High-dimensional isotropic scaling dynamics of Muon and SGD

Guangyuan Wang GUANGYUAN.WANG@MAIL.MCGILL.CA
Elliot Paquette ELLIOT.PAQUETTE@MCGILL.CA
McGill University

Atish Agarwala THETISH@GOOGLE.COM
Google DeepMind

Abstract

Recent developments in neural network optimization have brought a renewed interest to non-diagonal preconditioning methods. Momentum Orthogonalized by Newton-Schulz (Muon) is a promising algorithm which uses approximate orthogonalization of matrix-valued updates to efficiently traverse poorly conditioned loss landscapes. However, the theoretical underpinnings of Muon's performance, particularly in high-dimensional regimes, remain underexplored. This paper investigates the isotropic scaling dynamics of Muon compared to SGD in a matrix-valued linear regression setting. We derive risk recursion equations for both optimizers under isotropic data assumptions, and find the correct scaling rules for increasing batch size with dimension for efficient training. Our work suggests that in the high dimensional limit, Muon's default normalization by the Frobenius norm may not be sufficient to maintain its nonlinear properties.

1. Introduction

The training of modern large-scale neural networks demands optimizers that can efficiently navigate high-dimensional, non-convex loss landscapes. Many settings, including those involving transformer based models, require the use of some kind of explicit or implicit preconditioning — the most commonly used being AdamW [10, 11], which is generally considered robust and hardware efficient. In recent years researchers have attempted to move beyond diagonal preconditioning methods like AdamW towards non-diagonal preconditioners which take into account larger structure in the gradients [8].

One recent optimizer that has garnered significant interests is Muon, a method which takes advantage of the fact that most parameters are matrix valued [9]. Muon uses Newton-Schulz (NS) iterations to approximately orthogonalize SGD-momentum updates [3]. Empirical results suggest that Muon can significantly accelerate convergence, but there is not yet a good quantitative theoretical understanding of the convergence properties of this algorithm.

We analyze Muon's training dynamics in a setting where we can analytically predict full learning curves in certain high dimensional limits. We introduce a matrix-valued linear regression problem motivated by the gradient structure in real neural networks. Under the assumption of isotropic data, we derive closed-form recursion relations for the risk for SGD and Muon. We leveraged free probability theory [12] to approximate high-dimensional gradient moments for Muon, allowing us to account for the nonlinear effects of NS iterations. We found that the normalization scheme in Muon requires sufficiently large batch size to train well, but at such large batch, the nonlinear nature of the NS iterations vanishes. Our results suggest that future high-dimensional theoretical analysis of Muon is indeed tractable, and suggests that alternative normalization strategies may be needed for Muon for training very large models.

The paper is organized as follows: Section 1 provides background on Muon and its relation to prior optimizers. Section 2 defines notation and assumptions. Sections 3 and 5 present our theoretical results, followed by a discussion of their implications. Proofs can be found in Appendix C.

2. Preliminaries

Muon relies on the fact that the neural network parameters in a single layer are matrix-valued, and tries to exploit this matrix structure to improve optimization. Therefore any theoretical analysis of Muon is only interesting if the problem has the appropriate structure. Motivated by this, we consider the problem of minimizing a stochastic risk function with matrix-valued parameters,

$$\min_{W\in\mathbb{R}^{N_{\text{out}}\times N_{\text{in}}}} \left\{\mathscr{R}(W) := \mathbb{E}_{(x_{\text{in}},x_{\text{out}})}\mathscr{L}(W;(x_{\text{in}},x_{\text{out}}))\right\}, \tag{1}$$

given input and output characteristic vectors $(x_{\text{in}}, x_{\text{out}})$, and $\mathscr{L}$ is the mean squared error (MSE)

$$\mathscr{L}(W;(x_{\text{in}},x_{\text{out}})) = \tfrac{1}{2}\left(f(x_{\text{in}},x_{\text{out}}) - y(W,x_{\text{in}},x_{\text{out}})\right)^2, \tag{2}$$

where $f(x_{\text{in}}, x_{\text{out}})$ is the target value and $y(W, x_{\text{in}}, x_{\text{out}})$ is the model prediction.

In this work, we focus on a matrix-valued linear regression model $y(W, x_{\text{in}}, x_{\text{out}}) := x_{\text{out}}^\top W x_{\text{in}}$, where $y \in \mathbb{R}$ is a scalar output, $W \in \mathbb{R}^{N_{\text{out}}\times N_{\text{in}}}$ is the parameter matrix. In iteration t, we minimize the stochastic risk in (1) by parameterizing the algorithm using

$$\text{student } W_t \in \mathbb{R}^{N_{\text{out}}\times N_{\text{in}}} \quad \text{with fixed teacher} \quad W^\star := \operatorname{argmin}_W \mathscr{R}(W) \in \mathbb{R}^{N_{\text{out}}\times N_{\text{in}}} \tag{3}$$

used to generate noiseless targets $f(x_{\text{in}}, x_{\text{out}}) := x_{\text{out}}^\top W^\star x_{\text{in}}$. Here, $x_{\text{in}} \in \mathbb{R}^{N_{\text{in}}}$ is an input feature vector, and $x_{\text{out}} \in \mathbb{R}^{N_{\text{out}}}$ is an output feature vector, both are drawn from isotropic distributions satisfying $\mathbb{E}x_{\text{in}}^{\otimes 2} = \sigma_1^2 \operatorname{Id}_{N_{\text{in}}}$ and $\mathbb{E}x_{\text{out}}^{\otimes 2} = \sigma_2^2 \operatorname{Id}_{N_{\text{out}}}$. Define $T := \mathbb{E}(x_{\text{out}} \otimes x_{\text{in}})^{\otimes 2}$. Then, $T_{ijk\ell} = \delta_{ik}\delta_{j\ell}$. This implies that the risk (i.e., the loss (2) averaged over $(x_{\text{in}}, x_{\text{out}})$) is

$$\mathscr{R}(W) := \mathbb{E}_{(x_{\text{in}},x_{\text{out}})}\mathscr{L}(W) = \tfrac{1}{2}\mathbb{E}\left\langle (x_{\text{out}}\otimes x_{\text{in}})^{\otimes 2}, (W - W^\star)^{\otimes 2}\right\rangle = \tfrac{1}{2}\|W - W^\star\|_{\text{F}}^2. \tag{4}$$

To analyze optimization algorithms, it is important to first examine the structure of the gradient. For a batch of B samples, the stochastic gradient is calculated analytically in the implementation as

$$G = \frac{1}{B}\sum_{i=1}^{B}(y_{\text{predicted}}^{(i)} - y_{\text{target}}^{(i)}) \cdot x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \tag{5}$$

where $\otimes$ denotes the tensor product. This can be rewritten in matrix form as $G = B^{-1}(Z \cdot x_{\text{out}})^\top \cdot x_{\text{in}}$ where $x_{\text{in}} \in \mathbb{R}^{B\times N_{\text{in}}}$ is a matrix containing the batch of x_{in} vectors (each row is a sample), $x_{\text{out}} \in \mathbb{R}^{B\times N_{\text{out}}}$ is a matrix containing the batch of x_{out} vectors (each row is a sample), and $Z \in \mathbb{R}^{B\times B}$ is a diagonal matrix with entries $Z_{ii} = (y_{\text{predicted}}^{(i)} - y_{\text{target}}^{(i)})$, the per-sample residuals. The resulting gradient $G \in \mathbb{R}^{N_{\text{out}}\times N_{\text{in}}}$ is a matrix with the same shape as the parameter matrix W. This structure is similar to gradients in fully connected layers of neural networks, where $(y_{\text{predicted}}^{(i)} - y_{\text{target}}^{(i)})x_{\text{out}}^{(i)}$ is replaced by the derivative $\partial\mathscr{L}/\partial h$ for an activation vector h in an intermediate layer.

The key insight is that the distribution and correlation of the inputs directly affect the singular value spectrum of the gradient. When the batch size B is comparable to the dimensions of the parameter matrix (N_{in} and N_{out}), higher order moments of the distributions become important. In this work we will focus on the i.i.d. input case; we expect these effects are even more important for anisotropic data distributions.

3. Optimization Algorithms

3.1. Standard Stochastic Gradient Descent (SGD)

For each batch of data $(x_{\text{in}}, x_{\text{out}})$ with batch size B, SGD updates the parameters as

$$W_{t+1} = W_t - \eta \cdot \nabla_W \mathscr{L}(W_t; (x_{\text{in}}, x_{\text{out}})) = W_t - \eta \cdot \frac{1}{B} \sum_{i=1}^{B} \nabla_W \mathscr{L}(W_t; (x_{\text{in}}^{(i)}, x_{\text{out}}^{(i)})), \tag{6}$$

where η is the learning rate, $x_{\text{in}} = \{x_{\text{in}}^{(i)}\}_{i=1}^{B}, x_{\text{out}} = \{x_{\text{out}}^{(i)}\}_{i=1}^{B}$, and the gradient is averaged over all samples in the batch. To compare with Muon, we also consider *normalized SGD* where the gradient is scaled by its Frobenius norm — that is, the replacement $G_{t+1} \leftarrow G_{t+1}/\|G_{t+1}\|_{\text{F}}$, where $\|G_{t+1}\|_{\text{F}} = \sqrt{\langle G_{t+1}, G_{t+1}\rangle}$. The parameter update then becomes $W_{t+1} = W_t - \eta_t G_{t+1}/\|G_{t+1}\|_{\text{F}}$.

3.2. Momentum Orthogonalized by Newton-Schulz (Muon)

Muon, as defined by [9], applies a Newton-Schulz (NS) iteration to approximately orthogonalize the gradient matrix. This prevents the network from learning only in a few dominant directions and ensures isotropic updates. For a batch of data $(x_{\text{in}}, x_{\text{out}})$ of size B, we compute the batch gradient $G_{t+1} \in \mathbb{R}^{N_{\text{out}} \times N_{\text{in}}}$ at iteration $t+1$ as

$$G_{t+1} = \frac{1}{B} \sum_{i=1}^{B} \nabla_W \mathscr{L}(W_t; (x_{\text{in}}^{(i)}, x_{\text{out}}^{(i)})) = \frac{1}{B} \sum_{i=1}^{B} x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \langle x_{\text{out}}^{(i)}, (W_t - W^\star) x_{\text{in}}^{(i)} \rangle. \tag{7}$$

Then, we form the momentum buffer $M_{t+1} = \mu M_t = (1-\mu) G_{t+1}$ with $\mu \in [0,1)$. The momentum step may be implemented with other averaging conventions; our analysis below uses this canonical form. Recall that for $p \in \mathbb{R}\backslash\{0\}$ and bounded linear operator $R \in \mathscr{L}(\mathscr{H})$, we define the *Schatten p-norm* of R as $\|R\|_p := \operatorname{Tr}(|R|^p)^{1/p}$. This extends to $\|R\|_\infty := \lambda_{\max}(|R|)$, where $\lambda_{\max}(|R|)$ is the largest eigenvalue of $|R|$. The gradient is normalized by its Schatten p-norm before NS iteration; i.e., $\widetilde{M}_t \leftarrow M_t/\|M_t\|_p = M_t/\operatorname{Tr}(|M_t|^p)^{1/p}$. In particular, the standard implementation of Muon [9] sets $p = 2$, which is equivalent to normalizing the gradient by its Frobenius norm $\|M_t\|_{\text{F}}$. This is the case we analyze in the main text; in Appendix D.3 we present the scaling analysis for the case $p = \infty$, where $\|M_t\|_\infty$ is the operator norm of the gradient.

The NS iteration then approximates the orthogonalization operation

$$\operatorname{Ortho}(\Xi) = \underset{O \in \mathbb{R}^{N_{\text{out}} \times N_{\text{in}}} : \|O\|_2 \leq 1}{\operatorname{argmin}} \left\{ \|O - \Xi\|_{\text{F}} : \text{either } O^\top O = \operatorname{Id} \text{ or } OO^\top = \operatorname{Id} \right\}. \tag{8}$$

Along with the momentum step, this is equivalent to replacing $\widetilde{M}_t$ with the nearest semi-orthogonal matrix $UV^\top$ from its singular value decomposition $\widetilde{M}_t = U\Sigma V^\top$. Specifically, we can view the one-step NS iteration as using a quintic polynomial in Σ to approximate the matrix sign function,

$$O_{t+1} = \Phi_5(\Sigma; a, b, c) \widetilde{M}_{t+1}, \quad \text{where} \quad \Phi_5(\Sigma; a, b, c) = a \operatorname{Id} + b\Sigma\Sigma^\top + c(\Sigma\Sigma^\top)^2, \tag{9}$$

with fixed hyperparameters a, b, c. One can apply K Newton-Schulz iterations to project the momentum buffer onto the nearest semi-orthogonal matrix. Denote the K-step NS iterates by $\operatorname{NS}_K(\cdot)$ and the output after K steps by $O_{t+1} := O_{t+1}^{(K)} = \operatorname{NS}_K(\widetilde{M}_{t+1})$. Now, NS_K is an explicit, iterative map that converges rapidly to the semi-orthogonal factor of its argument. Finally, we update the parameters with the orthogonalized momentum $W_{t+1} = W_t - \eta O_{t+1}$.

4. Main results

In this section, we present the core theoretical results for the convergence of SGD and Muon in a matrix-valued optimization setting with isotropic input and output data. We analyze the expected risk dynamics under large-dimensional regimes, where the batch size B, input dimension N_{in}, and output dimension N_{out} scale to infinity with ratios to be determined later. Theorem 2 establishes a risk recursion for SGD with normalized gradients, extending classical analyses to matrix-valued problems. Theorem 4 derives a similar recursion for the Muon update, leveraging free probability techniques to approximate gradient moments in high dimensions. These results provide insights into the interplay between batch size, problem dimensions, and convergence behavior.

We emphasize that prior theoretical analyses rarely address matrix-valued objectives under isotropic data, despite their relevance to modern architectures. As a consequence, even the behavior of SGD in this regime is not fully understood. Establishing a precise baseline for SGD with normalized gradients is thus a prerequisite for evaluating Muon, enabling a controlled comparison that isolates the role of Newton-Schulz orthogonalization in shaping the risk dynamics.

In what follows, the filtration $\mathscr{F}_t = \sigma(W_0, G_1, \ldots, G_t)$ captures the history up to step t. Let the risk be defined as in (4), where $W, W^\star \in \mathbb{R}^{N_{\text{out}} \times N_{\text{in}}}$, $x_{\text{out}} \in \mathbb{R}^{N_{\text{out}}}$, $x_{\text{in}} \in \mathbb{R}^{N_{\text{in}}}$. Assume $B, N_{\text{in}}, N_{\text{out}} \to \infty$ with ratios to be determined later. The input and output vectors $x_{\text{in}}, x_{\text{out}}$ are i.i.d. with $\mathbb{E}[x_{\text{in}}^{\otimes 2}] = \sigma_1^2 \operatorname{Id}_{N_{\text{in}}}$ and $\mathbb{E}[x_{\text{out}}^{\otimes 2}] = \sigma_2^2 \operatorname{Id}_{N_{\text{out}}}$ for $\sigma_1, \sigma_2 > 0$. The stochastic gradient G_t satisfies $\mathbb{E}[\|G_t\|_{\text{F}}^2 | \mathscr{F}_t] \leq C\mathscr{R}(W_{t-1})$ for some constant $C > 0$. The learning rate $\eta_t > 0$ satisfies $\eta_t = O(1/L)$.

Theorem 1 (SGD risk recursion with unnormalized gradient) *The expected SGD risk given as in (1) at iteration $t+1$, conditioned on the filtration $\mathscr{F}_t$, satisfies the finite difference equation*

$$\mathbb{E}\left[\mathscr{R}(W_{t+1}) | \mathscr{F}_t\right] = \Big(1 - 2\eta_t + \eta_t^2 B^{-1}\left(B + 3 + N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}})\right)\Big)\mathscr{R}(W_t). \tag{10}$$

Theorem 2 (SGD risk recursion with normalized gradient) *Denote the constant $\kappa := B/(B + N_{\text{in}}N_{\text{out}})$. For large batch size B and problem dimensions $N_{\text{in}}, N_{\text{out}}$, the expected risk of streaming SGD at iteration $t+1$, conditioned on $\mathscr{F}_t$, satisfies the finite difference equation*

$$\mathbb{E}\left[\mathscr{R}(W_{t+1}) | \mathscr{F}_t\right] = \mathscr{R}(W_t) - \eta_t \sqrt{2\kappa}\sqrt{\mathscr{R}(W_t)} + \frac{\eta_t^2}{2} + O\left(\eta_t \frac{\sqrt{\mathscr{R}(W_t)}}{\sqrt{B^2\kappa^{-1}}}\right). \tag{11}$$

Remark 3 *This result extends the classical SGD convergence analyses [5, 7] to the matrix-valued setting with isotropic input and output data and the fully connected normalized gradient structure.*

One interesting batch size scaling regime for SGD is the *large batch regime* $B = \alpha N_{\text{in}}N_{\text{out}}$ for an $O(1)$ constant α (does not scale with N_{in} or N_{out}). This is the largest regime we can take, which trains efficiently (in terms of flops or number of data points processed) and shows universal behavior for training at a reasonable speed with non-scaling step size. In this regime for vanilla SGD (Theorem 1), the risk recursion shows a linear decay rate modulated by $1 - \eta_t + \eta_t^2(1 + \alpha^{-1})$, implying convergence in $O(1)$ steps. For normalized SGD (Theorem 2), the optimal learning rate scales as $\eta_t = O(\sqrt{\kappa})$; in this regime, $\sqrt{\kappa} = \sqrt{2\alpha/(\alpha+1)} = O(1)$ again leading to convergence in a dimension-independent number of steps. the asymptotic risk $R_\infty \sim \frac{\eta_\infty^2}{8\kappa}$ is also well-behaved.

An alternative batch size scaling regime for SGD is the *batch-fan proportional regime* where batch is proportional to the matrix widths — $N_{\text{in}}/B = \phi$ and $N_{\text{out}}/B = \psi$ for $O(1)$ constants ϕ and

ψ. For unnormalized SGD (Theorem 1), the risk recursion yields a linear decay rate modulated by the constant $\kappa^{-1} = B\phi\psi + o(1)$. The optimal learning rate in this case is proportional to κ, so the learning rate is $O(B^{-1})$. In this regime learning takes $O(B)$ steps to reach a target loss value; this implies that $O(B^2) = O(N_{\text{in}}N_{\text{out}})$ samples must be processed. This is in fact just as sample efficient as the large batch regime, which trains using $O(1)$ steps but $O(N_{\text{in}}N_{\text{out}})$ samples per step. For normalized SGD, the limiting risk at long times is given by $\kappa^{-1}\eta_\infty^2/8$. In this setting getting a dimensionally-independent value for the risk requires $\eta \propto B^{-1/2}$. This once again implies $O(B)$ steps for convergence, again matching the sample efficiency of the large batch regime. Detailed analysis of both regimes can be found in Appendix C.3.

Similarly, Theorem 4 below develops an analogous recursion for Muon and reveals how the adaptive preconditioning built into Muon modifies the contraction rates of the expected risk.

Theorem 4 (Muon risk recursion on isotropic data) *Let the one-step Muon update be given by*

$$W_{t+1} = W_t - \eta G_{t+1}, \quad G_{t+1} = \big(a\,\mathrm{Id} + b(G_tG_t^\top) + c(G_tG_t^\top)^2\big)G_t, \tag{12}$$

where the quintic polynomial $\Phi_5(\Sigma; a, b, c) = a\Sigma + b\Sigma^3 + c\Sigma^5$ approximates the matrix sign function in the limit. Assume that the dimensions $B, N_{\text{in}}, N_{\text{out}} \to \infty$ with fixed ratios $B/N_{\text{in}} \to \phi$, $B/N_{\text{out}} \to \psi$, and the gradient moments $\mathbb{E}[\langle \Delta_t, (G_tG_t^\top)^q G_t\rangle | \mathscr{F}_t]$ can be approximated by their free probability limits, dominated by non-crossing partitions, as described in Proposition 9. For each sample i, let the quadratic $z_t^{(i)} := \langle x_{\text{out}}^{(i)}, \Delta_t x_{\text{in}}^{(i)}\rangle$ be the per-sample residual inner product. Under isotropic assumptions, these are approximately i.i.d. Gaussian with variance $\sigma_{\Delta_t}^2$. Then, in the joint large-dimensional limit, the expected risk at iteration $t+1$, conditioned on $\mathscr{F}_t$, is

$$\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] = \tfrac{1}{2}\left(\|\Delta_t\|_{\mathrm{F}}^2 - 2\eta\mathbb{E}[\langle \Delta_t, G_{t+1}\rangle|\mathscr{F}_t] + \eta^2\mathbb{E}\left[\|G_{t+1}\|_{\mathrm{F}}^2|\mathscr{F}_t\right]\right) \tag{13}$$

$$= \mathscr{R}(W_t) - \eta\mathscr{D}(\mathscr{R}(W_t)) + \tfrac{1}{2}\eta^2\mathscr{V}(\mathscr{R}(W_t)). \tag{14}$$

The drift and variance terms are given by, respectively,

$$\begin{aligned}
&\mathscr{D}(\mathscr{R}(W_t)) := \mathbb{E}[\langle \Delta_t, G_{t+1}\rangle|\mathscr{F}_t] \\
&= \frac{4a\mathscr{R}(W_t)}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^{1/2}} + \frac{bN_{\text{in}}N_{\text{out}}}{B^2}\frac{\mathbb{E}z^2}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^{3/2}} \\
&\quad + \left(\frac{N_{\text{in}}N_{\text{out}}}{B^3}(\mathbb{E}z^2)^2 + \frac{N_{\text{in}}N_{\text{out}}^2}{B^4}(\mathbb{E}z^4) + \frac{N_{\text{out}}^2}{B^3}(\mathbb{E}z^2)^2\right)\frac{2c\mathscr{R}(W_t)}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^{5/2}}(1+o(1)),
\end{aligned}$$

and

$$\begin{aligned}
\mathscr{V}(\mathscr{R}(W_t)) := \mathbb{E}\left[\|G_{t+1}\|_{\mathrm{F}}^2|\mathscr{F}_t\right] = \Bigg(&a^2 + \frac{2ab\mathbb{E}\,\mathrm{Tr}((G_tG_t^\top)^2)}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^2} + \frac{(b^2+2ac)\mathbb{E}\,\mathrm{Tr}((G_tG_t^\top)^3)}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^3} \\
&+ \frac{2bc\mathbb{E}\,\mathrm{Tr}((G_tG_t^\top)^4)}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^4} + \frac{c^2\mathbb{E}\,\mathrm{Tr}((G_tG_t^\top)^5)}{(\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top))^5}\Bigg)(1+o(1)),
\end{aligned} \tag{15}$$

where the higher order gradient moments $\mathbb{E}\,\mathrm{Tr}((G_tG_t^\top)^q)$, $1 \le q \le 5$, are given by (102).

A detailed analysis of the dynamics under the different batch size scaling regimes can be found in Appendix D.2; we summarize the key points here.

In the isotropic setting, the large batch regime $B = \alpha N_{\text{in}} N_{\text{out}}$ quickly leads to uninteresting dynamics for Muon due to the fact that the spectrum of $GG^\top$ degenerates to a point, and all the moments $\mathbb{E}[\text{Tr}((G_t G_t^\top)^p)]$ are proportional to $\mathbb{E}[\text{Tr}(G_t G_t^\top)]$. This may not be true in the anisotropic case; see discussion.

The batch-fan regime avoids issues with the spectrum of $GG^\top$, but another issue arises: the quintic polynomial Φ_5 degenerates to its first-order approximation as dimension becomes large. This means that at very large model size, the dynamics of Muon degenerates to normalized SGD. The effects of the non-linear terms can be rescued with alternative normalization schemes; for example using the operator norm $p = \infty$, the third order term contributes even at infinite width (Appendix D.3). We hypothesize that this normalization issue may still be present in many anisotropic settings as well.

The cost of the NS iterations are manageable in both scaling regimes. The per-sample gradient computational cost often will scale at least as $O(BN_{\text{in}}N_{\text{out}})$ (particularly when examining real models, e.g. transformer blocks on long sequences). The NS iterations scale closer to $O(N_{\text{in}}N_{\text{out}}^2 + N_{\text{out}}N_{\text{in}}^2)$, which is the same order as the per-sample gradients for the batch-fan proportional regime, and subleading in the large batch regime. Correctly identifying computational bottlenecks in practice can be more difficult and relies on careful analysis of parallelization strategies and hardware utilization. Regardless, our work suggests that in the isotropic regime the batch-fan scaling is more promising but requires rethinking of the normalization in Muon to scale to very large matrix sizes.

Finally, we note that in transformer architectures the Muon update enjoys an additional structural advantage: the NS orthogonalization acts only on the parameter matrix itself and is independent of the sequence length. As a result, the computational overhead of NS does not grow with context size. This makes the method particularly appealing in large-sequence transformer regimes.

5. Discussion

Our analysis provides theoretical insights into the convergence behavior of SGD and the Muon optimizer in the matrix-valued setting with isotropic data and highlights the role of orthogonalization in balancing gradient singular values. One key finding is that for SGD, the large batch ($B = \alpha N_{\text{in}} N_{\text{out}}$) and batch fan proportional ($N_{\text{in}}/B = \phi$, $N_{\text{out}}/B = \psi$) regimes both have similar dynamics and computational efficiency; however, for Muon, in the isotropic case the large batch regime degenerates quickly. It is not clear under what conditions this degeneracy occurs for the anisotropic case; this is a topic for future study.

We also found that even the batch fan proportional regime loses the non-linear information from NS at large enough N due to the normalization scheme, and eventually degenerates to (normalized) SGD. This suggests that at very large model sizes alternative normalization schemes may indeed be necessary to maintain good, predictable performance. We conjecture that this issue is more fundamental and persists over a large variety of data distributions.

Our results highlight the importance of our matrix flavored linear regression model; the standard high-dimensional linear regression does not have the structure to probe these behaviors. Our work suggests that theoretical analysis of Muon is indeed tractable using methods from random matrix theory. The next step is to repeat this study in anisotropic settings, where Muon is expected to actually outperform SGD. Studies in these more realistic settings may uncover actionable insights about Muon and suggest potential improvements to the algorithm.

References

[1] Sebastien Abou Assaly and Lucas Benigni. Eigenvalue distribution of the hadamard product of sample covariance matrices in a quadratic regime. *arXiv preprint arXiv:2502.12374*, 2025.

[2] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.

[3] Jeremy Bernstein. Old dog, new trick: A matrix perspective on optimizer development. *arXiv preprint*, 2024.

[4] Daniel Birmajer, Juan B Gil, and Michael D Weiner. Colored partitions of a convex polygon by noncrossing diagonals. *Discrete Mathematics*, 340(4):563–571, 2017.

[5] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

[6] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024.

[7] Robert M. Gower, Markus Schmidt, and Peter Richtárik. Sgd: General analysis and improved rates. *arXiv preprint*, 2019.

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

[9] Jordan Keller. Muon: A minimal algorithm for unsupervised online learning. `https://kellerjordan.github.io/posts/muon/`, 2024. Accessed: 2025-08-04.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*, 2017.

[12] James A. Mingo and Roland Speicher. *Free Probability and Random Matrices*, volume 35 of *Fields Institute Monographs*. Springer, New York, 2017. ISBN 978-1-4939-6941-8. doi: 10.1007/978-1-4939-6942-5.

[13] N. J. A. Sloane. The on-line encyclopedia of integer sequences. `http://oeis.org`, 2025.

Appendix A. Algorithms pseudocode

Algorithm 1: Isotropic Muon

```
Data: $\eta > 0$, $\mu \in [0, 1)$, $(x_{\text{in}}, x_{\text{out}})$, $a$, $b$, $c$
Result: Optimized parameters $W_T$
$M_{-1} \leftarrow 0, W_{-1}, W^\star \sim N(0, \mathrm{Id}_{N_{\text{out}} \times N_{\text{in}}})$
for $t = 0$ to $T - 1$ do
    $G_t \leftarrow \nabla_W \mathscr{L}(W_t + \mu M_{t-1}; (x_{\text{in}}, x_{\text{out}}))$
    $M_t \leftarrow \mu M_{t-1} + (1 - \mu) G_t$
    $\widetilde{M}_t \leftarrow M_t / \|M_t\|_p$
    $O_t \leftarrow \textsc{NewtonSchulz}(\widetilde{M}_t; a, b, c)$
    $W_{t+1} \leftarrow W_t - \eta O_t$
end
return $W_T$
```

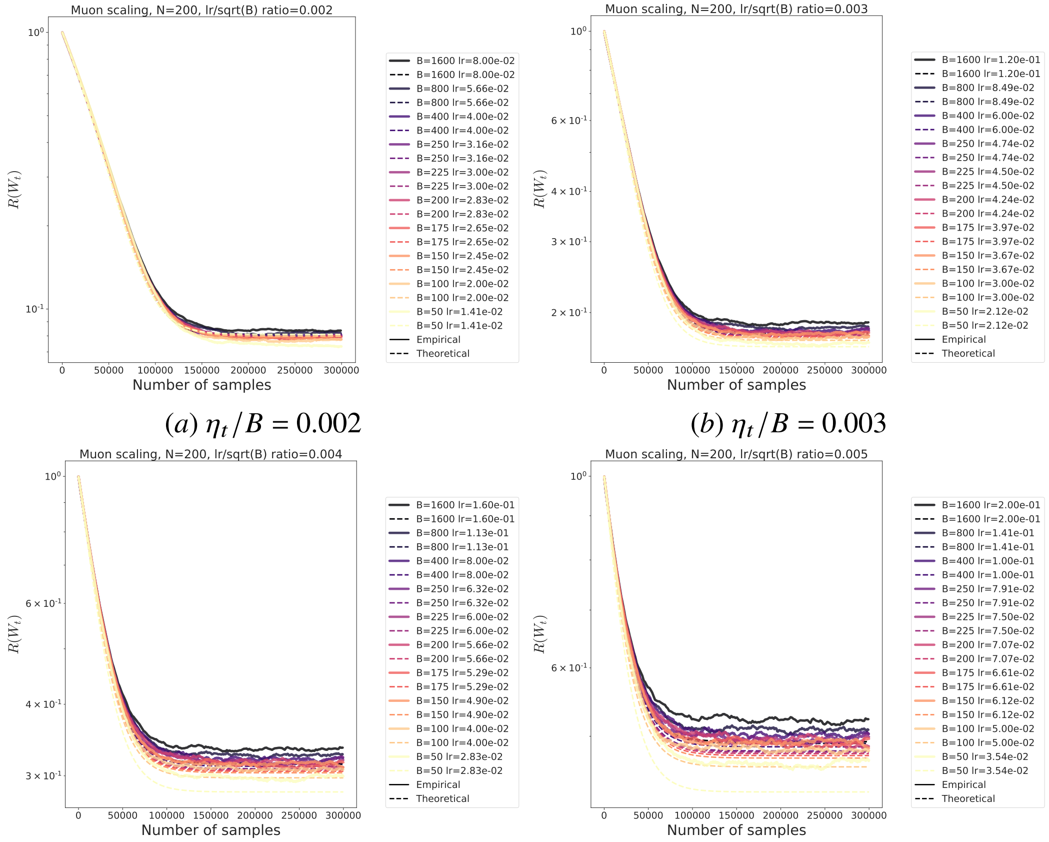
Algorithm 2: Isotropic streaming SGD

```
Data: $\eta > 0$, $\mu \in [0, 1)$, $(x_{\text{in}}, x_{\text{out}})$
Result: Optimized parameters $W_T$
$M_{-1} \leftarrow 0, W_{-1}, W^\star \sim N(0, \mathrm{Id}_{N_{\text{out}} \times N_{\text{in}}})$
for $t = 0$ to $T - 1$ do
    $G_t \leftarrow \nabla_W \mathscr{L}(W_{t-1}; (x_{\text{in}}, x_{\text{out}}))$
    $M_T \leftarrow \mu M_{t-1} + (1 - \mu) G_t$
    $W_{t+1} \leftarrow W_t - \eta M_T$ ;
end
return $W_T$
```

Appendix B. Plots and experimental analysis

To investigate the scaling behavior of SGD and Muon, we perform a sweep over different batch sizes B and learning rates η_t such that the ratio η_t / B remains constant. This allows us to isolate the effect of scaling N, the system size, or model dimensionality, while holding the effective learning rate per sample fixed.

Experiments are conducted for $N = 200, 300, 500$, and the performance of both algorithms is evaluated in terms of convergence speed and final loss reached. The results indicate that as N increases, Muon maintains a relatively consistent convergence while SGD degrades more. Notably, maintaining a fixed η_t / B exposes differences in how each algorithm handles gradient noise and curvature scaling with model size. For larger N, Muon exhibits better robustness to batch-size-induced variance, suggesting that Muon's adaptive components scale more favorably under fixed effective learning rate conditions. These findings highlight the importance of considering algorithmic stability and noise sensitivity when scaling model size, even under normalized optimization hyperparameters like fixed η_t / B.



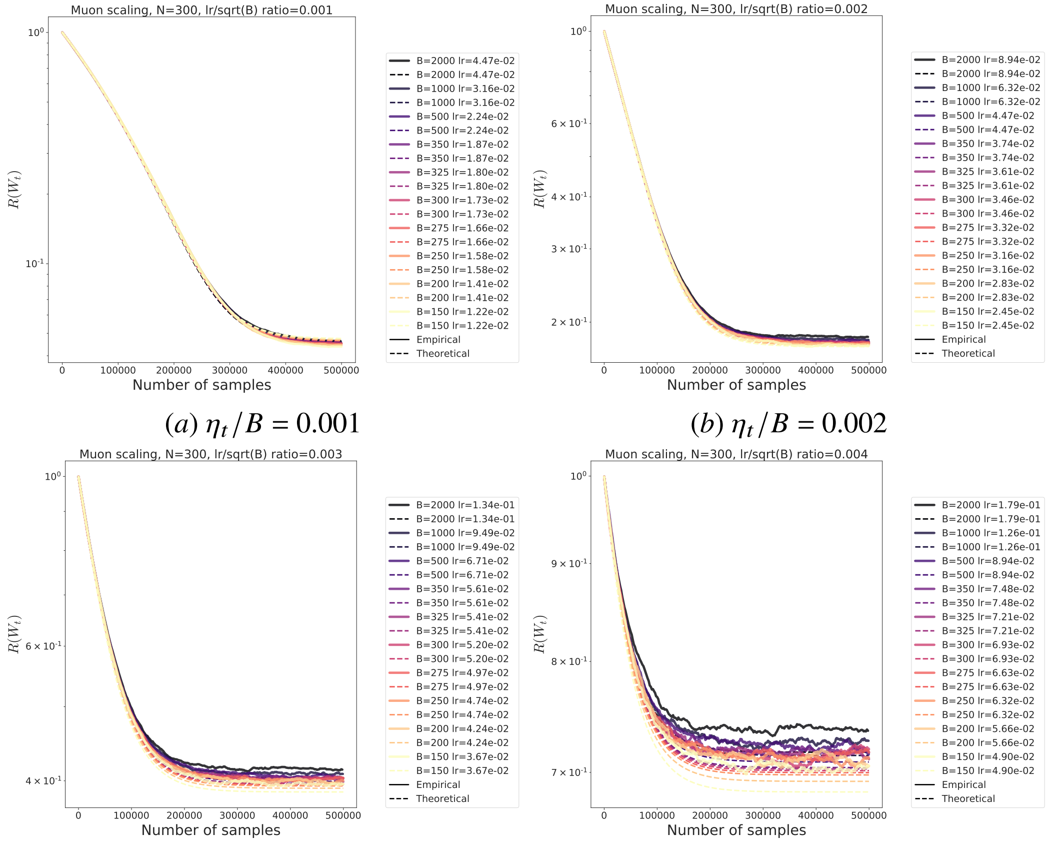
(a) $\eta_t/B = 0.002$

(b) $\eta_t/B = 0.003$

(c) $\eta_t/B = 0.004$

(d) $\eta_t/B = 0.005$

Figure 1: Muon scaling, sweeping over fixed η_t/B ratio ($N = 200$)



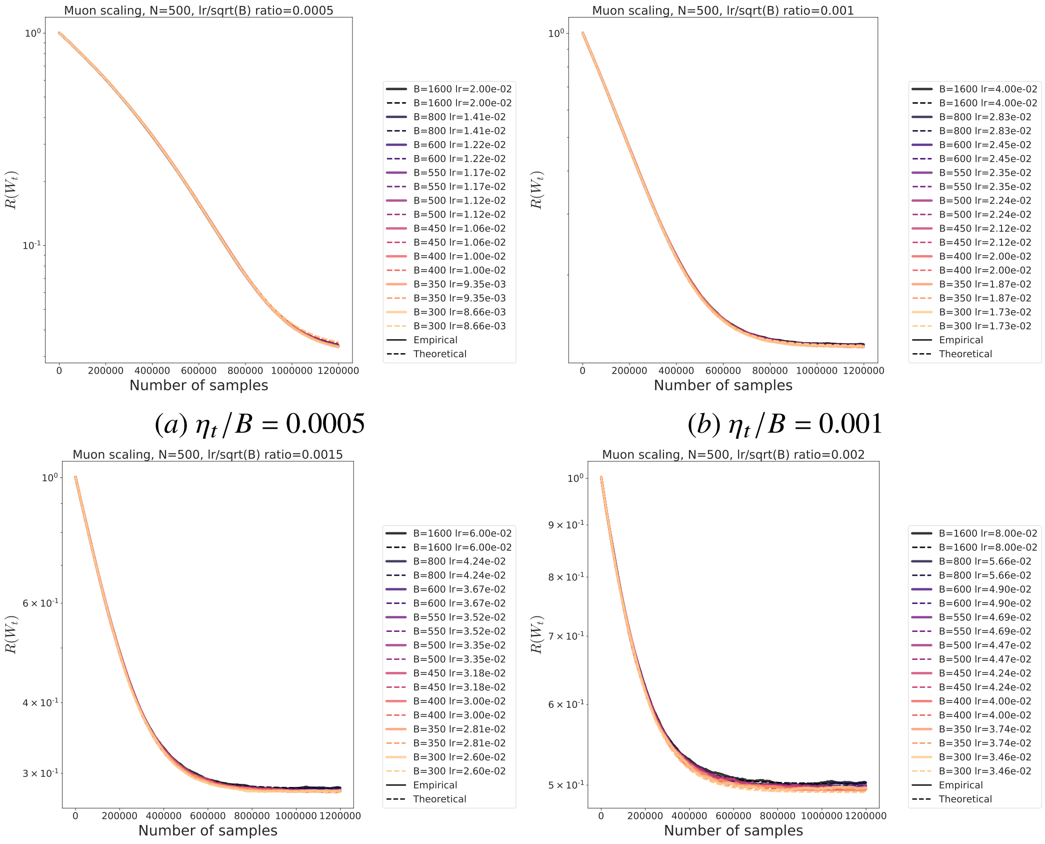
(a) $\eta_t/B = 0.001$

(b) $\eta_t/B = 0.002$

(c) $\eta_t/B = 0.003$

(d) $\eta_t/B = 0.004$

Figure 2: Muon scaling, sweeping over fixed η_t/B ratio ($N = 300$)



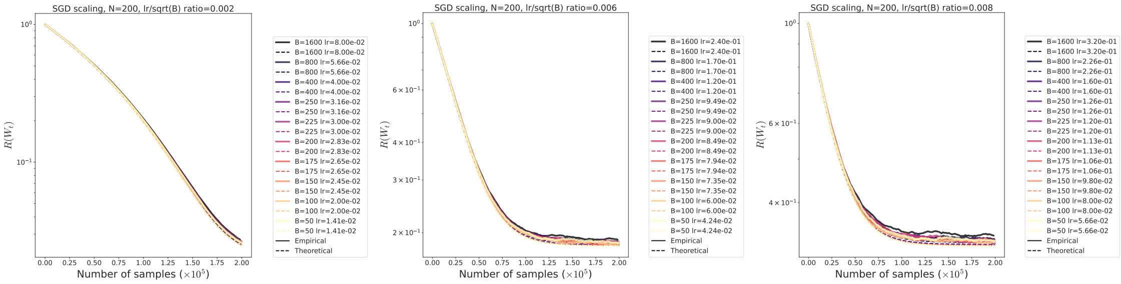
(a) $\eta_t/B = 0.0005$

(b) $\eta_t/B = 0.001$

(c) $\eta_t/B = 0.0015$

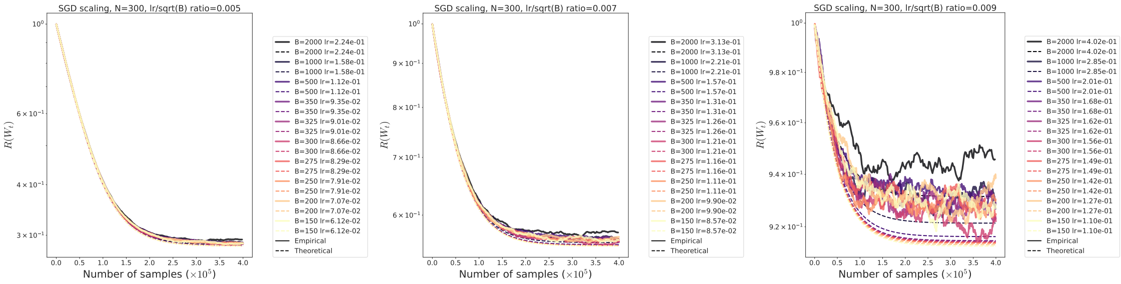
(d) $\eta_t/B = 0.003$

Figure 3: Muon scaling, sweeping over fixed η_t/B ratio ($N = 500$)



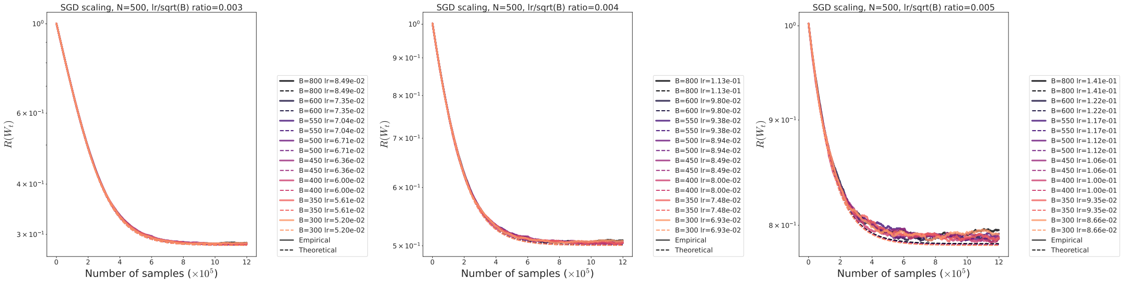
(*a*) $\eta_t/B = 0.002$ (*b*) $\eta_t/B = 0.006$ (*c*) $\eta_t/B = 0.008$

Figure 4: Normalized SGD scaling, sweeping over fixed η_t/B ratio ($N = 200$)



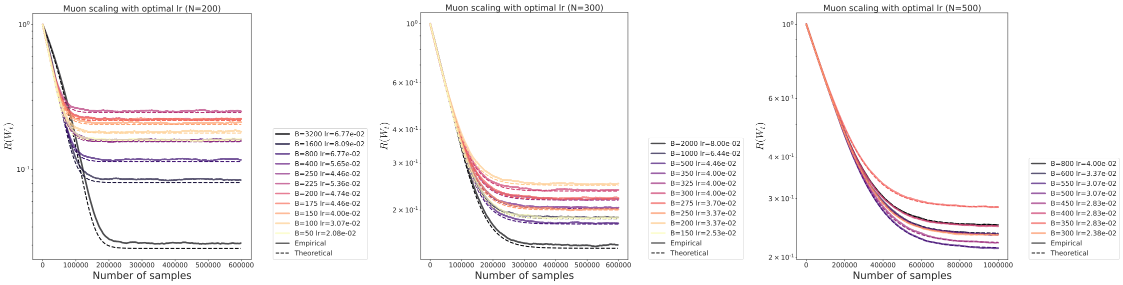
(*a*) $\eta_t/B = 0.005$ (*b*) $\eta_t/B = 0.007$ (*c*) $\eta_t/B = 0.009$

Figure 5: Normalized SGD scaling, sweeping over fixed η_t/B ratio ($N = 300$)



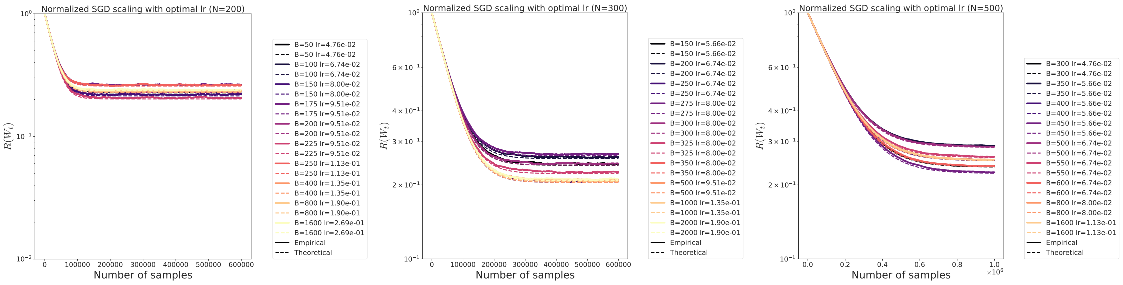
(*a*) $\eta_t/B = 0.003$ (*b*) $\eta_t/B = 0.004$ (*c*) $\eta_t/B = 0.005$

Figure 6: Normalized SGD scaling, sweeping over fixed η_t/B ratio ($N = 500$)



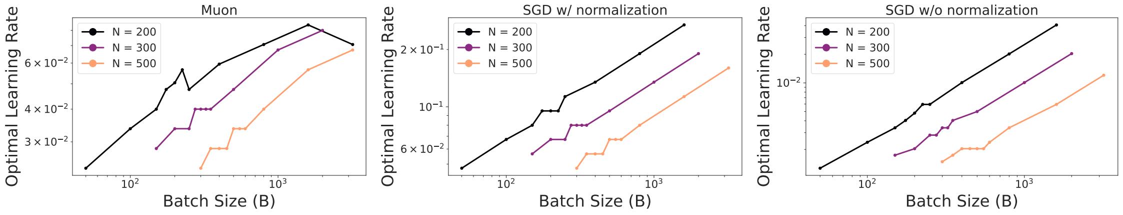
(*a*) $N = 200$ (*b*) $N = 300$ (*c*) $N = 500$

Figure 7: Muon scaling, sweeping logarithmically over batch sizes with optimal learning rates



(*a*) $N = 200$ (*b*) $N = 300$ (*c*) $N = 500$

Figure 8: SGD scaling, sweeping logarithmically over batch sizes with optimal learning rates



(*a*) Muon (*b*) SGD w/ normalization (*c*) SGD w/o normalization

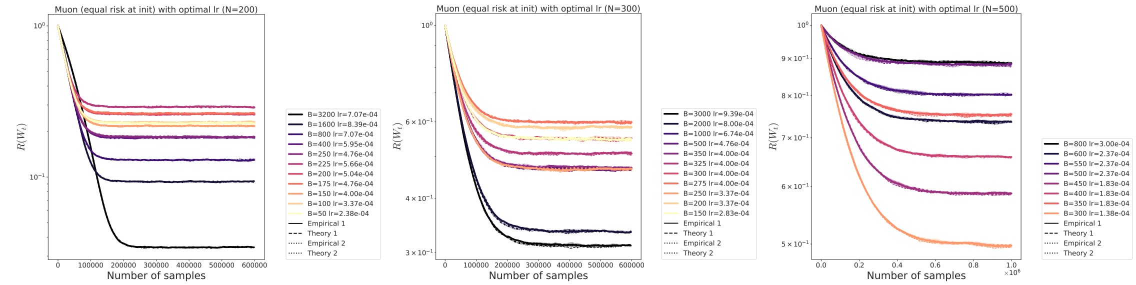
Figure 9: Scaling of optimal learning rates as a function of batch sizes for all three algorithms.

We optimize the learning rate for each fixed N and each algorithm via grid search, and pick the learning rate that gives the lowest loss after some fixed number of samples processed. This gives good early time learning properties for SGD and for Muon. We sweep logarithmically over $2^{0.25}$. Then, for each algorithm, we plot the set of learning curves across B for these optimal learning rates with $num\_samples = B \times steps$ on the x-axis. Both x- and y-axes scales are held fixed in each column of Figure 7 and Figure 8.

Figure 7 and Figure 8 show that, near the crossover region $B \sim N$, SGD behaves more stably and is better collapsed, while Muon diverges more easily while only slightly varying batch sizes inside the crossover region. Overall, Muon shows greater variability in the final convergent loss values than SGD, when the batch size B is large (e.g., when reaching $16 \times N$). Initially, both algorithms are exponentially decaying with the steepest descent possible. However, Muon and SGD converge differently even when tuned optimally, where SGD plateaus earlier while Muon continues decreasing longer, indicating different asymptotic behaviors. It is also interesting that we observe different optimal ratios between Muon and SGD: with a fixed $\eta/\sqrt{B}$ ratio, SGD handles larger effective step sizes without diverging or incurring high risk. They achieve different optimal risk values at the same B and N values. In general, Muon almost always reaches optimal performance at larger B compared to SGD, which indicates that the position of optimal batch size with respect to a fixed N differs between SGD and Muon.

Moreover, Figure 9 shows the optimal learning rate versus batch size for Muon and normalized SGD. The Muon curves vary much more in final loss reached as a function of B/N, while SGD curves show fairly linear scaling trend even at larger batch sizes compared to muon, which already starts to level off.

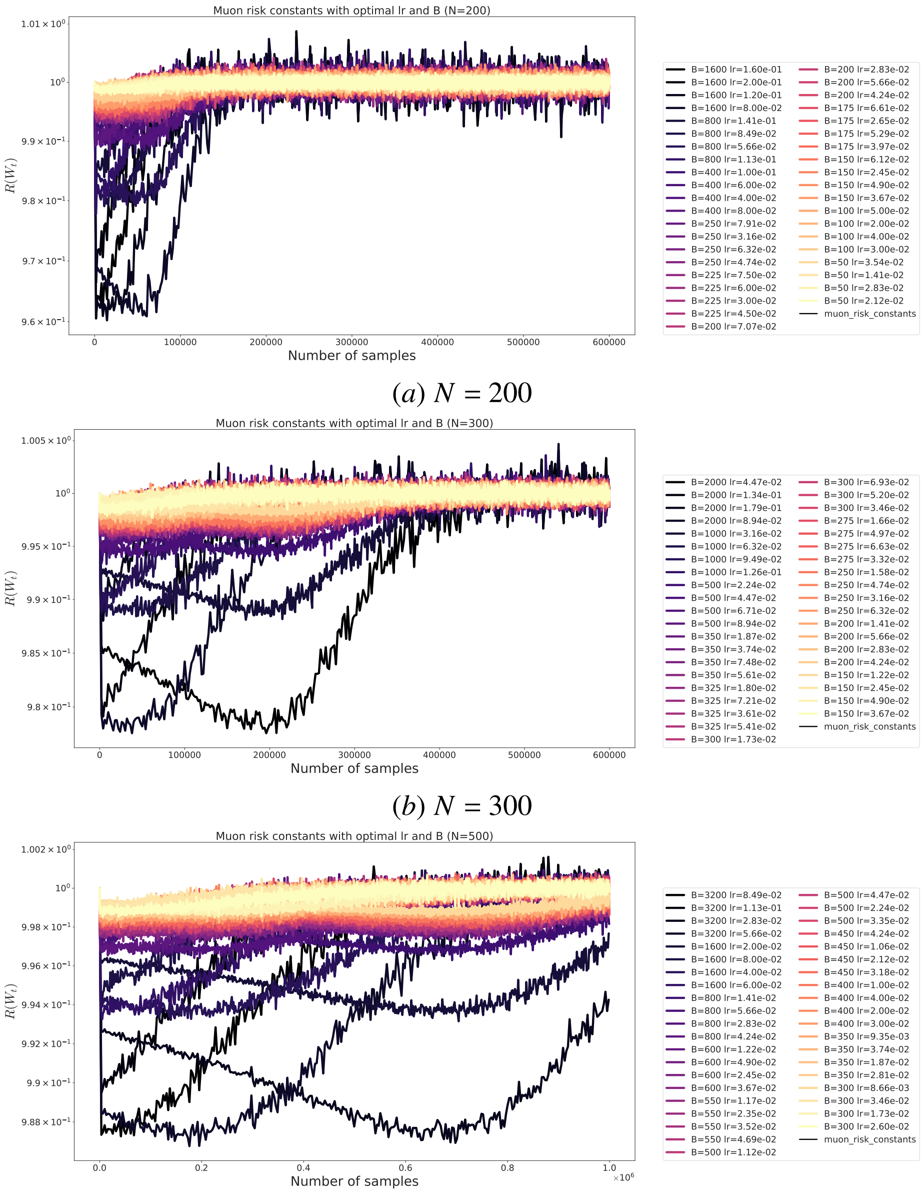
The above phenomena clearly indicate that there are behaviors of Muon that cannot be approximated by SGD at any learning rate or batch size. Muon deviates from the small batch size universal regime in a way different from SGD, in addition to having different learning curves. In particular, the effects around $B = N$, as we transition from $B < N$ to $B > N$, is much stronger for muon as compared to SGD.



(a) $N = 200$ (b) $N = 300$ (c) $N = 500$

Figure 10: Muon scaling with different $W_0, W^\star$ setup while preserving $\|W_t - W^\star\|_{\mathrm{F}}$, swept over optimal learning rates for each batch size.

In Figure 11, we test two different initializations for Muon with the same starting risk and plot loss curves, where we initialize W_1 as an i.i.d. random matrix, W_2 as a random matrix with singular values that are all 1. We scale W_2 by some factor α so that $\|W_1 - W^\star\| = \|\alpha W_2 - W^\star\|$. Note that we use the same $W^\star$ for both simulations. We average over 5 seeds for the sampling randomness of the trajectory while keeping $W_1, \alpha W_2, W^\star$ the same for all runs. This indicates that matrix statistics do not matter much in the isotropic case, as long as the initial expected loss is kept equal.

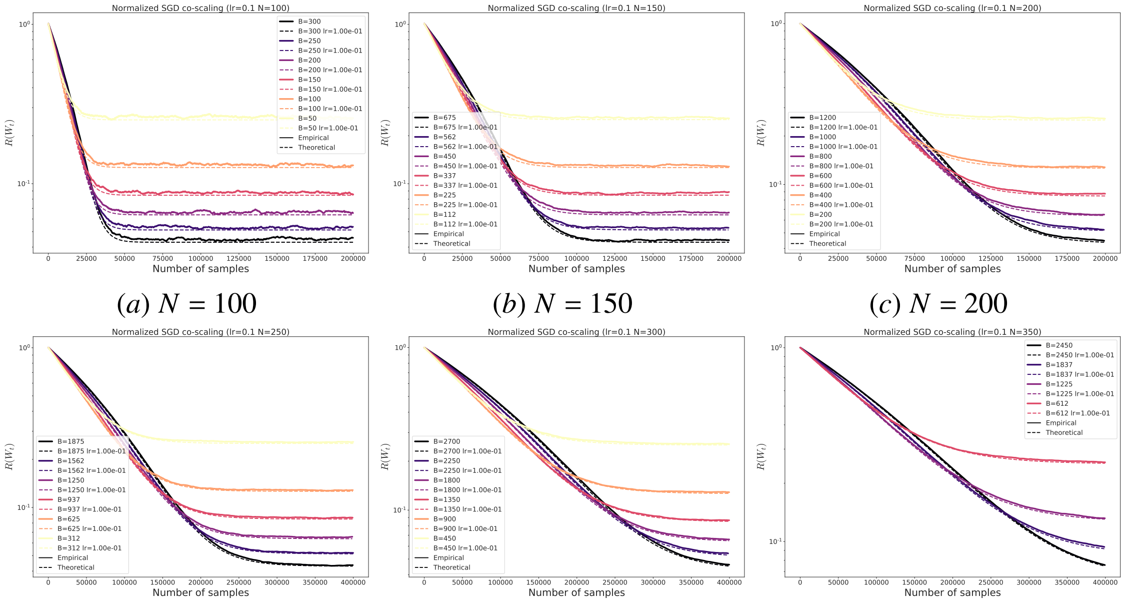


(a) $N = 200$

(b) $N = 300$

(c) $N = 500$

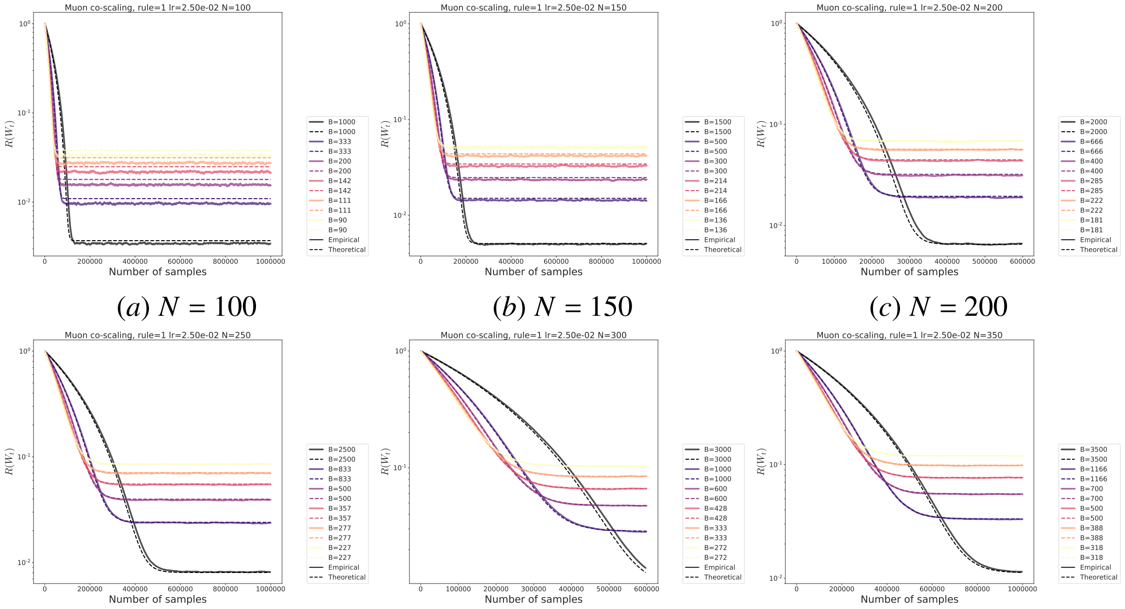
Figure 11: Muon risk update ratio $\mathscr{R}(W_{t+1})/\mathscr{R}(W_t)$, swept over various optimal learning rates for each batch size. It can be seen that the Muon risk recurrence does not follow the same affine linear pattern in the SGD risk update as in (10).



(a) $N = 100$ (b) $N = 150$ (c) $N = 200$

(d) $N = 250$ (e) $N = 300$ (f) $N = 350$

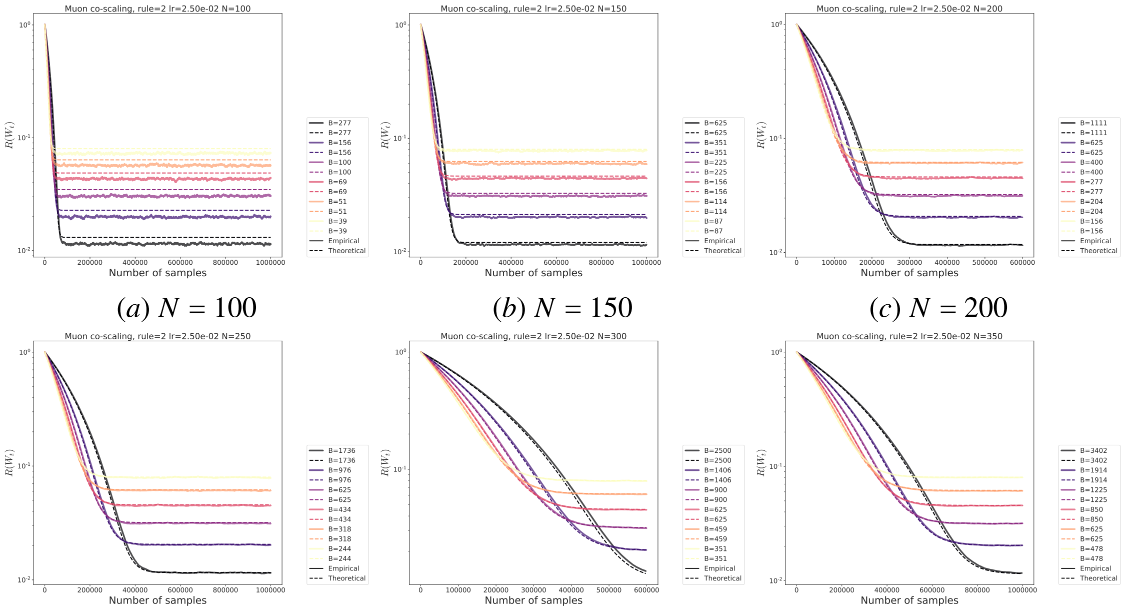
Figure 12: Normalized SGD co-scaling with fixed $B/N_{\text{in}}N_{\text{out}}$.



(a) $N = 100$ (b) $N = 150$ (c) $N = 200$

(d) $N = 250$ (e) $N = 300$ (f) $N = 350$

Figure 13: Muon co-scaling (rule 1) with $N_{\text{in}}/B = \phi, N_{\text{out}}/B = \psi$.



(*a*) $N = 100$

(*b*) $N = 150$

(*c*) $N = 200$

(*d*) $N = 250$

(*e*) $N = 300$

(*f*) $N = 350$

Figure 14: Muon co-scaling (rule 2) with $N_{\text{in}}/\sqrt{B} = \phi$, $N_{\text{out}}/\sqrt{B} = \psi$.

Appendix C. SGD risk, isotropic case

C.1. Basic SGD risk recursion on isotropic data, Theorem 1

Proof Let the SGD risk be $\mathscr{R}(W_t) := \frac{1}{2}\mathbb{E}\langle x_{\text{out}}, (W_t - W^\star)x_{\text{in}}\rangle$. If $\Xi := \mathbb{E}(x_{\text{out}} \otimes x_{\text{in}})^{\otimes 2}$, then $\Xi_{i,j,k,\ell} = \delta_{ik}\delta_{j\ell}$, and

$$\begin{aligned}\mathscr{R}(W_t) &= \tfrac{1}{2}\mathbb{E}\left\langle (x_{\text{out}} \otimes x_{\text{in}})^{\otimes 2}, (W_t - W^\star)^{\otimes 2}\right\rangle = \tfrac{1}{2}\left\langle \Upsilon, (W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \tfrac{1}{2}\sum_{i,j}(W_t - W^\star)_{ij}(W_t - W^\star)_{k,\ell}\delta_{ik}\delta_{j\ell} = \tfrac{1}{2}\|W_t - W^\star\|_{\text{F}}^2.\end{aligned} \tag{16}$$

where the first equality follows from $\langle v, Aw\rangle = \langle v \otimes w, A\rangle$ and $\langle a, b\rangle \otimes \langle c, d\rangle = \langle a \otimes c, b \otimes d\rangle$. Denote by $\mathscr{F}_t := \sigma\big(W_s, ((x_{\text{in}})_s^{(i)}, (x_{\text{out}})_s^{(i)})_{i=1}^B : s \le t\big)$ the natural filtration up to time t in the SGD process, where $((x_{\text{in}})_s^{(i)}, (x_{\text{out}})_s^{(i)})_{i=1}^B$ is a batch of i.i.d. input-output pairs sampled at each iteration to compute the stochastic gradient, and W_t is the weight matrix at iteration t which is $\mathscr{F}_t$-measurable. The expected risk is then

$$\begin{aligned}\mathbb{E}(\mathscr{R}(W_{t+1}|\mathscr{F}_t)) &= \tfrac{1}{2}\mathbb{E}(\|W_{t+1} - W^\star\|_{\text{F}}^2|\mathscr{F}_t) \\ &= \tfrac{1}{2}\mathbb{E}\left[\left\|W_t - W^\star - \eta_t B^{-1}\sum_{i=1}^{B} x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}\langle x_{\text{out}}^{(i)}, (W_t - W^\star)x_{\text{in}}^{(i)}\rangle\right\|_{\text{F}}^2\middle|\mathscr{F}_t\right] \\ &= \tfrac{1}{2}\left(\|W_t - W^\star\|_{\text{F}}^2 - 2\eta_t\langle W_t - W^\star, \mathbb{E}\left[G_{t+1}|\mathscr{F}_t\right]\rangle + \eta_t^2\mathbb{E}\left[\langle G_{t+1}, G_{t+1}\rangle|\mathscr{F}_t\right]\right).\end{aligned} \tag{17}$$

In the third equality, we use the definition of the gradient $G_{t+1} := B^{-1}\sum_{i=1}^B x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}\langle x_{\text{out}}^{(i)}, (W_t - W^\star)x_{\text{in}}^{(i)}\rangle$. Expanding the inner product of G_{t+1} gives two cases, where we have squared terms with the same index i and cross-product terms with different indices $i \neq j$, i.e.,

$$\mathbb{E}\left(\langle G_{t+1}, G_{t+1}\rangle|\mathscr{F}_t\right) = B^{-2}\big(1_{i=j}\sum_{i\in[B]}\mathbb{E}\Big[\|G_{t+1}\|_{\text{F}}^2\Big|\mathscr{F}_t\Big] + 1_{i\neq j}\sum_{i,j\in[B]}\mathbb{E}\Big[\|G_{t+1}\|_{\text{F}}^2\Big|\mathscr{F}_t\Big]\big). \tag{18}$$

To evaluate the second sum in (18), note that

$$\begin{aligned}&\mathbb{E}\langle x_{\text{out}}^{(i)}, x_{\text{out}}^{(j)}\rangle\langle x_{\text{in}}^{(i)}, x_{\text{in}}^{(j)}\rangle\langle x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}, W_t - W^\star\rangle\langle x_{\text{out}}^{(j)} \otimes x_{\text{in}}^{(j)}, W_t - W^\star\rangle \\ &= \mathbb{E}\left\langle x_{\text{out}}^{(i)} \otimes x_{\text{out}}^{(j)} \otimes x_{\text{in}}^{(i)} \otimes x_{\text{in}}^{(j)} \otimes x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \otimes x_{\text{out}}^{(j)} \otimes x_{\text{in}}^{(j)}, \text{Id} \otimes \text{Id} \otimes (W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \left\langle \delta_{15}\delta_{27}\delta_{36}\delta_{48}, \delta_{12}\delta_{34}(W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \sum_{i,j,k,\ell}(W_t - W^\star)_{ij}(W_t - W^\star)_{k,\ell}.\end{aligned} \tag{19}$$

Definition 5 (Non-crossing pairings [12]) *Let π be a partition of $[n]$. If there exists $i < j < k < \ell$ such that i, k are in one block V of π and j, ℓ are in another block W of π, we say that V and W* cross. *If no pair of blocks of π cross, then we say that π is* non-crossing. *The set of non-crossing partitions of $[n]$ is denoted $\mathscr{NC}(n)$. The set of non-crossing pairings of $[n]$ is denoted $\mathscr{NC}_2(n)$.*

Example 1

(i) If we set $i = 5, j = 6, k = 7, \ell = 8$, then with paired partitions in the form $\pi = \{(1,5),(2,7),(3,6),(4,8)\} \in \mathscr{NC}_2(8) = \mathscr{P}_2(8)$, the above sum immediately simplifies to

$$\sum_{i,j,k,\ell} (W_t - W^\star)_{ij} (W_t - W^\star)_{k,\ell} \delta_{ik} \delta_{j\ell} = \sum_{i,j} (W_t - W^\star)_{ij}^2 = \|W_t - W^\star\|_{\mathrm{F}}^2. \tag{20}$$

(ii) We form non-crossing partition $\pi = \{(5,7),(1,2),(6,3,4,8)\} \in \mathscr{NC}(8)$ with the same index assignments

$$\delta_{12}\delta_{57}\delta_{34}\delta_{68}\delta_{12}\delta_{34}(W_t - W^\star)^{\otimes 2} = N_{\text{out}} \|W_t - W^\star\|_{\mathrm{F}}^2, \tag{21}$$

where the N_{out} factor is due to the fact that indices $1, 2$ are isolated and none of i, j, k, ℓ equals them. Similarly, the isolation of $(3, 4)$ will contribute a factor of N_{in}.

Theorem 6 (Wick) *Let $X_1, X_2, \ldots, X_n$ be a Gaussian family. Then we have for any $k \in \mathbb{N}$ and $1 \le i(1), i(2), \ldots, i(k) \le n$ that*

$$\mathbb{E}\left[\prod_{i=1}^{k} X_{i(j)}\right] = \sum_{\pi \in \mathscr{P}_2(k)} \prod_{\{r,s\} \in \pi} \mathbb{E}\left[X_{i(r)} X_{i(s)}\right]. \tag{22}$$

Here, $\mathscr{P}_2(k)$ denotes the set of all pairings of the set $\{1, \ldots, k\}$.

Definition 7 (Tensor contractions [6]) *Let $\mathscr{A}, \mathscr{O}$ be finite-dimensional real vector spaces, which we equip with inner products and so are finite-dimensional Hilbert spaces. Recall that as a vector space $\mathscr{A} \otimes \mathscr{O}$ is all (finite) linear combinations of simple tensors, i.e., those of the form $a \otimes b$ where $a \in \mathscr{A}$ and $b \in \mathscr{O}$. This becomes an algebra, allowing scalars to commute, i.e., for $c \in \mathbb{R}$,*

$$c(a \otimes b) = (ca) \otimes b = a \otimes (cb) \tag{23}$$

and by allowing $\otimes$ to distribute over addition,

$$(a + b) \otimes c = (a \otimes c) + (b \otimes c) \quad \text{and} \quad a \otimes (b + c) = (a \otimes b) + (a \otimes c). \tag{24}$$

General tensor contractions generalize matrix multiplication and dot products. We will use the inner product $\langle \cdot, \cdot \rangle$ operator in various ways to describe this contraction. Each $\mathscr{A}$ and $\mathscr{O}$ carries with it an inner product, and so $\mathscr{A} \otimes \mathscr{O}$ has a natural inner product which for simple tensors is defined by

$$\langle a \otimes b, c \otimes d \rangle_{\mathscr{A} \otimes \mathscr{O}} = \langle a, c \rangle_{\mathscr{A}} \langle b, d \rangle_{\mathscr{O}}. \tag{25}$$

This is extended to the full space $\mathscr{A} \otimes \mathscr{O}$ by bilinearity. This, for example, can be connected to the Frobenius inner product. If we represent an element $A \in \mathbb{R}^d \otimes \mathbb{R}^\ell$ in the orthonormal basis $\{e_i \otimes e_j\}$ as $A = \sum_{i,j} A_{ij} e_i \otimes e_j$, then we have the identification

$$\langle A, B \rangle_{\mathscr{A} \otimes \mathscr{O}} = \sum_{i,j} A_{ij} B_{ij} = \operatorname{Tr}\left(AB^\top\right). \tag{26}$$

In particular, the dot products written above extend naturally to

$$(\mathscr{A} \otimes \mathscr{O})^{\otimes 2} := (\mathscr{A} \otimes \mathscr{O}) \otimes (\mathscr{A} \otimes \mathscr{O}) \cong \mathscr{A}^{\otimes 2} \otimes \mathscr{O}^{\otimes 2} \tag{27}$$

where the last isomorphism corresponds to reshaping the tensor to have its ambient directions listed first, and its observable directions second. Furthermore, tensor computations naturally give rise to an inner product on higher tensor products, which we define first for simple tensors, $t_i := (a_i \otimes o_i)$ *for* $i = 1, 2, 3, 4$*,*

$$\begin{aligned}\langle t_1 \otimes t_2, t_3 \otimes t_4\rangle_{(\mathscr{A}\otimes\mathscr{O})^{\otimes 2}} &= \langle t_1, t_3\rangle_{\mathscr{A}\otimes\mathscr{O}}\, \langle t_2, t_4\rangle_{\mathscr{A}\otimes\mathscr{O}} \\ &= \langle a_1, a_3\rangle_{\mathscr{A}}\, \langle a_2, a_4\rangle_{\mathscr{A}}\, \langle o_1, o_3\rangle_{\mathscr{O}}\, \langle o_2, o_4\rangle_{\mathscr{O}}\,. \end{aligned} \tag{28}$$

This is once more extended by multi-linearity.

From Theorem 6, we know that

$$\mathbb{E}(X^{\otimes 2n}) = \sum_{\text{pairings}} \prod_{\{u,v\}\in\text{pairing}} \mathbb{E}(X_u \otimes X_v), \tag{29}$$

we have that And thus in the case $i = j$, the first sum in (18) evaluates to

$$\begin{aligned}&\mathbb{E}\langle x_{\text{out}}^{(i)}, x_{\text{out}}^{(i)}\rangle\langle x_{\text{in}}^{(i)}, x_{\text{in}}^{(i)}\rangle\langle x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}, W_t - W^\star\rangle\langle x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}, W_t - W^\star\rangle \\ &= \mathbb{E}\left\langle (x_{\text{out}}^{(i)})^{\otimes 4} \otimes (x_{\text{in}}^{(i)})^{\otimes 4}, \text{Id}^{\otimes 2} \otimes (W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \left\langle (\delta_{12}\delta_{57} + \delta_{15}\delta_{27} + \delta_{17}\delta_{25})(\delta_{34}\delta_{68} + \delta_{36}\delta_{48} + \delta_{38}\delta_{46}), \delta_{12}\delta_{34}(W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \big(N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}}) + 4\big)\|W_t - W^\star\|_{\text{F}}^2.\end{aligned} \tag{30}$$

In summary,

$$\begin{aligned}&\eta_t^2\mathbb{E}(\langle G_{t+1}, G_{t+1}\rangle|\mathscr{F}_t) \\ &= \frac{\eta_t^2}{B^2}\left(B(B-1)\|W_t - W^\star\|_{\text{F}}^2 + B\big(N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}}) + 4\big)\|W_t - W^\star\|_{\text{F}}^2\right) \\ &= \frac{\eta_t^2}{B}(B + 3 + N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}}))\|W_t - W^\star\|_{\text{F}}^2.\end{aligned} \tag{31}$$

Similarly, we can expand the second term in (17):

$$\begin{aligned}&\left\langle W_t - W^\star, \mathbb{E}\left[(x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)})\langle x_{\text{out}}^{(i)}, (W_t - W^\star)x_{\text{in}}^{(i)}\rangle\middle|\mathscr{F}_t\right]\right\rangle \\ &= \left\langle W_t - W^\star, \mathbb{E}\left[(x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)})\langle x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}, W_t - W^\star\rangle\middle|\mathscr{F}_t\right]\right\rangle \\ &= \mathbb{E}\left[\left\langle x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)}, W_t - W^\star\right\rangle^2\middle|\mathscr{F}_t\right] \\ &= \left\langle \mathbb{E}\left[(x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)})^{\otimes 2}\middle|\mathscr{F}_t\right], \text{Id}^{\otimes 2} \otimes (W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \left\langle \delta_{13}\delta_{24}, \delta_{12}\delta_{34}(W_t - W^\star)^{\otimes 2}\right\rangle \\ &= \sum_{i,j,k,\ell}(W_t - W^\star)_{ij}(W_t - W^\star)_{k,\ell}\delta_{ik}\delta_{j\ell}\delta_{ij}\delta_{k,\ell} \\ &= \|W_t - W^\star\|_{\text{F}}^2,\end{aligned} \tag{32}$$

and therefore

$$\begin{aligned}
&2\eta_t \langle W_t - W^\star, \mathbb{E}(G_{t+1}|\mathscr{F}_t)\rangle \\
&= \frac{2\eta_t}{B}\sum_{i=1}^{B} \mathbb{E}\left[\left\langle W_t - W^\star, x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \left\langle x_{\text{out}}^{(i)}, (W_t - W^\star) x_{\text{in}}^{(i)} \right\rangle\right\rangle\middle|\mathscr{F}_t\right] \\
&= 2\eta_t \|W_t - W^\star\|_{\text{F}}^2.
\end{aligned} \tag{33}$$

Combining everything yields

$$\begin{aligned}
&\mathbb{E}(\mathscr{R}(W_{t+1}|\mathscr{F}_t)) \\
&= \tfrac{1}{2}\Big(\|W_t - W^\star\|_{\text{F}}^2 - 2\eta_t\|W_t - W^\star\|_{\text{F}}^2 \\
&\quad + \eta_t^2 B^{-1}\left(B + 3 + N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}})\right)\|W_t - W^\star\|_{\text{F}}^2\Big) \\
&= \tfrac{1}{2}\left(1 - 2\eta_t + \eta_t^2 B^{-1}\left(B + 3 + N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}})\right)\right)\|W_t - W^\star\|_{\text{F}}^2 \\
&= \mathbb{E}(\mathscr{R}(W_t)|\mathscr{F}_t)\left(1 - 2\eta_t + \eta_t^2 B^{-1}\left(B + 3 + N_{\text{in}}N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}})\right)\right),
\end{aligned} \tag{34}$$

where $\frac{1}{2}\|W_t - W^\star\|_{\text{F}}^2 = \mathscr{R}(W_t) = \mathbb{E}(\mathscr{R}(W_t)|\mathscr{F}_t)$ as $\mathscr{R}(W_t)$ is $\mathscr{F}_t$-measurable.

■

C.2. SGD risk updates with normalized gradient, Theorem 2

Proof To derive the SGD risk equation with a normalized gradient, we start with the given SGD risk and modify the gradient update to account for normalization by the Frobenius norm. The risk is defined as $\mathscr{R}(W_t) = \frac{1}{2}\mathbb{E}\langle x_{\text{out}}, (W_t - W^\star)x_{\text{in}}\rangle = \frac{1}{2}\|W_t - W^\star\|_{\text{F}}^2$, and the natural filtration is $\mathscr{F}_t = \sigma(W_s, ((x_{\text{in}})_s^{(i)}, (x_{\text{out}})_s^{(i)})_{i=1}^B : s \le t)$. The unnormalized gradient is $G_{t+1} = B^{-1}\sum_{i=1}^B x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \langle x_{\text{out}}^{(i)}, (W_t - W^\star)x_{\text{in}}^{(i)}\rangle$. For the normalized case, the gradient is scaled by its Frobenius norm via $G_{t+1} \leftarrow \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}$, where $\|G_{t+1}\|_{\text{F}} = \sqrt{\langle G_{t+1}, G_{t+1}\rangle}$. The parameter update becomes $W_{t+1} = W_t - \eta_t \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}$. We need to compute the expected risk at the next iteration, $\mathbb{E}(\mathscr{R}(W_{t+1})|\mathscr{F}_t) = \frac{1}{2}\mathbb{E}(\|W_{t+1} - W^\star\|_{\text{F}}^2|\mathscr{F}_t)$. Substituting the update rule

$$W_{t+1} - W^\star = W_t - W^\star - \eta_t \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}. \tag{35}$$

Taking the Frobenius norm squared and the conditional expectation,

$$\begin{aligned}
&\mathbb{E}(\|W_{t+1} - W^\star\|_{\text{F}}^2|\mathscr{F}_t)\\
&= \mathbb{E}\left[\left\|W_t - W^\star - \eta_t \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}\right\|_{\text{F}}^2 \middle| \mathscr{F}_t\right]\\
&= \mathbb{E}\left[\|W_t - W^\star\|_{\text{F}}^2 - 2\eta_t\left\langle W_t - W^\star, \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}\right\rangle + \eta_t^2\left\|\frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}\right\|_{\text{F}}^2 \middle| \mathscr{F}_t\right]\\
&= \|W_t - W^\star\|_{\text{F}}^2 - 2\eta_t\mathbb{E}\left[\left\langle W_t - W^\star, \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}\right\rangle \middle| \mathscr{F}_t\right] + \eta_t^2.
\end{aligned} \tag{36}$$

Thus, the expected risk is

$$\mathbb{E}(\mathscr{R}(W_{t+1})|\mathscr{F}_t) = \frac{1}{2}\left(\|W_t - W^\star\|_{\text{F}}^2 - 2\eta_t\mathbb{E}\left[\left\langle W_t - W^\star, \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}\right\rangle \middle| \mathscr{F}_t\right] + \eta_t^2\right). \tag{37}$$

Since $\mathscr{R}(W_t) = \frac{1}{2}\|W_t - W^\star\|_{\text{F}}^2$ is $\mathscr{F}_t$-measurable, we have

$$\mathbb{E}(\mathscr{R}(W_{t+1})|\mathscr{F}_t) = \mathscr{R}(W_t) - \eta_t\mathbb{E}\left[\left\langle W_t - W^\star, \frac{G_{t+1}}{\|G_{t+1}\|_{\text{F}}}\right\rangle \middle| \mathscr{F}_t\right] + \frac{\eta_t^2}{2}. \tag{38}$$

To proceed, we approximate by assuming that for large batch sizes B, the gradient G_{t+1} behaves like a Gaussian random variable due to the central limit theorem, and we use properties of Gaussian distributions to handle the normalization. Assume G_{t+1} is approximately Gaussian with mean $\mathbb{E}[G_{t+1}|\mathscr{F}_t] = B^{-1}\sum_{i=1}^B \mathbb{E}[x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \langle x_{\text{out}}^{(i)}, (W_t - W^\star)x_{\text{in}}^{(i)}\rangle|\mathscr{F}_t] = W_t - W^\star$ (from the unnormalized derivation), and covariance determined by σ^2. For a Gaussian vector $Z \sim N(\mu, \Sigma)$ in a Hilbert space, the expectation $\mathbb{E}\left[\frac{\langle a, Z\rangle}{\|Z\|}\right]$ can be approximated as

$$\mathbb{E}\left[\frac{\langle a, Z\rangle}{\|Z\|}\right] \sim \frac{\langle a, \mu\rangle}{\sqrt{\mathbb{E}[\|Z\|^2]}}. \tag{39}$$

Here, $\mu = W_t - W^\star$, and $\langle W_t - W^\star, \mu\rangle = \|W_t - W^\star\|_{\text{F}}^2$. We would need to evaluate

$$\mathbb{E}\left[\frac{\langle W_t - W^\star, G_{t+1}\rangle}{\|G_{t+1}\|_{\text{F}}} \middle| \mathscr{F}_t\right], \tag{40}$$

where W_t is $\mathscr{F}_t$-measurable, and $\mathscr{F}_t = \sigma(W_s, ((x_{\rm in})_s^{(i)}, (x_{\rm out})_s^{(i)})_{i=1}^B : s \leq t)$. From the unnormalized derivation, we know that $\mathbb{E}[G_{t+1}|\mathscr{F}_t] = W_t - W^\star$ and that

$$\mathbb{E}[\langle G_{t+1}, G_{t+1}\rangle|\mathscr{F}_t] = \tfrac{1}{B}(B + 3 + N_{\rm in}N_{\rm out} + 2(N_{\rm in} + N_{\rm out}))\|W_t - W^\star\|_{\rm F}^2. \tag{41}$$

Define $\kappa = B + 3 + N_{\rm in}N_{\rm out} + 2(N_{\rm in} + N_{\rm out})$ and denote $\Delta_t = W_t - W^\star$ and $\sigma^2 = \mathbb{E}[\|G_{t+1}\|_{\rm F}^2|\mathscr{F}_t] = \frac{\kappa}{B}\|\Delta_t\|_{\rm F}^2$. Since $G_{t+1} = B^{-1}\sum_{i=1}^B Z_i$, where $Z_i = x_{\rm out}^{(i)} \otimes x_{\rm in}^{(i)}\langle x_{\rm out}^{(i)}, (W_t - W^\star)x_{\rm in}^{(i)}\rangle$, and the Z_i are i.i.d. given $\mathscr{F}_t$, we apply the central limit theorem for large B. Each $Z_i \in \mathbb{R}^{N_{\rm out}\times N_{\rm in}} \cong \mathbb{R}^{N_{\rm out}N_{\rm in}}$, and we treat G_{t+1} as approximately Gaussian in the tensor space with mean $\mu = \mathbb{E}[G_{t+1}|\mathscr{F}_t] = \Delta_t$ and variance $\mathrm{Var}(G_{t+1}|\mathscr{F}_t) = B^{-1}\,\mathrm{Var}(Z_i|\mathscr{F}_t)$. For large B, the central limit theorem suggests

$$\sqrt{B}(G_{t+1} - \mu) \xrightarrow{\mathscr{D}} N(0, \Sigma), \tag{42}$$

where $\Sigma = \mathrm{Cov}(Z_i|\mathscr{F}_t)$ is the covariance tensor of Z_i. The variance of the Frobenius norm is

$$\sigma^2 = \mathbb{E}[\|G_{t+1}\|_{\rm F}^2|\mathscr{F}_t] - \|\mathbb{E}[G_{t+1}|\mathscr{F}_t]\|_{\rm F}^2 = \frac{\kappa}{B}\|\Delta_t\|_{\rm F}^2 - \|\Delta_t\|_{\rm F}^2 = \frac{\kappa - B}{B}\|\Delta_t\|_{\rm F}^2. \tag{43}$$

As $B \to \infty$, assuming $N_{\rm in}, N_{\rm out}$ are fixed or grow slower than B, we have that

$$\lim_{B\to\infty} \kappa/B = 1 + B^{-1}\left(N_{\rm in}N_{\rm out} + 2(N_{\rm in} + N_{\rm out}) + 3\right) = 1 + O\left(B^{-1}N_{\rm in}N_{\rm out}\right). \tag{44}$$

Thus $\sigma^2 = \|\Delta_t\|_{\rm F}^2 \cdot O\left(B^{-1}N_{\rm in}N_{\rm out}\right)$. To approximate

$$\mathbb{E}\left[\frac{\langle \Delta_t, G_{t+1}\rangle}{\|G_{t+1}\|_{\rm F}}\middle|\mathscr{F}_t\right], \tag{45}$$

we first let $X = \langle \Delta_t, G_{t+1}\rangle$ and $Y = \|G_{t+1}\|_{\rm F}$. Then,

$$X = \left\langle \Delta_t, \frac{1}{B}\sum_{i=1}^B Z_i\right\rangle, \quad Y = \sqrt{\left\langle \frac{1}{B}\sum_{i=1}^B Z_i, \frac{1}{B}\sum_{j=1}^B Z_j\right\rangle}. \tag{46}$$

From the unnormalized case, $\mathbb{E}[X|\mathscr{F}_t] = \|\Delta_t\|_{\rm F}^2$ and $\mathbb{E}[Y^2|\mathscr{F}_t] = \frac{\kappa}{B}\|\Delta_t\|_{\rm F}^2$. Thus, we can approximate $G_{t+1} \xrightarrow{\mathscr{D}} N(\Delta_t, \frac{\Sigma}{B})$. Consider furthermore the projection

$$X = \langle \Delta_t, G_{t+1}\rangle \xrightarrow{\mathscr{D}} N\left(\|\Delta_t\|_{\rm F}^2, B^{-1}\langle \Delta_t, \Sigma\Delta_t\rangle\right). \tag{47}$$

The variance of X is

$$\mathrm{Var}(X|\mathscr{F}_t) = B^{-1}\mathbb{E}[\langle Z_i, \Delta_t\rangle^2|\mathscr{F}_t] - B^{-1}\mathbb{E}[\langle Z_i, \Delta_t\rangle|\mathscr{F}_t]^2. \tag{48}$$

Since $\mathbb{E}[\langle Z_i, \Delta_t\rangle|\mathscr{F}_t] = \|\Delta_t\|_{\rm F}^2$, and $\mathbb{E}[\langle Z_i, \Delta_t\rangle^2|\mathscr{F}_t] = \mathbb{E}[\langle x_{\rm out}^{(i)} \otimes x_{\rm in}^{(i)}, \Delta_t\rangle^2\langle x_{\rm out}^{(i)}, (\Delta_t)x_{\rm in}^{(i)}\rangle^2|\mathscr{F}_t]$, we can reuse the unnormalized result for the case $i = j$,

$$\mathbb{E}[\langle Z_i, \Delta_t\rangle^2|\mathscr{F}_t] = (N_{\rm in}N_{\rm out} + 2(N_{\rm in} + N_{\rm out}) + 4)\|\Delta_t\|_{\rm F}^2. \tag{49}$$

Thus

$$\mathrm{Var}(X|\mathscr{F}_t) = B^{-1}[(N_{\rm in}N_{\rm out} + 2(N_{\rm in} + N_{\rm out}) + 4)\|\Delta_t\|_{\rm F}^2 - \|\Delta_t\|_{\rm F}^4]. \tag{50}$$

For large B, the norm $\|G_{t+1}\|_{\mathrm{F}} = \sqrt{\mathbb{E}[Y^2|\mathscr{F}_t]} = \sqrt{\frac{\kappa}{B}}\|\Delta_t\|_{\mathrm{F}}$. We apply the approximation

$$\mathbb{E}\left[\tfrac{X}{Y}|\mathscr{F}_t\right] = \frac{\mathbb{E}[X|\mathscr{F}_t]}{\sqrt{\mathbb{E}[Y^2|\mathscr{F}_t]}} + O\left(\frac{\mathrm{Var}(Y)}{\mathbb{E}[Y|\mathscr{F}_t]^3}\right). \tag{51}$$

This is justified by the central limit theorem and the fact that Y concentrates around its mean for large B. And since $\mathbb{E}[Y^2] = \frac{\kappa}{B}\|\Delta_t\|_{\mathrm{F}}^2$, for an approximately Gaussian G_{t+1}, the variance of $Y^2 = \|G_{t+1}\|_{\mathrm{F}}^2$ is

$$\begin{aligned}\mathrm{Var}(Y^2|\mathscr{F}_t) &= \mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^4|\mathscr{F}_t] - (\mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^2|\mathscr{F}_t])^2\\ &= \mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^4|\mathscr{F}_t] - \left(\tfrac{\kappa}{B}\|\Delta_t\|_{\mathrm{F}}^2\right)^2.\end{aligned} \tag{52}$$

We expand the first term,

$$\|G_{t+1}\|_{\mathrm{F}}^4 = \left(\sum_{i,j}(G_{t+1})_{ij}^2\right)^2 = \sum_{i,j,k,\ell}(G_{t+1})_{ij}(G_{t+1})_{k\ell}(G_{t+1})_{ij}(G_{t+1})_{k\ell}. \tag{53}$$

Since $G_{t+1} = B^{-1}\sum_{i=1}^{B} Z_i$, we have

$$(G_{t+1})_{ij} = \frac{1}{B}\sum_{m=1}^{B}(Z_m)_{ij}, \quad \|G_{t+1}\|_{\mathrm{F}}^4 = \frac{1}{B^4}\sum_{i,j,k,\ell}\sum_{m,n,p,q=1}^{B}(Z_m)_{ij}(Z_n)_{k\ell}(Z_p)_{ij}(Z_q)_{k\ell}. \tag{54}$$

Taking the expectation,

$$\mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^4|\mathscr{F}_t] = \frac{1}{B^4}\sum_{i,j,k,\ell}\sum_{m,n,p,q=1}^{B}\mathbb{E}[(Z_m)_{ij}(Z_n)_{k\ell}(Z_p)_{ij}(Z_q)_{k\ell}|\mathscr{F}_t]. \tag{55}$$

Since the Z_i are i.i.d., the expectation is non-zero only when the indices pair appropriately. We use the Wick theorem (as in the unnormalized case) for the Gaussian approximation,

$$\mathbb{E}[(Z_m)_{ij}(Z_n)_{kl}(Z_p)_{ij}(Z_q)_{kl}|\mathscr{F}_t] = \sum_{\pi\in\mathscr{P}_2(4)}\prod_{\{r,s\}\in\pi}\mathbb{E}[(Z_r)_{ab}(Z_s)_{cd}|\mathscr{F}_t], \tag{56}$$

where indices (a,b), (c,d) correspond to the paired terms. The pairings are:

- $\{(m,n),(p,q)\}$: $\mathbb{E}[(Z_m)_{ij}(Z_n)_{kl}]\mathbb{E}[(Z_p)_{ij}(Z_q)_{kl}]$.
- $\{(m,p),(n,q)\}$: $\mathbb{E}[(Z_m)_{ij}(Z_p)_{ij}]\mathbb{E}[(Z_n)_{kl}(Z_q)_{kl}]$.
- $\{(m,q),(n,p)\}$: $\mathbb{E}[(Z_m)_{ij}(Z_q)_{kl}]\mathbb{E}[(Z_n)_{kl}(Z_p)_{ij}]$.

Summing over indices and considering contributions,

$$\mathbb{E}[(Z_i)_{ij}(Z_i)_{kl}|\mathscr{F}_t] = (\delta_{ik}\delta_{jl})(N_{\mathrm{in}}N_{\mathrm{out}} + 2(N_{\mathrm{in}} + N_{\mathrm{out}}) + 4)\|\Delta_t\|_{\mathrm{F}}^2, \tag{57}$$

For large B, the dominant terms come from pairings where indices align. The second moments are

$$\mathbb{E}[(Z_i)_{ab}(Z_j)_{cd}|\mathscr{F}_t] = \begin{cases}\mathbb{E}[(Z_i)_{ab}(Z_i)_{cd}|\mathscr{F}_t] & \text{if } i = j,\\ (\Delta_t)_{ab}(\Delta_t)_{cd} & \text{if } i \neq j,\end{cases} \tag{58}$$

where

$$(Z_i)_{ab} = (x_{\text{out}}^{(i)})_a (x_{\text{in}}^{(i)})_b \langle x_{\text{out}}^{(i)}, \Delta_t x_{\text{in}}^{(i)} \rangle. \tag{59}$$

From the unnormalized derivation (case $i = j$):

$$\mathbb{E}[\langle Z_i, \Delta_t \rangle^2 | \mathscr{F}_t] = (N_{\text{in}} N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}}) + 4) \|\Delta_t\|_{\text{F}}^2. \tag{60}$$

For the pairing $\{(i, j), (k, l)\}$, we compute:

$$\mathbb{E}[(Z_i)_{ab}(Z_j)_{ab}(Z_k)_{cd}(Z_l)_{cd} | \mathscr{F}_t] = \begin{cases} \mathbb{E}[(Z_i)_{ab}^2 (Z_i)_{cd}^2 | \mathscr{F}_t] & \text{if } i = j = k = l, \\ \mathbb{E}[(Z_i)_{ab}^2 | \mathscr{F}_t] \mathbb{E}[(Z_k)_{cd}^2 | \mathscr{F}_t] & \text{if } i = j, k = l, i \neq k, \\ (\Delta_t)_{ab}^2 (\Delta_t)_{cd}^2 & \text{if } i = k, j = l, i \neq j, \\ (\Delta_t)_{ab}^2 (\Delta_t)_{cd}^2 & \text{if } i = l, j = k, i \neq j, \\ (\Delta_t)_{ab} (\Delta_t)_{ab} (\Delta_t)_{cd} (\Delta_t)_{cd} & \text{if all distinct.} \end{cases} \tag{61}$$

Summing over indices, we can case into the following scenarios.

(i) $i = j = k = l$ (contributes B terms):

$$\sum_{a,b,c,d} \mathbb{E}[(Z_i)_{ab}^2 (Z_i)_{cd}^2 | \mathscr{F}_t]. \tag{62}$$

Using the Gaussian moment for Z_i:

$$\mathbb{E}[(Z_i)_{ab}^2 (Z_i)_{cd}^2 | \mathscr{F}_t] = \mathbb{E}[(Z_i)_{ab}^2 | \mathscr{F}_t] \mathbb{E}[(Z_i)_{cd}^2 | \mathscr{F}_t] + 2\mathbb{E}[(Z_i)_{ab} (Z_i)_{cd} | \mathscr{F}_t]^2. \tag{63}$$

Since

$$\mathbb{E}[(Z_i)_{ab} (Z_i)_{cd} | \mathscr{F}_t] = (\delta_{ac} \delta_{bd}) (N_{\text{in}} N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}}) + 4) \|\Delta_t\|_{\text{F}}^2, \tag{64}$$

summing over all indices gives a factor proportional to $N_{\text{out}} N_{\text{in}}$,

$$\sum_{a,b,c,d} \mathbb{E}[(Z_i)_{ab} (Z_i)_{cd} | \mathscr{F}_t]^2 = \sum_{a,b} (N_{\text{in}} N_{\text{out}} + 2(N_{\text{in}} + N_{\text{out}}) + 4)^2 \|\Delta_t\|_{\text{F}}^4 \sim (N_{\text{in}} N_{\text{out}})^2 \|\Delta_t\|_{\text{F}}^4. \tag{65}$$

(ii) $i = j, k = l, i \neq k$ (contributes $B(B - 1)$):

$$\sum_{a,b,c,d} \mathbb{E}[(Z_i)_{ab}^2 | \mathscr{F}_t] \mathbb{E}[(Z_k)_{cd}^2 | \mathscr{F}_t] \sim (N_{\text{in}} N_{\text{out}})^2 \|\Delta_t\|_{\text{F}}^4. \tag{66}$$

(iii) $i = k, j = l$ or $i = l, j = k$ (contributes $2B(B - 1)$):

$$\sum_{a,b,c,d} (\Delta_t)_{ab}^2 (\Delta_t)_{cd}^2 = \|\Delta_t\|_{\text{F}}^4. \tag{67}$$

(iv) i, j, k, ℓ all distinct (contributes $B(B - 1)(B - 2)(B - 3)$):

$$\sum_{a,b,c,d} (\Delta_t)_{ab}^2 (\Delta_t)_{cd}^2 = \|\Delta_t\|_{\text{F}}^4. \tag{68}$$

Combining all four cases, after summing, the fourth moment scales as the square of the second moment $\mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^4|\mathscr{F}_t] \sim 3\left(\frac{\kappa}{B}\|\Delta_t\|_{\mathrm{F}}^2\right)^2$. Thus, $\mathrm{Var}(Y^2|\mathscr{F}_t) \sim 2\left(\frac{\kappa}{B}\|\Delta_t\|_{\mathrm{F}}^2\right)^2$. Then it immediately follows that $\mathrm{Var}(Y) \sim \frac{\mathrm{Var}(Y^2)}{2\mathbb{E}[Y^2]} \sim \frac{\kappa}{B}\|\Delta_t\|_{\mathrm{F}}^2$. Since $\mathbb{E}[Y] \sim \sqrt{\frac{\kappa}{B}}\|\Delta_t\|_{\mathrm{F}}$, the error term is

$$\frac{\mathrm{Var}(Y)}{\mathbb{E}[Y]^3} = O\left(\frac{\sigma^2/(N_{\mathrm{out}}N_{\mathrm{in}})}{(\sigma\sqrt{\kappa/B})^3}\right) = O\left(\frac{B^{3/2}}{\sigma\kappa^{3/2}N_{\mathrm{out}}N_{\mathrm{in}}}\right). \tag{69}$$

For large B, $\kappa \sim B$. Hence,

$$\mathbb{E}\left[\frac{\langle\Delta_t, G_{t+1}\rangle}{\|G_{t+1}\|_{\mathrm{F}}}\middle|\mathscr{F}_t\right] = \frac{\|\Delta_t\|_{\mathrm{F}}^2}{\sqrt{\frac{\kappa}{B}}\|\Delta_t\|_{\mathrm{F}}} + O\left(\frac{\|\Delta_t\|_{\mathrm{F}}}{\sqrt{B\kappa}}\right) = \sqrt{\frac{B}{\kappa}}\|\Delta_t\|_{\mathrm{F}} + O\left(\frac{\|\Delta_t\|_{\mathrm{F}}}{\sqrt{B\kappa}}\right). \tag{70}$$

Now, the expected risk is

$$\begin{aligned}\mathbb{E}(\mathscr{R}(W_{t+1})|\mathscr{F}_t) &= \mathscr{R}(W_t) - \eta_t\mathbb{E}\left[\frac{\langle\Delta_t, G_{t+1}\rangle}{\|G_{t+1}\|_{\mathrm{F}}}\middle|\mathscr{F}_t\right] + \frac{\eta_t^2}{2} \\ &= \mathscr{R}(W_t) - \eta_t\left(\sqrt{\frac{B}{\kappa}}\|\Delta_t\|_{\mathrm{F}} + O\left(\frac{\|\Delta_t\|_{\mathrm{F}}}{\sqrt{B\kappa}}\right)\right) + \frac{\eta_t^2}{2} \\ &= \mathscr{R}(W_t) - \eta_t\sqrt{\frac{B}{\kappa}}\sqrt{2\mathscr{R}(W_t)} + O\left(\eta_t\frac{\sqrt{\mathscr{R}(W_t)}}{\sqrt{B\kappa}}\right) + \frac{\eta_t^2}{2}\end{aligned} \tag{71}$$

where the last equality follows from the fact that $\mathscr{R}(W_t) = \frac{1}{2}\|\Delta_t\|_{\mathrm{F}}^2$ and $\|\Delta_t\|_{\mathrm{F}} = \sqrt{2\mathscr{R}(W_t)}$. As $B, N_{\mathrm{out}}, N_{\mathrm{in}} \to \infty$, the error term is $o(\eta_t\sqrt{\mathscr{R}(W_t)})$, assuming $N_{\mathrm{out}}N_{\mathrm{in}} \gg B^{-1/2}$. Thus, the leading-order approximation is

$$\mathbb{E}(\mathscr{R}(W_{t+1})|\mathscr{F}_t) = \mathscr{R}(W_t) - \eta_t\sqrt{\frac{2B}{\kappa}}\sqrt{\mathscr{R}(W_t)} + \frac{\eta_t^2}{2} + o(\eta_t\sqrt{\mathscr{R}(W_t)}). \tag{72}$$

This shows the nonlinear dependence on the risk due to normalization. ■

C.3. Asymptotic behavior of the SGD risk

Next, we show that $B \propto N_{\text{in}}N_{\text{out}}$ is the appropriate scaling limit with universal behavior for training at a reasonable speed in SGD with a normalized gradient, and compute the final loss as $t \longrightarrow \infty$. In other words, we want to ensure that the training dynamics achieve a balance between convergence speed and stability, and to determine the asymptotic risk in terms of B, N_{in}, N_{out}, and η_t. For large B, we approximate $\kappa \sim B/(B + N_{\text{in}}N_{\text{out}})$. The leading-order recursion is

$$\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] \sim \mathscr{R}(W_t) - \eta_t\sqrt{\frac{2B}{B + N_{\text{in}}N_{\text{out}}}}\sqrt{\mathscr{R}(W_t)} + \frac{\eta_t^2}{2}. \tag{73}$$

To analyze training speed, we need the risk to decrease at a reasonable rate, meaning the descent term dominates the noise term and leads to convergence in a practical number of iterations. The nonlinear term $\sqrt{\mathscr{R}(W_t)}$ suggests a different convergence behavior compared to unnormalized SGD, where the risk decreases linearly. The coefficient of the nonlinear term should be neither too small (which slows convergence) nor too large (which could destabilize the update due to large steps). Let $B = \alpha N_{\text{in}}N_{\text{out}}$, where α is a constant, and analyze the coefficient

$$\sqrt{\frac{2B}{B + N_{\text{in}}N_{\text{out}}}} = \sqrt{\frac{2\alpha N_{\text{in}}N_{\text{out}}}{\alpha N_{\text{in}}N_{\text{out}} + N_{\text{in}}N_{\text{out}}}} = \sqrt{\frac{2\alpha}{\alpha + 1}} = O(1). \tag{74}$$

for the recursion to yield effective descent and ensure that the step size $\eta_t\sqrt{\frac{2\alpha}{\alpha+1}}\sqrt{\mathscr{R}(W_t)}$ is significant relative to $\mathscr{R}(W_t)$. We examine different scaling regimes:

- ***B* ≪ $\boldsymbol{N_{\text{in}}N_{\text{out}}}$:** The descent term becomes $-\eta_t\sqrt{2B/(N_{\text{in}}N_{\text{out}})\mathscr{R}(W_t)}$. If B is small in the sense that $B = o(N_{\text{in}}N_{\text{out}})$, this makes the descent term negligible unless $\eta_t \gg 1$, which leads to slow convergence and risks instability since the noise term $\eta_t^2/2$ grows quadratically.

- ***B* ≫ $\boldsymbol{N_{\text{in}}N_{\text{out}}}$:** The recursion becomes $\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] \sim \mathscr{R}(W_t) - \eta_t\sqrt{2\mathscr{R}(W_t)} + \frac{\eta_t^2}{2}$. While this maximizes the descent term, a very large B is computationally expensive, as it requires processing many samples per iteration, which may not be practical for large-scale problems. Even though the error term $O\left(\eta_t\sqrt{\mathscr{R}(W_t)}/\sqrt{B\kappa}\right)$ is small, the computational cost outweighs the marginal gain in descent rate.

- ***B* ∝ $\boldsymbol{N_{\text{in}}N_{\text{out}}}$:** Assume that $B = \alpha N_{\text{in}}N_{\text{out}}$, the descent term coefficient $\sqrt{\frac{2\alpha}{\alpha+1}} \in (0, \sqrt{2})$. This scaling ensures the descent term is $O(\eta_t\sqrt{\mathscr{R}(W_t)})$, providing significant progress per iteration without excessive computational cost. We can approximate $\kappa \sim B + N_{\text{in}}N_{\text{out}} = (\alpha + 1)N_{\text{in}}N_{\text{out}}$, $\sqrt{B\kappa} \sim \sqrt{\alpha(\alpha + 1)}N_{\text{in}}N_{\text{out}}$. The error term scales as $O\left(\eta_t\sqrt{\mathscr{R}(W_t)}/\sqrt{B\kappa}\right) \sim O\left(\eta_t\sqrt{\mathscr{R}(W_t)}/(\sqrt{\alpha(\alpha + 1)}N_{\text{in}}N_{\text{out}})\right)$, which is $o(\sqrt{\mathscr{R}(W_t)})$ for large $N_{\text{in}}N_{\text{out}}$. Thus, the approximation is tight, and $B \propto N_{\text{in}}N_{\text{out}}$ balances computational efficiency and convergence speed as batch size co-scales with problem size.

To find the final loss as $t \longrightarrow \infty$ in the $B \propto N_{\text{in}}N_{\text{out}}$ regime, we assume a constant learning rate $\eta_t = \eta = O(1/L)$ for simplicity. Let $R_t = \mathscr{R}(W_t)$. We take the expectation over all iterations while assuming the risk converges to a steady state $R_\infty = \lim_{t\to\infty} \mathbb{E}[R_t|\mathscr{F}_t]$, i.e., $\mathbb{E}[R_{t+1}] = \mathbb{E}[R_t] = R_\infty$. Then, the SGD risk recursion in the iteration limit is

$$R_\infty \sim R_\infty - \eta\sqrt{\frac{2B}{B + N_{\text{in}}N_{\text{out}}}}\mathbb{E}\left[\sqrt{R_\infty}\right] + \frac{\eta^2}{2}. \tag{75}$$

Since $\sqrt{R_t}$ concentrates around its mean for large B, we approximate $\mathbb{E}[\sqrt{R_\infty}] \sim \sqrt{R_\infty}$ to get

$$0 \sim -\eta\sqrt{\frac{2B}{B + N_{\text{in}}N_{\text{out}}}}\sqrt{R_\infty} + \frac{\eta^2}{2}. \tag{76}$$

Solving with $B = \alpha N_{\text{in}}N_{\text{out}}$ gives

$$R_\infty = \left(\frac{\eta}{2\sqrt{\frac{2B}{B+N_{\text{in}}N_{\text{out}}}}}\right)^2 = \frac{\eta^2(B + N_{\text{in}}N_{\text{out}})}{8B} = \frac{\eta^2(\alpha + 1)}{8\alpha}. \tag{77}$$

Appendix D. Muon risk updates

D.1. Proof of Theorem 4

Proof Define $\Delta_t := W_t - W^\star$. Recall the one-step Muon update

$$\begin{aligned} G_{t+1} &:= aG_t + b(G_tG_t^\top)G_t + c(G_tG_t^\top)^2G_t \\ &= \left(a\,\mathrm{Id} + b(G_tG_t^\top + c(G_tG_t^\top)^2\right)G_t \\ &= U\Phi_5(\Sigma; a, b, c)V^\top \end{aligned} \tag{78}$$

where $\Phi_5(\Sigma; a, b, c) := a\Sigma + b\Sigma^3 + c\Sigma^5$ is some quintic polynomial with fixed hyperparameters a, b, c such that $\lim_{N\to\infty} \Phi_5^N = 1$. The expected risk in the isotropic case is

$$\begin{aligned} \mathbb{E}(\mathscr{R}(W_{t+1})|\mathscr{F}_t) &= \tfrac{1}{2}\mathbb{E}(\|W_{t+1} - W^\star\|_{\mathrm{F}}^2|\mathscr{F}_t) \\ &= \tfrac{1}{2}\mathbb{E}\left(\|W_t - W^\star - \eta G_{t+1}\|_{\mathrm{F}}^2|\mathscr{F}_t\right) \\ &= \tfrac{1}{2}\left(\|\Delta_t\|_{\mathrm{F}}^2 - 2\eta\mathbb{E}(\langle\Delta_t, G_{t+1}\rangle|\mathscr{F}_t) + \eta^2\mathbb{E}(\langle G_{t+1}, G_{t+1}\rangle|\mathscr{F}_t)\right). \end{aligned} \tag{79}$$

We further expand the expected risk in terms of $(G_tG_t^\top)^qG_t$ and Δ_t. Note that

$$\begin{aligned} \mathbb{E}(\langle\Delta_t, G_{t+1}\rangle|\mathscr{F}_t) = a\mathbb{E}\left(\langle\Delta_t, G\rangle\,|\mathscr{F}_t\right) \\ + b\mathbb{E}\left[\left\langle\Delta_t, (G_tG_t^\top)G\right\rangle\middle|\mathscr{F}_t\right] + c\mathbb{E}\left[\left\langle\Delta_t, (G_tG_t^\top)^2G_t\right\rangle\middle|\mathscr{F}_t\right] \end{aligned} \tag{80}$$

and

$$\begin{aligned} &\mathbb{E}\langle\Delta_t, (G_tG_t^\top)^qG_t\rangle = \mathbb{E}\,\mathrm{Tr}\left[\Delta_t^\top(G_tG_t^\top)^qG\right] \\ &= \frac{1}{B^{2q+1}}\mathbb{E}\sum_{i_1,j_1,\ldots,i_q,j_q,k\in[B]}\mathrm{Tr}\left[\Delta_t^\top\Big(\prod_{m\in[q]}G_{i_m}G_{j_m}^\top\Big)G_k\right] \\ &= \frac{1}{B^{2q+1}}\mathbb{E}\sum_{i_1,j_1,\ldots,i_q,j_q,k\in[B]}\mathrm{Tr}\left[\Delta_t^\top\prod_{m\in[q]}\left(x_{\mathrm{in}}^{i_m}\otimes x_{\mathrm{out}}^{i_m}\right)\left(x_{\mathrm{out}}^{j_m}\otimes x_{\mathrm{in}}^{j_m}\right)\left(x_{\mathrm{in}}^{k}\otimes x_{\mathrm{out}}^{k}\right)\times\right. \\ &\qquad\left.\times\prod_{n\in[q]}\left\langle x_{\mathrm{in}}^{i_n}, \Delta_t x_{\mathrm{out}}^{i_n}\right\rangle\left\langle x_{\mathrm{out}}^{j_n}, \Delta_t x_{\mathrm{in}}^{j_n}\right\rangle\left\langle x_{\mathrm{in}}^{k}, \Delta_t x_{\mathrm{out}}^{k}\right\rangle\right] \\ &= \frac{1}{B^{2q+1}}\mathbb{E}\sum_{\substack{i_1,i_2,\ldots,i_q,\\ j_1,j_2,\ldots,j_q\geq 1}}\sum_{k\geq 1}\prod_{\substack{\ell=1\\ i_{q+1}=k}}^{q}\left\langle x_{\mathrm{in}}^{i_\ell}, \Delta_t^\top x_{\mathrm{out}}^{i_\ell}\right\rangle\left\langle x_{\mathrm{in}}^{j_\ell}, \Delta_t^\top x_{\mathrm{out}}^{j_\ell}\right\rangle\times \\ &\qquad\times\left\langle x_{\mathrm{out}}^{i_\ell}, x_{\mathrm{out}}^{j_\ell}\right\rangle\left\langle x_{\mathrm{in}}^{j_\ell}, x_{\mathrm{in}}^{i_{\ell+1}}\right\rangle\left\langle x_{\mathrm{in}}^{k}, \Delta_t^\top x_{\mathrm{out}}^{k}\right\rangle\left\langle x_{\mathrm{in}}^{i_1}, \Delta_t^\top x_{\mathrm{out}}^{k}\right\rangle \\ &= \frac{1}{B^{2q+1}}\sum_{\substack{i_1,\ldots,i_q,\\ j_1,\ldots,j_q\geq 1}}\sum_{k\geq 1}\sum_{\substack{a_1,c_1,f_1,\ldots,\\ a_q,c_q,f_q\in[N_{\mathrm{in}}]\\ a_{q+1},a_{q+2}=1}}\sum_{\substack{b_1,d_1,e_1\ldots,\\ b_q,d_q,e_q\in[N_{\mathrm{out}}]\\ b_{q+1},b_{q+2}=1}}\Big(\prod_{\ell=1}^{q}[\Delta_t]_{b_\ell a_\ell}[\Delta_t]_{d_\ell c_\ell}\Big)(\Delta_t)_{b_{q+1}a_{q+1}}(\Delta_t)_{b_{q+2}a_{q+2}}\times \\ &\qquad\times\mathbb{E}\Big[(x_{\mathrm{in}})_{a_{q+1}}^{k}(x_{\mathrm{out}})_{b_{q+1}}^{k}(x_{\mathrm{in}})_{a_{q+2}}^{i_1}(x_{\mathrm{out}})_{b_{q+2}}^{k}\times \\ &\qquad\quad\times\prod_{\ell\in[q]}(x_{\mathrm{in}})_{a_\ell}^{i_\ell}(x_{\mathrm{out}})_{b_\ell}^{i_\ell}(x_{\mathrm{in}})_{c_\ell}^{j_\ell}(x_{\mathrm{out}})_{d_\ell}^{j_\ell}(x_{\mathrm{out}})_{e_\ell}^{i_\ell}(x_{\mathrm{out}})_{e_\ell}^{j_\ell}(x_{\mathrm{in}})_{f_\ell}^{j_\ell}(x_{\mathrm{in}})_{f_\ell}^{i_{\ell+1}}\Big], \end{aligned} \tag{81}$$

where we employed tensor simplification rules such as

$$\begin{aligned}\frac{1}{B}\sum_{i=1}^{B} e^i \otimes x_{\mathrm{in}}^i &= \left(\frac{1}{B}\right)^{2k} \sum_{i_1,j_1,\ldots,i_k,j_k\geq 1} \langle x_{\mathrm{in}}^{i_1}, x_{\mathrm{in}}^{j_1}\rangle\langle x_{\mathrm{in}}^{i_2}, x_{\mathrm{in}}^{j_2}\rangle \cdots \langle x_{\mathrm{in}}^{i_k}, x_{\mathrm{in}}^{j_k}\rangle \delta_{j_1 i_2}\delta_{j_2 i_3}\cdots\delta_{j_k i_1} \\ &= \left(\frac{1}{B}\right)^{2k} \sum_{i_1,\ldots,i_k\geq 1} \langle x_{\mathrm{in}}^{i_1}, x_{\mathrm{in}}^{i_2}\rangle\langle x_{\mathrm{in}}^{i_2}, x_{\mathrm{in}}^{i_3}\rangle \cdots \langle x_{\mathrm{in}}^{i_k}, x_{\mathrm{in}}^{i_1}\rangle,\end{aligned} \tag{82}$$

given either of x_{in} or x_{out} is orthonormal, and for contractions with cyclic indices, $\mathbb{E}\langle x_{\mathrm{in}}^1, x_{\mathrm{in}}^2\rangle\,\langle x_{\mathrm{in}}^2, x_{\mathrm{in}}^1\rangle = N_{\mathrm{in}}$ contributes one factor of N_{in}, and the same holds for more contractions with cyclic indices like $\mathbb{E}\langle x_{\mathrm{in}}^1, x_{\mathrm{in}}^2\rangle\langle x_{\mathrm{in}}^2, x_{\mathrm{in}}^3\rangle\langle x_{\mathrm{in}}^3, x_{\mathrm{in}}^1\rangle = N_{\mathrm{in}}$. On the other hand, contractions with paired indices satisfy $\mathbb{E}\langle x_{\mathrm{in}}^1, x_{\mathrm{in}}^1\rangle\langle x_{\mathrm{in}}^2, x_{\mathrm{in}}^2\rangle = N_{\mathrm{in}}^2$. To approximate the higher-order mixed gradient moments, we introduce the notion of *freeness* as a non-commutative analogue of the classical notion of independence in probability theory—*free independence*.

Definition 8 ($C^\star$-probability space and non-commutative random variables [12]) *In general we refer to a pair $(\mathscr{A}, \varphi)$, consisting of a unital algebra $\mathscr{A}$ and a unital linear functional $\varphi : \mathscr{A} \to \mathbb{C}$ with $\varphi(1) = 1$, as a non-commutative probability space. If $\mathscr{A}$ is a $\star$-algebra and φ is a state, i.e., in addition to $\varphi(1) = 1$ also positive (which means $\varphi(a^\star a) \geq 0$ for all $a \in \mathscr{A}$), then we call $(\mathscr{A}, \varphi)$ a $\star$-probability space. If $\mathscr{A}$ is a $C^\star$-algebra and φ a state, $(\mathscr{A}, \varphi)$ is a $C^\star$*-probability space. *Elements of $\mathscr{A}$ are called non-commutative random variables.*

Proposition 9 *In the large-dimensional limit where $B, N_{\mathrm{in}}, N_{\mathrm{out}} \to \infty$ with fixed ratios $B/N_{\mathrm{in}} = \phi$ and $B/N_{\mathrm{out}} = \psi$, the normalized trace $\mathbb{E}\left[B^{-1}\operatorname{Tr}(\Delta_t^\top (GG^\top)^q G)\right]$ converges in probability to the free probability moment*

$$\tau\left[\Delta_t^\top \left(GG^\top\right)^q G\right] = \lim_{B\to\infty} B^{-1}\mathbb{E}\operatorname{Tr}\left[\Delta_t^\top \left(GG^\top\right)^q G\right], \tag{83}$$

where τ is the normalized trace in a $C^\star$-probability space, and G is treated as a free random variable with a distribution determined by the isotropic covariance $\mathbb{E}x_{\mathrm{in}}^{\otimes 2} = \mathbb{E}x_{\mathrm{out}}^{\otimes 2} = \mathrm{Id}$. The expectation of the trace is approximated by the contribution of non-crossing pairings of the indices $i_1, \ldots, i_q, j_1, \ldots, j_q, k$ given by the free cumulant expansion of the moment, where

$$\begin{aligned} B \times \tau\left[\Delta_t^\top \left(GG^\top\right)^q G\right] &= \mathbb{E}\operatorname{Tr}\left[\Delta_t^\top \left(GG^\top\right)^q G\right] \\ &= \frac{1+o(1)}{B^{2q+1}}\mathbb{E} \sum_{\substack{i_1,i_2,\ldots,i_q,\\ j_1,j_2,\ldots,j_q=1}}^{B} \sum_{\substack{k=1\\ k\neq i,j}}^{B} \nu(i_q, j_q, i_1) \prod_{\ell=1}^{q-1} \mu(i_\ell, j_\ell, i_{\ell+1}), \end{aligned} \tag{84}$$

where

$$\begin{aligned} \mu(\alpha,\beta,\gamma) &:= \langle x_{\mathrm{in}}^\alpha, \Delta_t^\top x_{\mathrm{out}}^\alpha\rangle\langle x_{\mathrm{in}}^\beta, \Delta_t^\top x_{\mathrm{out}}^\beta\rangle\langle x_{\mathrm{out}}^\alpha, x_{\mathrm{out}}^\beta\rangle\langle x_{\mathrm{in}}^\beta, x_{\mathrm{in}}^\gamma\rangle \\ \nu(\alpha,\beta,\gamma) &:= \mu(\alpha,\beta,\gamma)\langle \Delta_t x_{\mathrm{in}}^\gamma, \Delta_t x_{\mathrm{in}}^\beta\rangle / \langle x_{\mathrm{in}}^\beta, x_{\mathrm{in}}^\gamma\rangle. \end{aligned} \tag{85}$$

This reflects the dominance of non-crossing partitions in the free limit, weighted by the ratios ϕ, ψ.

Proof In the $C^\star$-probability space $(\mathscr{A}, \varphi)$, let $\varphi(A) = \frac{1}{N_{\text{out}}}\operatorname{Tr}(A)$ and define $b_\ell = x_{\text{out}}^\ell \otimes x_{\text{in}}^\ell$, $b_\ell^\star = x_{\text{in}}^\ell \otimes x_{\text{out}}^\ell$, so the trace satisfies

$$\begin{aligned} &\varphi\left(\Delta_t^\top \left(GG^\top\right)^q G\right) \\ &= \varphi\left(\Delta_t^\top \left(\frac{1}{B^2}\sum_{\ell,m=1}^{B} b_\ell b_m^\star \varphi(b_\ell^\star \Delta_t)\varphi(b_m \Delta_t^\top)\right)^q \frac{1}{B}\sum_{n=1}^{B} b_n \varphi(b_n^\star \Delta_t)\right) \\ &= \frac{1}{B^{2q+1}} \sum_{i_1,j_1,\dots,i_q,j_q,k=1}^{B} \varphi\left(\Delta_t^\top \prod_{\ell=1}^{q}(b_{i_\ell} b_{j_\ell}^\star) b_k\right) \prod_{\ell=1}^{q} \varphi(b_{i_\ell}^\star \Delta_t)\varphi(b_{j_\ell}\Delta_t^\top)\varphi(b_k^\star \Delta_t). \end{aligned} \tag{86}$$

The free cumulant expansion for the moment is

$$\varphi\left(\Delta_t^\top b_{i_1} b_{j_1}^\star \cdots b_{i_q} b_{j_q}^\star b_k\right) = \sum_{\pi \in \mathscr{NC}(2q+2)} \kappa_\pi(\Delta_t^\top, b_{i_1}, b_{j_1}^\star, \dots, b_{i_q}, b_{j_q}^\star, b_k). \tag{87}$$

The dominant non-crossing partition pairs indices to maximize cycles. Consider the partition $\pi = \{(1, 2q+2), (2,3), (4,5), \dots, (2q, 2q+1)\}$. For $(1, 2q+2)$, we pair $\Delta_t^\top$ with b_k, which contributes $\kappa_2(\Delta_t^\top, b_k)$. For $(2\ell, 2\ell+1)$, we pair $b_{i_\ell}, b_{j_\ell}^\star$, contributing $\kappa_2(b_{i_\ell}, b_{j_\ell}^\star)$. Thus

$$\kappa_2(b_{i_\ell}, b_{j_\ell}^\star) = \varphi(b_{i_\ell} b_{j_\ell}^\star) - \varphi(b_{i_\ell})\varphi(b_{j_\ell}^\star) = \frac{1}{N_{\text{out}}}\mathbb{E}[\langle x_{\text{in}}^{i_\ell}, x_{\text{in}}^{j_\ell}\rangle\langle x_{\text{out}}^{i_\ell}, x_{\text{out}}^{j_\ell}\rangle] = \delta_{i_\ell j_\ell} N_{\text{in}}, \tag{88}$$

and for $\Delta_t^\top, b_k$,

$$\varphi(\Delta_t^\top b_k) = \frac{1}{N_{\text{out}}}\operatorname{Tr}(\Delta_t^\top(x_{\text{out}}^k \otimes x_{\text{in}}^k)) = \frac{1}{N_{\text{out}}}\langle x_{\text{in}}^k, \Delta_t^\top x_{\text{out}}^k\rangle, \quad \mathbb{E}[\varphi(\Delta_t^\top b_k)] = 0. \tag{89}$$

Thus $\kappa_2(\Delta_t^\top, b_k) \approx \varphi(\Delta_t^\top b_k)$, and the scalar terms are $\varphi(b_{i_\ell}^\star \Delta_t) = \frac{1}{N_{\text{out}}}\langle x_{\text{out}}^{i_\ell}, \Delta_t x_{\text{in}}^{i_\ell}\rangle$, $\varphi(b_{j_\ell}\Delta_t^\top) = \frac{1}{N_{\text{out}}}\langle x_{\text{in}}^{j_\ell}, \Delta_t^\top x_{\text{out}}^{j_\ell}\rangle$. The expectation becomes

$$\mathbb{E}\left[\prod_{\ell=1}^{q}\delta_{i_\ell j_\ell} N_{\text{in}} \cdot \langle x_{\text{in}}^k, \Delta_t^\top x_{\text{out}}^{i_1}\rangle \prod_{\ell=1}^{q}\langle x_{\text{out}}^{i_\ell}, \Delta_t x_{\text{in}}^{i_\ell}\rangle\langle x_{\text{in}}^{j_\ell}, \Delta_t^\top x_{\text{out}}^{j_\ell}\rangle\langle x_{\text{out}}^k, \Delta_t x_{\text{in}}^k\rangle\right]. \tag{90}$$

Summing over $i_\ell = j_\ell$, (84) becomes

$$\begin{aligned} &\varphi\left(\Delta_t^\top \left(GG^\top\right)^q G\right) \\ &= \frac{1}{B^{2q+1}} \sum_{i_1,\dots,i_q,k=1}^{B} N_{\text{in}}^q \mathbb{E}\left[\langle x_{\text{in}}^k, \Delta_t^\top x_{\text{out}}^{i_1}\rangle \prod_{\ell=1}^{q}\langle x_{\text{out}}^{i_\ell}, \Delta_t x_{\text{in}}^{i_\ell}\rangle^2 \langle x_{\text{out}}^k, \Delta_t x_{\text{in}}^k\rangle\right]. \end{aligned} \tag{91}$$

The expectation involves $2q+1$ terms of the form $\langle x_{\text{out}}^\ell, \Delta_t x_{\text{in}}^\ell\rangle$ and one $\langle x_{\text{in}}^k, \Delta_t^\top x_{\text{out}}^{i_1}\rangle$. Using Isserlis' theorem for Gaussian variables, we know $\mathbb{E}[\langle x_{\text{out}}^{i_\ell}, \Delta_t x_{\text{in}}^{i_\ell}\rangle^2] = \operatorname{Tr}(\Delta_t \mathrm{Id}_{N_{\text{in}}} \Delta_t^\top \mathrm{Id}_{N_{\text{out}}}) = \operatorname{Tr}(\Delta_t \Delta_t^\top)$. For the cross terms, $\mathbb{E}[\langle x_{\text{in}}^k, \Delta_t^\top x_{\text{out}}^{i_1}\rangle\langle x_{\text{out}}^k, \Delta_t x_{\text{in}}^k\rangle] = \mathbb{E}[\langle x_{\text{out}}^{i_1}, \Delta_t x_{\text{in}}^k\rangle\langle x_{\text{out}}^k, \Delta_t x_{\text{in}}^k\rangle] = \langle \Delta_t x_{\text{in}}^k, \Delta_t x_{\text{in}}^k\rangle = \langle x_{\text{in}}^k, \Delta_t^\top \Delta_t x_{\text{in}}^k\rangle$ and $\mathbb{E}[\langle x_{\text{in}}^k, \Delta_t^\top \Delta_t x_{\text{in}}^k\rangle] = \operatorname{Tr}(\Delta_t^\top \Delta_t \mathrm{Id}_{N_{\text{in}}}) = \operatorname{Tr}(\Delta_t^\top \Delta_t)$. Thus summing (84) gives

$$\begin{aligned} &\frac{N_{\text{in}}^q}{B^{2q+1}} \sum_{i_1,\dots,i_q,k=1}^{B} \delta_{i_1 k}[\operatorname{Tr}(\Delta_t \Delta_t^\top)]^q \operatorname{Tr}(\Delta_t^\top \Delta_t) \\ &= \frac{N_{\text{in}}^q}{B^{2q+1}} B^q [\operatorname{Tr}(\Delta_t \Delta_t^\top)]^{q+1} = \frac{1}{B^{q+1}} N_{\text{in}}^q [\operatorname{Tr}(\Delta_t \Delta_t^\top)]^{q+1}. \end{aligned} \tag{92}$$

Since $N_{\text{in}} \sim \phi B$, the normalized trace is scales as

$$\frac{1}{B}\mathbb{E}\operatorname{Tr}[\Delta_t^\top \left(GG^\top\right)^q G] = O\left(\frac{\phi^q}{B^{q+2}}[\operatorname{Tr}(\Delta_t\Delta_t^\top)]^{q+1}\right). \tag{93}$$

The subleading terms come from matched k indices. WLOG assume $k = i_1$, then (84) turns into

$$\sum_{i_1,\dots,i_q,k=1}^{B} \delta_{k,i_1}\mathbb{E}\left[\langle x_{\text{out}}^{i_1}, \Delta_t x_{\text{in}}^{i_1}\rangle \prod_{\ell=1}^{q}\langle x_{\text{out}}^{i_\ell}, \Delta_t x_{\text{in}}^{i_\ell}\rangle^2 \langle x_{\text{in}}^{i_1}, x_{\text{in}}^{j_1}\rangle \cdots \langle x_{\text{out}}^{j_q}, x_{\text{out}}^{i_1}\rangle\right]. \tag{94}$$

This introduces a cycle $x_{\text{in}}^{i_1} \to x_{\text{in}}^{j_1} \to \cdots \to x_{\text{out}}^{j_q} \to x_{\text{out}}^{i_1}$, which mandates $i_1 = j_1 = \cdots = j_q$. The expectation then becomes

$$N_{\text{in}}\mathbb{E}\left[\langle x_{\text{out}}^{i_1}, \Delta_t x_{\text{in}}^{i_1}\rangle^{2q+1}\right] = O\left((2q+1)!![\operatorname{Tr}(\Delta_t\Delta_t^\top)]^{(2q+1)/2}\right). \tag{95}$$

Summing over i_1 and normalizing gives

$$\frac{N_{\text{in}}}{B^{2q+1}}B(2q+1)!![\operatorname{Tr}(\Delta_t\Delta_t^\top)]^{(2q+1)/2} = O\left(\frac{\phi(2q+1)!!}{B^{2q}}[\operatorname{Tr}(\Delta_t\Delta_t^\top)]^{\frac{2q+1}{2}}\right). \tag{96}$$

This is subleading by a factor of B^{-1}, as the unmatched case scales as $B^{-(q+2)}$.

The proof is then completed by noting that if the index k is not matched with any i_u or j_v for $1 \le u, v \le q$, then following Theorem 9, the last three factors $\langle x_{\text{in}}^{j_q}, x_{\text{in}}^{k}\rangle\langle x_{\text{in}}^{k}, \Delta_t^\top x_{\text{out}}^{k}\rangle\langle x_{\text{in}}^{i_1}, \Delta_t^\top x_{\text{out}}^{k}\rangle$ can be contracted twice via the leading Wick pairings in the Gaussian expectation: first, pairing the two x_{in}^{k} terms across the inner products yields an identity operator expectation $\mathbb{E}[x_{\text{in}}^{k}(x_{\text{in}}^{k})^\top] = \text{Id}$, reducing to $\langle x_{\text{in}}^{j_q}, \Delta^\top x_{\text{out}}^{k}\rangle\langle x_{\text{in}}^{i_1}, \Delta^\top x_{\text{out}}^{k}\rangle$; second, pairing the two remaining $\Delta^\top x_{\text{out}}^{k}$ terms yields $\mathbb{E}[(\Delta^\top x_{\text{out}}^{k})(\Delta^\top x_{\text{out}}^{k})^\top] = \Delta^\top\Delta$, resulting in $\langle x_{\text{in}}^{j_q}, \Delta^\top\Delta x_{\text{in}}^{i_1}\rangle = \langle \Delta_t x_{\text{in}}^{i_1}, \Delta_t x_{\text{in}}^{j_q}\rangle + o(1)$. This closes the chain in a non-crossing manner. ■

We make use of the cases where $q = 1, 2$ in Theorem 9 to arrive at

$$\begin{aligned}
&\mathbb{E}(\langle \Delta_t, G_{t+1}\rangle|\mathscr{F}_t) \\
&= a\mathbb{E}\left(\langle \Delta_t, G\rangle\,|\mathscr{F}_t\right) + b\mathbb{E}\left(\left\langle \Delta_t, (G_tG_t^\top)G\right\rangle\middle|\mathscr{F}_t\right) + c\mathbb{E}\left(\left\langle \Delta_t, (G_tG_t^\top)^2G_t\right\rangle\middle|\mathscr{F}_t\right) \\
&= 2a\|\Delta_t\|_{\text{F}}^2 + \tfrac{b\mathbb{E}z^2}{B^3}(BN_{\text{in}}N_{\text{out}}) \\
&\quad + \tfrac{c}{B^5}\left((\mathbb{E}z^2)^2B^2\operatorname{Tr}(\Delta_t^\top\Delta)N_{\text{in}}N_{\text{out}} + (\mathbb{E}z^4)B\operatorname{Tr}(\Delta_t^\top\Delta)N_{\text{in}}N_{\text{out}}^2 + (\mathbb{E}z^2)^2B^2\operatorname{Tr}(\Delta_t^\top\Delta)N_{\text{out}}^2\right).
\end{aligned} \tag{97}$$

Lemma 10 (Gradient re-normalization) *Let $\mathfrak{G}_q := \operatorname{Tr}(GG_t^\top)^q$ and $\bar{\mathfrak{G}}_q := \operatorname{Tr}(GG_t^\top)^qG_t$. Then, in the limits of $B \to \infty, B/N_{\text{in}} \to \phi, B/N_{\text{out}} \to \psi$, we have that $\mathbb{E}\left[\mathfrak{G}_q\mathfrak{G}_1^{-q}\right] \sim \mathbb{E}\mathfrak{G}_q(\mathbb{E}\mathfrak{G}_1)^{-q}$ and that $\mathbb{E}\left[\bar{\mathfrak{G}}_q\mathfrak{G}^{-q-1/2}\right] \sim \mathbb{E}\bar{\mathfrak{G}}_q(\mathbb{E}\mathfrak{G}_1)^{-q-1/2}$.*

After gradient normalization, high-dimensional concentration from the lemma implies

$$\begin{aligned}
\mathbb{E}\left[\langle \Delta_t, G_{t+1}\rangle|\mathscr{F}_t\right] &= 2a\frac{\|\Delta_t\|_{\text{F}}^2}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^{1/2}} + \frac{b\mathbb{E}z^2}{B^3}\frac{BN_{\text{in}}N_{\text{out}}}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^{3/2}} \\
&+ \frac{c}{B^5}\left(\frac{(\mathbb{E}z^2)^2B^2\operatorname{Tr}(\Delta_t^\top\Delta)N_{\text{in}}N_{\text{out}} + (\mathbb{E}z^4)B\operatorname{Tr}(\Delta_t^\top\Delta)N_{\text{in}}N_{\text{out}}^2 + (\mathbb{E}z^2)^2B^2\operatorname{Tr}(\Delta_t^\top\Delta)N_{\text{out}}^2}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^{5/2}}\right).
\end{aligned} \tag{98}$$

On the other hand, the last term expands into

$$\begin{aligned}
&\mathbb{E}\langle G_{t+1}, G_{t+1}\rangle \\
&= a^2\mathbb{E}\langle G_t, G_t\rangle + b^2\mathbb{E}\langle (G_tG_t^\top)G_t, (G_tG_t^\top)G_t\rangle + c^2\mathbb{E}\langle (G_tG_t^\top)^2G, (G_tG_t^\top)^2G_t\rangle \\
&\quad + 2ab\mathbb{E}\langle G_t, (G_tG_t^\top)G_t\rangle + 2ac\mathbb{E}\langle G_t, (G_tG_t^\top)^2G_t\rangle + 2bc\mathbb{E}\langle (G_tG_t^\top)G_t, (G_tG_t^\top)^2G_t\rangle \\
&= a^2\mathbb{E}\operatorname{Tr}(G_tG_t^\top) + 2ab\mathbb{E}\operatorname{Tr}(G_tG_t^\top)^2 + (b^2+2ac)\mathbb{E}\operatorname{Tr}(G_tG_t^\top)^3 + 2bc\mathbb{E}\operatorname{Tr}(G_tG_t^\top)^4 + c^2\mathbb{E}\operatorname{Tr}(G_tG_t^\top)^5.
\end{aligned} \tag{99}$$

To compute the expected traces in these equations, we adapt the combinatorial counting approach in [1, 2, 12], which we describe in detail below.

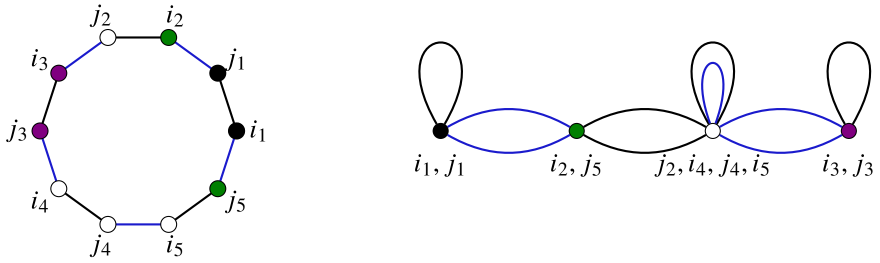


Figure 15: An admissible vertex partitioning (left) and its associated cactus graph (right). We identify vertices with the same color and join them with black and blue edges representing contributions of N_{in} and N_{out}, respectively. On the right, we count that there are four i, j identifications with four black and three blue edges. Thus, the unnormalized tracial moment reads $B^4N_{\text{in}}^4N_{\text{out}}^3$ in that order. By topological symmetry, there are ten ways of obtaining the same contribution up to relabeling the i, j indices. If we flip the edge colors, we get another ten contributions of $B^4N_{\text{in}}^3N_{\text{out}}^4$. We additionally weigh these contributions with the z moments, which can be exactly described by the number of outgoing edges for each identification. Thus, the total contribution of this configuration is $10(\mathbb{E}z^2)^3(\mathbb{E}z^4)(B^4N_{\text{in}}^4N_{\text{out}}^3 + B^4N_{\text{in}}^3N_{\text{out}}^4)$.

Our goal is to write down the formulation for any moment $q \geq 1$ in terms of the statistics of these diagrams. To this end, we decouple the Gaussian factors in (81) and (84) using the above incident graph representations. As a first step, we represent the indices $i_1, j_1, \ldots, i_q, j_q$ (and k for the second trace) as vertices in a cycle graph with $2q$ vertices for $\operatorname{Tr}(G_tG_t^\top)^q$, alternating between i_ℓ and j_ℓ, connected by edges representing inner products, e.g., $\langle x_{\text{out}}^{i_\ell}, x_{\text{out}}^{j_\ell}\rangle$ or $\langle x_{\text{in}}^{j_\ell}, x_{\text{in}}^{i_{\ell+1}}\rangle$. For the second trace, an additional vertex k and edges involving Δ_t are included, as shown in Figure 17. The cactus graph is formed by partitioning the indices into blocks (identifications), where each block corresponds to a vertex in the cactus graph, with edges representing inner products, weighted by N_{in}, N_{out}, or $\operatorname{Tr}(\Delta_t^\top\Delta_t)$, corresponding to inner products or Δ_t-related terms. We focus on non-crossing partitions with even-sized blocks denoted by

$$\mathscr{NC}^{\text{even}}(1, \ldots, n) := \{\pi \in \mathscr{NC}(1, \ldots, n) | \text{every block of } \pi \text{ has even size}\},$$

as these yield the leading-order terms due to maximal index summations. Let $C_1^{\ell_i}(\pi)$ denote the number of black cycles of length ℓ_i and $C_2^{\ell_j}(\pi)$ denote the number of blue cycles of length ℓ_j

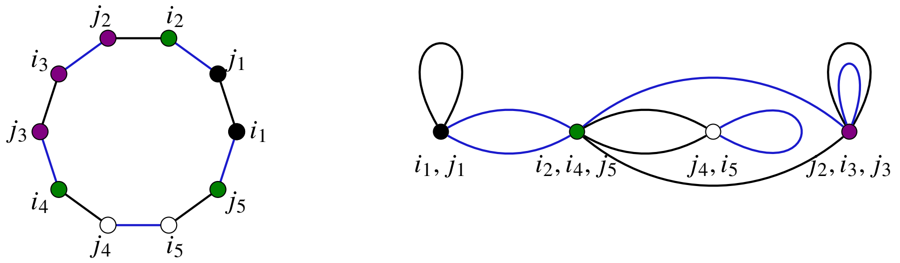


Figure 16: A non-admissible matching and its corresponding cactus graph. We see that there is a cycle $\{i_2, i_4, j_5\} \leftrightarrow \{j_2, i_3, j_3\}$ with heterogeneously decorated edge colors. In addition, we observe that these two groups are odd-sized partitions. This breaks the perfect matching condition and introduces moments of lower order.

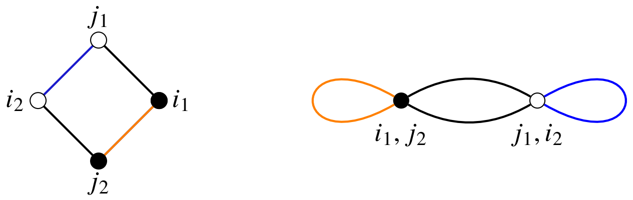


Figure 17: To compute terms like $\langle \Delta_t, (G_tG_t^\top)^q G_t\rangle$ (for $q \le 2$) in Theorem 9, we just need one extra modification to the admissible matching and its cactus graph. By contracting the chain of inner products $\langle x_{\text{in}}^{j_q}, x_{\text{in}}^{k}\rangle\langle x_{\text{in}}^{k}, \Delta_t^\top x_{\text{out}}^{k}\rangle\langle x_{\text{in}}^{i_1}, \Delta_t^\top x_{\text{out}}^{k}\rangle$ in (84), we can replace one of the N_{in}-edges (colored in orange) with $\mathrm{Tr}(\Delta_t\Delta_t^\top)$ in the cactus graph. The moment contribution in this case is thus $B^2\,\mathrm{Tr}(\Delta_t^\top\Delta_t)N_{\text{in}}N_{\text{out}}$.

in the corresponding cactus graph representation of partitioning π. Define N as the number of identifications we partitioned $i_1, j_1, \dots, i_q, j_q$ into. Let the number of outgoing edges at the k-th cactus graph vertex be E_k. Then, by Figure 15, we have that for $q \ge 1$,

$$\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top)^q = \frac{1+o(1)}{B^{2q}} \sum_{\pi\in\mathscr{NC}^{\text{even}}(1,\dots,2q)} \prod_{k=1}^{N} \mathbb{E}z^{E_k} B^N N_{\text{in}}^{\sum_i C_1^{\ell_i}(\pi)} N_{\text{out}}^{\sum_j C_2^{\ell_j}(\pi)}, \tag{100}$$

and following Figure 17, we can write (after replacing one copy of N_{in} by $\mathrm{Tr}(\Delta_t^\top\Delta_t)$)

$$\mathbb{E}\left\langle \Delta_t, (G_tG_t^\top)^q G_t\right\rangle = \frac{1+o(1)}{B^{2q+1}} \sum_{\pi\in\mathscr{NC}^{\text{even}}(1,\dots,2q)} \prod_{k=1}^{N} \mathbb{E}z^{E_k} B^N \|\Delta_t\|_{\mathrm{F}}^2 N_{\text{in}}^{\sum_i C_1^{\ell_i}(\pi)-1} N_{\text{out}}^{\sum_j C_2^{\ell_j}(\pi)}. \tag{101}$$

Calculating $\mathbb{E}z^p$ for arbitrary powers p is also straightforward: Theorem 6 gives us the even moments of one Gaussian random variable X in the form $\mathbb{E}[X^{2m}] = \#\mathscr{P}_2(2m)\cdot\sigma^{2m} = (2m-1)!!\mathbb{E}[X^2]$, where $\#\mathscr{P}_2(2m)$ denotes the number of pairings of $2m$ elements. Denote $\sigma_{\Delta_t} := \mathrm{Var}\, z$, then $z := \langle x_{\text{out}}^{(i)}, \Delta_t x_{\text{in}}^{(i)}\rangle$ is a Gaussian scalar random variable, and we have the first few even moments

$\mathbb{E}z^2 = \sigma^2_{\Delta_t}$, $\mathbb{E}z^4 = 3(\sigma^2_{\Delta_t})^2$, $\mathbb{E}z^6 = 15(\sigma^2_{\Delta_t})^3$, $\mathbb{E}z^8 = 105(\sigma^2_{\Delta_t})^4$, and $\mathbb{E}z^{10} = 945(\sigma^2_{\Delta_t})^5$. The first to fifth-order tracial moments are listed below for reference:

$$\mathbb{E}\operatorname{Tr}(G_tG_t^\top) = B^{-2}\mathbb{E}z^2(BN_{\text{in}}N_{\text{out}}), \tag{102}$$

$$\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^2) = B^{-4}\left((\mathbb{E}z^2)^2B^2N_{\text{in}}N_{\text{out}}^2 + (\mathbb{E}z^2)^2B^2N_{\text{in}}^2N_{\text{out}} + \mathbb{E}z^4BN_{\text{in}}^2N_{\text{out}}^2\right), \tag{103}$$

$$\begin{aligned}\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^3) = B^{-6}\Big(&(\mathbb{E}z^2)^3 3B^3N_{\text{in}}^2N_{\text{out}}^2 + \mathbb{E}z^2\mathbb{E}z^4 3B^2N_{\text{in}}^3N_{\text{out}}^2 + \mathbb{E}z^2\mathbb{E}z^4 3B^2N_{\text{in}}^2N_{\text{out}}^3\\ &+\mathbb{E}z^6BN_{\text{in}}^3N_{\text{out}}^3 + (\mathbb{E}z^2)^3B^3N_{\text{in}}^3N_{\text{out}} + (\mathbb{E}z^2)^3B^3N_{\text{in}}N_{\text{out}}^3\Big)(1+o(1)),\end{aligned} \tag{104}$$

$$\begin{aligned}&\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^4)\\ &= B^{-8}\Big(\mathbb{E}z^2\mathbb{E}z^6 4B^2N_{\text{in}}^4N_{\text{out}}^3 + \mathbb{E}z^2\mathbb{E}z^6 4B^2N_{\text{in}}^3N_{\text{out}}^4 + (\mathbb{E}z^2)^4B^4N_{\text{in}}^4N_{\text{out}} + (\mathbb{E}z^2)^4B^4N_{\text{in}}N_{\text{out}}^4\\ &+ (\mathbb{E}z^2)^4 2B^4N_{\text{in}}^3N_{\text{out}}^2 + (\mathbb{E}z^2)^4 2B^4N_{\text{in}}^2N_{\text{out}}^3 + (\mathbb{E}z^2)^4 4B^4N_{\text{in}}^2N_{\text{out}}^3 + (\mathbb{E}z^2)^4 4B^4N_{\text{in}}^3N_{\text{out}}^2\\ &+ (\mathbb{E}z^2)^2\mathbb{E}z^4 2B^3N_{\text{in}}^4N_{\text{out}}^2 + (\mathbb{E}z^2)^2\mathbb{E}z^4 2B^3N_{\text{in}}^2N_{\text{out}}^4 + (\mathbb{E}z^2)^2\mathbb{E}z^4 4B^3N_{\text{in}}^4N_{\text{out}}^2\\ &+ (\mathbb{E}z^2)^2\mathbb{E}z^4 4B^3N_{\text{in}}^2N_{\text{out}}^4 + (\mathbb{E}z^2)^2\mathbb{E}z^4 8B^3N_{\text{in}}^3N_{\text{out}}^3 + (\mathbb{E}z^4)^2 2B^2N_{\text{in}}^3N_{\text{out}}^4\\ &+(\mathbb{E}z^4)^2 2B^2N_{\text{in}}^4N_{\text{out}}^3 + \mathbb{E}z^8BN_{\text{in}}^4N_{\text{out}}^4 + (\mathbb{E}z^2)^2\mathbb{E}z^4 8B^3N_{\text{in}}^3N_{\text{out}}^3\Big)(1+o(1)),\end{aligned} \tag{105}$$

$$\begin{aligned}&\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^5)\\ &= B^{-10}\Big(B^5N_{\text{in}}^5N_{\text{out}}(\mathbb{E}z^2)^5 + B^5N_{\text{in}}N_{\text{out}}^5(\mathbb{E}z^2)^5 + 20B^5N_{\text{in}}^3N_{\text{out}}^3(\mathbb{E}z^2)^5\\ &+ 10B^5N_{\text{in}}^4N_{\text{out}}^2(\mathbb{E}z^2)^5 + 10B^5N_{\text{in}}^2N_{\text{out}}^4(\mathbb{E}z^2)^5 + 10B^4N_{\text{in}}^2N_{\text{out}}^5(\mathbb{E}z^2)^3\mathbb{E}z^4\\ &+ 10B^4N_{\text{in}}^5N_{\text{out}}^2(\mathbb{E}z^2)^3\mathbb{E}z^4 + 20B^4N_{\text{in}}^3N_{\text{out}}^4(\mathbb{E}z^2)^3\mathbb{E}z^4 + 20B^4N_{\text{in}}^4N_{\text{out}}^3(\mathbb{E}z^2)^3\mathbb{E}z^4\\ &+ 10B^3N_{\text{in}}^5N_{\text{out}}^3\mathbb{E}z^2(\mathbb{E}z^4)^2 + 5B^3N_{\text{in}}^5N_{\text{out}}^3(\mathbb{E}z^2)^2\mathbb{E}z^6 + 10B^3N_{\text{in}}^3N_{\text{out}}^5\mathbb{E}z^2(\mathbb{E}z^4)^2\\ &+ 5B^3N_{\text{in}}^3N_{\text{out}}^5(\mathbb{E}z^2)^2\mathbb{E}z^6 + 25B^3N_{\text{in}}^4N_{\text{out}}^4\mathbb{E}z^2(\mathbb{E}z^4)^2 + 25B^3N_{\text{in}}^4N_{\text{out}}^4(\mathbb{E}z^2)^2\mathbb{E}z^6\\ &+ 5B^2N_{\text{in}}^4N_{\text{out}}^5\mathbb{E}z^4\mathbb{E}z^6 + 5B^2N_{\text{in}}^4N_{\text{out}}^5\mathbb{E}z^2\mathbb{E}z^8 + 5B^2N_{\text{in}}^5N_{\text{out}}^4\mathbb{E}z^4\mathbb{E}z^6\\ &+ 5B^2N_{\text{in}}^5N_{\text{out}}^4\mathbb{E}z^2\mathbb{E}z^8 + BN_{\text{in}}^5N_{\text{out}}^5\mathbb{E}z^{10} + 10B^4N_{\text{in}}^4N_{\text{out}}^3(\mathbb{E}z^2)^3\mathbb{E}z^4\\ &+ 10B^4N_{\text{in}}^3N_{\text{out}}^4(\mathbb{E}z^2)^3\mathbb{E}z^4 + 5B^4N_{\text{in}}^4N_{\text{out}}^3(\mathbb{E}z^2)^3\mathbb{E}z^4 + 5B^4N_{\text{in}}^3N_{\text{out}}^4(\mathbb{E}z^2)^3\mathbb{E}z^4\\ &+ 5B^4N_{\text{in}}^4N_{\text{out}}^3(\mathbb{E}z^2)^3\mathbb{E}z^4 + 5B^4N_{\text{in}}^3N_{\text{out}}^4(\mathbb{E}z^2)^3\mathbb{E}z^4 + 10B^4N_{\text{in}}^3N_{\text{out}}^4(\mathbb{E}z^2)^3\mathbb{E}z^4\\ &+10B^4N_{\text{in}}^4N_{\text{out}}^3(\mathbb{E}z^2)^3\mathbb{E}z^4 + 5B^3N_{\text{in}}^3N_{\text{out}}^5(\mathbb{E}z^2)^2\mathbb{E}z^6 + 5B^3N_{\text{in}}^5N_{\text{out}}^3(\mathbb{E}z^2)^2\mathbb{E}z^6\Big)(1+o(1)).\end{aligned} \tag{106}$$

An interesting observation is that, for moments $q \geq 1$, the number of terms in the moment formula progresses following the sequence $\binom{3q}{q}/(2q+1)$ which, according to [13], enumerates non-crossing trees and colored partitions of a convex polygon by non-crossing diagonals [4]. Finally, normalizing the gradient square term by the Frobenius norm of G_t yields

$$\begin{aligned}\mathbb{E}\langle G_{t+1}, G_{t+1}\rangle = \frac{a^2\mathbb{E}\operatorname{Tr}(G_tG_t^\top)}{\mathbb{E}\operatorname{Tr}(G_tG_t^\top)} + \frac{2ab\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^2)}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^2} + \frac{(b^2+2ac)\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^3)}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^3}\\ + \frac{2bc\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^4)}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^4} + \frac{c^2\mathbb{E}\operatorname{Tr}((G_tG_t^\top)^5)}{(\mathbb{E}\operatorname{Tr}(G_tG_t^\top))^5}.\end{aligned} \tag{107}$$

We thus finished simplifying (81). ■

D.2. Asymptotic behavior of the Muon risk: Gradient normalization by the Frobenius norm

Regime 1: $N_{\mathrm{in}}/B = \phi, N_{\mathrm{out}}/B = \psi$. Recall that Muon updates the parameters as $W_{t+1} = W_t - \eta G_{t+1}$, where the transformed gradient is $G_{t+1} = \left(a\,\mathrm{Id} + b(G_t G_t^\top) + c(G_t G_t^\top)^2\right) G_t$, and $G_t = \frac{1}{B}\sum_{i=1}^{B} x_{\mathrm{out}}^{(i)} \otimes x_{\mathrm{in}}^{(i)} \langle x_{\mathrm{out}}^{(i)}, (W_t - W^\star) x_{\mathrm{in}}^{(i)} \rangle / \|G_t\|_{\mathrm{F}}$. The risk is $\mathscr{R}(W_t) = \frac{1}{2}\mathbb{E}\langle x_{\mathrm{out}}, (W_t - W^\star) x_{\mathrm{in}} \rangle = \frac{1}{2}\|\Delta_t\|_{\mathrm{F}}^2$. The scaling rule implies that $N_{\mathrm{in}} = \phi B$, $N_{\mathrm{out}} = \psi B$, $N_{\mathrm{in}} N_{\mathrm{out}} = \phi\psi B^2$. The variable $z = \langle x_{\mathrm{out}}^{(i)}, \Delta_t x_{\mathrm{in}}^{(i)} \rangle$ is Gaussian with variance and higher even moments

$$\begin{aligned}
\sigma_{\Delta_t}^2 &= \mathbb{E}[z^2] = \frac{\|\Delta_t\|_{\mathrm{F}}^2}{N_{\mathrm{in}} N_{\mathrm{out}}} = \frac{2\mathscr{R}(W_t)}{N_{\mathrm{in}} N_{\mathrm{out}}}, \\
\mathbb{E}[z^4] &= 3(\sigma_{\Delta_t}^2)^2 = 3\left(\frac{2\mathscr{R}(W_t)}{N_{\mathrm{in}} N_{\mathrm{out}}}\right)^2 = \frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2 B^4}, \\
\mathbb{E}[z^6] &= 15(\mathbb{E}[z^2])^3 = 15\left(\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\right)^3 = \frac{120\mathscr{R}(W_t)^3}{\phi^3\psi^3 B^6}, \\
\mathbb{E}[z^8] &= 105(\mathbb{E}[z^2])^4 = 105\left(\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\right)^4 = \frac{1680\mathscr{R}(W_t)^4}{\phi^4\psi^4 B^8}, \\
\mathbb{E}[z^{10}] &= 945(\mathbb{E}[z^2])^5 = 945\left(\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\right)^5 = \frac{30240\mathscr{R}(W_t)^5}{\phi^5\psi^5 B^{10}}.
\end{aligned} \tag{108}$$

For the drift term, we compute each subterm as follows. The first two terms are

$$2a\frac{\|\Delta_t\|_{\mathrm{F}}^2}{(\mathbb{E}[\mathrm{Tr}(G_t(G_t)^\top)])^{1/2}} = 2a\frac{2\mathscr{R}(W_t)}{\sqrt{\frac{2\mathscr{R}(W_t)}{B}}} = 2a\sqrt{2\mathscr{R}(W_t)B} \tag{109}$$

and

$$\begin{aligned}
b\frac{N_{\mathrm{in}} N_{\mathrm{out}}}{B^2}\mathbb{E}[z^2]\frac{1}{(\mathbb{E}[\mathrm{Tr}(G_t(G_t)^\top)])^{3/2}} &= b\frac{\phi\psi B^2}{B^2}\cdot\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\cdot\frac{1}{\left(\frac{2\mathscr{R}(W_t)}{B}\right)^{3/2}} \\
&= b\phi\psi\cdot\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\cdot\frac{B^{3/2}}{(2\mathscr{R}(W_t))^{3/2}} = b\frac{\sqrt{B}}{\sqrt{2\mathscr{R}(W_t)}},
\end{aligned} \tag{110}$$

while the third approximated term is

$$\begin{aligned}
&c\left(\frac{N_{\mathrm{in}} N_{\mathrm{out}}}{B^3}(\mathbb{E}[z^2])^2 + \frac{N_{\mathrm{in}} N_{\mathrm{out}}^2}{B^4}\mathbb{E}[z^4] + \frac{N_{\mathrm{out}}^2}{B^3}(\mathbb{E}[z^2])^2\right)\frac{\|\Delta_t\|_{\mathrm{F}}^2}{(\mathbb{E}[\mathrm{Tr}(G_t(G_t)^\top)])^{5/2}} \\
&= c\cdot\frac{4\mathscr{R}(W_t)^2}{B^5}\cdot\frac{1+3\psi^2+\phi}{\phi^2\psi^2}\cdot\frac{2\mathscr{R}(W_t)}{B^{5/2}}\cdot\frac{B^{5/2}}{(2\mathscr{R}(W_t))^{5/2}} \\
&= c\frac{2\sqrt{2}\mathscr{R}(W_t)^{1/2}}{B^{5/2}}\cdot\frac{1+3\psi^2+\phi}{\phi^2\psi^2}
\end{aligned} \tag{111}$$

And thus the total drift is

$$\mathbb{E}[\langle \Delta_t, G_{t+1} \rangle | \mathscr{F}_t] = 2a\sqrt{2\mathscr{R}(W_t)B} + b\frac{\sqrt{B}}{\sqrt{2\mathscr{R}(W_t)}} + c\frac{2\sqrt{2}\mathscr{R}(W_t)^{1/2}}{B^{5/2}}\cdot\frac{1+3\psi^2+\phi}{\phi^2\psi^2}(1+o(1)). \tag{112}$$

Since G_t is normalized, $\mathbb{E}[\mathrm{Tr}(G_t G_t^\top)] = 1$. We further compute higher moments. Substitute in the z moments $(\mathbb{E}[z^2])^2 = 4\mathscr{R}(W_t)^2(\phi\psi B^2)^{-2}$, $\mathbb{E}[z^4] = 12\mathscr{R}(W_t)^2(\phi\psi B^2)^{-2}$, $N_{\mathrm{in}}N_{\mathrm{out}}^2 = \phi\psi^2 B^3$, $N_{\mathrm{in}}^2 N_{\mathrm{out}} = \phi^2\psi B^3$, and $N_{\mathrm{in}}^2 N_{\mathrm{out}}^2 = \phi^2\psi^2 B^4$, then since $\mathscr{R}(W_t) = O(\phi\psi B^2)$,

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}((G_t G_t^\top)^2)] &= \frac{1}{B^4}\left((\mathbb{E}[z^2])^2 B^2 N_{\mathrm{in}} N_{\mathrm{out}}^2 + (\mathbb{E}[z^2])^2 B^2 N_{\mathrm{in}}^2 N_{\mathrm{out}} + \mathbb{E}[z^4] B N_{\mathrm{in}}^2 N_{\mathrm{out}}^2\right) \\
&= \frac{1}{B^4}\left(\frac{4\mathscr{R}(W_t)^2}{\phi^2\psi^2 B^4}\cdot B^2\phi\psi^2 B^3 + \frac{4\mathscr{R}(W_t)^2}{\phi^2\psi^2 B^4}\cdot B^2\phi^2\psi B^3 + \frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2 B^4}\cdot B\phi^2\psi^2 B^4\right) \\
&= \frac{4\mathscr{R}(W_t)^2}{B^4}\cdot\frac{B^5}{\phi\psi B^4}(1+\phi+3\phi\psi) = \frac{4\mathscr{R}(W_t)^2}{\phi\psi B^3}(1+\phi+3\phi\psi) = O(B),
\end{aligned} \tag{113}$$

and so this term is subleading in the sense that

$$\frac{\mathbb{E}[\mathrm{Tr}((G_t G_t^\top)^2)]}{(\mathbb{E}[\mathrm{Tr}(G_t G_t^\top)])^2} \sim \frac{4\mathscr{R}(W_t)^2}{\phi\psi B^3}\frac{1+\phi+3\phi\psi}{\left(\frac{2\mathscr{R}(W_t)}{B}\right)^2} = O(B^{-1}) \tag{114}$$

after gradient normalization. Similarly, for $q = 3$ we have that

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}((G_t(G_t)^\top)^3)] &= \frac{1}{B^6}\Big(3(\mathbb{E}[z^2])^3 B^3 N_{\mathrm{in}}^2 N_{\mathrm{out}}^2 + 3\mathbb{E}[z^2]\mathbb{E}[z^4] B^2 N_{\mathrm{in}}^3 N_{\mathrm{out}}^2 + 3\mathbb{E}[z^2]\mathbb{E}[z^4] B^2 N_{\mathrm{in}}^2 N_{\mathrm{out}}^3 \\
&\quad + \mathbb{E}[z^6] B N_{\mathrm{in}}^3 N_{\mathrm{out}}^3 + (\mathbb{E}[z^2])^3 B^3 N_{\mathrm{in}}^3 N_{\mathrm{out}} + (\mathbb{E}[z^2])^3 B^3 N_{\mathrm{in}} N_{\mathrm{out}}^3\Big)(1+o(1)) \\
&= \frac{1}{B^6}\left(3\left(\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\right)^3 B^3\phi^2 B^2\psi^2 B^2 + 3\cdot\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\cdot\frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2 B^4}\cdot B^2\phi^3 B^3\psi^2 B^2\right. \\
&\quad + 3\cdot\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\cdot\frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2 B^4}\cdot B^2\phi^2 B^2\psi^3 B^3 + \frac{120\mathscr{R}(W_t)^3}{\phi^3\psi^3 B^6}\cdot B\phi^3 B^3\psi^3 B^3 \\
&\quad \left. + \left(\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\right)^3 B^3\phi^3 B^3\psi B + \left(\frac{2\mathscr{R}(W_t)}{\phi\psi B^2}\right)^3 B^3\phi B\psi^3 B^3\right)(1+o(1)) \\
&= \frac{8\mathscr{R}(W_t)^3}{B^5}\left(\frac{3}{\phi\psi} + 9\left(\frac{1}{\psi}+\frac{1}{\phi}\right) + 15 + \left(\frac{1}{\psi^2}+\frac{1}{\phi^2}\right)\right) = O(B),
\end{aligned} \tag{115}$$

and so after normalization of the gradient, this term is also subleading

$$\frac{\mathbb{E}[\mathrm{Tr}((G_t G_t^\top)^3)]}{(\mathbb{E}[\mathrm{Tr}(G_t G_t^\top)])^3} \sim \frac{8\mathscr{R}(W_t)^3}{B^5}\frac{3\phi^{-1}\psi^{-1} + 9\psi^{-1} + 9\phi^{-1} + 15 + \psi^{-2} + \phi^{-2}}{\left(\frac{2\mathscr{R}(W_t)}{B}\right)^3} = O(B^{-2}). \tag{116}$$

The contributions from moments $q = 4, 5$ are suppressed by higher powers of B. Thus, in the high-dimensional limit of batch size, the variance term is dominated by

$$\mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^2 | \mathscr{F}_t] \sim a^2(1+o(1)). \tag{117}$$

Then, the risk recursion in the limit is

$$\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] = \mathscr{R}(W_t) - \eta\left(2a\sqrt{2\mathscr{R}(W_t)B} + \frac{b\sqrt{B}}{\sqrt{2\mathscr{R}(W_t)}} + \frac{2c\sqrt{2\mathscr{R}(W_t)}}{B^{5/2}}\cdot\frac{1+3\psi^2+\phi}{\phi^2\psi^2}\right) + \frac{\eta^2 a^2}{2}(1+o(1)). \tag{118}$$

Assuming $\mathscr{R}(W_t) = O(\phi\psi B^2)$, we can scale the a-term as $O(\sqrt{B^3}) = O(B^{3/2})$, b-term as $O(B^{1/2})$, and c-term as $O(B^{-1/2})$. The a-term dominates, giving

$$\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] \sim \mathscr{R}(W_t) - 2\eta a\sqrt{2\mathscr{R}(W_t)B} + \frac{\eta^2 a^2}{2}. \tag{119}$$

In the continuous limit, let $R(t) = \mathbb{E}[\mathscr{R}(W_t)]$, with $t \to \eta t$, setting $\mathscr{R}(W_{t+1}) - \mathscr{R}(W_t) = 0$, we see that at equilibrium,

$$a\sqrt{\tfrac{2B}{u_\infty}} = \tfrac{a^2\eta}{4u_\infty}, \quad u_\infty = \tfrac{a\eta}{4\sqrt{2B}}, \quad R_\infty = u_\infty^2 = \tfrac{a^2\eta^2}{32B}. \tag{120}$$

The convergence rate near equilibrium can be obtained by linearizing around u_∞. If we let $u = u_\infty + v$, then

$$\begin{aligned} \frac{\mathrm{d}v}{\mathrm{d}t} &\sim -a\sqrt{\frac{2B}{u_\infty + v}} + \frac{a^2\eta}{4(u_\infty + v)} \sim -a\sqrt{\frac{2B}{u_\infty}} \cdot \frac{v}{2u_\infty} - \frac{a^2\eta v}{4u_\infty^2} \\ &= -\left(\frac{a\sqrt{2B}}{2 \cdot \frac{a\eta}{4\sqrt{2B}}} + \frac{a^2\eta}{4 \cdot \frac{a^2\eta^2}{32B}}\right) v = -\frac{4\sqrt{2B}}{\eta} v = O(\sqrt{B}/\eta). \end{aligned} \tag{121}$$

which grows with B, indicating faster convergence for larger batch sizes, while the asymptotic risk $R_\infty = \frac{a^2\eta^2}{32B}$ decreases with B. The scaling $N_{\text{in}} = \phi B$, $N_{\text{out}} = \psi B$ implies $B^2 \propto N_{\text{in}}N_{\text{out}}$. This is computationally expensive. The nonlinear descent term $-2a\eta\sqrt{2\mathscr{R}(W_t)B}$ suggests Muon benefits from larger B, but the computational cost of NS iteration (matrix operations scaling with $N_{\text{out}}N_{\text{in}}$) makes smaller B desirable. The risk $R_\infty \sim \frac{a^2\eta^2}{32B}$ depends on B, indicating non-universal behavior unless $\eta \propto \sqrt{B}$, which may destabilize training. It is thus crucial to carefully find the best intermediate B size provided dimension parameters $N_{\text{in}}, N_{\text{out}}$ and learning rate η in this regime.

Regime 2: $N_{\text{in}}/\sqrt{B} = \phi, N_{\text{out}}/\sqrt{B} = \psi$. Under this scaling, $N_{\text{in}} = \phi\sqrt{B}$, $N_{\text{out}} = \psi\sqrt{B}$, so $N_{\text{in}}N_{\text{out}} = \phi\psi B$. The moments of z are

$$\begin{aligned} \mathbb{E}[z^2] = \frac{2\mathscr{R}(W_t)}{\phi\psi B}, \quad \mathbb{E}[z^4] = \frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2B^2}, \quad \mathbb{E}[z^6] = \frac{120\mathscr{R}(W_t)^3}{\phi^3\psi^3B^3}, \\ \mathbb{E}[z^8] = \frac{1680\mathscr{R}(W_t)^4}{\phi^4\psi^4B^4}, \quad \mathbb{E}[z^{10}] = \frac{30240\mathscr{R}(W_t)^5}{\phi^5\psi^5B^5}. \end{aligned} \tag{122}$$

For the unnormalized gradient,

$$\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top) = \frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}[z_i^2]N_{\text{in}}N_{\text{out}} = \frac{B \cdot \frac{2\mathscr{R}(W_t)}{\phi\psi B} \cdot \phi\psi B}{B^2} = \frac{2\mathscr{R}(W_t)}{B}, \tag{123}$$

Then, the unnormalized second moment expands into

$$\begin{aligned} \mathbb{E}\,\mathrm{Tr}(G_tG_t^\top)^2 &= \frac{1}{B^4}\left((\mathbb{E}[z^2])^2B^2N_{\text{in}}N_{\text{out}}^2 + (\mathbb{E}[z^2])^2B^2N_{\text{in}}^2N_{\text{out}} + \mathbb{E}[z^4]BN_{\text{in}}^2N_{\text{out}}^2\right) \\ &= \frac{1}{B^4}\left(\left(\frac{2\mathscr{R}(W_t)}{\phi\psi B}\right)^2 B^2\phi\psi B\psi\sqrt{B} + \left(\frac{2\mathscr{R}(W_t)}{\phi\psi B}\right)^2 B^2\phi^2\sqrt{B}\psi\sqrt{B} + \frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2B^2} \cdot B\phi^2B\psi^2B\right) \\ &= \frac{4\mathscr{R}(W_t)^2}{B^3}\left(\psi\sqrt{B} + \phi\sqrt{B} + \frac{3}{\phi\psi}\right) = O(B^{-1/2}). \end{aligned} \tag{124}$$

similarly for the unnormalized third moment,

$$\begin{aligned}
\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top)^3 &= \frac{1}{B^6}\Big(3(\mathbb{E}[z^2])^3B^3N_{\mathrm{in}}^2N_{\mathrm{out}}^2 + 3\mathbb{E}[z^2]\mathbb{E}[z^4]B^2N_{\mathrm{in}}^3N_{\mathrm{out}}^2 + 3\mathbb{E}[z^2]\mathbb{E}[z^4]B^2N_{\mathrm{in}}^2N_{\mathrm{out}}^3 \\
&\quad + \mathbb{E}[z^6]BN_{\mathrm{in}}^3N_{\mathrm{out}}^3 + (\mathbb{E}[z^2])^3B^3N_{\mathrm{in}}^3N_{\mathrm{out}} + (\mathbb{E}[z^2])^3B^3N_{\mathrm{in}}N_{\mathrm{out}}^3\Big)\,(1+o(1)) \\
&= \frac{1}{B^6}\Bigg(3\left(\frac{2\mathscr{R}(W_t)}{\phi\psi B}\right)^3B^3\phi^2B\psi^2B + 3\cdot\frac{2\mathscr{R}(W_t)}{\phi\psi B}\cdot\frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2B^2}\cdot B^2\phi^3B^{3/2}\psi^2B \\
&\quad + 3\cdot\frac{2\mathscr{R}(W_t)}{\phi\psi B}\cdot\frac{12\mathscr{R}(W_t)^2}{\phi^2\psi^2B^2}\cdot B^2\phi^2B\psi^3B^{3/2} + \frac{120\mathscr{R}(W_t)^3}{\phi^3\psi^3B^3}\cdot B\phi^3B^{3/2}\psi^3B^{3/2} \\
&\quad + \left(\frac{2\mathscr{R}(W_t)}{\phi\psi B}\right)^3B^3\phi^3B^{3/2}\psi\sqrt{B} + \left(\frac{2\mathscr{R}(W_t)}{\phi\psi B}\right)^3B^3\phi\sqrt{B}\psi^3B^{3/2}\Bigg)\,(1+o(1)) \\
&= \frac{8\mathscr{R}(W_t)^3}{B^4}\left(\frac{3}{\phi\psi} + \frac{9}{\sqrt{B}}\left(\frac{1}{\phi}+\frac{1}{\psi}\right) + \frac{15}{B} + \left(\frac{1}{\phi^2}+\frac{1}{\psi^2}\right)\right)(1+o(1)) = O(B^{-1}).
\end{aligned} \tag{125}$$

After normalization,

$$\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top)^2 \sim \frac{\frac{4\mathscr{R}(W_t)^2}{B^3}\left(\psi\sqrt{B}+\phi\sqrt{B}+\frac{3}{\phi\psi}\right)}{\left(\frac{2\mathscr{R}(W_t)}{B}\right)^2} = O(B^{-1/2}), \tag{126}$$

and

$$\mathbb{E}\,\mathrm{Tr}(G_tG_t^\top)^3 \sim \frac{\frac{8\mathscr{R}(W_t)^3}{B^4}\left(\frac{3}{\phi\psi}+\frac{9}{\sqrt{B}}\left(\frac{1}{\phi}+\frac{1}{\psi}\right)+\frac{15}{B}+\left(\frac{1}{\phi^2}+\frac{1}{\psi^2}\right)\right)}{\left(\frac{2\mathscr{R}(W_t)}{B}\right)^3} = O(B^{-1}). \tag{127}$$

Higher moments ($q = 4, 5$) are suppressed by $O(B^{-3/2})$ and higher. Now, the total drift is

$$\mathbb{E}[\langle\Delta_t, G_{t+1}\rangle|\mathscr{F}_t] = 2a\sqrt{2\mathscr{R}(W_t)B} + b\frac{\sqrt{B}}{\sqrt{2\mathscr{R}(W_t)}} + c\frac{2\sqrt{2}\mathscr{R}(W_t)^{1/2}}{B^{3/2}}\left(\frac{1}{\phi\psi}+\frac{3\psi\sqrt{B}}{\phi}+\frac{1}{\phi^2\psi}\right)(1+o(1)). \tag{128}$$

Assuming $\mathscr{R}(W_t) = O(\phi\psi B)$, the a-term is $O(B^{3/4})$, the b-term is $O(B^{1/4})$, while the c-term is $O(B^{1/4})$. Since the a-term dominates, the Muon risk simplifies to

$$\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] \sim \mathscr{R}(W_t) - 2\eta a\sqrt{2\mathscr{R}(W_t)B} + \frac{\eta^2a^2}{2}. \tag{129}$$

Finally, by taking $t \to \infty$, we have the limiting risk $R_\infty = \frac{a^2\eta^2}{32B}$ with convergence rate $\frac{\mathrm{d}v}{\mathrm{d}t} = -\left(\frac{4\sqrt{2B}}{\eta}\right)v = O(\sqrt{B}/\eta)$. The NS iteration scales as $O(\phi\psi B)$, significantly reducing computational cost and making Rule 2 more efficient than Rule 1.

D.3. Gradient normalization by the Schatten p-norm when $p = \infty$

Regime $N_{\text{in}}/B = \phi, N_{\text{out}}/B = \psi$. With $N_{\text{in}} = \phi B$, $N_{\text{out}} = \psi B$, $N_{\text{in}} N_{\text{out}} = \phi\psi B^2$, and $\mathscr{R}(W_t) = O(\phi\psi B^2)$. The operator norm $\|\bar{G}_t\|_\infty$ is the largest singular value of the unnormalized gradient $\bar{G}_t$. For a random matrix of the form $\bar{G}_t = \frac{1}{B}\sum_{i=1}^{B} x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} z_i$, with $x_{\text{out}}^{(i)}$, $x_{\text{in}}^{(i)}$ isotropic and $z_i \sim N(0, \frac{2\mathscr{R}(W_t)}{\phi\psi B^2})$, we approximate $\sigma_{\max}$ in the high-dimensional limit. Assuming $\phi, \psi = O(1)$, and $N_{\text{in}}, N_{\text{out}} \propto B$, the matrix behaves like a random matrix with i.i.d. entries scaled by z_i. The operator norm of a random $N_{\text{out}} \times N_{\text{in}}$ matrix with entries $\sim N(0, \sigma^2/N_{\text{in}})$ is

$$\sigma_{\max} \sim \sigma(\sqrt{N_{\text{in}}} + \sqrt{N_{\text{out}}}). \tag{130}$$

Here, the entries of $\bar{G}_t$ have variance

$$\sigma^2 := \mathbb{E}\left[\left((x_{\text{out}}^{(i)})_k (x_{\text{in}}^{(i)})_l z_i\right)^2\right] = \sigma_1^2\sigma_2^2 \cdot \frac{2\mathscr{R}(W_t)}{\phi\psi B^2}, \tag{131}$$

and so

$$\mathbb{E}[\|\bar{G}_t\|_\infty] \sim \frac{\sqrt{2\sigma_1^2\sigma_2^2\mathscr{R}(W_t)}}{\sqrt{\phi\psi B^2}}(\sqrt{\phi B} + \sqrt{\psi B}) = \sqrt{2\sigma_1^2\sigma_2^2\mathscr{R}(W_t)}\frac{\sqrt{\phi} + \sqrt{\psi}}{\sqrt{\phi\psi B}}. \tag{132}$$

For simplicity, we WLOG assume that $\sigma_1^2 = \sigma_2^2 = 1$, so

$$\mathbb{E}[\|\bar{G}_t\|_\infty] \sim \sqrt{\tfrac{2\mathscr{R}(W_t)}{B}}(\sqrt{\phi} + \sqrt{\psi}). \tag{133}$$

Then,

$$\begin{aligned}
&\mathbb{E}[\langle \Delta_t, G_{t+1}\rangle | \mathscr{F}_t] \\
&\sim \frac{1}{\mathbb{E}[\|\bar{G}_t\|_\infty | \mathscr{F}_t]}\left(a\mathbb{E}[\langle \Delta_t, \bar{G}_t\rangle | \mathscr{F}_t] + \frac{b\mathbb{E}[\langle \Delta_t, \bar{G}_t(\bar{G}_t)^\top \bar{G}_t\rangle | \mathscr{F}_t]}{\mathbb{E}[\|\bar{G}_t\|_\infty^2 | \mathscr{F}_t]} + \frac{c\mathbb{E}[\langle \Delta_t, (\bar{G}_t(\bar{G}_t)^\top)^2 \bar{G}_t\rangle | \mathscr{F}_t]}{\mathbb{E}[\|\bar{G}_t\|_\infty^4 | \mathscr{F}_t]}\right) \\
&\sim \frac{2a\sqrt{\mathscr{R}(W_t)B}}{\sqrt{\phi} + \sqrt{\psi}} + \frac{2b\mathscr{R}(W_t)}{B(\sqrt{\phi} + \sqrt{\psi})^2} + \cdots
\end{aligned} \tag{134}$$

The third term involves higher moments but scales as $O(B^{-3})$, so it is subdominant. Thus, the total drift is

$$\mathbb{E}[\langle \Delta_t, G_{t+1}\rangle | \mathscr{F}_t] \sim \frac{2a\sqrt{\mathscr{R}(W_t)B}}{\sqrt{\phi} + \sqrt{\psi}} + \frac{2b\mathscr{R}(W_t)}{B(\sqrt{\phi} + \sqrt{\psi})^2} + O(B^{-3}). \tag{135}$$

Next, we compute the variance contributions:

$$
\begin{aligned}
&\mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^2|\mathscr{F}_t] \\
&\sim \frac{1}{\mathbb{E}[\|\bar{G}_t\|_\infty^2|\mathscr{F}_t]}\left(a^2\mathbb{E}[\|\bar{G}_t\|_{\mathrm{F}}^2|\mathscr{F}_t] + 2ab\mathbb{E}[\langle \bar{G}_t, \bar{G}_t(\bar{G}_t)^\top\bar{G}_t\rangle|\mathscr{F}_t] + (b^2+2ac)\mathbb{E}[\langle \bar{G}_t, (\bar{G}_t(\bar{G}_t)^\top)^2\bar{G}_t\rangle|\mathscr{F}_t]\right) \\
&= \frac{a^2\mathbb{E}[\|\bar{G}_t\|_{\mathrm{F}}^2|\mathscr{F}_t]}{\mathbb{E}[\|\bar{G}_t\|_\infty^2|\mathscr{F}_t]} + \frac{2ab\mathbb{E}[\mathrm{Tr}((\bar{G}_t(\bar{G}_t)^\top)^2)]}{\mathbb{E}[\|\bar{G}_t\|_\infty^4|\mathscr{F}_t]} + \frac{(b^2+2ac)\mathbb{E}[\mathrm{Tr}((\bar{G}_t(\bar{G}_t)^\top)^3)]}{\mathbb{E}[\|\bar{G}_t\|_\infty^6|\mathscr{F}_t]} \\
&\sim \frac{a^2\cdot\frac{2\mathscr{R}(W_t)}{B}}{\frac{2\mathscr{R}(W_t)}{B}(\sqrt{\phi}+\sqrt{\psi})^2} + \frac{2ab\cdot\frac{4\mathscr{R}(W_t)^2}{\phi\psi B^3}(1+\phi+3\phi\psi)}{\left(\sqrt{\frac{2\mathscr{R}(W_t)}{B}}(\sqrt{\phi}+\sqrt{\psi})\right)^4} \\
&\quad + \frac{(b^2+2ac)\cdot\frac{8\mathscr{R}(W_t)^3}{B^5}\left(\frac{3}{\phi\psi}+9\left(\frac{1}{\psi}+\frac{1}{\phi}\right)+15+\frac{1}{\psi^2}+\frac{1}{\phi^2}\right)}{\left(\sqrt{\frac{2\mathscr{R}(W_t)}{B}}(\sqrt{\phi}+\sqrt{\psi})\right)^6} \\
&= \frac{a^2}{(\sqrt{\phi}+\sqrt{\psi})^2} + \frac{8ab(1+\phi+3\phi\psi)}{\phi\psi(\sqrt{\phi}+\sqrt{\psi})^4B^2\sqrt{\mathscr{R}(W_t)}} + \frac{8(b^2+2ac)\left(\frac{3}{\phi\psi}+9\left(\frac{1}{\psi}+\frac{1}{\phi}\right)+15+\frac{1}{\psi^2}+\frac{1}{\phi^2}\right)}{\sqrt{2}(\sqrt{\phi}+\sqrt{\psi})^6\mathscr{R}(W_t)^{3/2}B^2}
\end{aligned}
\tag{136}
$$

Higher-order terms ($q = 4, 5$) are suppressed by higher powers of B. With $\mathscr{R}(W_t) = O(\phi\psi B^2)$, the variance terms scale as $O(1)$, $O(B^{-1/2})$, and $O(B^{-2})$, respectively:

$$
\mathbb{E}[\|G_{t+1}\|_{\mathrm{F}}^2|\mathscr{F}_t] \sim \frac{a^2}{(\sqrt{\phi}+\sqrt{\psi})^2} + \frac{8ab(1+\phi+3\phi\psi)}{\phi\psi(\sqrt{\phi}+\sqrt{\psi})^4B^2\sqrt{\mathscr{R}(W_t)}} + O(B^{-2}). \tag{137}
$$

Unlike the Frobenius norm case shown in Appendix D.2, the variance retains higher-order contributions, as desired. The final risk recursion is therefore

$$
\begin{aligned}
\mathbb{E}[\mathscr{R}(W_{t+1})|\mathscr{F}_t] \sim \mathscr{R}(W_t) &- \eta\left(\frac{2a\sqrt{\mathscr{R}(W_t)B}}{\sqrt{\phi}+\sqrt{\psi}} + \frac{2b\mathscr{R}(W_t)}{B(\sqrt{\phi}+\sqrt{\psi})^2}\right) \\
&+ \frac{\eta^2}{2}\left(\frac{a^2}{(\sqrt{\phi}+\sqrt{\psi})^2} + \frac{8ab(1+\phi+3\phi\psi)}{\phi\psi(\sqrt{\phi}+\sqrt{\psi})^4B^2\sqrt{\mathscr{R}(W_t)}}\right)(1+o(1)).
\end{aligned}
\tag{138}
$$

With $\mathscr{R}(W_t) = O(\phi\psi B^2)$, the drift terms scale as $O(B^{3/2})$ for the a-term and $O(B)$ for the b-term. The variance terms scale as $O(\eta^2)$ for the first term and $O(\eta^2 B^{-1/2})$ for the second term. The a-term in the drift dominates, and the first variance term dominates for fixed η. At equilibrium and in the limit of t,

$$
R_\infty \sim \tfrac{\eta^2 a^2\phi\psi}{32B}. \tag{139}
$$

with convergence rate near equilibrium

$$
\tfrac{\mathrm{d}v}{\mathrm{d}t} \sim -\tfrac{4\sqrt{2B}}{\eta(\sqrt{\phi}+\sqrt{\psi})^2}v, \tag{140}
$$

indicating a rate of $O(\sqrt{B}/\eta)$.