Introduction to Muon

Shaocong Ma

April 1, 2026

Outline

Motivation: Why a New Optimizer?

- Theoretically motivated by Shampoo (Bernstein & Newhouse 2024, Gupta et al. 2018).
- Empirical observation: The updates produced by both SGD-momentum and Adam for the 2D parameters are dominated by a few directions.
- **Idea:** orthogonalization can help in updating underrepresented but informative directions.

Outline

Algorithm Design: Muon

Muon applies to **2D parameters** (weight matrices) of hidden layers.

Two-step recipe:

1. Compute the SGD-momentum (Nesterov) update $\mathbf{G}$ as usual.
2. **Orthogonalize $\mathbf{G}$** via a Newton-Schulz iteration to obtain the nearest semi-orthogonal matrix.

Update Rule

$$\mathbf{G}_t = \mu\,\mathbf{G}_{t-1} + \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}_{t-1})$$

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta \cdot \mathrm{NewtonSchulz5}(\mathbf{G}_t)$$

Scalar/vector parameters and input/output layers still use AdamW.

Algorithm Design: Orthogonalization

Given the SVD of the momentum matrix $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$:

$$\text{Orthogonalize}(\mathbf{G}) \;=\; \mathbf{U}\mathbf{V}^\top$$

- Replaces all singular values with 1.
- Preserves the *directions* $(\mathbf{U}, \mathbf{V})$ but removes magnitude imbalance.

Key Point

We do *not* compute the SVD directly (too slow). Instead we approximate $\mathbf{U}\mathbf{V}^\top$ via a fast Newton-Schulz iteration (NewtonSchulze5).

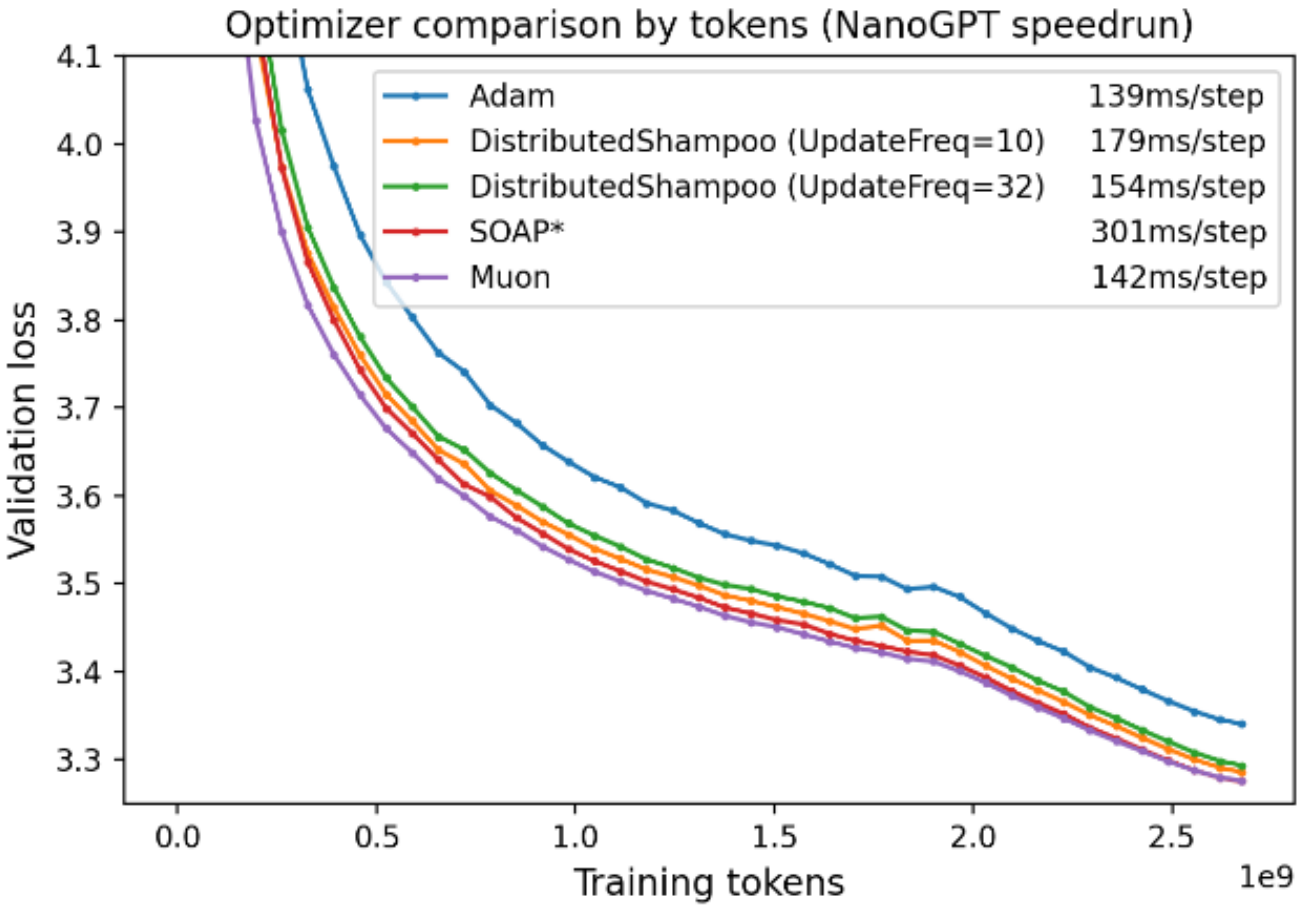
Algorithm Design: Muon

Algorithm 2 Muon

Require: Learning rate η, momentum μ

1: Initialize $B_0 \leftarrow 0$
2: **for** $t = 1, \ldots$ **do**
3: Compute gradient $G_t \leftarrow \nabla_\theta \mathcal{L}_t(\theta_{t-1})$
4: $B_t \leftarrow \mu B_{t-1} + G_t$
5: $O_t \leftarrow \text{NewtonSchulz5}(B_t)$
6: Update parameters $\theta_t \leftarrow \theta_{t-1} - \eta O_t$
7: **end for**
8: **return** θ_t

Algorithm Performance: Muon



*SOAP is under active development. Future versions will significantly improve the wallclock overhead.

Outline

Newton-Schulz (NS) Iteration

Goal: Given $\mathbf{G} \in \mathbb{R}^{m \times n}$ (with $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$), find $\mathbf{U}\mathbf{V}^\top$.

Method	**Precision**	**Speed**
Full SVD ($\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$)	Exact	Slow
Newton-Schulz iteration	**OK**	**Fast**

Derivation: We start from the equality

$$\text{Orthogonalize}(\mathbf{G}) \;=\; \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-1/2}$$

(Easy to check: Orthogonalize($\mathbf{G}$) is a fixed point.)

Newton-Schulz (NS) Iteration

Derivation: We start from the equality

$$\text{Orthogonalize}(\mathbf{G}) \;=\; \mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-1/2}$$

1. Taylor expansion on $t^{-1/2}$:

$$t^{-1/2} = 1 - \frac{1}{2}(t-1) + \frac{3}{8}(t-1)^2 - \frac{5}{16}(t-1)^3 + \dots$$

2. Iteration:

$$\text{Orthogonalize}(\mathbf{G}) \;\approx\; \frac{15}{8}\mathbf{G} - \frac{5}{4}\mathbf{G}(\mathbf{G}^\top \mathbf{G}) + \frac{3}{8}\mathbf{G}(\mathbf{G}^\top \mathbf{G})^2.$$

$$\mathbf{G}_{t+1} \leftarrow \frac{15}{8}\mathbf{G}_t - \frac{5}{4}\mathbf{G}_t(\mathbf{G}_t^\top \mathbf{G}_t) + \frac{3}{8}\mathbf{G}_t(\mathbf{G}_t^\top \mathbf{G}_t)^2$$

Newton-Schulz (NS) Iteration

Practical approach: Consider the iteration

$$\mathbf{G}_{t+1} \leftarrow a\,\mathbf{G}_t - b\,\mathbf{G}_t(\mathbf{G}_t^\top \mathbf{G}_t) + c\,\mathbf{G}_t(\mathbf{G}_t^\top \mathbf{G}_t)^2$$

Tuned coefficients:

$$\boxed{a = 3.4445, \quad b = -4.7750, \quad c = 2.0315}$$

- Maximizes convergence rate for small singular values.
- Empirically good.
- **5 steps** suffice in practice.

Outline

Remaining Open Questions

Remaining open questions in Keller Jordan's blog:

1. Will Muon scale to larger trainings? (e.g., 20B+ parameters for 1T+ tokens)

   Solved in:

   - Liu, Jingyuan, et al. "Muon is scalable for LLM training." arXiv preprint arXiv:2502.16982 (2025). **(3B/16B params for 5.7T tokens)**
   - Team, Kimi, et al. "Kimi k2: Open agentic intelligence." arXiv preprint arXiv:2507.20534 (2025). **(32B params for 15.5T tokens)**

2. Will it be possible to properly distribute the Newton-Schulz iterations used by Muon across a large-scale GPU cluster?

   Solved in:

   - Liu, Jingyuan, et al. "Muon is scalable for LLM training." arXiv preprint arXiv:2502.16982 (2025). **(Distributed Muon)**

Remaining Open Questions

3. Is it possible that Muon works only for pretraining, and won't work for finetuning or reinforcement learning workloads?

   Solved in:

   - Team, Kimi, et al. "Kimi k2: Open agentic intelligence." arXiv preprint arXiv:2507.20534 (2025). **(Used Muon in RL fine-tuning)**

Outline

Trick 1: Weight Decay

1. Performance gains diminish when scale up to train a larger model with more tokens.
   Solution: Weight Decay

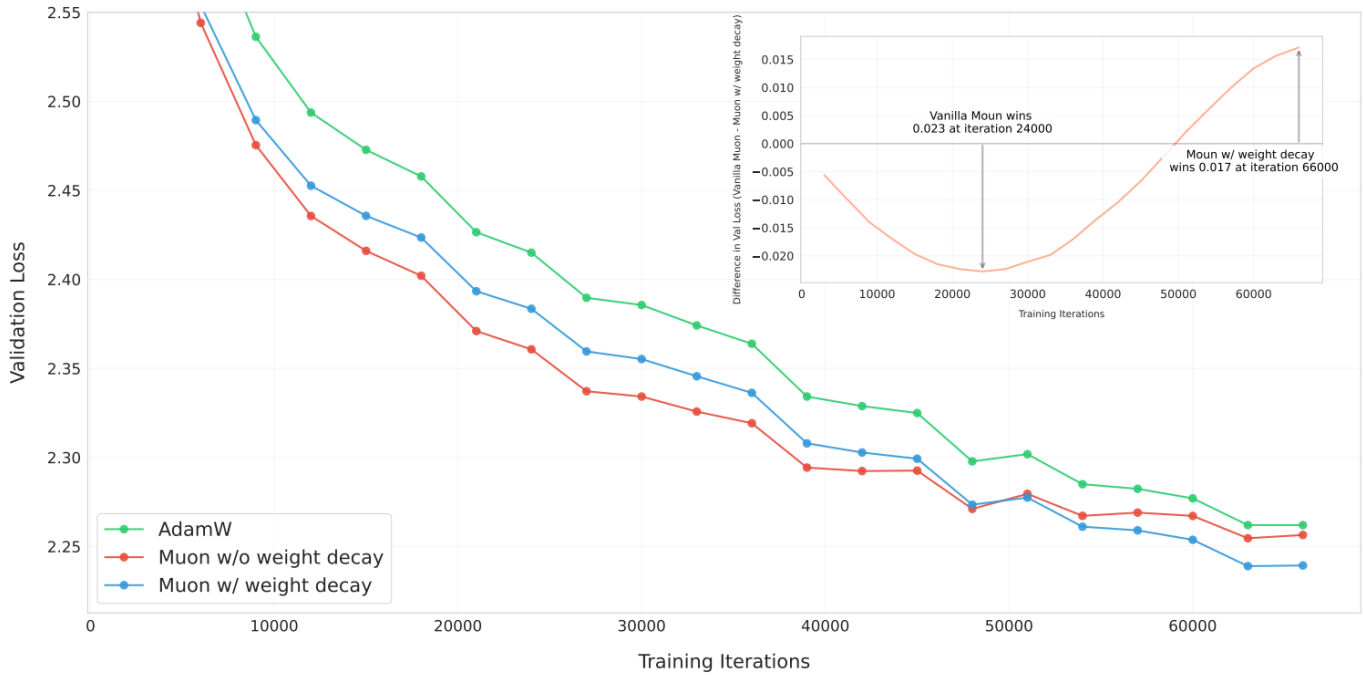
Update Rule

$$\mathbf{G}_t = \mu\,\mathbf{G}_{t-1} + \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}_{t-1})$$

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta \cdot \mathrm{NewtonSchulz5}(\mathbf{G}_t + \lambda\mathbf{W}_{t-1})$$

Trick 1: Empirical Observations

1. Performance gains diminish when scale up to train a larger model with more tokens.



Outline

Trick 2: Consistent update RMS (Root Mean Square)

1. Muon's update RMS varies depending on the shape of the parameters.
   - **Lemma 1**: For a full-rank matrix of shape $[A, B]$, the theoretical update RMS is $1/\max(A, B)$.
2. **Problem**:
   - Large $\max(A, B)$ (e.g., dense MLP) $\rightarrow$ updates too small, limiting model capacity.
   - Small $\max(A, B)$ (e.g., KV heads) $\rightarrow$ updates too large, causing instability.
3. **Solution:** Scale the update for each matrix by its $\sqrt{\max(A, B)}$.

Shape-Adjusted Update Rule

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta_t \left(\mathrm{NewtonSchulz5}(\mathbf{G}_t) \cdot \sqrt{\max(A, B)} + \lambda \mathbf{W}_{t-1} \right)$$

Outline

Trick 3: Matching update RMS of AdamW

1. In practice, Muon handles matrices, while AdamW handles non-matrix parameters (RMSNorm, LM head, embeddings).
2. **Goal**: Share hyper-parameters (learning rate η_t, weight decay λ) between matrix and non-matrix parameters.
3. Empirical observation: AdamW's update RMS is usually around 0.2 to 0.4.
4. **Solution:** Scale Muon's update RMS to match AdamW's baseline of 0.2.

Final Adjusted Update Rule

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta_t \left(0.2 \cdot \mathrm{NewtonSchulz5}(\mathbf{G}_t) \cdot \sqrt{\max(A, B)} + \lambda \mathbf{W}_{t-1} \right)$$

Outline

Performance of Muon

Table 4: Comparison of different models at around 1.2T tokens.

	Benchmark (Metric)	DSV3-Small	Moonlight-A@1.2T	Moonlight@1.2T
	Activated Params†	2.24B	2.24B	2.24B
	Total Params†	15.29B	15.29B	15.29B
	Training Tokens	1.33T	1.2T	1.2T
	Optimizer	AdamW	AdamW	Muon
English	MMLU	53.3	60.2	**60.4**
	MMLU-pro	-	26.8	**28.1**
	BBH	41.4	**45.3**	43.2
	TriviaQA	-	57.4	**58.1**
Code	HumanEval	26.8	29.3	**37.2**
	MBPP	36.8	49.2	**52.9**
Math	GSM8K	31.4	43.8	**45.0**
	MATH	10.7	16.1	**19.8**
	CMath	-	57.8	**60.2**
Chinese	C-Eval	-	57.2	**59.9**
	CMMLU	-	58.2	**58.8**

† The reported parameter counts exclude the embedding parameters.

Figure: For the same model, using Muon to pre-train is better.

Performance of Muon

Table 7: Comparison of Adam and Muon optimizers applied to the SFT of the Qwen2.5-7B pretrained model.

Benchmark (Metric)	**# Shots**	**Adam-SFT**	**Muon-SFT**
Pretrained Model	-	Qwen2.5-7B	
MMLU (EM)	0-shot (CoT)	**71.4**	70.8
HumanEval (Pass@1)	0-shot	**79.3**	77.4
MBPP (Pass@1)	0-shot	**71.9**	71.6
GSM8K (EM)	5-shot	**89.8**	85.8

Figure: Muon may not be good for SFT.

Outline

Relationship to Shampoo

Shampoo (Gupta et al. 2018)

Update Rule

Let $\mathbf{G}_t = \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_{t-1})$.

$$\boldsymbol{L}_t = \beta \boldsymbol{L}_{t-1} + \mathbf{G}_t \mathbf{G}_t^\top$$
$$\boldsymbol{R}_t = \beta \boldsymbol{R}_{t-1} + \mathbf{G}_t^\top \mathbf{G}_t$$
$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta_t \boldsymbol{L}_t^{-1/4} \mathbf{G}_t \boldsymbol{R}_t^{-1/4}$$

Muon Update Rule (simplified)

Let $\mathbf{G}_t = \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_{t-1})$.

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta_t \text{Orthogonalize}(\mathbf{G}_t)$$

Relationship to Shampoo

Shampoo Update Rule ($\beta = 0$)

Let $\mathbf{G}_t = \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_{t-1})$.

$$\begin{aligned}
\mathbf{W}_t &= \mathbf{W}_{t-1} - \eta_t (\mathbf{G}_t \mathbf{G}_t^\top)^{-1/4} \mathbf{G}_t (\mathbf{G}_t^\top \mathbf{G}_t)^{-1/4} \\
&= \mathbf{W}_{t-1} - \eta_t (\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top)^{-1/4} \mathbf{G}_t (\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top)^{-1/4} \\
&= \mathbf{W}_{t-1} - \eta_t (\mathbf{U}\mathbf{\Sigma}^{-1/2}\mathbf{U}^\top)\mathbf{U}\Sigma\mathbf{V}^\top(\mathbf{V}\mathbf{\Sigma}^{-1/2}\mathbf{V}^\top) \\
&= \mathbf{W}_{t-1} - \eta_t \mathbf{U}\mathbf{V}^\top
\end{aligned}$$

Same as Muon Update Rule (simplified).

Muon Update Rule (simplified)

Let $\mathbf{G}_t = \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_{t-1})$.

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta_t \text{Orthogonalize}(\mathbf{G}_t)$$

Outline

Relationship to Stochastic Spectral Descent

Ref: Bernstein, Jeremy, and Laker Newhouse. "Old optimizer, new norm: An anthology." arXiv preprint arXiv:2409.20325 (2024).

Steepest Descent:

$$w_{t+1} = \arg\min_{w} \left(\frac{\|w - w_t\|^2}{2\eta_t} + \mathcal{L}(w) \right)$$

First-Order Approximation of $\mathcal{L}(w)$:

$$w_{t+1} = \arg\min_{w} \left(\frac{\|w - w_t\|^2}{2\eta_t} + \nabla_{w_t} \mathcal{L}(w_t)^\top (w - w_t) \right)$$

Different norms will result in different algorithms. E.g. SGD corresponds to the ℓ_2-norm.

Relationship to Stochastic Spectral Descent

Muon

$$\mathbf{G}_t = \mu\,\mathbf{G}_{t-1} + \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}_{t-1})$$

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \eta \cdot \mathrm{NewtonSchulz5}(\mathbf{G}_t)$$

The matrix 2-norm corresponds to Muon (with taking $\mu = 0$).

Matrix Form:

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} \left(\frac{\|\mathbf{W} - \mathbf{W}_t\|^2}{2\eta_t} + \mathrm{Tr}(\nabla_{\mathbf{W}_t}\mathcal{L}(\mathbf{W}_t)^\top(\mathbf{W} - \mathbf{W}_t)) \right)$$

Equivalent form: Let $\Delta\mathbf{W}_{t+1} = \mathbf{W}_{t+1} - \mathbf{W}_t$ and $\mathbf{G}_t = \nabla_{\mathbf{W}_t}\mathcal{L}(\mathbf{W}_t)$. Let $\gamma = \|\Delta\mathbf{W}_{t+1}\|$ and $\mathbf{\Phi} = -\Delta\mathbf{W}_{t+1}/\|\Delta\mathbf{W}_{t+1}\|$.

$$\min_{\gamma \geq 0} \left(\frac{\gamma^2}{2\eta_t} - \gamma \left(\max_{\|\mathbf{\Phi}\|=1} \mathrm{Tr}(\mathbf{G}_t^\top \mathbf{\Phi}) \right) \right)$$

Relationship to Stochastic Spectral Descent

Let $\Delta \mathbf{W}_{t+1} = \mathbf{W}_{t+1} - \mathbf{W}_t$ and $\mathbf{G}_t = \nabla_{\mathbf{W}_t} \mathcal{L}(\mathbf{W}_t)$. Let $\gamma = \|\Delta \mathbf{W}_{t+1}\|$ and $\mathbf{\Phi} = -\Delta \mathbf{W}_{t+1} / \|\Delta \mathbf{W}_{t+1}\|$.

$$\min_{\gamma \geq 0} \left(\frac{\gamma^2}{2\eta_t} - \gamma \left(\max_{\|\mathbf{\Phi}\|=1} \operatorname{Tr}(\mathbf{G}_t^\top \mathbf{\Phi}) \right) \right)$$

- Optimal γ:

$$\gamma^* = \eta_t \left(\max_{\|\mathbf{\Phi}\|=1} \operatorname{Tr}(\mathbf{G}_t^\top \mathbf{\Phi}) \right).$$

- Optimal $\mathbf{\Phi}$:

$$\mathbf{\Phi}^* = \text{Orthogonalize}(\mathbf{G})$$

- Update rule:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \cdot \text{Orthogonalize}(\mathbf{G}_t)$$

Relationship to Stochastic Spectral Descent

How to find the optimal $\mathbf{\Phi}$?

$$\mathbf{\Phi}^* = \arg \max_{\|\mathbf{\Phi}\|=1} \mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi})$$

- Find an upper bound. Let $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

$$\mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi}) = \mathrm{Tr}(\sum_{i=1}^{r} \sigma_i v_i u_i^\top \mathbf{\Phi}) = \sum_{i=1}^{r} \sigma_i u_i^\top \mathbf{\Phi} v_i.$$

  Here: (1) Linearity of Tr, and (2) $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

Relationship to Stochastic Spectral Descent

How to find the optimal $\mathbf{\Phi}$?

$$\mathbf{\Phi}^* = \arg \max_{\|\mathbf{\Phi}\|=1} \mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi})$$

- Find an upper bound.

$$\mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi}) = \sum_{i=1}^{r} \sigma_i u_i^\top \mathbf{\Phi} v_i \leq \sum_{i=1}^{r} \sigma_i \cancel{\|u_i\|} \|\mathbf{\Phi} v_i\|.$$

By the definition of matrix 2-norm: $\|\mathbf{\Phi}\| := \max_{\|v\|=1} \|\mathbf{\Phi} v\| = 1$.
Therefore,

$$\|\mathbf{\Phi} v_i\| \leq 1.$$

$$\mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi}) \leq \sum_{i=1}^{r} \sigma_i.$$

Relationship to Stochastic Spectral Descent

How to find the optimal $\mathbf{\Phi}$?

$$\mathbf{\Phi}^* = \arg \max_{\|\mathbf{\Phi}\|=1} \mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi})$$

- Equality condition.

$$\mathrm{Tr}(\mathbf{G}^\top \mathbf{\Phi}) \leq \sum_{i=1}^{r} \sigma_i.$$

The equality holds if and only if $u_i^\top \mathbf{\Phi} v_i = 1$ for all i.

$$\mathbf{\Phi}^* = \mathbf{U}\mathbf{V}^\top = \mathrm{Orthogonalize}(\mathbf{G})$$

satisfies these equalities.

References:

- Keller Jordan's blog:
  https://kellerjordan.github.io/posts/muon/
- Muon is Scalable for LLM Training
  https://arxiv.org/abs/2502.16982
- https://spaces.ac.cn/archives/10592
- https://spaces.ac.cn/archives/10739