Steepest Descent on Manifolds: 7. Analytical Solution for Stiefel

Su Jianlin

August 17, 2026

For orthogonal manifolds, our previous conclusion was: in the square matrix case, we can completely write out the analytical solution for the corresponding steepest descent (refer to *Steepest Descent on Manifolds: 2. Muon + Orthogonal*); however, for the non-square Stiefel manifold, its steepest descent problem requires solving a system of nonlinear equations, and its analytical solution cannot be written directly (refer to *Steepest Descent on Manifolds: 3. Muon + Stiefel*).

But recently, the paper *Muon on the Stiefel Manifold Admits an Exact Closed-Form Update* overturned this assertion, pointing out that an explicit solution can likewise be written for the Stiefel manifold, without needing to solve a system of equations. This article will restate the entire derivation process using my own approach.

Problem Review

We will not restate the problem background in too much detail, and directly give the problem to be solved: let $\boldsymbol{W} \in \mathbb{R}^{n\times m}(n \geq m)$ be the current parameters, and $\boldsymbol{G}$ be the gradient of the objective function at $\boldsymbol{W}$. We are looking for the steepest descent direction $\boldsymbol{\Phi} \in \mathbb{R}^{n\times m}$ with step size η under the orthogonality constraint, i.e.,

$$\max_{\boldsymbol{\Phi}} \operatorname{tr}(\boldsymbol{G}^{\top}\boldsymbol{\Phi}) \qquad \text{s.t.} \qquad \|\boldsymbol{\Phi}\|_2 \leq 1,\ \boldsymbol{W}^{\top}\boldsymbol{W} = \boldsymbol{I}_m,\ (\boldsymbol{W} - \eta\boldsymbol{\Phi})^{\top}(\boldsymbol{W} - \eta\boldsymbol{\Phi}) = \boldsymbol{I}_m \tag{1}$$

Following the "first-order approximation is sufficient" principle, we expand the last constraint and omit the η^2 term, simplifying it to

$$\max_{\boldsymbol{\Phi}} \operatorname{tr}(\boldsymbol{G}^{\top}\boldsymbol{\Phi}) \qquad \text{s.t.} \qquad \|\boldsymbol{\Phi}\|_2 \leq 1,\ \boldsymbol{W}^{\top}\boldsymbol{W} = \boldsymbol{I}_m,\ \boldsymbol{W}^{\top}\boldsymbol{\Phi} + \boldsymbol{\Phi}^{\top}\boldsymbol{W} = \boldsymbol{0} \tag{2}$$

where $\|\cdot\|_2$ is the spectral norm, and $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. Here, changing $\|\boldsymbol{\Phi}\|_2 \leq 1$ to $\|\boldsymbol{\Phi}\|_2 = 1$ yields a completely equivalent result (the optimal point of a linear objective function can always be attained on the boundary), but here we retain $\leq$ because

$\|\boldsymbol{\Phi}\|_2 \leq 1$ is a convex set, which provides more room for maneuver during the argument. Our previous solution result was

$$\boldsymbol{\Phi} = \mathrm{msign}(\boldsymbol{G} + \boldsymbol{W}\boldsymbol{X}) \tag{3}$$

where $\boldsymbol{X} \in \mathbb{R}^{m\times m}$ is an antisymmetric matrix, and it satisfies the following matrix equation:

$$\boldsymbol{W}^\top \mathrm{msign}(\boldsymbol{G} + \boldsymbol{W}\boldsymbol{X}) + \mathrm{msign}(\boldsymbol{G} + \boldsymbol{W}\boldsymbol{X})^\top \boldsymbol{W} = \boldsymbol{0} \tag{4}$$

The solution for the case $n = m$ was already obtained in *Steepest Descent on Manifolds: 2. Muon + Orthogonal*, but for the standard Stiefel manifold with $n > m$, solving this equation is non-trivial, and even numerical solving is somewhat troublesome. The assertion that "the Stiefel case has no analytical solution" arose from this.

Weakened Version

The key to breaking the deadlock is to remove the equality constraint $\boldsymbol{W}^\top\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top\boldsymbol{W} = \boldsymbol{0}$ through some explicit representation. To this end, we first introduce a weakened version of the original problem: let $\boldsymbol{\Phi} = \boldsymbol{S}\boldsymbol{W}$, where $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ is any antisymmetric matrix satisfying $\|\boldsymbol{S}\|_2 \leq 1$. Then we consider solving

$$\max_{\boldsymbol{S}} \mathrm{tr}(\boldsymbol{G}^\top\boldsymbol{S}\boldsymbol{W}) \qquad \text{s.t.} \qquad \|\boldsymbol{S}\|_2 \leq 1 \tag{5}$$

The reason this is a weakening of the original problem is that the $\boldsymbol{\Phi}$ set up in this way obviously satisfies the two constraints of the original proposition:

$$1) \quad \|\boldsymbol{\Phi}\|_2 = \|\boldsymbol{S}\boldsymbol{W}\|_2 \leq \|\boldsymbol{S}\|_2\|\boldsymbol{W}\|_2 \leq 1 \tag{6}$$

$$2) \quad \boldsymbol{W}^\top\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top\boldsymbol{W} = \boldsymbol{W}^\top\boldsymbol{S}\boldsymbol{W} + \boldsymbol{W}^\top\boldsymbol{S}^\top\boldsymbol{W} = \boldsymbol{W}^\top(\boldsymbol{S} + \boldsymbol{S}^\top)\boldsymbol{W} = \boldsymbol{0} \tag{7}$$

So the new problem is solved on a subset of the feasible region of the original problem, and the maximum value obtained will not exceed the maximum value of the original problem. Furthermore, using the trace identity and the antisymmetry of $\boldsymbol{S}$, the objective function can be transformed into

$$\mathrm{tr}(\boldsymbol{G}^\top\boldsymbol{S}\boldsymbol{W}) = \mathrm{tr}(\boldsymbol{W}\boldsymbol{G}^\top\boldsymbol{S}) = -\mathrm{tr}(\boldsymbol{W}\boldsymbol{G}^\top\boldsymbol{S}^\top) = -\mathrm{tr}(\boldsymbol{G}\boldsymbol{W}^\top\boldsymbol{S}) = \mathrm{tr}([\boldsymbol{W}\boldsymbol{G}^\top]_{\text{skew}}\boldsymbol{S}) \tag{8}$$

where $[\boldsymbol{X}]_{\text{skew}} = (\boldsymbol{X} - \boldsymbol{X}^\top)/2$. So the new problem is equivalent to

$$\max_{\boldsymbol{S}} \mathrm{tr}([\boldsymbol{W}\boldsymbol{G}^\top]_{\text{skew}}\boldsymbol{S}) \qquad \text{s.t.} \qquad \|\boldsymbol{S}\|_2 \leq 1,\ \boldsymbol{S} + \boldsymbol{S}^\top = \boldsymbol{0} \tag{9}$$

This is already in the form of the standard Muon problem, and from the Muon result we can directly obtain

$$\boldsymbol{S} = \mathrm{msign}([\boldsymbol{G}\boldsymbol{W}^\top]_{\text{skew}}) \qquad \Rightarrow \qquad \boldsymbol{\Phi} = \mathrm{msign}([\boldsymbol{G}\boldsymbol{W}^\top]_{\text{skew}})\boldsymbol{W} \tag{10}$$

In fact, this is exactly the form of the solution for the square matrix case, and it is also the left rotation in *Steepest Descent on Manifolds: 6. Muon + Double Rotation*. One thing worth explaining here is, why do we apply Muon's msign only after a series of identity transformations? This is mainly to utilize the conclusion that "the msign of an antisymmetric matrix is still an antisymmetric matrix", thereby ensuring that the final result satisfies the antisymmetry requirement of $\boldsymbol{S}$.

Exact Equivalence

Next, we will prove that the weakened problem we solved in the previous section actually shares the exact same optimal solution as the original problem!

The proof idea is also very direct. The reason the weakened problem is "weakened" is that the current parameterization $\boldsymbol{\Phi} = \boldsymbol{S}\boldsymbol{W}$ is only a subset of the feasible region of the original problem, which might miss some exploration areas and thus fail to reach the optimal solution of the original problem. But if we can prove that for any feasible $\boldsymbol{\Phi}$ in the original problem, we can find an $\boldsymbol{S}$ satisfying the conditions such that $\boldsymbol{\Phi} = \boldsymbol{S}\boldsymbol{W}$, then it shows that their exploration spaces are exactly the same, and thus their optimal solutions are also identical.

The ensuing proof is carried out in three steps. (Note: In hindsight, given appropriate hints, K3 can also independently complete the derivation and proof, refer here and here.)

Solving the Equation

The first thing to do is, given $\boldsymbol{\Phi}, \boldsymbol{W}$, solve the equation

$$\boldsymbol{\Phi} = \boldsymbol{S}\boldsymbol{W}, \qquad \boldsymbol{S} + \boldsymbol{S}^\top = \boldsymbol{0} \tag{11}$$

Without loss of generality, considering only $n > m$, we can complete $\boldsymbol{W}$ into an $n \times n$ orthogonal matrix $\boldsymbol{U} = [\boldsymbol{W}, \boldsymbol{W}_\perp]$. Since $\boldsymbol{S}$ is antisymmetric, $\boldsymbol{U}^\top\boldsymbol{S}\boldsymbol{U}$ must also be, so we can write $\boldsymbol{S}$ as

$$\boldsymbol{S} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}\boldsymbol{U}^\top = [\boldsymbol{W}, \boldsymbol{W}_\perp]\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}\begin{bmatrix}\boldsymbol{W}^\top \\ \boldsymbol{W}_\perp^\top\end{bmatrix} \tag{12}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times m}, \boldsymbol{C} \in \mathbb{R}^{(n-m)\times(n-m)}$ are both antisymmetric matrices, and $\boldsymbol{B} \in \mathbb{R}^{(n-m)\times m}$. Then, right-multiplying both sides by $\boldsymbol{W}$ yields

$$\boldsymbol{\Phi} = \boldsymbol{S}\boldsymbol{W} = [\boldsymbol{W}, \boldsymbol{W}_\perp]\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_m \\ \boldsymbol{0}\end{bmatrix} = [\boldsymbol{W}, \boldsymbol{W}_\perp]\begin{bmatrix}\boldsymbol{A} \\ \boldsymbol{B}\end{bmatrix} \tag{13}$$

Left-multiplying both sides by $\boldsymbol{U}^\top$, we can directly read off $\boldsymbol{A} = \boldsymbol{W}^\top\boldsymbol{\Phi}, \boldsymbol{B} = \boldsymbol{W}_\perp^\top\boldsymbol{\Phi}$, meaning $\boldsymbol{A}, \boldsymbol{B}$ can both be uniquely solved, leaving $\boldsymbol{C}$ as the remaining degree of freedom.

Note that $\boldsymbol{A}$ is an antisymmetric matrix, which raises a compatibility condition

$$\boldsymbol{0} = \boldsymbol{A} + \boldsymbol{A}^\top = \boldsymbol{W}^\top\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top\boldsymbol{W} \tag{14}$$

which is exactly one of the conditions of the original problem. So far, all results are self-consistent.

Spectral Norm

The remaining question now is whether we can find an antisymmetric matrix $\boldsymbol{C}$ such that $\|\boldsymbol{S}\|_2 \leq 1$. The answer is affirmative, but the proof requires a "Parrott's Lemma" (the original paper cites Davis–Kahan–Weinberger, but research shows Parrott's lemma is earlier):

> Given a block matrix $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{B} & \boldsymbol{D}\end{bmatrix}$, fixing $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, we can always find a matrix $\boldsymbol{D}$ such that
>
> $$\left\|\begin{bmatrix}\boldsymbol{A} & \boldsymbol{C} \\ \boldsymbol{B} & \boldsymbol{D}\end{bmatrix}\right\|_2 = \max\left\{\left\|\begin{bmatrix}\boldsymbol{A} \\ \boldsymbol{B}\end{bmatrix}\right\|_2, \left\|\begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}\end{bmatrix}\right\|_2\right\} \tag{15}$$

Returning to our problem. To make $\|\boldsymbol{S}\|_2 \leq 1$, we only need $\left\|\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}\right\|_2 \leq 1$. For this block matrix, we have

$$\left\|\begin{bmatrix}\boldsymbol{A} \\ \boldsymbol{B}\end{bmatrix}\right\|_2 = \left\|\begin{bmatrix}\boldsymbol{W}^\top\boldsymbol{\Phi} \\ \boldsymbol{W}_\perp^\top\boldsymbol{\Phi}\end{bmatrix}\right\|_2 = \|\boldsymbol{U}^\top\boldsymbol{\Phi}\|_2 = \|\boldsymbol{\Phi}\|_2 \leq 1 \tag{16}$$

As for $[\boldsymbol{A}, -\boldsymbol{B}^\top] = [-\boldsymbol{A}^\top, -\boldsymbol{B}^\top]$, which is the negative transpose of $\begin{bmatrix}\boldsymbol{A} \\ \boldsymbol{B}\end{bmatrix}$, the spectral norm remains unchanged (and likewise does not exceed 1). Therefore, according to Parrott's lemma, we can find a matrix $\boldsymbol{C}$ such that $\|\boldsymbol{S}\|_2 = \left\|\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}\right\|_2 \leq 1$.

Antisymmetry

However, things are not over yet. We also have an antisymmetry requirement for $\boldsymbol{C}$, which Parrott's lemma does not guarantee.

Fortunately, this patch is not difficult. Let the matrix given by Parrott's lemma be $\boldsymbol{C}_0$, we directly take $\boldsymbol{C} = [\boldsymbol{C}_0]_{\text{skew}}$. Since the remaining parts inherently satisfy the antisymmetry requirement, we have

$$\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & [\boldsymbol{C}_0]_{\text{skew}}\end{bmatrix} = \frac{1}{2}\left(\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}_0\end{bmatrix} - \begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}_0\end{bmatrix}^\top\right) \tag{17}$$

The key point here is that antisymmetrization does not increase the spectral norm: by the triangle inequality, we get

$$\left\|\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & [\boldsymbol{C}_0]_{\text{skew}}\end{bmatrix}\right\|_2 \leq \frac{1}{2}\left\|\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}_0\end{bmatrix}\right\|_2 + \frac{1}{2}\left\|\begin{bmatrix}\boldsymbol{A} & -\boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C}_0\end{bmatrix}^\top\right\|_2 \leq \frac{1}{2} + \frac{1}{2} = 1 \tag{18}$$

So $[\boldsymbol{C}_0]_{\text{skew}}$ is the antisymmetric matrix we are ultimately looking for.

At this point, we have proved the result we initially hoped to prove: for every $\boldsymbol{\Phi}$ in the original problem, we can find an $n \times n$, antisymmetric matrix $\boldsymbol{S}$ with a spectral norm not exceeding 1, such that $\boldsymbol{\Phi} = \boldsymbol{S}\boldsymbol{W}$. Thus, the exploration space of the weakened version is identical to the exploration space of the original problem. In other words, Equation (10) is the exact solution to the original problem!

Efficient Computation

From a theoretical perspective, the problem has been perfectly resolved; but from a practical perspective, there is still an efficiency issue worth discussing. Muon requires performing msign on an $n \times m$ matrix, whereas the exact solution (10) on the Stiefel manifold requires performing msign on an $n \times n$ matrix. If computed directly, it is obviously much more expensive than Muon when $n \gg m$, so it is necessary to explore schemes that save computation.

Actually, it is not hard to imagine. First, we rewrite the core operation into a low-rank form:

$$\text{msign}([\boldsymbol{G}\boldsymbol{W}^\top]_{\text{skew}}) = \text{msign}(\boldsymbol{G}\boldsymbol{W}^\top - \boldsymbol{W}\boldsymbol{G}^\top) = \text{msign}\left(\begin{bmatrix}\boldsymbol{G} & \boldsymbol{W}\end{bmatrix}\underbrace{\begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}_m \\ -\boldsymbol{I}_m & \boldsymbol{0}\end{bmatrix}}_{\boldsymbol{J}}\begin{bmatrix}\boldsymbol{G}^\top \\ \boldsymbol{W}^\top\end{bmatrix}\right) \tag{19}$$

It can be seen that the rank of $[\boldsymbol{G}\boldsymbol{W}^\top]_{\text{skew}}$ is at most $2m$. Assuming $n \geq 2m$, we first perform a QR decomposition on $[\boldsymbol{G}, \boldsymbol{W}]$ to get $\boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q} \in \mathbb{R}^{n\times 2m}$ satisfies $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_{2m}$, and $\boldsymbol{R} \in \mathbb{R}^{2m\times 2m}$ is an upper triangular matrix. Then the part to be calculated becomes $\text{msign}(\boldsymbol{Q}\boldsymbol{R}\boldsymbol{J}\boldsymbol{R}^\top\boldsymbol{Q}^\top)$. Note that the msign operation has covariance with respect to orthogonal matrices, i.e.,

$$\text{msign}(\boldsymbol{Q}\boldsymbol{R}\boldsymbol{J}\boldsymbol{R}^\top\boldsymbol{Q}^\top) = \boldsymbol{Q}\,\text{msign}(\boldsymbol{R}\boldsymbol{J}\boldsymbol{R}^\top)\boldsymbol{Q}^\top \tag{20}$$

So in reality, we only need to perform msign once on the $2m \times 2m$ sized matrix $\boldsymbol{R}\boldsymbol{J}\boldsymbol{R}^\top$, which provides a significant speedup for scenarios where $n \gg 2m$. However, if n and m do not differ by more than an order of magnitude, it is more convenient to calculate using the original formula, since QR decomposition also incurs a cost.

In addition, how to switch to a momentum scenario is also worth discussing. We have two conceivable schemes: the first is $\boldsymbol{M} = \text{EMA}(\boldsymbol{G})$, in which case we only need to replace

$\boldsymbol{G}$ with $\boldsymbol{M}$, leaving the rest unchanged; the other is $\boldsymbol{M} = \text{EMA}(\boldsymbol{G}\boldsymbol{W}^\top)$, in which case $[\boldsymbol{G}\boldsymbol{W}^\top]_{\text{skew}}$ is replaced by $[\boldsymbol{M}]_{\text{skew}}$. This form can retain more information, at the cost of occupying more space. Moreover, after a long-term moving average, even if $n \gg 2m$, $\boldsymbol{M}$ typically will not be low-rank, and thus cannot be accelerated.

Of course, under mainstream model architectures, scenarios where $n \gg 2m$ are not common (only in the case of Per-Head Muon, the matrix for each Head barely qualifies), so the practical significance of this acceleration trick is limited, and we can just briefly understand it.

Open Problem

Here the author raises an open question: if we remove the orthogonality constraint and only consider steepest descent on the tangent manifold, what would its optimal solution be? That is, considering

$$\max_{\boldsymbol{\Phi}} \text{tr}(\boldsymbol{G}^\top\boldsymbol{\Phi}) \qquad \text{s.t.} \qquad \|\boldsymbol{\Phi}\|_2 \leq 1,\ \boldsymbol{W}^\top\boldsymbol{\Phi} + \boldsymbol{\Phi}^\top\boldsymbol{W} = \boldsymbol{0} \tag{21}$$

This problem can be viewed as a generalization of "Stiefel-version Muon" to general matrices; it does not depend on the orthogonality of $\boldsymbol{W}$ itself, and thus is expected to be applicable to more general scenarios. The result the author can currently obtain is that Equation (3) and Equation (4) are still applicable, meaning the optimal solution has the following form:

$$\boldsymbol{\Phi} = \text{msign}(\boldsymbol{G} + \boldsymbol{W}\boldsymbol{X}) \tag{22}$$

where $\boldsymbol{X} \in \mathbb{R}^{m\times m}$ is an antisymmetric matrix satisfying the following matrix equation:

$$\boldsymbol{W}^\top \text{msign}(\boldsymbol{G} + \boldsymbol{W}\boldsymbol{X}) + \text{msign}(\boldsymbol{G} + \boldsymbol{W}\boldsymbol{X})^\top\boldsymbol{W} = \boldsymbol{0} \tag{23}$$

As for how to solve it when $\boldsymbol{W}^\top\boldsymbol{W} \neq \boldsymbol{I}_m$, it remains unknown for the time being (assuming $n = m$ seems unable to simplify it either). Furthermore, the construction of the weakened problem also holds for it, requiring only a slight modification:

$$\boldsymbol{\Phi}_{\text{weak}} = \text{msign}([\boldsymbol{G}\boldsymbol{W}^\top]_{\text{skew}})\boldsymbol{W}/\|\boldsymbol{W}\|_2 \tag{24}$$

However, because the earlier equivalence proof strongly depends on $\boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I}_m$, it appears that this weak solution is genuinely just a weak solution, rather than the optimal solution. Everyone is welcome to provide more progress

Article Summary

In this article, we successfully found an analytical solution for the manifold steepest descent of "Muon + Stiefel", bringing this problem to a satisfactory conclusion. From now on, regardless of whether it is a square matrix or not, the Muon optimizer under orthogonal constraints can be calculated analytically, and solving it itself will no longer be a difficulty.