High-dimensional isotropic scaling dynamics of Muon and SGD

McGill

Google DeepMind

Guangyuan Wang [1] Elliot Paquette [1] Atish Agarwala [2]

[1]McGill University [2]Google DeepMind

Motivation

- **Muon (Momentum Orthogonalized by Newton-Schulz)** is a new optimizer gaining traction for training large models (e.g., Transformers).
- It uses **approximate orthogonalization** via Newton-Schulz iterations.
- **The Problem:** Theoretical understanding in high-dimensional regimes is lacking. Does it actually learn differently from SGD in the limit?
- **Our Contribution:** We derive exact risk recursion dynamics for Muon and SGD under isotropic data assumptions using **Free Probability**.

Problem Setup

- Standard vector regression misses structural correlations of neural net layers. We use a **Matrix-Valued Linear Regression** with MSE loss:

$$\min_{W\in\mathbb{R}^{N_{\text{out}}\times N_{\text{in}}}} \left\{ \mathscr{R}(W) := \mathbb{E}_{(x_{\text{in}},x_{\text{out}})} \frac{1}{2} \left[x_{\text{out}}^\top W^\star x_{\text{in}} - y(W, x_{\text{in}}, x_{\text{out}}) \right]^2 \right\},$$

  where $y = x_{\text{out}}^\top W x_{\text{in}}$ with isotropic Gaussian inputs $x_{\text{in}}, x_{\text{out}}$ and risk $\mathscr{R}(W) = \frac{1}{2}\|W - W^\star\|_{\text{F}}^2$ with $W^\star := \operatorname{argmin}_W \mathscr{R}(W) \in \mathbb{R}^{N_{\text{out}}\times N_{\text{in}}}$.
- **Theorem (SGD risk recursion with unnormalized gradient).** The expected SGD risk at iteration $t+1$, conditioned on the filtration $\mathscr{F}_t$, satisfies the finite difference equation

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

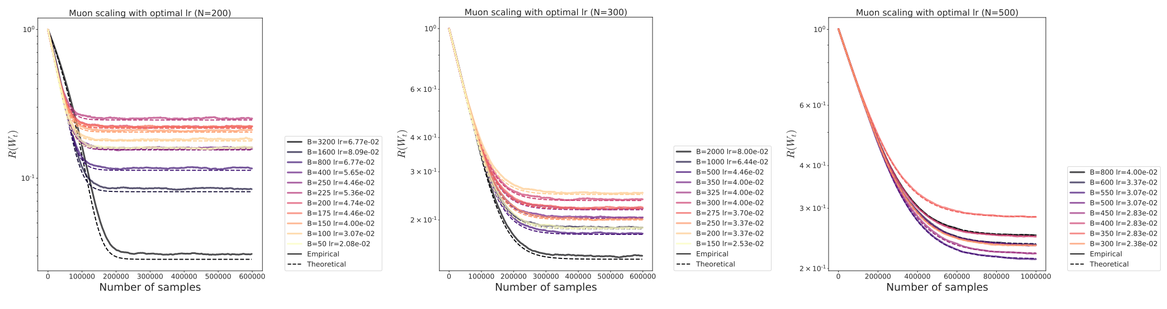
- **Theorem (SGD risk recursion with normalized gradient).** Denote the constant $\kappa := B/(B + N_{\text{in}}N_{\text{out}})$. For large batch size B and problem dimensions $N_{\text{in}}, N_{\text{out}}$, the expected risk of streaming SGD at iteration $t+1$, conditioned on $\mathscr{F}_t$, satisfies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

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

- Muon 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

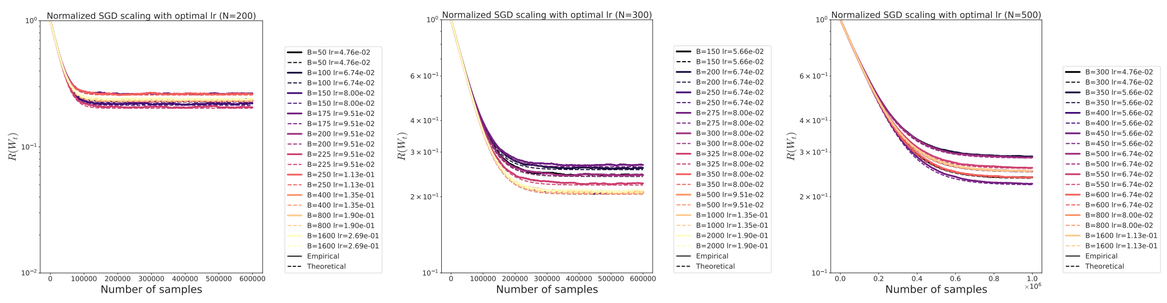
$$G_{t+1} = \frac{1}{B}\sum_{i=1}^{B} \nabla_W \mathscr{L}(W_t; (x_{\text{in}}^{(i)}, x_{\text{out}}^{(i)})) = \frac{1}{B}\sum_{i=1}^{B} x_{\text{out}}^{(i)} \otimes x_{\text{in}}^{(i)} \langle x_{\text{out}}^{(i)}, (W_t - W^\star) x_{\text{in}}^{(i)} \rangle.$$

Results & Analysis



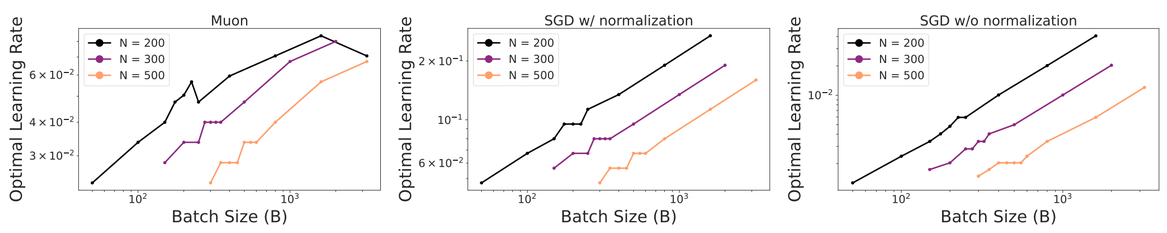
(a) $N = 200$ (b) $N = 300$ (c) $N = 500$

Figure 1. Muon scaling, sweeping logarithmically over batch sizes with optimal lr.



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

Figure 2. SGD scaling, sweeping logarithmically over batch sizes with optimal lr.



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

Figure 3. Scaling of optimal learning rates as a function of batch sizes.

Muon Dynamics & Theory

- **The Muon update:** Muon orthogonalizes the gradient G (normalized by **Frobenius norm**) using a Newton-Schulz (NS) polynomial Φ_5:

$$G_{t+1} = \Phi_5(G_t) = (a\text{Id} + bG_tG_t^\top + c(G_tG_t^\top)^2)G_t.$$

- **Challenge:** Computing expectations of high-order matrix terms like $\mathbb{E}\,\text{Tr}((GG^\top)^kG)$ in high dimensions.
- **Tool:** Free Probability Theory.
  - To solve the high-dimensional moments, we treat matrices as **non-commutative random variables**.
  - We use **cactus graphs** (combinatorial planar diagrams) to compute trace moments.
  - In the limit $B, N \to \infty$, only **non-crossing partitions** contribute to the expectation.

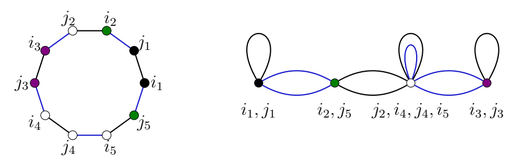


Figure 4. An admissible vertex partitioning (left) and its associated cactus graph (right). The total contribution of this configuration is $10(\mathbb{E}z^2)^3(\mathbb{E}z^4)(B^4N_{\text{in}}^4N_{\text{out}}^3 + B^4N_{\text{in}}^3N_{\text{out}}^4)$.

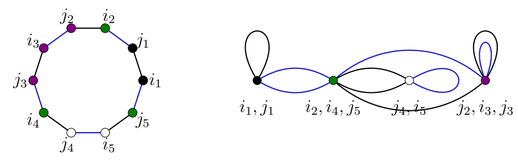


Figure 5. A non-admissible partitioning and its corresponding cactus graph. This breaks the perfect matching condition and introduces moments of lower order.

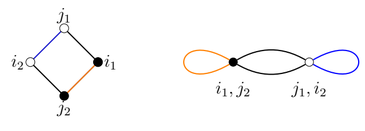


Figure 6. To compute terms like $\langle \Delta_t, (G_tG_t^\top)^qG_t\rangle$ (for $q \le 2$), we just need an extra contraction to the admissible matching and its cactus graph. The moment contribution in this case is $B^2\,\text{Tr}(\Delta_t^\top\Delta_t)N_{\text{in}}N_{\text{out}}$.

- **Theorem (Muon risk recursion).** 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 $\mathbb{E}[\langle \Delta_t, (G_tG_t^\top)^qG_t\rangle|\mathscr{F}_t]$ can be approximated by their free probability limits, dominated by non-crossing partitions. Then

$$\mathbb{E}[\mathscr{R}_{t+1}] = \mathscr{R}_t \underbrace{-\eta\mathscr{D}(\mathscr{R}_t)}_{\text{Drift}} + \underbrace{\tfrac{1}{2}\eta^2\mathscr{V}(\mathscr{R}_t)}_{\text{Variance}}.$$

1. **Drift Term:** Driven by the a, b, c coefficients of the NS polynomial.

$$\mathscr{D}(\mathscr{R}(W_t)) = \frac{4a\mathscr{R}(W_t)}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^{1/2}} + \frac{bN_{\text{in}}N_{\text{out}}}{B^2}\frac{\mathbb{E}z^2}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^{3/2}} + \left(\frac{N_{\text{in}}N_{\text{out}}}{B^3}(\mathbb{E}z^2)^2 + \frac{N_{\text{in}}N_{\text{out}}^2}{B^4}(\mathbb{E}z^4) + \frac{N_{\text{out}}^2}{B^3}(\mathbb{E}z^2)^2\right)\frac{2c\mathscr{R}(W_t)(1+o(1))}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^{5/2}}.$$

2. **Variance Term:** Stabilized by the normalization.

$$\mathscr{V}(\mathscr{R}(W_t)) = \left(a^2 + \frac{2ab\mathbb{E}\,\text{Tr}((G_tG_t^\top)^2)}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^2} + \frac{(b^2+2ac)\mathbb{E}\,\text{Tr}((G_tG_t^\top)^3)}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^3} + \frac{2bc\mathbb{E}\,\text{Tr}((G_tG_t^\top)^4)}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^4} + \frac{c^2\mathbb{E}\,\text{Tr}((G_tG_t^\top)^5)}{(\mathbb{E}\,\text{Tr}(G_tG_t^\top))^5}\right)(1+o(1)).$$

3. We focus on **non-crossing partitions with even-sized blocks** $\mathscr{NC}^{\text{even}}(n)$, as these yield the **leading-order terms** due to maximal index summations. Let $C_1^{\ell_i}(\pi)$ denote the number of **black ℓ_i-cycles** and $C_2^{\ell_j}(\pi)$ denote the number of **blue ℓ_j-cycles** in the cactus graph representation of partitioning π. Define N as the **number of identifications** we partitioned $i_1, j_1, \ldots, i_q, j_q$ into. Let the **number of outgoing edges** at the k-th cactus graph vertex be E_k, then

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

Muon Scaling Regimes

- **Main result: The "degeneracy" of Muon.** How does Muon behave as width N and batch size B grow?
- **Large batch regime ($B \propto N^2$):** The spectrum of $GG^\top$ degenerates to a point. The a-term dominates and the non-linear NS terms vanish and Muon degenerates to Normalized SGD:

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

  The orthogonalization does nothing in this isotropic limit.
- **Batch-fan proportional ($B \propto N$):** The spectrum remains rich. With Frobenius normalization, the non-linear terms still decay as $N \to \infty$. **Normalization by operator norm ($p = \infty$)** is required to preserve the benefits of Muon in this regime.