0^3)}\Big) + \frac{1}{d_k^2}\underbrace{\big(\mathbb{E}[c_2\beta_1] + \mathbb{E}[c_2\beta_2]\big)}_{\Theta(d_2^2 d_v^2 d_0^6)},$$
$$c_K = \frac{1}{d_k}\Big(\underbrace{\mathbb{E}[c_2\beta_1]}_{\Theta(d_2^2 d_v^2 d_0^6)} + \underbrace{d_v\gamma\,\|\mathbf{\Phi}\|_{\mathrm{F}}^2}_{\Theta(d_v d_0^3)}\Big) + \frac{1}{d_k^2}\underbrace{\mathbb{E}[(c_1^2{+}c_2)\beta_1]}_{\Theta(d_2^2 d_v^2 d_0^6)},$$

where $\gamma = \Theta(d_0^3)$ by Lemma 33. All terms are positive, and since $d_2, d_v, d_k, d_0 \geq 1$ the second and third terms are at most a constant multiple of the first; hence

$$c_Q \asymp c_K = \Theta\big(d_2^2 d_v^2 d_0^6/d_k\big), \qquad c_K - c_Q = \frac{\mathbb{E}[c_1^2\beta_1] - \mathbb{E}[c_2\beta_2]}{d_k^2} = O\Big(\frac{d_2^2 d_v^2 d_0^6}{d_k^2}\Big) = O\Big(\frac{1}{d_k}\Big)\,c_Q.$$

Step 3: c_V. By (68) with $\mathsf{A}_3 = \mathsf{A}_1 + \mathsf{A}_2$,

$$c_V = d_2(d_2{+}d_v{+}1)\Big(\mathsf{A}_1 + \frac{\mathsf{A}_3}{d_k}\Big) + \gamma\,\|\mathbf{\Phi}\|_{\mathrm{F}}^2, \qquad d_2\mathsf{A}_1 \leq d_2\Big(\mathsf{A}_1 + \frac{\mathsf{A}_3}{d_k}\Big) \leq d_2(\mathsf{A}_1 + \mathsf{A}_3),$$

so c_V is affine in d_v with slope between $d_2\mathsf{A}_1$ and $d_2(\mathsf{A}_1{+}\mathsf{A}_3)$ and intercept between $d_2(d_2{+}1)\mathsf{A}_1$ and $d_2(d_2{+}1)(\mathsf{A}_1{+}\mathsf{A}_3) + \gamma\,\|\mathbf{\Phi}\|_{\mathrm{F}}^2$; since $\mathsf{A}_1, \mathsf{A}_2 = \Theta(d_0^6)$ and $\gamma = \Theta(d_0^3)$ (Lemma 33), $c_V = \Theta\big(d_2(d_2{+}d_v)\,d_0^6\big)$.

Step 4: $\bar{\alpha}_O$. Let $\bar{\alpha}_O^{\infty} := d_v(d_v{+}1)\mathsf{A}_1 + d_v\mathsf{A}_2$ be the d_k-free part of (95). Then

$$\bar{\alpha}_O - \bar{\alpha}_O^{\infty} = \frac{1}{d_k}\Big[d_v(d_v{+}1)\mathsf{A}_3 + 2d_v\mathsf{A}_1\Big] \leq \frac{1}{d_k}\Big[d_v(d_v{+}1)(\mathsf{A}_1{+}\mathsf{A}_2) + 2d_v(d_v{+}1)\mathsf{A}_1\Big]$$
$$= \frac{3\mathsf{A}_1 + \mathsf{A}_2}{\mathsf{A}_1}\cdot\frac{d_v(d_v{+}1)\mathsf{A}_1}{d_k} \leq \frac{3\mathsf{A}_1 + \mathsf{A}_2}{\mathsf{A}_1}\cdot\frac{\bar{\alpha}_O^{\infty}}{d_k},$$

so that

$$\bar{\alpha}_O^{\infty} \leq \bar{\alpha}_O \leq \Big(1 + \frac{3\mathsf{A}_1 + \mathsf{A}_2}{\mathsf{A}_1 d_k}\Big)\bar{\alpha}_O^{\infty}, \qquad d_v^2\mathsf{A}_1 \leq \bar{\alpha}_O^{\infty} \leq 2d_v^2\mathsf{A}_1 + d_v^2\mathsf{A}_2, \tag{97}$$

and therefore, with $\mathsf{A}_1, \mathsf{A}_2 = \Theta(d_0^6)$, $\bar{\alpha}_O = \Theta(d_v^2 d_0^6)$.

Step 5: the $\boldsymbol{W}_O$ ratio. By (69), for $a \neq b$ with $(\mathbf{\Phi}\mathbf{\Phi}^\top)_{ab} \neq 0$,

$$\mathbb{E}[(\boldsymbol{G}_O\boldsymbol{G}_O^\top)_{aa}] = \bar{\alpha}_O + d_v\gamma(\mathbf{\Phi}\mathbf{\Phi}^\top)_{aa} = \Theta(d_v^2 d_0^6),$$
$$\mathbb{E}[(\boldsymbol{G}_O\boldsymbol{G}_O^\top)_{ab}] = d_v\gamma(\mathbf{\Phi}\mathbf{\Phi}^\top)_{ab} = \Theta(d_v d_0^3),$$

and dividing gives the ratio $\Theta(d_v d_0^3)$. Finally, Steps 3 and 4 show that d_k enters c_V and $\bar{\alpha}_O$ only through the factors $1 + \mathsf{A}_3/(\mathsf{A}_1 d_k) \in [1,\, 1 + \mathsf{A}_3/\mathsf{A}_1]$ and the bracket in (97), both bounded above and below; the same is then true of the ratio. ■

E.6 Discussion

Remark 3 (Scope of the isotropy). *This remark records how far the exact zeros of Theorem 4 for $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$ extend beyond the setting of Assumption 8. The proof of Lemma 34 uses only that f depends on $(\boldsymbol{W}_Q, \boldsymbol{W}_K)$ through $\boldsymbol{W}_Q^\top \boldsymbol{W}_K$ and on $(\boldsymbol{W}_V, \boldsymbol{W}_O)$ through $\boldsymbol{W}_O \boldsymbol{W}_V$, and that the law of the weights is invariant under the sign and permutation maps. Hence the isotropy of $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V$ holds conditionally on every dataset, for any differentiable loss, for any initialization whose law has these symmetries, and, applying the sign matrix to stacked heads, for multi-head attention. RoPE inserts a position-dependent matrix between queries and keys and is not covered. For these three matrices $\boldsymbol{P} \propto \boldsymbol{I}$ in expectation, so orthogonalization degenerates to the scalar normalization $\boldsymbol{G}/\|\boldsymbol{G}\|_2$.*

Remark 4 (Role of data randomness). *This remark identifies which part of the argument actually uses the Gaussian tokens, since the answer differs from the ReLU setting. Unlike the ReLU setting, where Gaussian data creates the cancellation (κ_1, κ_2 are input integrals), here the data enters no cancellation: the zeros are weight-space symmetries holding per dataset, and (86) is an identity for every dataset. For instance, since $\alpha_O(\boldsymbol{C}) \geq d_v^2 c_2^3$ and $c_2 \geq \lambda_{\max}(\boldsymbol{C})^2$, (86) gives $\mathbb{E}[(\boldsymbol{G}_O \boldsymbol{G}_O^\top)_{aa} \,|\, \boldsymbol{X}]/\mathbb{E}[(\boldsymbol{G}_O \boldsymbol{G}_O^\top)_{ab} \,|\, \boldsymbol{X}] \geq d_v \lambda_{\max}(\boldsymbol{C})^2/(\|\boldsymbol{\varphi}_a\| \, \|\boldsymbol{\varphi}_b\|)$ deterministically, with $\boldsymbol{\varphi}_a$ the rows of $\boldsymbol{\Phi}$. Gaussian tokens only translate c_1, c_2, c_4 into (T, d_0) via Lemma 33; what is genuinely needed is independence of the data from the initialization, since a data-dependent initialization breaks the symmetry argument.*

Remark 5 (Sign of $c_K - c_Q$). *This remark settles which of the query and key constants is larger, a question that Corollary 3 leaves open because it only bounds their difference. By (68) and (94), $c_K - c_Q = d_2 d_v \big[(d_2 d_v + 2)\Delta_1 + (d_2 + d_v + 1)\Delta_2\big]/d_k^2$ with $\Delta_1 := \mathbb{E}[c_1^2 c_4] - \mathbb{E}[c_2^3]$ and $\Delta_2 := \mathbb{E}[c_1^2 c_4] + \mathbb{E}[c_1^2 c_2^2] - 2\mathbb{E}[c_2 c_4]$. When $\min(T, d_0) = 1$, $\boldsymbol{C}$ has rank one, all four moments equal $\mathbb{E}[(\chi^2_{\max(T,d_0)})^6]$, and $c_K = c_Q$. For every integer $2 \leq T, d_0 \leq 512$ the polynomials of Lemma 33 give $\Delta_1, \Delta_2 > 0$, hence $c_K > c_Q$ for all d_2, d_v in that range.*

Remark 6 (Fan-in scaling). *This remark checks how the conclusions change under the standard fan-in initialization in place of the unit-variance entries of Assumption 8. With entry variances $1/d_0$ ($\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$) and $1/d_v$ ($\boldsymbol{W}_O$), each application of Lemma 27 picks up the corresponding variance and the d_v factor of the $\boldsymbol{W}_O$ ratio cancels: the ratio becomes $\Theta(1)$ in d_k, d_v and d_0 alike, its size being set by the sequence length through the data moments alone, so under fan-in scaling the sequence length, not the hidden dimension, drives diagonalization, consistent with the near-diagonal $\boldsymbol{G}\boldsymbol{G}^\top$ observed along GPT-2 and LLaMA trajectories by Deng et al. (2026).*

F Asymptotic Approximation between Orthogonalization and 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 $\operatorname{diag}(\boldsymbol{V}_t \boldsymbol{V}_t^\top)_{ii} = \|(\boldsymbol{V}_t)_{i,:}\|_2^2$,

$$\big[(\boldsymbol{V}_t \boldsymbol{V}_t^\top)^{-1/2} \boldsymbol{V}_t\big]_{i,:} = \frac{(\boldsymbol{V}_t)_{i,:}}{\|(\boldsymbol{V}_t)_{i,:}\|_2}. \tag{98}$$

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 (Ma et al., 2025; Scetbon et al., 2025; Glentis et al., 2025; Wen et al., 2025; Pethick et al., 2025; Deng et al., 2026); 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 (98) together with Theorems 1 to 4: at a Gaussian initialization, as the hidden dimension grows, Muon's $\Theta(mn \min(m, n))$ orthogonalization collapses to a $\Theta(mn)$ row-wise ℓ_2 normalization with the same preconditioner structure, and for the query, key and value weights of attention the collapse is exact in expectation. Deng et al. (2026) 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 hidden-dimension dependence in Theorem 3 and Theorem 4. Carrying the analysis past initialization, and to deep nonlinear networks, is left for future work.

G Simulation of Theoretical Results

We numerically verify the dimension-scaling predictions of Theorems 1 to 4 under the Gaussian initialization of Section 3.2. For each model, we perform two experiments. First, at a representative configuration, we estimate the expected gradient self outer-product and examine its diagonal and off-diagonal structure. Second, we vary the relevant model dimensions and measure how the diagonal magnitude, off-diagonal magnitude, and their ratio change.

To summarize the structure of a matrix $\boldsymbol{P} \in \mathbb{R}^{m\times m}$, we define the mean diagonal magnitude, mean off-diagonal magnitude, and diagonal-dominance ratio as

$$D(\boldsymbol{P}) = \frac{1}{m}\sum_{i=1}^{m}|P_{ii}|, \qquad \mathrm{OD}(\boldsymbol{P}) = \frac{1}{m(m-1)}\sum_{i\neq j}|P_{ij}|, \qquad R(\boldsymbol{P}) = \frac{D(\boldsymbol{P})}{\mathrm{OD}(\boldsymbol{P})}. \tag{99}$$

Given S independent initializations with gradients $\boldsymbol{G}^{(1)}, \ldots, \boldsymbol{G}^{(S)}$, we estimate the expected gradient self outer-product by

$$\widehat{\mathbb{E}}[\boldsymbol{G}\boldsymbol{G}^\top] = \frac{1}{S}\sum_{s=1}^{S}\boldsymbol{G}^{(s)}(\boldsymbol{G}^{(s)})^\top. \tag{100}$$

We use $\widehat{D}$, $\widehat{\mathrm{OD}}$, and $\widehat{R}$ to denote the corresponding empirical estimates of $D(\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top])$, $\mathrm{OD}(\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top])$, and $R(\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top])$, respectively.

G.1 Symmetric Matrix Factorization

Setup. We next verify the predictions of Theorem 1 and Corollary 1 for symmetric matrix factorization. As defined in (16), the gradient is

$$\boldsymbol{G} = (\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{M}^\star)\boldsymbol{W},$$

where $\boldsymbol{W} \in \mathbb{R}^{d\times k}$ and $\boldsymbol{M}^\star \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite. Under Assumption 5, Theorem 1 gives

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top] = \alpha_3\boldsymbol{I}_d - 2\alpha_2\boldsymbol{M}^\star + k(\boldsymbol{M}^\star)^2, \tag{101}$$

where $\alpha_2 = k(k+d+1)$ and $\alpha_3 = k^3 + 3(d+1)k^2 + (d^2+3d+4)k$. Corollary 1 gives the corresponding diagonal and off-diagonal entries,

$$\begin{aligned}\mathbb{E}[(\boldsymbol{G}\boldsymbol{G}^\top)_{ii}] &= \alpha_3 - 2\alpha_2\boldsymbol{M}^\star_{ii} + k\big((\boldsymbol{M}^\star)^2\big)_{ii},\\ \mathbb{E}[(\boldsymbol{G}\boldsymbol{G}^\top)_{ij}] &= -2\alpha_2\boldsymbol{M}^\star_{ij} + k\big((\boldsymbol{M}^\star)^2\big)_{ij}, \qquad i\neq j.\end{aligned} \tag{102}$$

For fixed d and a fixed target with nonzero off-diagonal structure, this gives $D(\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]) = \Theta(k^3)$ and $\mathrm{OD}(\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]) = \Theta(k^2)$, and hence $R(\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]) = \Theta(k)$.

Empirical structure of the expected gradient Gram. We first examine the structure of the expected gradient Gram at a representative hidden dimension. We set $d = 25$, $k = 80$, and construct a fixed rank-5 target $\boldsymbol{M}^\star = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} \in \mathbb{R}^{25\times 5}$ has orthonormal columns obtained from the QR factorization of a Gaussian matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(3, 2.5, 2, 1.5, 1)$. We draw $S = 10{,}000$ independent initializations $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(S)}$, with i.i.d. entries from $\mathcal{N}(0,1)$. For each $s \in \{1, \ldots, S\}$, let $\boldsymbol{G}^{(s)}$ denote the gradient corresponding to $\boldsymbol{W}^{(s)}$. Using (100), we estimate $\mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]$ from these 10,000 initializations.

Figure 7 shows a strongly diagonal expected gradient Gram, with substantially smaller off-diagonal entries. This is consistent with the diagonal-dominance behavior predicted by Theorem 1.

Hidden-dimension scaling. We next vary k over 15 values from 5 to 640, while keeping $d = 25$ and $\boldsymbol{M}^\star$ fixed, using $S = 10{,}000$ independent initializations at each hidden dimension.

To reduce Monte Carlo noise in the smaller off-diagonal entries, we subtract a term whose expected off-diagonal contribution is zero. By Lemma 4, setting $\boldsymbol{M}^\star = \boldsymbol{0}$ gives $\boldsymbol{G}(\boldsymbol{0})\boldsymbol{G}(\boldsymbol{0})^\top = \boldsymbol{S}^3$, where

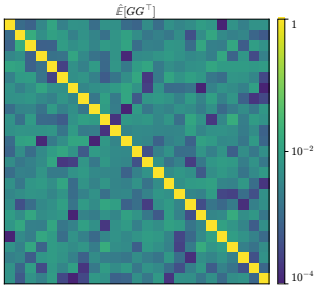


Figure 7: Empirical expected gradient self outer-product for symmetric matrix factorization with $d = 25$, $k = 80$, and a fixed rank-5 target, estimated over 10,000 independent initializations. Each heatmap is normalized by its empirical mean diagonal magnitude. The panel uses the logarithmic color scale as shown.

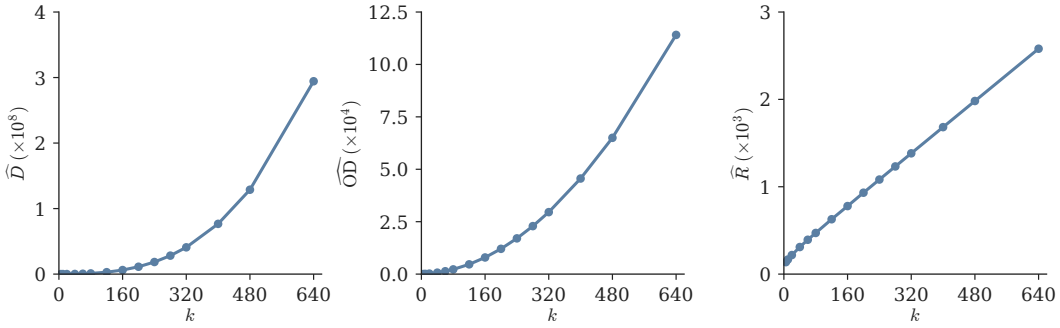


Figure 8: Hidden-dimension scaling for symmetric matrix factorization with $d = 25$ and a fixed rank-5 target. From left to right, the panels show $\widehat{D}$, $\widehat{\mathrm{OD}}$, and $\widehat{R} = \widehat{D}/\widehat{\mathrm{OD}}$ as the factorization hidden dimension k is varied. Empirical estimates use 10,000 independent initializations at each hidden dimension; $\widehat{\mathrm{OD}}$ uses the paired estimator in (103).

$\boldsymbol{S} = \boldsymbol{W}\boldsymbol{W}^\top$. Moreover, Lemma 7 gives $\mathbb{E}[\boldsymbol{S}^3] = \alpha_3 \boldsymbol{I}_d$, so this term has zero expected off-diagonal entries. We therefore define

$$\boldsymbol{\Delta}_{\mathrm{MF}}^{(s)} = \boldsymbol{G}^{(s)}(\boldsymbol{M}^\star)\boldsymbol{G}^{(s)}(\boldsymbol{M}^\star)^\top - \boldsymbol{G}^{(s)}(\boldsymbol{0})\boldsymbol{G}^{(s)}(\boldsymbol{0})^\top. \tag{103}$$

For $i \neq j$,

$$\mathbb{E}\Big[(\boldsymbol{\Delta}_{\mathrm{MF}}^{(s)})_{ij}\Big] = \mathbb{E}\big[(\boldsymbol{G}\boldsymbol{G}^\top)_{ij}\big] .$$

Thus, the paired difference removes the $\boldsymbol{S}^3$ contribution while preserving the expected off-diagonal entries.

For each hidden dimension, we compute $\widehat{D}$ from the full empirical average (100). For $\widehat{\mathrm{OD}}$, we first average $\boldsymbol{\Delta}_{\mathrm{MF}}^{(s)}$ over the same S initializations and then apply $\mathrm{OD}(\cdot)$. We then compute $\widehat{R} = \widehat{D}/\widehat{\mathrm{OD}}$.

Figure 8 shows that the diagonal magnitude grows faster with k than the off-diagonal magnitude, while $\widehat{R}$ grows approximately linearly with k. This is consistent with the theoretical predictions $D = \Theta(k^3)$, $\mathrm{OD} = \Theta(k^2)$, and $R = \Theta(k)$.

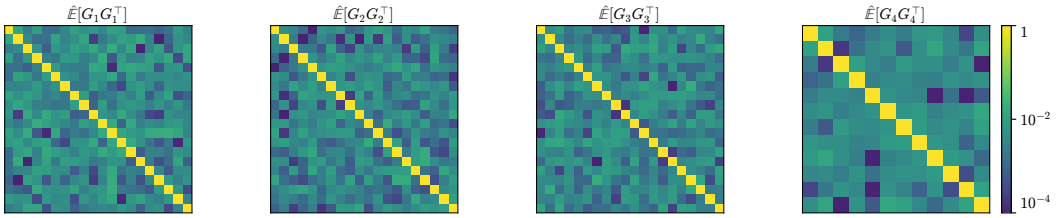


Figure 9: Empirical expected gradient self outer-products for the depth-4 linear network with baseline dimensions $[512, 80, 80, 80, 512]$, estimated from 10,000 independent initializations. The four panels show $\widehat{\mathbb{E}}[\boldsymbol{G}_i\boldsymbol{G}_i^\top]$ for $i = 1, \ldots, 4$. Each heatmap is normalized by its empirical mean diagonal magnitude. A common logarithmic color scale is used across all panels.

G.2 Deep Linear Network

Setup. We next verify the predictions of Theorem 2 and Corollary 2 for the deep linear network. Under the Gaussian initialization of Assumption 2, Theorem 2 gives

$$\mathbb{E}[\boldsymbol{G}_i\boldsymbol{G}_i^\top] = V_i s_i \boldsymbol{I}_{d_i} + \boldsymbol{T}_i, \qquad \boldsymbol{T}_i = \begin{cases} \left(\prod_{j=1}^{L-1} d_j\right) \boldsymbol{\Phi}\boldsymbol{\Phi}^\top, & i = L, \\ \left(\prod_{\substack{1 \le j \le L-1 \\ j \ne i}} d_j\right) \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_i}, & i < L. \end{cases} \tag{104}$$

Corollary 2 therefore gives zero expected off-diagonal entries for every hidden layer $i < L$, while for the output layer

$$\mathbb{E}[(\boldsymbol{G}_L\boldsymbol{G}_L^\top)_{ab}] = \left(\prod_{j=1}^{L-1} d_j\right) (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)_{ab}, \qquad a \ne b. \tag{105}$$

Empirical structure of the expected gradient Gram. We first examine the empirical expected gradient Grams at a representative configuration. We use depth $L = 4$ with dimensions $[d_0, d_1, d_2, d_3, d_4] = [512, 80, 80, 80, 512]$ and a fixed target $\boldsymbol{\Phi} \in \mathbb{R}^{512\times512}$ with entries sampled from $\mathcal{N}(0, 1)$. We use $S = 10{,}000$ independent Gaussian weight initializations and estimate the expected Gram for each $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_4$ using (100).

Figure 9 shows strongly diagonal expected Grams for $\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3$, consistent with Corollary 2, which gives zero expected off-diagonal entries for the hidden layers. The output-layer Gram retains visible off-diagonal structure, consistent with the $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ term in Theorem 2.

Dimension scaling. We next vary each hidden dimension d_1, d_2, d_3 individually while holding the remaining dimensions fixed at the baseline values above. Each hidden dimension is varied over a 12-point grid $\{5, 10, 20, 40, 60, 80, 120, 160, 200, 240, 280, 320\}$, and we evaluate $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_4$. A fixed 512×512 target is sampled with entries from $\mathcal{N}(0, 1)$. We use $S = 10{,}000$ independent initializations for every configuration.

Because the expected off-diagonal entries are much smaller than the diagonal entries, direct Monte Carlo estimation of OD has high variance. We therefore use a paired construction to reduce this Monte Carlo variance while preserving the expected off-diagonal entries. From (29) and (30), for fixed weights, the gradient is affine in $\boldsymbol{\Phi}$. For each $s \in \{1, \ldots, S\}$, let $\boldsymbol{G}_i^{(s)}(\boldsymbol{\Phi})$ denote the layer-i gradient for the s-th weight initialization and target $\boldsymbol{\Phi}$. Using the same weights, we evaluate the gradient at $\boldsymbol{\Phi}$, $-\boldsymbol{\Phi}$, and $\mathbf{0}$, and define

$$\boldsymbol{\Delta}_i^{(s)} = \frac{1}{2}\left[\boldsymbol{G}_i^{(s)}(\boldsymbol{\Phi})\boldsymbol{G}_i^{(s)}(\boldsymbol{\Phi})^\top + \boldsymbol{G}_i^{(s)}(-\boldsymbol{\Phi})\boldsymbol{G}_i^{(s)}(-\boldsymbol{\Phi})^\top\right] - \boldsymbol{G}_i^{(s)}(\mathbf{0})\boldsymbol{G}_i^{(s)}(\mathbf{0})^\top. \tag{106}$$

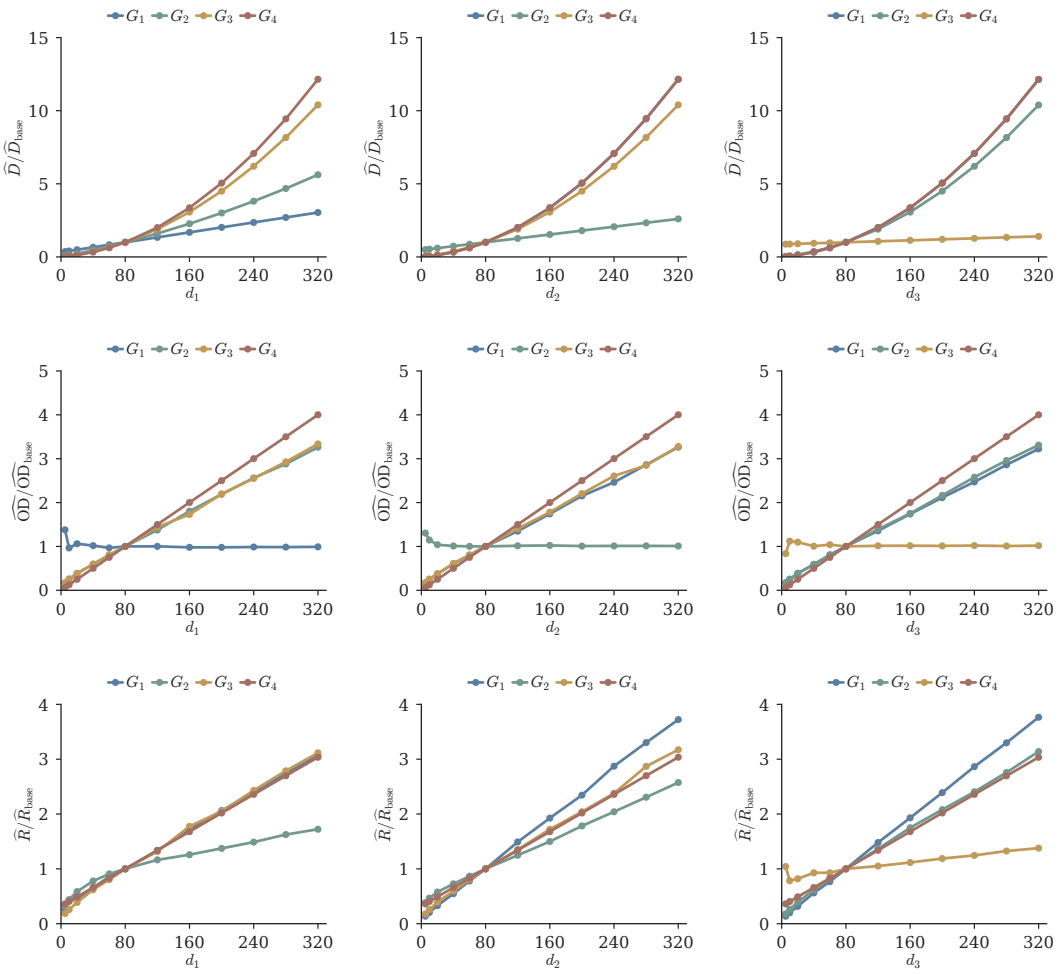


Figure 10: Empirical dimension scaling for the depth-4 linear network. Columns vary d_1, d_2, and d_3 individually while holding the remaining dimensions fixed at their baseline values. Rows show $\widehat{D}$, $\widehat{\text{OD}}$, and $\widehat{R} = \widehat{D}/\widehat{\text{OD}}$ for $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_4$, each normalized by its value at the baseline configuration $[d_0, d_1, d_2, d_3, d_4] = [512, 80, 80, 80, 512]$ to facilitate comparison across layers with different absolute scales. Empirical estimates use 10,000 independent initializations per configuration. The off-diagonal statistic is estimated using (106).

The symmetric combination cancels the target-independent and target-linear terms. The remaining term is

$$\boldsymbol{\Delta}_i^{(s)} = (\boldsymbol{A}_i^{(s)})^\top \boldsymbol{\Phi} (\boldsymbol{B}_i^{(s)})^\top \boldsymbol{B}_i^{(s)} \boldsymbol{\Phi}^\top \boldsymbol{A}_i^{(s)}. \tag{107}$$

By Lemma 15,

$$\mathbb{E}[\boldsymbol{\Delta}_i^{(s)}] = \boldsymbol{T}_i. \tag{108}$$

Since the remaining term $V_i s_i \boldsymbol{I}_{d_i}$ in Theorem 2 is diagonal, (106) has exactly the same expected off-diagonal entries as the full gradient Gram.

For each configuration, we compute $\widehat{D}$ from the full empirical expected Gram. For $\widehat{\text{OD}}$, we first average $\boldsymbol{\Delta}_i^{(s)}$ over the same S initializations and then apply $\text{OD}(\cdot)$. We then compute $\widehat{R} = \widehat{D}/\widehat{\text{OD}}$.

Figure 10 shows the layer-dependent scaling predicted by Theorem 2 and Corollary 2. When d_j is varied, $\widehat{D}$ grows more slowly for the corresponding gradient $\boldsymbol{G}_j$ than for the other hidden-layer gradients. This is consistent with $D = \Theta(d_j)$ for $\boldsymbol{G}_j$, compared with $D = \Theta(d_j^2)$ for $\boldsymbol{G}_i$, $i \neq j$, when the remaining dimensions are fixed.

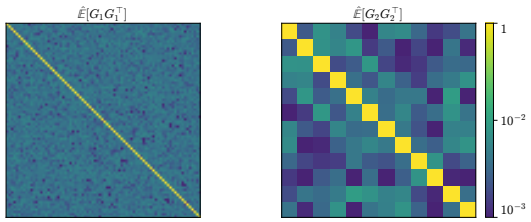


Figure 11: Empirical expected gradient self outer-products for the two-layer ReLU network with dimensions $[d_0, d_1, d_2] = [512, 80, 12]$, estimated from 10,000 independent weight initializations. The panels show $\widehat{\mathbb{E}}[\boldsymbol{G}_1\boldsymbol{G}_1^\top]$ and $\widehat{\mathbb{E}}[\boldsymbol{G}_2\boldsymbol{G}_2^\top]$. Each heatmap is normalized by its empirical mean diagonal magnitude. A common logarithmic color scale is used across all panels.

Corollary 2 gives zero expected off-diagonal entries for $i < L$, and the remaining nonzero values are finite-sample Monte Carlo variances. For the output layer, the normalized $\widehat{D}$ curves show approximately quadratic growth, while the normalized $\widehat{\mathrm{OD}}$ curves grow approximately linearly when any one of d_1, d_2, d_3 is varied. This is consistent with $D = \Theta(d_j^2)$, $\mathrm{OD} = \Theta(d_j)$, and hence $R = \Theta(d_j)$.

G.3 Two-Layer ReLU Network

Setup. We next verify the predictions of Theorem 3 for the two-layer ReLU network $f(\boldsymbol{x}) = \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x})$, with hidden dimension d_1. Unlike the hidden layers of the deep linear network, the expected gradient self outer-products are not exactly diagonal.

Under the Gaussian initialization of Assumption 3, Theorem 3 gives, for $\boldsymbol{G}_1$,

$$\mathbb{E}\big[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{aa}\big] = \frac{d_0 d_2^2}{4} + d_0 d_2\left((d_1-1)\kappa_1 + \frac{1}{2}\right) + \frac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}, \tag{109}$$

while for $a \neq c$,

$$\mathbb{E}\big[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{ac}\big] = d_0 d_2 \kappa_2 + O(d_2\sqrt{d_0}), \tag{110}$$

where κ_1 and κ_2 are defined in Theorem 3 and depend only on d_0. Holding d_0 and d_2 fixed while varying d_1 therefore gives

$$D\big(\mathbb{E}[\boldsymbol{G}_1\boldsymbol{G}_1^\top]\big) = \Theta(d_1), \qquad \mathrm{OD}\big(\mathbb{E}[\boldsymbol{G}_1\boldsymbol{G}_1^\top]\big) = \Theta(1), \qquad R\big(\mathbb{E}[\boldsymbol{G}_1\boldsymbol{G}_1^\top]\big) = \Theta(d_1). \tag{111}$$

For $\boldsymbol{G}_2$, Theorem 3 gives

$$\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top] = \alpha\boldsymbol{I}_{d_2} + \frac{d_1}{4}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top, \qquad \alpha = \Theta(d_1^2). \tag{112}$$

Thus, for fixed d_0, d_2, and a fixed target with nonzero off-diagonal structure,

$$D\big(\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top]\big) = \Theta(d_1^2), \qquad \mathrm{OD}\big(\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top]\big) = \Theta(d_1), \qquad R\big(\mathbb{E}[\boldsymbol{G}_2\boldsymbol{G}_2^\top]\big) = \Theta(d_1). \tag{113}$$

Empirical structure of the expected gradient Gram. We first examine the empirical expected gradient Grams at a representative configuration. We use $[d_0, d_1, d_2] = [512, 80, 12]$ and a fixed target $\boldsymbol{\Phi} \in \mathbb{R}^{12\times 512}$ whose entries are sampled from $\mathcal{N}(0, 9)$. For each of $S = 10{,}000$ independent Gaussian weight initializations, we estimate the population gradients using 128,000 independent inputs sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$, and estimate the expected gradient Grams using (100).

Figure 11 shows diagonally dominant empirical expected Grams for both $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$, with smaller but nonzero off-diagonal entries. This is consistent with Theorem 3, which predicts asymptotic diagonal dominance.

Dimension scaling. We set $d_0 = 512$ and $d_2 = 12$, and vary $d_1 \in \{5, 10, 20, 40, 60, 80, 120, 160, 200, 240, 280, 320, 400, 480, 640\}$. The same fixed target is used throughout the sweep, and each population gradient is estimated from 128,000 Gaussian inputs for each of $S = 10{,}000$ independent weight initializations.

Because the expected off-diagonal entries are much smaller than the diagonal entries, direct Monte Carlo estimation of OD is noisy. We therefore use the following expressions to reduce this Monte Carlo variance while preserving the expected off-diagonal entries.

For $\boldsymbol{G}_1$, (58) gives

$$\boldsymbol{G}_1 = \boldsymbol{G}_1^{(1)} - \boldsymbol{G}_1^{(2)}. \tag{114}$$

By (59), the cross terms vanish in expectation, and Lemma 24 gives

$$\boldsymbol{G}_1^{(2)} = \frac{1}{2}\boldsymbol{W}_2^\top \boldsymbol{\Phi}, \qquad \mathbb{E}\Big[\boldsymbol{G}_1^{(2)}(\boldsymbol{G}_1^{(2)})^\top\Big] = \frac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}\boldsymbol{I}_{d_1}. \tag{115}$$

Hence,

$$\mathbb{E}[\boldsymbol{G}_1\boldsymbol{G}_1^\top] = \mathbb{E}\Big[\boldsymbol{G}_1^{(1)}(\boldsymbol{G}_1^{(1)})^\top\Big] + \frac{\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2}{4}\boldsymbol{I}_{d_1}. \tag{116}$$

The second term is diagonal, so all expected off-diagonal entries come from $\boldsymbol{G}_1^{(1)}$. Moreover, setting $\boldsymbol{\Phi} = \boldsymbol{0}$ in (58) gives $\boldsymbol{G}_1(\boldsymbol{0}) = \boldsymbol{G}_1^{(1)}$. For each $s \in \{1, \dots, S\}$, we therefore define

$$\boldsymbol{\Delta}_1^{(s)} = \boldsymbol{G}_1^{(s)}(\boldsymbol{0})\boldsymbol{G}_1^{(s)}(\boldsymbol{0})^\top. \tag{117}$$

For $a \neq c$,

$$\mathbb{E}\Big[(\boldsymbol{\Delta}_1^{(s)})_{ac}\Big] = \mathbb{E}\big[(\boldsymbol{G}_1\boldsymbol{G}_1^\top)_{ac}\big]. \tag{118}$$

Thus, (117) preserves the expected off-diagonal entries while removing the diagonal target contribution.

For $\boldsymbol{G}_2$, Lemma 22 gives $\boldsymbol{G}_2 = \boldsymbol{W}_2\boldsymbol{H} - \boldsymbol{\Phi}\boldsymbol{C}^\top$. For each $s \in \{1, \dots, S\}$, let $\boldsymbol{G}_2^{(s)}(\boldsymbol{\Phi})$ denote the output-layer gradient for the s-th weight initialization and target $\boldsymbol{\Phi}$. Using the same weights and input samples, we evaluate the gradient at $\boldsymbol{\Phi}$, $-\boldsymbol{\Phi}$, and $\boldsymbol{0}$, and define

$$\boldsymbol{\Delta}_2^{(s)} = \frac{1}{2}\Big[\boldsymbol{G}_2^{(s)}(\boldsymbol{\Phi})\boldsymbol{G}_2^{(s)}(\boldsymbol{\Phi})^\top + \boldsymbol{G}_2^{(s)}(-\boldsymbol{\Phi})\boldsymbol{G}_2^{(s)}(-\boldsymbol{\Phi})^\top\Big] - \boldsymbol{G}_2^{(s)}(\boldsymbol{0})\boldsymbol{G}_2^{(s)}(\boldsymbol{0})^\top. \tag{119}$$

Substituting Lemma 22 into (119) cancels the terms independent of $\boldsymbol{\Phi}$ and linear in $\boldsymbol{\Phi}$, leaving

$$\boldsymbol{\Delta}_2^{(s)} = \boldsymbol{\Phi}(\boldsymbol{C}^{(s)})^\top\boldsymbol{C}^{(s)}\boldsymbol{\Phi}^\top. \tag{120}$$

By Lemma 23, $\boldsymbol{C} = \frac{1}{2}\boldsymbol{W}_1$, and Lemma 16 gives $\mathbb{E}[\boldsymbol{W}_1^\top\boldsymbol{W}_1] = d_1\boldsymbol{I}_{d_0}$. Hence

$$\mathbb{E}[\boldsymbol{\Delta}_2^{(s)}] = \frac{d_1}{4}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top. \tag{121}$$

Since the remaining term $\alpha\boldsymbol{I}_{d_2}$ in (112) is diagonal, (119) has exactly the same expected off-diagonal entries as the full $\boldsymbol{G}_2\boldsymbol{G}_2^\top$.

For both layers, we compute $\widehat{D}$ from the full empirical expected Gram. For $\boldsymbol{G}_1$, we compute $\widehat{\mathrm{OD}}$ from the zero-target empirical Gram. For $\boldsymbol{G}_2$, we first average $\boldsymbol{\Delta}_2^{(s)}$ over the same S initializations and then apply $\mathrm{OD}(\cdot)$. We then compute $\widehat{R} = \widehat{D}/\widehat{\mathrm{OD}}$.

Figure 12 shows the different hidden-dimension scalings predicted by Theorem 3. For $\boldsymbol{G}_1$, $\widehat{D}$ grows approximately linearly with d_1, while $\widehat{\mathrm{OD}}$ remains approximately constant, consistent with $D = \Theta(d_1)$ and $\mathrm{OD} = \Theta(1)$ in (111). For $\boldsymbol{G}_2$, $\widehat{D}$ grows approximately quadratically with d_1, while $\widehat{\mathrm{OD}}$ grows approximately linearly, consistent with (113). Accordingly, $\widehat{R}$ increases with d_1 for both layers, consistent with the predicted $R = \Theta(d_1)$.

G.4 Single-Layer Attention

Setup. Finally, we verify the predictions of Theorem 4 and Corollary 3 for single-layer attention defined in (8). We consider the gradients $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V, \boldsymbol{G}_O$ of the query, key, value, and output projections, respectively. Under the Gaussian initialization of Assumption 4, Theorem 4 gives

$$\begin{aligned}
\mathbb{E}[\boldsymbol{G}_Q\boldsymbol{G}_Q^\top] &= c_Q\boldsymbol{I}_{d_k},\\
\mathbb{E}[\boldsymbol{G}_K\boldsymbol{G}_K^\top] &= c_K\boldsymbol{I}_{d_k},\\
\mathbb{E}[\boldsymbol{G}_V\boldsymbol{G}_V^\top] &= c_V\boldsymbol{I}_{d_v},\\
\mathbb{E}[\boldsymbol{G}_O\boldsymbol{G}_O^\top] &= \bar{\alpha}_O\boldsymbol{I}_{d_2} + d_v\gamma\,\boldsymbol{\Phi}\boldsymbol{\Phi}^\top,
\end{aligned} \tag{122}$$

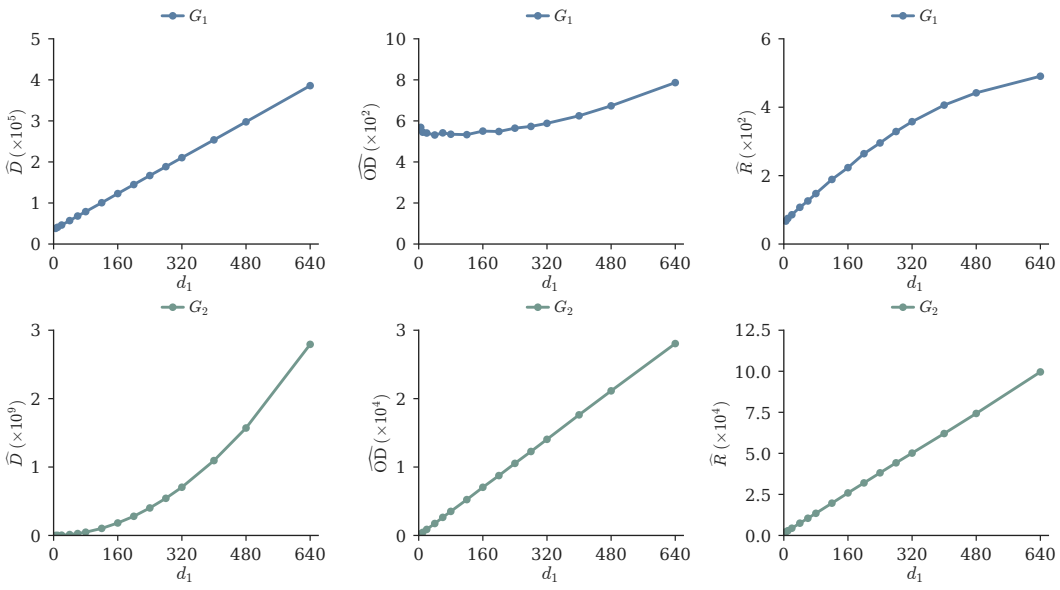


Figure 12: Hidden-dimension scaling for the two-layer ReLU network with $d_0 = 512$ and $d_2 = 12$. The top and bottom rows correspond to $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$, respectively, and show $\widehat{D}$, $\widehat{\mathrm{OD}}$, and $\widehat{R} = \widehat{D}/\widehat{\mathrm{OD}}$ as d_1 is varied. For $\boldsymbol{G}_1$, $\widehat{\mathrm{OD}}$ is estimated from the zero-target Gram; for $\boldsymbol{G}_2$, it is estimated using (119).

where $c_Q, c_K, c_V, \bar{\alpha}_O$, and γ are given explicitly in Theorem 4 and Appendix E.3. Thus, the expected gradient self outer-products for the query, key, and value projections are exactly diagonal,

$$\mathrm{OD}\big(\mathbb{E}[\boldsymbol{G}_p \boldsymbol{G}_p^\top]\big) = 0, \qquad p \in \{Q, K, V\}. \tag{123}$$

Combining Theorem 4 and Corollary 3, for a fixed target with $\mathrm{OD}(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top) > 0$, we obtain

$$R\big(\mathbb{E}[\boldsymbol{G}_O \boldsymbol{G}_O^\top]\big) = \Theta(d_v d_0^3). \tag{124}$$

Hence increasing the hidden dimension d_v or the token dimension d_0 makes the expected output-projection Gram increasingly diagonally dominant, while the hidden dimension d_k enters the ratio only through a bounded factor.

Empirical structure of the expected gradient Gram. We first examine the empirical expected gradient Grams at a representative configuration. We use linear attention with $T = 16$, $d_0 = 32$, $d_k = d_v = 20$, and $d_2 = 12$. The entries of the fixed target $\boldsymbol{\Phi} \in \mathbb{R}^{d_2 \times d_0}$ are sampled from $\mathcal{N}(0, 1)$. We use $S = 10{,}000$ independent initializations. For each initialization, the rows of $\boldsymbol{X} \in \mathbb{R}^{T \times d_0}$ are sampled independently from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_0})$, independently of the Gaussian weights, as specified in Assumption 4. We estimate the expected gradient Gram for each projection using (100).

Figure 13 shows strongly diagonal empirical expected Grams for the query, key, and value projections, consistent with the exact isotropy in Theorem 4. The output-projection Gram is also diagonally dominant but retains nonzero off-diagonal structure, consistent with the $d_v \gamma\, \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ term in (122).

Dimension scaling. We examine four dimension sweeps: $d_k = d_v$ jointly, d_k with $d_v = 80$, d_v with $d_k = 80$, and $d_0 = d_2$ jointly. For the first three sweeps, we set $T = 32$, $d_0 = 512$, and $d_2 = 12$, and use the 12-point grid $\{5, 10, 20, 40, 60, 80, 120, 160, 200, 240, 280, 320\}$. For the joint $d_0 = d_2$ sweep, we set $T = 16$ and $d_k = d_v = 20$, and use the 13-point grid $\{32, 64, 128, 192, 256, 320, 384, 448, 512, 640, 768, 896, 1024\}$. Further, for the joint $d_0 = d_2$ sweep, we rescale the target $\boldsymbol{\Phi}$ at each dimension to have the same Frobenius norm as the 32×32 reference target. This controls for the explicit target dependence in Theorem 4: the diagonal constants c_Q, c_K, c_V contain terms proportional to $\|\boldsymbol{\Phi}\|_{\mathrm{F}}^2$ in (68), while the output Gram contains the term $d_v \gamma\, \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ in (69). Holding $\|\boldsymbol{\Phi}\|_{\mathrm{F}}$ fixed therefore isolates the effect of increasing $d_0 = d_2$

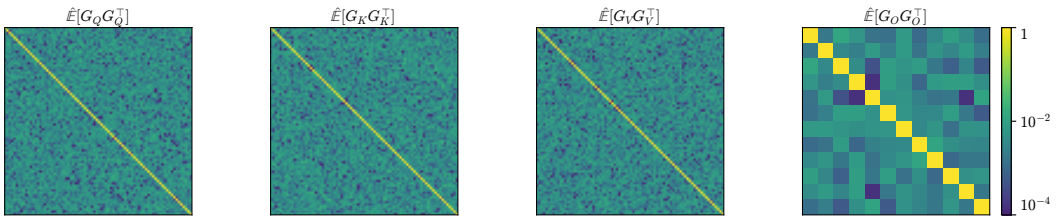


Figure 13: Empirical expected gradient self outer-products for single-layer attention with $T = 16$, $d_0 = 32$, $d_k = d_v = 20$, and $d_2 = 12$, estimated from 10,000 independent initializations. The four panels show $\widehat{\mathbb{E}}[\boldsymbol{G}_p \boldsymbol{G}_p^\top]$ for $p \in \{Q, K, V, O\}$. Each heatmap is normalized by its empirical mean diagonal magnitude. A common logarithmic color scale is used across all panels.

from the increase in target magnitude from the larger target matrix. For every configuration, we use $S = 10{,}000$ independent initializations.For each initialization, the rows of $\boldsymbol{X}$ are sampled independently from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d_0})$, independently of the weights.

Since the expected off-diagonal entries are much smaller than the diagonal entries, direct Monte Carlo estimation of OD is noisy. We therefore use the following expressions to reduce this Monte Carlo variance while preserving the expected off-diagonal entries. From (65) and (67), each gradient is affine in $\boldsymbol{\Phi}$ when the weights and $\boldsymbol{X}$ are fixed. For each $s \in \{1, \dots, S\}$, let $\boldsymbol{G}_p^{(s)}(\boldsymbol{\Phi})$ denote the gradient for projection $p \in \{Q, K, V, O\}$ under the s-th initialization and target $\boldsymbol{\Phi}$. Using the same weights and $\boldsymbol{X}$, we evaluate the gradient at $\boldsymbol{\Phi}$, $-\boldsymbol{\Phi}$, and $\mathbf{0}$, and define

$$\boldsymbol{\Delta}_p^{(s)} = \frac{1}{2}\left[\boldsymbol{G}_p^{(s)}(\boldsymbol{\Phi})\boldsymbol{G}_p^{(s)}(\boldsymbol{\Phi})^\top + \boldsymbol{G}_p^{(s)}(-\boldsymbol{\Phi})\boldsymbol{G}_p^{(s)}(-\boldsymbol{\Phi})^\top\right] - \boldsymbol{G}_p^{(s)}(\mathbf{0})\boldsymbol{G}_p^{(s)}(\mathbf{0})^\top. \tag{125}$$

The symmetric combination cancels the terms independent of $\boldsymbol{\Phi}$ and linear in $\boldsymbol{\Phi}$, leaving the contribution quadratic in the target. Combining this cancellation with the target-dependent terms in (68) and (69) gives

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Delta}_Q^{(s)}] &= \frac{d_v \gamma}{d_k} \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_k}, \\
\mathbb{E}[\boldsymbol{\Delta}_K^{(s)}] &= \frac{d_v \gamma}{d_k} \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_k}, \\
\mathbb{E}[\boldsymbol{\Delta}_V^{(s)}] &= \gamma \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 \boldsymbol{I}_{d_v}, \\
\mathbb{E}[\boldsymbol{\Delta}_O^{(s)}] &= d_v \gamma \boldsymbol{\Phi}\boldsymbol{\Phi}^\top.
\end{aligned} \tag{126}$$

Thus, $\boldsymbol{\Delta}_p^{(s)}$ in (125) has zero expected off-diagonal entries for $p \in \{Q, K, V\}$, exactly as in (123), while for $p = O$ it preserves the off-diagonal term in (122).

For each configuration, we compute $\widehat{D}$ from the full empirical expected Gram. For $\widehat{\mathrm{OD}}$, we first average $\boldsymbol{\Delta}_p^{(s)}$ over the same S initializations and then apply $\mathrm{OD}(\cdot)$. We then compute $\widehat{R} = \widehat{D}/\widehat{\mathrm{OD}}$.

Figure 14 shows the dimension-dependent diagonal scaling predicted by Theorem 4 and Corollary 3. When d_k is varied with d_v fixed, $\widehat{D}$ for $\boldsymbol{G}_Q$ and $\boldsymbol{G}_K$ decreases, consistent with $D = \Theta(d_2^2 d_v^2 d_0^6 / d_k)$, while $\boldsymbol{G}_V$ and $\boldsymbol{G}_O$ depend only weakly on d_k. Increasing d_v increases $\widehat{D}$ for all four projections; the stronger growth of $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_O$ relative to $\boldsymbol{G}_V$ is consistent with their quadratic versus linear dependence on d_v. When $d_k = d_v$ are increased jointly, $\widehat{D}$ grows approximately linearly for $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V$ and more rapidly for $\boldsymbol{G}_O$, again matching the predicted orders.

For Q, K, and V, the population off-diagonal entries are exactly zero by Theorem 4; the nonzero $\widehat{\mathrm{OD}}$ values are finite-sample Monte Carlo variances. For the output projection, $\widehat{\mathrm{OD}}$ grows approximately linearly with d_v and is nearly unchanged when only d_k is varied, while $\widehat{R}$ grows approximately linearly with d_v, consistent with Corollary 3, which gives $D = \Theta(d_v^2 d_0^6)$, $\mathrm{OD} = \Theta(d_v d_0^3)$, and $R = \Theta(d_v d_0^3)$. For the joint $d_0 = d_2 = m$ sweep, Corollary 3 gives $D = \Theta(m^8)$ for $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V$, while for $\boldsymbol{G}_O$, $D = \Theta(m^6)$, $\mathrm{OD} = \Theta(m^3)$, and hence $R = \Theta(m^3)$.

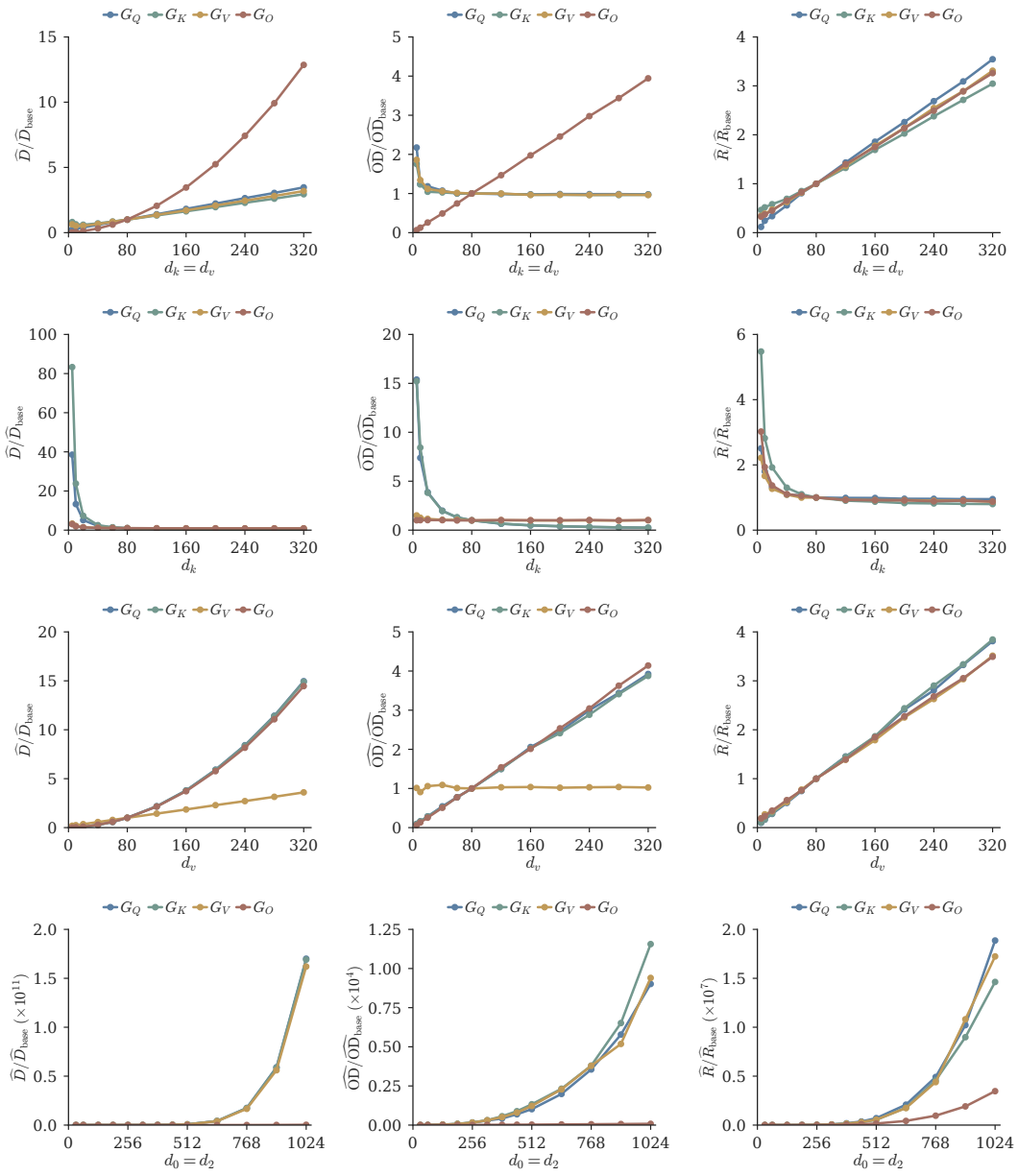


Figure 14: Empirical dimension scaling for single-layer attention. Rows vary $d_k = d_v$ jointly, d_k, d_v, and $d_0 = d_2$ (top to bottom). Columns show $\widehat{D}$, $\widehat{\text{OD}}$, and $\widehat{R} = \widehat{D}/\widehat{\text{OD}}$ for $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V$, and $\boldsymbol{G}_O$, each normalized by its value at the baseline configuration (first three rows: $T = 32$, $d_0 = 512$, $d_k = d_v = 80$, $d_2 = 12$; fourth row: $T = 16, d_0 = d_2 = 32, d_k = d_v = 20$), to facilitate comparison across projections with different absolute scales. Empirical estimates use 10,000 independent initializations per configuration; the off-diagonal statistics are estimated using (125).

H Empirical Evaluation on Large Language Models

We next test whether the diagonal dominance behavior of Theorem 4 persists in transformer language models. We consider Qwen3.5-0.8B-Base (Qwen Team, 2026), Granite-4.0-1B-Base (IBM Granite Team, 2025), and Llama-3.2-1B (Meta AI, 2024). We perform two experiments. First, we examine the expected gradient Grams of individual attention projections at random initializa-

tion. Second, we measure how their diagonal dominance changes with model width across the three models.

H.1 Gradient Gram Structure at Initialization

We first examine the gradient Gram of attention projections in Qwen3.5-0.8B-Base, Granite-4.0-1B-Base, and Llama-3.2-1B at random initialization. The models are instantiated from their released Hugging Face configurations (Wolf et al., 2020) without pretrained weights. We set the sequence length to $T = 256$, retain each model's released model width and head dimension, and set the number of key/value heads equal to the number of query heads. Native RoPE behavior is used throughout.

We analyze head 0 in block 3 for Qwen and in block 0 for Granite and Llama. For Qwen, DeltaNet blocks are excluded, so block 3 is the first full-attention block considered. For each projection $p \in \{Q, K, V, O\}$, we estimate

$$\widehat{\boldsymbol{P}}_p = \frac{1}{N_W N_X} \sum_{w=1}^{N_W} \sum_{x=1}^{N_X} \boldsymbol{G}_{p,w,x} \boldsymbol{G}_{p,w,x}^\top, \tag{127}$$

using $N_W = 25$ independent random initializations and $N_X = 64$ fixed FineWeb-Edu (Penedo et al., 2024) sequences. We use $N_W = 25$ independent random initializations obtained from each model's native Hugging Face initialization scheme, which uses zero-mean Gaussian initialization for the attention projection weights with model-specific scales. The same input sequences are reused across all initializations. For Qwen, the gating rows of the query projection are excluded when forming the query Gram.

Figure 15 shows the resulting empirical expected Grams. All four attention projections exhibit a clear diagonal component across all three models. This indicates that the diagonal structure predicted for the single-layer attention model in Theorem 4 also appears in deeper transformer architectures at random initialization.

H.2 Model Width Scaling

We next examine how diagonal dominance changes with model width. Using the same random initialization protocol as above, we vary the transformer hidden width (`hidden_size`) while holding the head dimension (`head_dim`) fixed at its released value. In the notation of our single-layer attention model, this corresponds to varying the residual input and output dimensions $d_0 = d_2$ while keeping $d_k = d_v$ fixed. The released head dimensions are 256 for Qwen3.5-0.8B-Base, 128 for Granite-4.0-1B-Base, and 64 for Llama-3.2-1B. The query-head count (`num_attention_heads`) is held fixed, and the number of key/value heads (`num_key_value_heads`) is set equal to the number of query heads to enable analogous dimension sweep in the models. The sequence length is fixed at $T = 256$.

We consider the model-width grid

$$\{128, 256, 512, 768, 1024, 1536, 2048, 3072, 4096, 6144, 8192, 12288, 16384\}.$$

At each width, we use 25 independent random initializations (obtained from each model's native Hugging Face initialization scheme, which uses zero-mean Gaussian initialization for the attention projection weights with model-specific scales) and the same 64 fixed FineWeb-Edu sequences. We analyze head 0 in block 3 for Qwen and in block 0 for Granite and Llama. For each projection $p \in \{Q, K, V, O\}$, we report the empirical diagonal-dominance ratio $\widehat{R}_p = R(\widehat{\boldsymbol{P}}_p)$, with $R(\cdot)$ defined in (99).

Figure 16 shows that $\widehat{R}$ increases with model width for all four attention projections across all three models. For $\boldsymbol{G}_O$, this is consistent with the $R = \Theta(d_v d_0^3)$ scaling in Corollary 3. The increase is strongest for $\boldsymbol{G}_Q$ and $\boldsymbol{G}_K$ in Qwen, and for $\boldsymbol{G}_V$ and $\boldsymbol{G}_O$ in Granite and Llama.

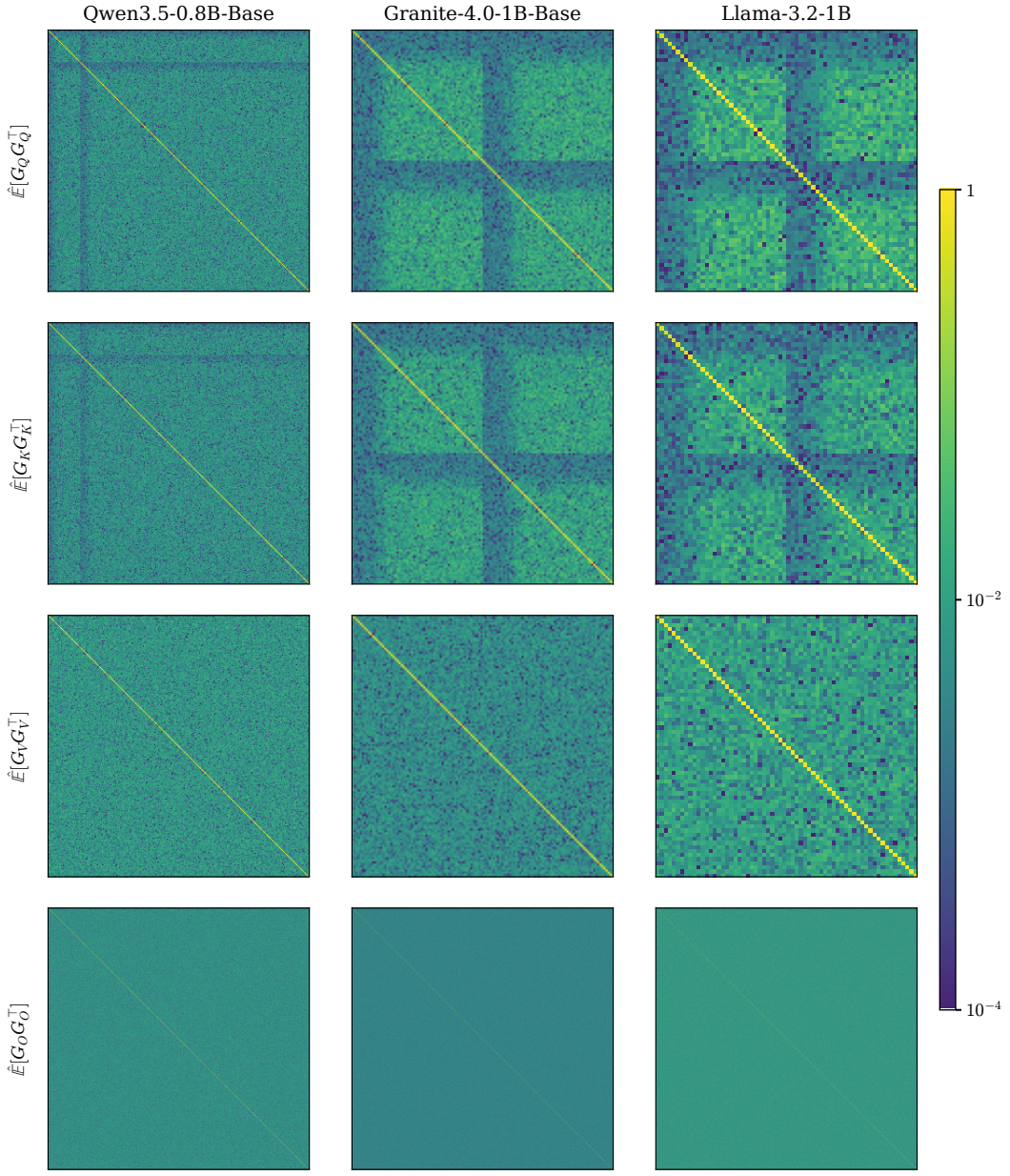


Figure 15: Empirical expected gradient self outer-products for Qwen3.5-0.8B-Base, Granite-4.0-1B-Base, and Llama-3.2-1B (columns), estimated from 25 independent random initializations and 64 fixed FineWeb-Edu sequences of length $T = 256$. The rows show $\widehat{\mathbb{E}}[\boldsymbol{G}_Q \boldsymbol{G}_Q^\top]$, $\widehat{\mathbb{E}}[\boldsymbol{G}_K \boldsymbol{G}_K^\top]$, $\widehat{\mathbb{E}}[\boldsymbol{G}_V \boldsymbol{G}_V^\top]$, and $\widehat{\mathbb{E}}[\boldsymbol{G}_O \boldsymbol{G}_O^\top]$. We use head 0 of block 3 for Qwen and head 0 of block 0 for Granite and Llama. Each heatmap is normalized by its empirical mean diagonal magnitude. A common logarithmic color scale is used across all panels.

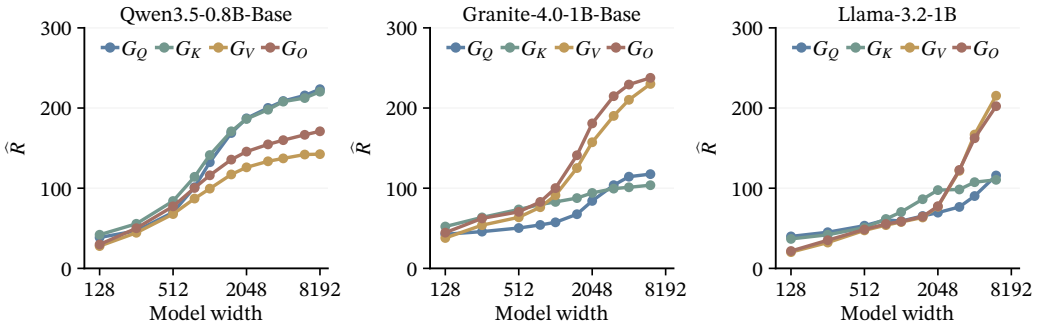


Figure 16: Model width scaling of the empirical diagonal-dominance ratio $\widehat{R}$ for $\boldsymbol{G}_Q, \boldsymbol{G}_K, \boldsymbol{G}_V$, and $\boldsymbol{G}_O$ in Qwen3.5-0.8B-Base, Granite-4.0-1B-Base, and Llama-3.2-1B. Model width (`hidden_size`) is varied while the head dimension (`head_dim`) is held fixed at its released value. We analyze head 0 in block 3 for Qwen and in block 0 for Granite and Llama. Each point is estimated from 25 independent random initializations and 64 fixed FineWeb-Edu sequences of length $T = 256$.