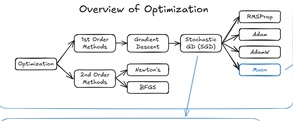
Optimization 101 and Muon

Biao Zhang
1zb.github.io

Overview of Optimization



Framework of Optimization

$$\min_w \{f(w)\}$$

Network · Parameters

$$w \leftarrow w + \lambda \cdot \Delta w$$

Parameters · Learning Rate · Direction

How to pick the best direction?

Gradient Descent

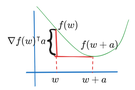
Before Update $f(w) \to f(w+a)$ After Update

Find update a, s.t. $f(w+a)$ is as small as possible

$$\min_a \{f(w+a)\} \approx \min_a \{f(w) + \nabla f(w)^\top a\}$$

(1st Order) Taylor Expansion

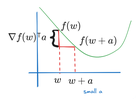
Another optimization problem · Can be ignored · Linear Change in Loss



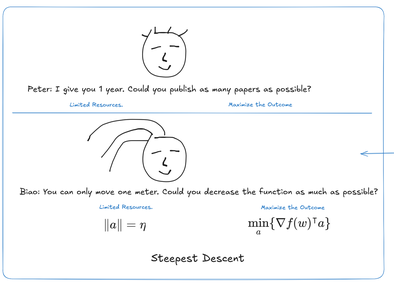
$$\min_a \{\nabla f(w)^\top a\}$$

$$\min_a \{\nabla f(w)^\top a : \|a\| = \eta\}$$

Size Constraint



Why?



Choices of Norms

Descent in Normed Spaces

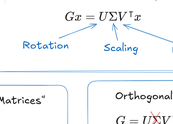
l_2 norm : $||x||_2 = \sqrt{\sum_{i=1}^N x_i^2}$

Solution

$$a^* = -\frac{\eta}{\|g\|_2} \cdot g$$

Negative · Gradient

$g = \nabla f(w)$ Gradient

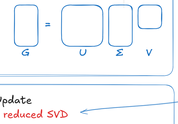


vanilla SGD

l_∞ norm : $||x||_\infty = \max_{i=1}^N |x_i|$

Solution

$$a^* = -\eta \cdot \text{sign}(g)$$



signSGD

Muon: From Vector to Matrix

Vanilla SGD
Vectorization

$$w = [W_1, W_2, W_3, W_4] \in \mathbb{R}^{(in\times d)+(d\times d)+(d\times d)+(d\times out)}$$

$$g = \nabla f(w)$$ Gradients

Muon
No Vectorization

$$G_1 = \nabla_{W_1} f(w)$$
$$G_2 = \nabla_{W_2} f(w)$$
$$G_3 = \nabla_{W_3} f(w)$$
$$G_4 = \nabla_{W_4} f(w)$$

Gradients with respect to each matrix

Vanilla SGD

$$\min_a \{g^\top a : \|a\| = \eta\}$$ Vector

Solution $\to$ $$a^* = -\frac{\eta}{\|g\|_2} \cdot g$$

Muon

$$\min_A \{\langle G, A\rangle_F : \|A\| = \eta\}$$ Matrix

Solution $\to$ $$A^* = -\eta \cdot UV^\top$$

$$\langle G, A\rangle_F = \text{trace}(G^\top A)$$

Trace = sum(⬔)

$$\|X\|_2 = \max_{x\neq 0} \frac{\|Xx\|_2}{\|x\|_2} = \sigma_{\max}(X)$$

Spectral Norm · Largest Singular Value

$$G = U\Sigma V^\top$$

Singular Value Decomposition (SVD) of Gradients

Geometric Meaning of SVD

$$Gx = U\Sigma V^\top x$$

Rotation · Scaling · Rotation

"Sign of Matrices"

$$\text{msign}(G) = UV^\top$$

Matrix Sign Function

$$\text{sign}(g) = \begin{cases} +1 & \text{if } g > 0 \\ 0 & \text{if } g = 0 \\ -1 & \text{if } g < 0 \end{cases}$$

Scalar Sign Function

Orthogonalizing the Update

$$G = U\Sigma V^\top \mapsto UV^\top$$

reduced SVD

Ignoring Scaling · Orthogonal

$$(UV^\top)^\top (UV^\top) = VU^\top UV^\top = VIV^\top = I$$

Why it is good?

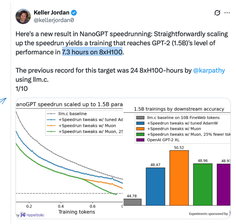
And for an empirically-flavored motivation, we observe that based on manual inspection, the updates produced by both SGD-momentum and Adam for the 2D parameters in transformer-based neural networks typically have very high condition number. That is, they are almost low-rank matrices, with the updates for all neurons being dominated by just a few directions. We speculate that orthogonalization effectively increases the scale of other "rare directions" which have small magnitude in the update but are nevertheless important for learning.

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

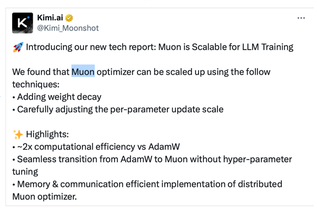
Approximation by Newton-Shultz Iterations

```python
# Pytorch code
def newtonschulz5(G, steps=5, eps=1e-7):
    assert G.ndim == 2
    a, b, c = (3.4445, -4.7750, 2.0315)
    X = G.bfloat16()
    X /= (X.norm() + eps)
    if G.size(0) > G.size(1):
        X = X.T
    for _ in range(steps):
        A = X @ X.T
        B = b * A + c * A @ A
        X = a * X + B @ X
    if G.size(0) > G.size(1):
        X = X.T
    return X
```

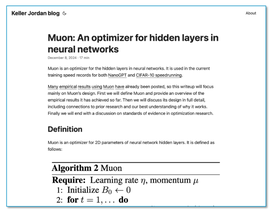
Author of Muon



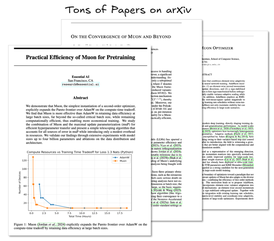
Training LLMs



Author's Blog



Tons of Papers on arXiv



Domain is Euclidean

$$f(w): \mathbb{R}^D \to \mathbb{R}^1$$

Update Direction

$$\min_a \{g^\top a : \|a\| = \eta\}$$

$$\min_w \{f(w) : w \in \mathbb{R}^D\}$$

$$a^* = -\frac{\eta}{\|g\|_2} \cdot g$$ Negative Gradients

What if the loss function has a different domain?

Domain is Hyperplane

$$f(w): H \to \mathbb{R}^1$$

Hyperplane $H = \{x \in \mathbb{R}^D : x^\top n = d\}$

$$\min_w \{f(w) : w \in H\}$$

Update Direction $\to$ $$\min_a \{g^\top a : \|a\| = \eta, a^\top n = 0\}$$

Subspace Constraint

$$w - \frac{\eta}{\|g\|_2} \cdot g \notin H$$

If w is on the hyperplane, then $w+a$ is also on the hyperplane.

$$a^* = -\frac{\eta}{\|g - nn^\top g\|_2} \cdot (g - nn^\top g)$$

$$w - \frac{\eta}{\|g - nn^\top g\|_2} \cdot (g - nn^\top g) \in H$$

Domain is Hypersphere

$$f(w): \mathcal{S}^D \to \mathbb{R}^1$$

Hypersphere $\mathcal{S}^D = \{x \in \mathbb{R}^{D+1} : \|x\| = 1\}$

$$\min_w \{f(w) : w \in \mathcal{S}^D\}$$

$$\min_a \{g^\top a : \|a\| = \eta, a^\top w = 0\}$$

Orthogonal (Tangent Space)

$$a^* = -\frac{\eta}{\|g - ww^\top g\|_2} \cdot (g - ww^\top g) \to \text{Retraction}(w, a^*) = \frac{1}{\sqrt{1+\eta^2}}(w + a^*$$

Retraction · Update

Manifold Optimization

Domain is Manifold

Optimization	Classical	Manifold
Direction	Euclidean	Tangent Space
Iteration	$w \leftarrow w + a^*$	$w \leftarrow \text{Retraction}(w + a^*$

Is there a "hypersphere" for matrices?

Muon

$$\min_A \{\langle G, A\rangle_F : \|A\| = \eta\}$$

$$A^* = -\eta \cdot UV^\top$$

	Spaces	Norm	Update
vanilla SGD	$\mathbb{R}^D$	L_2	Element-wise
Muon	$\mathbb{R}^{m\times n}$	Spectral	Matrix-wise

From Hypersphere to Orthogonal (Stiefel) Manifold

$\{x \in \mathbb{R}^D : x^\top x = 1\}$ — D × 1 — Vector — Hypersphere

$\{X \in \mathbb{R}^{N\times N} : X^\top X = XX^\top = I\}$ — N × N — Square Matrix — Orthogonal

$\{X \in \mathbb{R}^{M\times N} : X^\top X = I\}$ — M × N — Tall Matrix — Stiefel

Hypersphere

Domain (Manifold)

$$\{w \in \mathbb{R}^D : \|w\| = 1\}$$

Tangent Space

$$\mathcal{T}(w) = \{a : a^\top w = 0\}$$

$$\min_a \{g^\top a : \|a\| = \eta, a^\top w = 0\}$$

Tangent Constraint

Hypersphere SGD

Stiefel

Domain (Manifold)

$$\{W \in \mathbb{R}^{M\times N} : W^\top W = I\}$$

Tangent Space

$$\mathcal{T}(W) = \{A : A^\top W + W^\top A = 0\}$$

$$\min_A \{\langle G, A\rangle_F : \|A\| = \eta, A^\top W + W^\top A = 0\}$$

Tangent Constraint

Manifold Muon

Why?

From SGD to Muon → High Condition Number in Updates → Orthogonalizing the Updates

From Muon to Manifold Muon → High Condition Number in Weights → Orthogonalizing the Weights

Summary

https://kexue.fm/archives/11221

https://thinkingmachines.ai/blog/modular-manifolds/