Matrix Optimization and Muon: A Natural Perspective on Neural Network Training

Part 1: Numerical Linear Algebra

Maxim Rakhuba

HSE University

Fall into ML
October, 2025

Muon: high-level computational idea

Muon (**M**oment**U**m **O**rthogonalized by **N**ewton-Schulz)

`https://kellerjordan.github.io/posts/muon/`

Muon: high-level computational idea

Muon (**M**oment**U**m **O**rthogonalized by **N**ewton-Schulz)

`https://kellerjordan.github.io/posts/muon/`

Setting: 2D weight matrices in hidden layers (and conv kernels flattened).

Core move: *Orthogonalize* momentum M_t by a fast iteration (Newton–Schulz).

Muon: high-level computational idea

Muon (**M**oment**U**m **O**rthogonalized by **N**ewton-Schulz)
`https://kellerjordan.github.io/posts/muon/`

Setting: 2D weight matrices in hidden layers (and conv kernels flattened).
Core move: *Orthogonalize* momentum M_t by a fast iteration (Newton–Schulz).

Muon (layerwise, schematic)

Given gradient g_t and momentum $m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$:

(1) $M_t \leftarrow \mathtt{matricise}(m_t)$

(2) $\widehat{O}_t \leftarrow \mathtt{NewtonSchulz5}(M_t) \approx \arg\min_{O^\top O = I} \|O - M_t\|_F$

(3) $W_{t+1} \leftarrow W_t - \eta\, \widehat{O}_t$

A reminder: $\|A\|_F^2 = \sum_{i,j} |a_{ij}|^2 = \mathrm{Tr}(A^\top A)$.

Outline

Outline

Orthogonal Procrustes Problem: Formulation

Problem: Given $A, B \in \mathbb{R}^{m\times n}$, $m \geqslant n$, solve

$$\min\left\{\|A - BQ\|_F : Q^\top Q = I\right\}.$$

When $B = I$, we get the nearest orthogonal matrix problem from Muon.

Orthogonal Procrustes Problem: Formulation

Problem: Given $A, B \in \mathbb{R}^{m\times n}$, $m \geqslant n$, solve

$$\min\left\{\|A - BQ\|_F : Q^\top Q = I\right\}.$$

When $B = I$, we get the nearest orthogonal matrix problem from Muon.

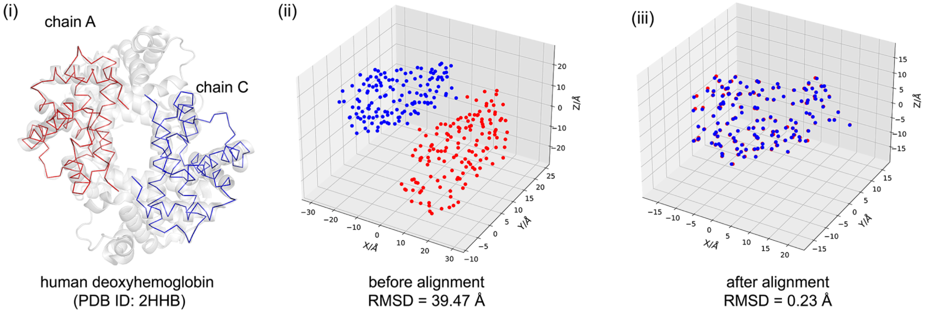


Figure: Protein structure alignment. Example from the documentation of `procrustes` library: `https://procrustes.readthedocs.io/_/downloads/en/latest/pdf/`

Orthogonal Procrustes Problem: Related Formulations

Description of Procrustes Methods

Procrustes problems arise when one wishes to find one or two transformations, $\mathbf{T} \in \mathbb{R}^{n \times n}$ and $\mathbf{S} \in \mathbb{R}^{m \times m}$, that make matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ (input matrix) resemble matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$ (target or reference matrix) as closely as possible:

$$\underbrace{\min}_{\mathbf{S},\mathbf{T}} \|\mathbf{SAT} - \mathbf{B}\|_F^2$$

Procrustes Type	S	T	Constraints
Generic [1]	$\mathbf{I}$	$\mathbf{T}$	None
Orthogonal [2, 3, 4]	$\mathbf{I}$	$\mathbf{Q}$	$\mathbf{Q}^{-1} = \mathbf{Q}^\dagger$
Rotational [4, 5, 6]	$\mathbf{I}$	$\mathbf{R}$	$\begin{cases} \mathbf{R}^{-1} = \mathbf{R}^\dagger \\ \lvert\mathbf{R}\rvert = 1 \end{cases}$
Symmetric [7, 8, 9]	$\mathbf{I}$	$\mathbf{X}$	$\mathbf{X} = \mathbf{X}^\dagger$
Permutation [10]	$\mathbf{I}$	$\mathbf{P}$	$\begin{cases} [\mathbf{P}]_{ij} \in \{0, 1\} \\ \sum_{i=1}^n [\mathbf{P}]_{ij} = \sum_{j=1}^n [\mathbf{P}]_{ij} = \end{cases}$
Two-sided Orthogonal [11]	$\mathbf{Q}_1^\dagger$	$\mathbf{Q}_2$	$\begin{cases} \mathbf{Q}_1^{-1} = \mathbf{Q}_1^\dagger \\ \mathbf{Q}_2^{-1} = \mathbf{Q}_2^\dagger \end{cases}$
Two-sided Orthogonal with One Transformation [12]	$\mathbf{Q}^\dagger$	$\mathbf{Q}$	$\mathbf{Q}^{-1} = \mathbf{Q}^\dagger$
Two-sided Permutation [13]	$\mathbf{P}_1^\dagger$	$\mathbf{P}_2$	$\begin{cases} [\mathbf{P}_1]_{ij} \in \{0, 1\} \\ [\mathbf{P}_2]_{ij} \in \{0, 1\} \\ \sum_{i=1}^n [\mathbf{P}_1]_{ij} = \sum_{j=1}^n [\mathbf{P}_1]_{ij} \\ \sum_{i=1}^n [\mathbf{P}_2]_{ij} = \sum_{j=1}^n [\mathbf{P}_2]_{ij} \end{cases}$
Two-sided Permutation with One Transformation [2, 14]	$\mathbf{P}^\dagger$	$\mathbf{P}$	$\begin{cases} [\mathbf{P}]_{ij} \in \{0, 1\} \\ \sum_{i=1}^n [\mathbf{P}]_{ij} = \sum_{j=1}^n [\mathbf{P}]_{ij} = \end{cases}$

Figure: The documentation of `procrustes` library: `https://procrustes.readthedocs.io/`

Orthogonal Procrustes Problem: Solution

$$\min\left\{\|A - BQ\|_F : Q^\top Q = I\right\}.$$

We have:

Orthogonal Procrustes Problem: Solution

$$\min\left\{\|A - BQ\|_F : Q^\top Q = I\right\}.$$

We have:

$$\|A - BQ\|_F^2 = \mathrm{Tr}\left((A - BQ)^\top (A - BQ)\right) =$$

Orthogonal Procrustes Problem: Solution

$$\min \left\{ \|A - BQ\|_F : Q^\top Q = I \right\}.$$

We have:

$$\begin{aligned} \|A - BQ\|_F^2 &= \mathrm{Tr}\left((A - BQ)^\top (A - BQ) \right) = \\ &= \mathrm{Tr}\left(A^\top A - Q^\top B^\top A - A^\top BQ + Q^\top B^\top BQ \right) = \end{aligned}$$

Orthogonal Procrustes Problem: Solution

$$\min \left\{ \|A - BQ\|_F : Q^\top Q = I \right\}.$$

We have:

$$\begin{aligned}
\|A - BQ\|_F^2 &= \mathrm{Tr}\left((A - BQ)^\top (A - BQ) \right) = \\
&= \mathrm{Tr}\left(A^\top A - Q^\top B^\top A - A^\top BQ + Q^\top B^\top BQ \right) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(Q^\top B^\top BQ) =
\end{aligned}$$

Orthogonal Procrustes Problem: Solution

$$\min\left\{\|A - BQ\|_F : Q^\top Q = I\right\}.$$

We have:

$$\begin{aligned}
\|A - BQ\|_F^2 &= \mathrm{Tr}\left((A - BQ)^\top (A - BQ)\right) = \\
&= \mathrm{Tr}\left(A^\top A - Q^\top B^\top A - A^\top BQ + Q^\top B^\top BQ\right) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(Q^\top B^\top BQ) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(B^\top B).
\end{aligned}$$

Orthogonal Procrustes Problem: Solution

$$\min\left\{\|A - BQ\|_F : Q^\top Q = I\right\}.$$

We have:

$$\begin{aligned}
\|A - BQ\|_F^2 &= \mathrm{Tr}\left((A - BQ)^\top (A - BQ)\right) = \\
&= \mathrm{Tr}\left(A^\top A - Q^\top B^\top A - A^\top BQ + Q^\top B^\top BQ\right) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(Q^\top B^\top BQ) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(B^\top B).
\end{aligned}$$

So we need to maximize $\mathrm{Tr}(Q^\top B^\top A)$. Compute SVD: $B^\top A = U\Sigma V^\top$

$$\mathrm{Tr}(Q^\top B^\top A) = \mathrm{Tr}(Q^\top U\Sigma V^\top) = \mathrm{Tr}((V^\top Q^\top U)\ \Sigma) \leqslant \sum_i \sigma_i(B^\top A),$$

Orthogonal Procrustes Problem: Solution

$$\min \left\{ \|A - BQ\|_F : Q^\top Q = I \right\}.$$

We have:

$$\begin{aligned}
\|A - BQ\|_F^2 &= \mathrm{Tr}\left((A - BQ)^\top (A - BQ) \right) = \\
&= \mathrm{Tr}\left(A^\top A - Q^\top B^\top A - A^\top BQ + Q^\top B^\top BQ \right) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(Q^\top B^\top BQ) = \\
&= \mathrm{Tr}(A^\top A) - 2\mathrm{Tr}(Q^\top B^\top A) + \mathrm{Tr}(B^\top B).
\end{aligned}$$

So we need to maximize $\mathrm{Tr}(Q^\top B^\top A)$. Compute SVD: $B^\top A = U \Sigma V^\top$

$$\mathrm{Tr}(Q^\top B^\top A) = \mathrm{Tr}(Q^\top U \Sigma V^\top) = \mathrm{Tr}((V^\top Q^\top U)\ \Sigma) \leqslant \sum_i \sigma_i(B^\top A),$$

with equality for $Q = UV^\top$.

Why Procrustes?

Setting $Q = UV^\top$ is like replacing $\Sigma \to I$: "cutting" $\sigma_i \geqslant 1$ and "stretching" $\sigma_i \leqslant 1$.

Figure: In Greek myth, Procrustes was a thief who forced his victims to fit his bed, either by stretching them or cutting their limbs.

Relation to the Polar Decomposition

Theorem (Polar decomposition)

For $A \in \mathbb{R}^{m \times n}$, $m \geqslant n$ there exists $Q \in \mathbb{R}^{m \times n}$ with orthonormal columns ($Q^\top Q = I$) and positive semidefinite $H \in \mathbb{R}^{n \times n}$ ($H = H^\top \succcurlyeq 0$):

$$A = QH.$$

Relation to the Polar Decomposition

Theorem (Polar decomposition)

For $A \in \mathbb{R}^{m \times n}$, $m \geqslant n$ there exists $Q \in \mathbb{R}^{m \times n}$ with orthonormal columns ($Q^\top Q = I$) and positive semidefinite $H \in \mathbb{R}^{n \times n}$ ($H = H^\top \succcurlyeq 0$):

$$A = QH.$$

Proof.

$$A = U\Sigma V^\top = (UV^\top)(V\Sigma V^\top) = QH.$$

□

Outline

Newton-Schulz iteration vs. SVD

$$UV^\top \in \operatorname*{arg\,min}_{Q:\ Q^\top Q = I} \|A - Q\|_F.$$

Methods

1. Using SVD: $\mathcal{O}(mn^2)$ FLOPs.
   - \+ High reliability, controlled accuracy.
   - − Large constant in $\mathcal{O}(\cdot)$.

Newton-Schulz iteration vs. SVD

$$UV^\top \in \operatorname*{arg\,min}_{Q:\ Q^\top Q = I} \|A - Q\|_F.$$

Methods

1. Using SVD: $\mathcal{O}(mn^2)$ FLOPs.
   - \+ High reliability, controlled accuracy.
   - − Large constant in $\mathcal{O}(\cdot)$.
2. Newton-Schulz iteration:

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

   - \+ GPU-friendly (matrix multiplications).
   - − Convergence can be an issue.

Newton–Schulz iteration convergence

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

Each step is a polynomial transform of singular values

For $X_k = U \Sigma_k V^\top$ (SVD):

$$X_{k+1} =$$

Newton–Schulz iteration convergence

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

Each step is a polynomial transform of singular values

For $X_k = U\Sigma_k V^\top$ (SVD):

$$X_{k+1} = \frac{1}{2}\left(3U\Sigma_k V^\top - U\Sigma_k V^\top V\Sigma_k U^\top U\Sigma_k V^\top\right) =$$

Newton–Schulz iteration convergence

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

Each step is a polynomial transform of singular values

For $X_k = U \Sigma_k V^\top$ (SVD):

$$X_{k+1} = \frac{1}{2}\left(3U\Sigma_k V^\top - U\Sigma_k V^\top V \Sigma_k U^\top U \Sigma_k V^\top\right) = \frac{1}{2} U(3\Sigma_k - \Sigma_k^3) V^\top = U\, p(\Sigma_k)\, V^\top,$$

Newton–Schulz iteration convergence

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

Each step is a polynomial transform of singular values

For $X_k = U\Sigma_k V^\top$ (SVD):

$$X_{k+1} = \frac{1}{2}\left(3U\Sigma_k V^\top - U\Sigma_k V^\top V\Sigma_k U^\top U\Sigma_k V^\top\right) = \frac{1}{2} U(3\Sigma_k - \Sigma_k^3)V^\top = U\, p(\Sigma_k)\, V^\top,$$

where $p(\sigma) = \frac{1}{2}\sigma(3 - \sigma^2)$.

Newton–Schulz iteration convergence

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

Each step is a polynomial transform of singular values

For $X_k = U \Sigma_k V^\top$ (SVD):

$$X_{k+1} = \frac{1}{2}\left(3U\Sigma_k V^\top - U\Sigma_k V^\top V \Sigma_k U^\top U \Sigma_k V^\top\right) = \frac{1}{2} U(3\Sigma_k - \Sigma_k^3) V^\top = U\, p(\Sigma_k)\, V^\top,$$

where $p(\sigma) = \frac{1}{2}\sigma(3 - \sigma^2)$. Thus,

$$X_{k+1} = U\, p(p(\ldots p(\Sigma_0)))\, V^\top.$$

Newton–Schulz iteration convergence

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^\top X_k), \quad X_0 = A.$$

Each step is a polynomial transform of singular values

For $X_k = U \Sigma_k V^\top$ (SVD):

$$X_{k+1} = \frac{1}{2} \left(3 U \Sigma_k V^\top - U \Sigma_k V^\top V \Sigma_k U^\top U \Sigma_k V^\top \right) = \frac{1}{2} U (3\Sigma_k - \Sigma_k^3) V^\top = U\, p(\Sigma_k)\, V^\top,$$

where $p(\sigma) = \frac{1}{2}\sigma(3 - \sigma^2)$. Thus,

$$X_{k+1} = U\, p(p(\dots p(\Sigma_0)))\, V^\top.$$

So we want

$$p(p(\dots p(x))) \to 1 \quad \text{on} \quad [\sigma_{\min}, \sigma_{\max}].$$

Short scalar convergence proof sketch

We just need to check that for $s_{k+1} = p(s_k)$

$$e_{k+1} = 1 - s_{k+1} \to 0.$$

Short scalar convergence proof sketch

We just need to check that for $s_{k+1} = p(s_k)$

$$e_{k+1} = 1 - s_{k+1} \to 0.$$

Using $s_k = 1 - e_k$, we have,

$$1 - e_{k+1} = p(1 - e_k),$$

Short scalar convergence proof sketch

We just need to check that for $s_{k+1} = p(s_k)$

$$e_{k+1} = 1 - s_{k+1} \to 0.$$

Using $s_k = 1 - e_k$, we have,

$$1 - e_{k+1} = p(1 - e_k),$$

which leads to

1. $e_{k+1} = \mathcal{O}(e_k^2)$ (quadratic convergence)
2. e_k approaches zero when $s_0 \leqslant \sqrt{3}$

Short scalar convergence proof sketch

We just need to check that for $s_{k+1} = p(s_k)$

$$e_{k+1} = 1 - s_{k+1} \to 0.$$

Using $s_k = 1 - e_k$, we have,

$$1 - e_{k+1} = p(1 - e_k),$$

which leads to

1. $e_{k+1} = \mathcal{O}(e_k^2)$ (quadratic convergence)
2. e_k approaches zero when $s_0 \leqslant \sqrt{3}$

Pre-scaling

To ensure $s_0 \leqslant \sqrt{3}$, one uses $X_0 = A/\|A\|_F$.

Muon's polynomial choice (practical acceleration)

Design space: odd polynomials

General odd matrix polynomial applied to X:

$$p(X) = aX + b\,XX^\top X + c\,(XX^\top)^2 X + \cdots$$

Muon's polynomial choice (practical acceleration)

Design space: odd polynomials

General odd matrix polynomial applied to X:

$$p(X) = aX + b\,XX^\top X + c\,(XX^\top)^2 X + \cdots$$

Again transforms singular values via a polynomial:

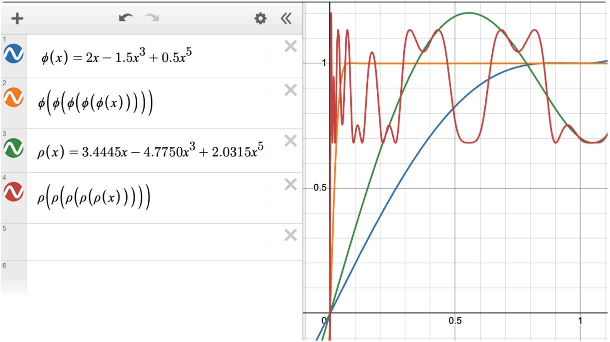
$$p(x) = ax + bx^3 + cx^5 + \cdots$$

Iterating $X \leftarrow p(X)$ should push singular values toward 1.

Muon inflates small singular values faster

Hand-tuned coefficients (used for 5 steps) that steepens the slope near 0:

$$\varphi(x) = 3.4445\,x \; - \; 4.7750\,x^3 \; + \; 2.0315\,x^5.$$



Source: Blog post: `https://kellerjordan.github.io/posts/muon/`

Is it possible to do better?

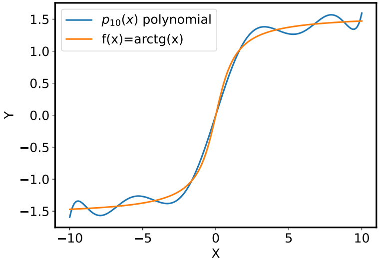
Theorem (Chebyshev)

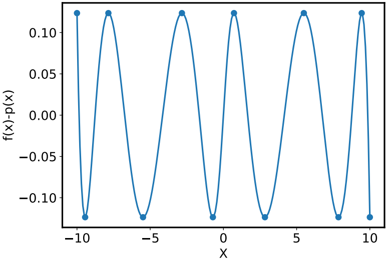
$f \in C[a,b]$ admits unique best approximation $p \in \mathbb{P}_n$:

$$p = \arg\min_{q \in \mathbb{P}_n} \|f - q\|_{C[a,b]}\,.$$

Moreover, p is best $\iff$ there exist $x_0 < x_1 < \cdots < x_{n+1}$ (alternance points):

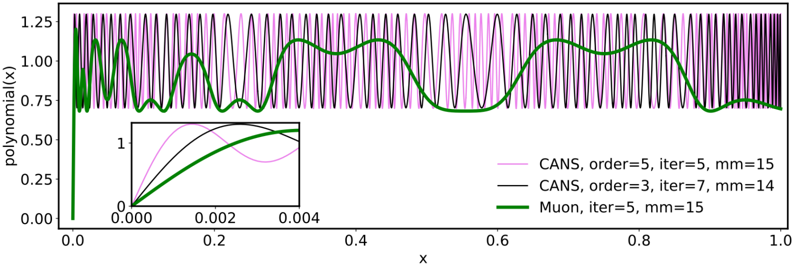
$$|p(x_j) - f(x_j)| = \|p - f\|_{C[a,b]}, \quad p(x_j) - f(x_j) = -(p(x_{j-1}) - f(x_{j-1}))$$





Is it possible to do better?

1. Grishina, Ekaterina, Matvey Smirnov, and Maxim Rakhuba. *Accelerating Newton-Schulz Iteration for Orthogonalization via Chebyshev-type Polynomials.* arXiv:2506.10935, 2025.



2. Amsel, Noah, et al. *The polar express: Optimal matrix sign methods and their application to the Muon algorithm.* arXiv:2505.16932, 2025.

Outline

Matrix Procrustes Problem

Newton-Schulz and Beyond

Another interpretation of Muon and Bonus Topic

Another interpretation of Muon

See arXiv:2502.07529

For a differentiable $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ at X:

$$f(X + S) = f(X) + \langle \nabla f(X), S\rangle + o(\|S\|).$$

Another interpretation of Muon

See arXiv:2502.07529

For a differentiable $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ at X:

$$f(X+S) = f(X) + \langle \nabla f(X), S\rangle + o(\|S\|).$$

Finding the steepest descent direction on a ball with radius η:

$$\min_{\|S\|_2 \leqslant \eta} \langle \nabla f(X), S\rangle\,, \quad \|A\|_2 = \sigma_1(A) \text{ (spectral norm)}.$$

Another interpretation of Muon

See arXiv:2502.07529

For a differentiable $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ at X:

$$f(X+S) = f(X) + \langle \nabla f(X), S\rangle + o(\|S\|).$$

Finding the steepest descent direction on a ball with radius η:

$$\min_{\|S\|_2 \leqslant \eta} \langle \nabla f(X), S\rangle\,, \quad \|A\|_2 = \sigma_1(A) \text{ (spectral norm).}$$

Solution is again the polar factor!

$$\langle \nabla f, S\rangle = \mathrm{Tr}(\nabla f^\top S) \geqslant$$

A reminder: $|\mathrm{Tr}(A^\top B)| \leqslant \sum_i \sigma_i(A)\sigma_i(B)$ (von Neumann trace inequality)

Another interpretation of Muon

See arXiv:2502.07529

For a differentiable $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ at X:

$$f(X+S) = f(X) + \langle \nabla f(X), S\rangle + o(\|S\|).$$

Finding the steepest descent direction on a ball with radius η:

$$\min_{\|S\|_2 \leqslant \eta} \langle \nabla f(X), S\rangle\,, \quad \|A\|_2 = \sigma_1(A) \text{ (spectral norm)}.$$

Solution is again the polar factor!

$$\langle \nabla f, S\rangle = \mathrm{Tr}(\nabla f^\top S) \geqslant -\sum_i \sigma_i(\nabla f)\sigma_i(S) \geqslant$$

A reminder: $|\mathrm{Tr}(A^\top B)| \leqslant \sum_i \sigma_i(A)\sigma_i(B)$ (von Neumann trace inequality)

Another interpretation of Muon

See arXiv:2502.07529

For a differentiable $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ at X:

$$f(X+S) = f(X) + \langle \nabla f(X), S\rangle + o(\|S\|).$$

Finding the steepest descent direction on a ball with radius η:

$$\min_{\|S\|_2 \leqslant \eta} \langle \nabla f(X), S\rangle\,, \quad \|A\|_2 = \sigma_1(A) \text{ (spectral norm).}$$

Solution is again the polar factor!

$$\langle \nabla f, S\rangle = \mathrm{Tr}(\nabla f^\top S) \geqslant -\sum_i \sigma_i(\nabla f)\sigma_i(S) \geqslant -\eta \sum_i \sigma_i(\nabla f).$$

A reminder: $|\mathrm{Tr}(A^\top B)| \leqslant \sum_i \sigma_i(A)\sigma_i(B)$ (von Neumann trace inequality)

Another interpretation of Muon

See arXiv:2502.07529

For a differentiable $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ at X:

$$f(X+S) = f(X) + \langle \nabla f(X), S\rangle + o(\|S\|).$$

Finding the steepest descent direction on a ball with radius η:

$$\min_{\|S\|_2 \leqslant \eta} \langle \nabla f(X), S\rangle\,, \quad \|A\|_2 = \sigma_1(A) \text{ (spectral norm)}.$$

Solution is again the polar factor!

$$\langle \nabla f, S\rangle = \operatorname{Tr}(\nabla f^\top S) \geqslant -\sum_i \sigma_i(\nabla f)\sigma_i(S) \geqslant -\eta \sum_i \sigma_i(\nabla f).$$

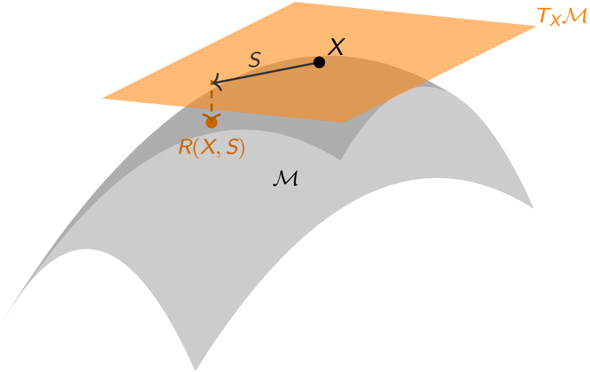
The equality is attained at $S = -\eta U V^\top$, where $\nabla f(X) = U\Sigma V^\top$.

A reminder: $|\operatorname{Tr}(A^\top B)| \leqslant \sum_i \sigma_i(A)\sigma_i(B)$ (von Neumann trace inequality)

Bonus Topic: Muon on Manifolds

We want to find a direction S on the tangent plane $T_X\mathcal{M}$:

$$\min_{\substack{\|S\|_2 \leqslant \eta \\ S \in T_X\mathcal{M}}} \langle \nabla f(X), S \rangle \,.$$



Bonus Topic: Muon on Manifolds

Applications

1. Jeremy Bernstein, "Modular Manifolds", Thinking Machines Lab: Connectionism, 2025 (`https://thinkingmachines.ai/blog/modular-manifolds/`):

$$\mathrm{St}_{m,n} = \{X \in \mathbb{R}^{m\times n} : X^\top X = I\} \quad \text{(Stiefel manifold)}$$

2. V. Bogachev, et al. "LoRA meets Riemannion: Muon Optimizer for Parametrization-independent Low-Rank Adapters." arXiv:2507.12142, 2025:

$$\mathcal{M}_r = \{X \in \mathbb{R}^{m\times n} : \mathrm{rank}(X) = r\} \quad \text{(fixed-rank manifold).}$$