

Slow Is Fast: 1. Is Muon the End of Shampoo?

Su Jianlin

September 28, 2026

As is well known, the Muon optimizer has been popular for quite a while now. Although it cannot be said to have completely replaced Adam, it has already become an optimizer that cannot be ignored and is extremely competitive. At the same time, in the past two years the Shampoo family of optimizers has also “refused to fall behind”: besides the original Shampoo, variants such as KL-Shampoo, OKLS, PSGD, and SOAP have emerged, all claiming to achieve better results than Muon.

However, this article will pour a bucket of cold water on this: for all Shampoo-like preconditioned optimizers based on Kronecker products, their ideal limit may well be Muon—the preconditioning matrices we build at great cost may all be “wasted effort.”

1 Average Convergence

We start from the average convergence conclusion obtained in “Making Deep Learning More Scientific (Part 10): Dismantling and Patching the Monotonicity Assumption”: under the convexity assumption, the update rule $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \mathbf{g}_t$ satisfies the following inequality

$$\frac{\sum_{t=1}^T \eta_t \mathbb{E}[L(\boldsymbol{\theta}_t) - L(\boldsymbol{\theta}^*)]}{\sum_{t=1}^T \eta_t} \leq \mathbb{E} \left[\frac{R^2 \Lambda_T}{2 \sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2}{2 \sum_{t=1}^T \eta_t} \right] \quad (1)$$

where

$$\Lambda_T \triangleq \lambda_{\max}(\mathbf{H}_1) + \sum_{t=2}^T [\lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1})]_+ \quad (2)$$

The meaning of the other notation will not be expanded upon here; readers can consult the original article. Intuitively, this convergence upper bound consists of two terms: the first term represents how fast the descent is, and the second term represents how much the convergence process jitters. In actual training, we need to choose appropriate η_t and $\mathbf{H}_t$ to balance the sizes of the two terms and make them as small as possible.

That is to say, we assume: **by minimizing the convergence upper bound, we can obtain a better optimizer.** There is a precedent for this approach: the article “Making Deep Learning More Scientific (Part 9): Classical Adaptive Gradient Algorithms”, based

on the assumptions $\mathbf{H}_t \succeq \mathbf{H}_{t-1}$ and constant η_t , replaced Λ_T with its upper bound $\text{tr}(\mathbf{H}_T)$, and then minimizing the upper bound yields $\mathbf{H}^* = \left(\sum_{t=1}^T \mathbf{g}_t \mathbf{g}_t^\top\right)^{1/2}$, which leads to the AdaGrad optimizer.

And this assumption is also the core starting point running through the derivation in this article.

2 Slowest Descent

Now we change perspective: suppose we have a ready-made, fixed matrix $\mathbf{H}_t$, and we transform it into $\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t$, where $\mathbf{O}_t$ is an orthogonal matrix. We hope to choose an appropriate $\mathbf{O}_t$ such that

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t (\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t)^{-1} \mathbf{g}_t \quad (3)$$

becomes a better optimizer. How do we judge whether it is good or not? According to the viewpoint of the previous section, we hope it can reduce the convergence upper bound. Note that orthogonal transformations do not change eigenvalues, so $\lambda_{\max}(\mathbf{O}_1^\top \mathbf{H}_1 \mathbf{O}_1) = \lambda_{\max}(\mathbf{H}_1)$, and then assuming $\mathbf{O}_t$ varies slowly with t , we have

$$\lambda_{\max}(\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t - \mathbf{O}_{t-1}^\top \mathbf{H}_{t-1} \mathbf{O}_{t-1}) \approx \lambda_{\max}(\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t - \mathbf{O}_t^\top \mathbf{H}_{t-1} \mathbf{O}_t) = \lambda_{\max}(\mathbf{H}_t - \mathbf{H}_{t-1}) \quad (4)$$

In this way, Λ_T is approximately unchanged under the replacement $\mathbf{H}_t \rightarrow \mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t$, so if we only want to optimize $\mathbf{O}_t$, then the objective function leaves only the second term

$$\frac{\sum_{t=1}^T \eta_t^2 \|\mathbf{g}_t\|_{(\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t)^{-1}}^2}{2 \sum_{t=1}^T \eta_t} \quad (5)$$

Since the $\mathbf{O}_t$ corresponding to each t can be different, we can completely minimize $\|\mathbf{g}_t\|_{(\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t)^{-1}}^2$ pointwise, which can also be written as

$$\min_{\mathbf{O}_t} \langle \mathbf{g}_t, (\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t)^{-1} \mathbf{g}_t \rangle \quad \text{s.t.} \quad \mathbf{O}_t^\top \mathbf{O}_t = \mathbf{I} \quad (6)$$

where $\boldsymbol{\varphi}_t = (\mathbf{O}_t^\top \mathbf{H}_t \mathbf{O}_t)^{-1} \mathbf{g}_t$ is the optimizer's update direction. We know that $\mathcal{L}(\boldsymbol{\theta} - \eta_t \boldsymbol{\varphi}_t) \approx \mathcal{L}(\boldsymbol{\theta}) - \eta_t \langle \mathbf{g}_t, \boldsymbol{\varphi}_t \rangle$, so another meaning of the above formula is that, under the first-order approximation, it minimizes the loss decrease at each step—that is, the slowest descent.

3 The Debate Between Fast and Slow

Readers who have read “The Muon Sequel: Why Did We Choose to Try Muon?”, “Steepest Descent on Manifolds: 1. SGD + Hypersphere”, “Beyond MuP: 2. Linear Layers and Steepest Descent” and similar articles should know that many optimizers are derived based

on the principle of steepest descent (i.e., fastest descent). Now here comes the slowest descent—don’t these two conflict?

In fact, this precisely represents two completely different viewpoints for looking at optimizers. Steepest descent is a first-order approximate, local greedy solution, expecting that taking the optimum at each step yields a globally good solution; it represents an optimistic assumption. Meanwhile, minimizing the convergence upper bound represents a global, pessimistic assumption, hoping that even in the worst case, we can still achieve some convergence property.

The reason the latter happens to take the counterintuitive form of the slowest descent is that here we restrict the optimization space to orthogonal transformations; the first term of the convergence upper bound is approximately negligible, leaving only the second term, whose meaning is precisely the slowest descent. This also reminds us that not all optimization spaces can consider the slowest descent; it is merely a special case within the orthogonal space, and the feasibility of other cases requires specific analysis.

At this point, the role of the slowest descent is, under the premise of guaranteeing descent, to make the optimization trajectory more stable, so that a larger learning rate can be used, or at the same learning rate to advance faster and more accurately. In the language of steepest descent, one would have to include assumptions about stability in the constraints on the update magnitude (for example, Muon chose the spectral norm).

In addition, for stochastic optimization scenarios, this term is also related to noise—decomposing $\mathbf{g}_t$ into “true value + random noise”, then $\mathbb{E}[\|\mathbf{g}_t\|_{\mathbf{H}_t^{-1}}^2]$ will yield a variance term—so the slowest descent also means reducing the influence of noise on the optimization process, thereby making the optimization more accurate.

4 Matrix Optimization

The previous sections discussed general principles; now we formally combine them with matrix optimization for the derivation. Consider a matrix parameter $\mathbf{W} \in \mathbb{R}^{n \times m}$, with gradient $\mathbf{G}$, and flattened vector $\mathbf{g} = \text{vec}(\mathbf{G}) \in \mathbb{R}^{nm}$; at this point $\mathbf{H}$ is an $nm \times nm$ matrix, which is rather taxing to think about.

For the matrix scenario, we consider preconditioning matrices that can be represented as Kronecker products, but here we do not intend to introduce the language of Kronecker products; instead we understand it directly from the perspective of linear operators, i.e., we directly understand $\mathbf{H}^{-1}\mathbf{g}$ as

$$\mathbf{H}^{-1}\mathbf{g} = \text{vec}(\mathbf{P}\mathbf{G}\mathbf{Q}) \tag{7}$$

That is to say, the optimizer’s update rule is $\mathbf{W} - \eta\mathbf{P}\mathbf{G}\mathbf{Q}$ (written with Kronecker products it is $\mathbf{H}^{-1} = \mathbf{Q} \otimes \mathbf{P}$), where $\mathbf{P} \in \mathbb{R}^{n \times n}$, $\mathbf{Q} \in \mathbb{R}^{m \times m}$ are two positive semidefinite preconditioning matrices; $\mathbf{P}, \mathbf{Q}$ can depend on the current gradient or even the historical gradients, so it can cover Muon and various Shampoo variants (momentum-free versions).

Now with the matrices $\mathbf{P}, \mathbf{G}, \mathbf{Q}$ fixed, we hope, by choosing appropriate orthogonal matrices $\mathbf{X} \in \mathbb{R}^{n \times n}$ and $\mathbf{Y} \in \mathbb{R}^{m \times m}$, to make the new update $\mathbf{X}^\top \mathbf{P} \mathbf{X} \mathbf{G} \mathbf{Y}^\top \mathbf{Q} \mathbf{Y}$ a better optimizer (written with Kronecker products it is $\mathbf{O} = \mathbf{Y} \otimes \mathbf{X}$), and the guiding principle is precisely the “orthogonal slowest descent” discussed in the previous sections, i.e.,

$$\min_{\mathbf{X}, \mathbf{Y}} \text{tr}(\mathbf{G}^\top \mathbf{X}^\top \mathbf{P} \mathbf{X} \mathbf{G} \mathbf{Y}^\top \mathbf{Q} \mathbf{Y}) \quad \text{s.t.} \quad \mathbf{X}^\top \mathbf{X} = \mathbf{I}_n, \quad \mathbf{Y}^\top \mathbf{Y} = \mathbf{I}_m \quad (8)$$

5 Preliminary Solution

The complete problem is slightly complicated. For simplicity, let us first solve the special case $\mathbf{Q} = \mathbf{I}_m$ to get a feel, where the problem simplifies to

$$\min_{\mathbf{X}} \text{tr}(\mathbf{G}^\top \mathbf{X}^\top \mathbf{P} \mathbf{X} \mathbf{G}) \quad \text{s.t.} \quad \mathbf{X}^\top \mathbf{X} = \mathbf{I}_n \quad (9)$$

Noting that $\text{tr}(\mathbf{G}^\top \mathbf{X}^\top \mathbf{P} \mathbf{X} \mathbf{G}) = \text{tr}(\mathbf{G} \mathbf{G}^\top \mathbf{X}^\top \mathbf{P} \mathbf{X})$, this is exactly the von Neumann trace inequality problem we derived in “The Rearrangement Inequality and Its Generalizations”, whose minimum value is

$$\sum_{i=1}^n \lambda_i(\mathbf{G} \mathbf{G}^\top) \lambda_{n+1-i}(\mathbf{P}) \quad (10)$$

where $\lambda_i(\cdot)$ denotes the i -th eigenvalue (from largest to smallest) of the corresponding matrix. Let the SVD of $\mathbf{G}$ be $\mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$ (the full form, i.e., $\mathbf{U}, \mathbf{V}$ are both square orthogonal matrices, with singular values arranged from largest to smallest), so the eigendecomposition of $\mathbf{G} \mathbf{G}^\top$ is $\mathbf{U} \mathbf{\Sigma} \mathbf{\Sigma}^\top \mathbf{U}^\top$; the $\mathbf{X}_*$ attaining the minimum satisfies $\mathbf{X}_*^\top \mathbf{P} \mathbf{X}_* = \mathbf{U} \mathbf{\Lambda}_P^\uparrow \mathbf{U}^\top$, where $\mathbf{\Lambda}_P^\uparrow$ is the diagonal matrix formed by the eigenvalues of $\mathbf{P}$ arranged from smallest to largest. At this point the optimal update is

$$\mathbf{X}_*^\top \mathbf{P} \mathbf{X}_* \mathbf{G} = (\mathbf{U} \mathbf{\Lambda}_P^\uparrow \mathbf{U}^\top)(\mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top) = \mathbf{U}(\mathbf{\Lambda}_P^\uparrow \mathbf{\Sigma}) \mathbf{V}^\top \quad (11)$$

When the author first saw this result, he was quite shocked. It tells us that no matter what the original $\mathbf{P}$ is, after orthogonal slowest descent, it can at most preserve the eigenvalues, while the eigenvector matrix completely disappears. The final update is $\mathbf{G}$ ’s left and right singular matrices sandwiching a diagonal matrix $\mathbf{\Lambda}_P^\uparrow \mathbf{\Sigma}$; this diagonal matrix is the eigenvalues of $\mathbf{P}$ multiplied in reverse order by the singular values of $\mathbf{G}$. This approach makes all singular values more uniform, consistent with the idea of Muon.

This result hints that no matter how expensive a preconditioner we build, the best outcome is merely “racing toward Muon.” Next we will see that this conclusion still holds for the general case $\mathbf{Q} \neq \mathbf{I}_m$.

6 The General Case

Let $\mathbf{A} = \mathbf{G}^\top \mathbf{X}^\top \mathbf{P} \mathbf{X} \mathbf{G}$, $\mathbf{B} = \mathbf{G} \mathbf{Y}^\top \mathbf{Q} \mathbf{Y} \mathbf{G}^\top$; by the invariance of the trace,

$$\text{tr}(\mathbf{G}^\top \mathbf{X}^\top \mathbf{P} \mathbf{X} \mathbf{G} \mathbf{Y}^\top \mathbf{Q} \mathbf{Y}) = \text{tr}(\mathbf{A} \mathbf{Y}^\top \mathbf{Q} \mathbf{Y}) = \text{tr}(\mathbf{B} \mathbf{X}^\top \mathbf{P} \mathbf{X}) \quad (12)$$

Let the optimal solution be $\mathbf{X}_*, \mathbf{Y}_*$, with the corresponding $\mathbf{A}, \mathbf{B}$ being $\mathbf{A}_*, \mathbf{B}_*$; fixing $\mathbf{X}_*, \mathbf{Y}_*$ respectively to solve for $\mathbf{Y}_*, \mathbf{X}_*$, according to the result of the previous section we can write the stationarity conditions

$$\mathbf{X}_*^\top \mathbf{P} \mathbf{X}_* = \mathbf{U}_B \Lambda_P^\uparrow \mathbf{U}_B^\top, \quad \mathbf{Y}_*^\top \mathbf{Q} \mathbf{Y}_* = \mathbf{U}_A \Lambda_Q^\uparrow \mathbf{U}_A^\top \quad (13)$$

where $\mathbf{U}_A \Lambda_A^\downarrow \mathbf{U}_A^\top, \mathbf{U}_B \Lambda_B^\downarrow \mathbf{U}_B^\top$ are respectively the eigendecompositions of $\mathbf{A}_*, \mathbf{B}_*$ (“ $\downarrow$ ” means the eigenvalues are sorted from largest to smallest). Our task is to analyze the properties of $\Phi = \mathbf{X}_*^\top \mathbf{P} \mathbf{X}_* \mathbf{G} \mathbf{Y}_*^\top \mathbf{Q} \mathbf{Y}_*$. The first conclusion is just like the previous section: **Φ has the form $\mathbf{U} \mathbf{D} \mathbf{V}^\top$, where $\mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$ is the SVD of $\mathbf{G}$, and $\mathbf{D}$ is some diagonal matrix.** The proof is not difficult: according to the stationarity conditions, direct computation gives

$$\mathbf{G}^\top \Phi = \mathbf{U}_A (\Lambda_A^\downarrow \Lambda_Q^\uparrow) \mathbf{U}_A^\top, \quad \Phi \mathbf{G}^\top = \mathbf{U}_B (\Lambda_B^\downarrow \Lambda_P^\uparrow) \mathbf{U}_B^\top \quad (14)$$

This shows that both $\mathbf{G}^\top \Phi$ and $\Phi \mathbf{G}^\top$ are symmetric matrices. Since $\mathbf{U}, \mathbf{V}$ are both square orthogonal matrices, any Φ can be uniquely written as $\Phi = \mathbf{U} \mathbf{M} \mathbf{V}^\top$ (i.e., $\mathbf{M} = \mathbf{U}^\top \Phi \mathbf{V}$). Substituting, we get $\mathbf{G}^\top \Phi = \mathbf{V} \mathbf{\Sigma}^\top \mathbf{M} \mathbf{V}^\top$ and $\Phi \mathbf{G}^\top = \mathbf{U} \mathbf{M} \mathbf{\Sigma}^\top \mathbf{U}^\top$; their symmetry is respectively equivalent to $\mathbf{\Sigma}^\top \mathbf{M}$ and $\mathbf{M} \mathbf{\Sigma}^\top$ being symmetric, which in component form reads

$$\sigma_i M_{i,j} = \sigma_j M_{j,i}, \quad \sigma_j M_{i,j} = \sigma_i M_{j,i} \quad (15)$$

From this we obtain $(\sigma_i^2 - \sigma_j^2) M_{i,j} = 0$. Assuming the singular values of $\mathbf{G}$ are pairwise distinct (otherwise, understand it as the limit of a matrix sequence), it follows that $M_{i,j} = 0$ when $i \neq j$, i.e., $\mathbf{M}$ is the diagonal matrix $\mathbf{D}$. This result again shows: even in the general case, the eigenvector matrix of the preconditioning matrix is still washed out by the orthogonal slowest descent, at most leaving the eigenvalues.

7 Tending Toward Equilibrium

Unfortunately, we cannot write down a closed-form solution for $\mathbf{D}$, but we can qualitatively derive: **the diagonal of $\mathbf{D}$ has a tendency toward equilibrium.** Specifically, according to $\Phi = \mathbf{U} \mathbf{D} \mathbf{V}^\top$, we can write

$$\mathbf{G}^\top \Phi = \mathbf{V} (\mathbf{\Sigma}^\top \mathbf{D}) \mathbf{V}^\top, \quad \Phi \mathbf{G}^\top = \mathbf{U} (\mathbf{D} \mathbf{\Sigma}^\top) \mathbf{U}^\top \quad (16)$$

Comparing with Equation (14), by the uniqueness of eigendecomposition (note: the uniqueness of eigendecomposition relies on the eigenvalues being pairwise distinct; otherwise we could construct counterexamples making the problem more complicated, but here we are mainly doing a heuristic derivation, so we won’t fuss over the details), we can conclude that $\mathbf{U}_A, \mathbf{U}_B$ differ from $\mathbf{V}, \mathbf{U}$ by at most a permutation matrix, so we can write

$$\mathbf{U}_A = \mathbf{V} \mathbf{\Pi}_Q, \quad \mathbf{U}_B = \mathbf{U} \mathbf{\Pi}_P \quad (17)$$

where Π_Q, Π_P are two permutation matrices, so

$$\begin{aligned}
\Phi &= X_*^\top P X_* G Y_*^\top Q Y_* = (U_B \Lambda_P^\uparrow U_B^\top) G (U_A \Lambda_Q^\uparrow U_A^\top) \\
&= (U \Pi_P \Lambda_P^\uparrow \Pi_P^\top U^\top) (U \Sigma V^\top) (V \Pi_Q \Lambda_Q^\uparrow \Pi_Q^\top V^\top) \quad (18) \\
&= U \left((\Pi_P \Lambda_P^\uparrow \Pi_P^\top) \Sigma (\Pi_Q \Lambda_Q^\uparrow \Pi_Q^\top) \right) V^\top
\end{aligned}$$

Thus $\text{tr}(G^\top \Phi) = \text{tr}(\Sigma^\top (\Pi_P \Lambda_P^\uparrow \Pi_P^\top) \Sigma (\Pi_Q \Lambda_Q^\uparrow \Pi_Q^\top))$; note that every term inside the parentheses is a diagonal matrix, so finding its minimum amounts to finding a pairing of the eigenvalues of $P, Q, G^\top G$ that minimizes their sum of products—this is essentially the three-sequence generalization of the rearrangement inequality. From “The Rearrangement Inequality and Its Generalizations”, the most ideal result is that after pairing, the componentwise products are all equal, which corresponds to $D \propto \Sigma^{-1}$ (taking the inverse entrywise on the diagonal, shape unchanged).

Of course, since we are only finding the minimum through pairing, it is hard to achieve equality of all products; generally we can only expect it to be an intermediate state between Σ (worst case) and Σ^{-1} (best case), and the geometric mean of these two is exactly the identity matrix! So this once again points to the orthogonalization operation that makes all singular values uniform!

8 Further Remarks

It must be declared that the author has no intention of praising Muon and trampling Shampoo. On the contrary, some time ago the author spent considerable time testing various Shampoo variants, with the original intention of finding an optimizer better than Muon. Unfortunately, after trying for a long time, the author found that no Shampoo variant could beat Muon. It happened that recently, while deriving convergence results for adaptive optimizers, the author combined the two lines of thinking, and ultimately this article’s conclusion emerged.

To be honest, this article’s conclusion relies on the “orthogonal slowest descent” framework, and that framework in turn relies on a series of assumptions, each link of which has questionable aspects. For example, using the convergence upper bound to guide optimizer design is itself a pessimistic, worst-case criterion; a smaller upper bound does not guarantee faster actual convergence. As another example, the assumption that $\mathbf{O}_t$ varies slowly and the first-order approximations everywhere are rather crude and cannot be made precisely quantitative.

More importantly, although minimizing the convergence upper bound looks more global, it is essentially also a greedy solution, and if we scrutinize it strictly, the very proposition cannot even hold! Because it takes the $\mathbf{g}_t$ sequence as given and then derives the remaining parameters. However, once the learning rate or other parameters are modified, the subse-

quent gradients change accordingly; there is simply no possibility of a given full gradient sequence.

Of course, if we accept these non-rigorous aspects and accept orthogonal slowest descent, then we can obtain the preceding heuristic conclusion, which tells us: we only need to focus on adjusting the spectrum (singular values) of the gradient or momentum, without spending great cost on complex preconditioning matrices—this is at least engineering-friendly. As for the Shampoo route, whether it works well is another matter, but the additional state variables and computation it introduces definitely carry a large engineering cost.

9 Article Summary

Starting from the average convergence upper bound of convex optimization, this article proposed the principle of “orthogonal slowest descent,” and applied it to Kronecker-product-based preconditioned optimizers (the Muon, Shampoo families), finding that no matter what the original preconditioning matrix is, orthogonal slowest descent will “wash away” its eigenvector matrix, at most leaving the eigenvalues to modify the singular value spectrum of the gradient (momentum) and equilibrate it—all of this is highly homogeneous with Muon—as the title says.