From Adam to Muon: Effective 'Deep' Optimizers and their Preconditioner

Yufei Gu

August 3, 2025

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

xLeaf Lab
Machine Learning Foundations

Table of Contents

1. From Preconditioned gradient methods to Adam

Preconditioned gradient methods

Preconditioning maintains a matrix, termed a preconditioner, to transform the gradient vector before it is used to take a step [3].

- Motivation: In poorly conditioned problems (e.g., long narrow valleys), standard gradient descent zig-zags and converges slowly.
- Goal: Normalize or rescale the space so that optimization is more "isotropic", i.e., behaves similarly in all directions.

However, full-matrix preconditioning in machine learning prohibits out-of-the-box use due to dimensionality issues.

From Second-order Optimization to Adam

In second-order optimization (e.g. Newton's method), the update rule is usually:

$$\theta_{t+1} = \theta_t - H^{-1}\nabla_\theta \mathcal{L} \tag{1}$$

- The optimizer preconditioner is the inverse Hessian H^{-1}.
- However, in deep learning, computing H^{-1} is expensive ($O(N^2)/O(N^3)$).
- In convex/quadratic settings, the expected squared gradient correlates with diagonal entries of the Hessian:

$$\mathbb{E}[g_t^2] \approx \text{diag}(H) \tag{2}$$

- Adam preconditioner estimates curvature (H^{-1}) separately per parameter with the diagonal matrix $\frac{1}{\sqrt{\hat{v}_t}+\epsilon}$, avoids off-diagonal estimates [5].

Adam with Decoupled Weight Decay Regularization [ICLR 2019]

Initial parameters θ_0, learning rate η , weight decay λ, exponential decay rates $\beta_1, \beta_2 \in [0, 1)$, and mini-batch gradient $\bar{g}_t = \sum_{i=0}^{B} g_{t,i}$ at step t.

$$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\bar{g}_t, \qquad v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\bar{g}_t{}^2$$

$$\hat{m}_t \leftarrow \frac{m_t}{1 - \beta_1^t}, \qquad \hat{v}_t \leftarrow \frac{v_t}{1 - \beta_2^t}$$

$$\theta_t \leftarrow \theta_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_{t-1} \right)$$

AdamW improves Adam with Weight Decay [8]:

- Control parameter norms through regularization.
- Better than L2 Regularization (add $\lambda\theta_{t-1}$ to g_t): decouples the optimal choice of weight decay factor from learning rate scheme.

Neural Networks (maybe) evolved to make Adam the best optimizer

Adam is considered the King of optimization algorithms (in 2020, and maybe even at present), but if we review the Adam paper:

- The Adam theory was weak; An error was discovered in the proof that Adam will not converge on certain one-dimensional stochastic convex functions.
- The exact same experiments would result in a surefire rejection in these days.
- Knowing that Adam will not always give you the best performance, Adam is considered the default optimizer for deep learning.

Adam is extremely good on deep neural networks and very poor at anything else, e.g., simple convex and non-convex problems [1].

Neural Networks (maybe) evolved to make Adam the best optimizer

Consider the applied deep learning community:

- The community is like a giant genetic algorithm: researchers are exploring the space of all variants of algorithms and architectures in a semi-random way, consistent success is kept, and the others are discarded.
- Usually, people try new architectures, keeping the optimization algorithm fixed (most of the time Adam).

Hypothesis: People kept evolving new architectures on which Adam works. All the particular choices of architectures in deep learning might have evloved to make Adam work better and better.

2. Evolving with Lion: Leveraging the Sign Function

Symbolic Discovery of Optimization Algorithms [NIPS 2023]

Could a 'better' optimizer than AdamW be out there, overlooked or undiscovered?

Symbolic Discovery of algorithms (by Google):

1. Defines a train function that takes model weight, gradient, and lr at the current step, and outputs the update to the weight (with extra variables).
2. Include mutations: Insert / Delete / Modify a random statement with randomly chosen functions and arguments.
   - Infinite and sparse search space: The best program among 2M randomly sampled programs is still significantly inferior to AdamW.
   - Search cost is reduced by evolution warm-start and restart, pruning, and low-cost proxies. Total cost is of ˜3K TPU V2 days ($40 – 90K today).

Symbolic Discovery of Optimization Algorithms [NIPS 2023] [Google]

Lion (Evolved Sign Momentum):

$$u_t \leftarrow \text{sign}((1-\beta_1)g_t + \beta_1 m_{t-1})$$
$$m_t \leftarrow (1-\beta_2)g_t + \beta_2 m_{t-1}$$
$$\theta_t \leftarrow \theta_{t-1} - \eta\,(u_t + \lambda\theta_{t-1})$$

- Sign update produces uniform magnitude across all dimensions, and adds noise to the updates as regularization.
- Surpasses Adam and Adafactor for a variety of models on different tasks: Transformer, ResNet, U-Net, and Hybrid [2].

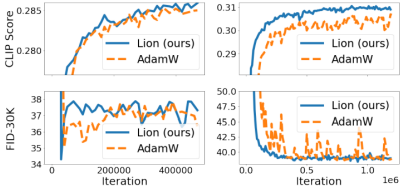


Figure 4: Imagen text-to-image 64^2 (**Left**) and the $64^2 \rightarrow 256^2$ diffusion models (**Right**).

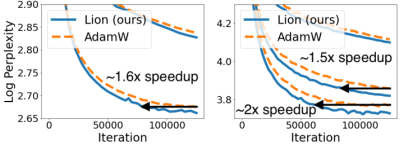


Figure 5: Log perplexity on Wiki-40B (**Left**) and PG-19 (**Right**). The speedup brought by Lion tends to increase with the model scale. The largest model on Wiki-40B is omitted as we observe severe overfitting.

Does Lion come without Cost? [Jianlin Su]

1. Lion does not suit small batch sizes (≤ 64): too much gradient noise.
2. Lion over-optimizes less-frequent token embedding weights:
   - $\text{sign}(m_t)$ being a constant;
   - Suggest fix: use AdamW instead.
3. Lion often produces a non-smooth loss curve in training.

Lion is still believed to have potential!

If $\beta_1 = \beta_2$ in Lion, we can save more running memory:

Tiger (Tight-first Optimizer): [11]

$$m_t \leftarrow (1-\beta) g_t + \beta m_{t-1}$$
$$\theta_t \leftarrow \theta_{t-1} - \eta \left(\text{sign}(m_t) + \lambda \theta_{t-1}\right)$$

Performance: Lion $\geq$ Tiger $\geq$ AdamW. [a]
Memory: Tiger $<$ Lion $<$ AdamW.

[a] Evaluated only on LLMs.

Cautious Optimizers: Improving Training with One Line of Code [arXiv]

Do not update unless the update direction is aligned with the current gradient [6].

$$
\begin{aligned}
\theta_t &\leftarrow \theta_{t-1} - \eta(u_t) \\
&\Downarrow \\
M_t &\leftarrow (u_t \times g_t > 0) \\
M_t &\leftarrow \frac{M_t \times |M_t|}{\sum M_t} \\
\theta_t &\leftarrow \theta_{t-1} - \eta(u_t \times M_t)
\end{aligned}
$$

- Cautious optimizer will not get stuck at non-stationary points of loss: momentum continues to accumulate and will eventually be updated.
- Cautious optimizer preserves the convergence guarantees of the base optimizer and is compatible to all momentum-based methods.

Cautious Optimizers: Improving Training with One Line of Code [arXiv]

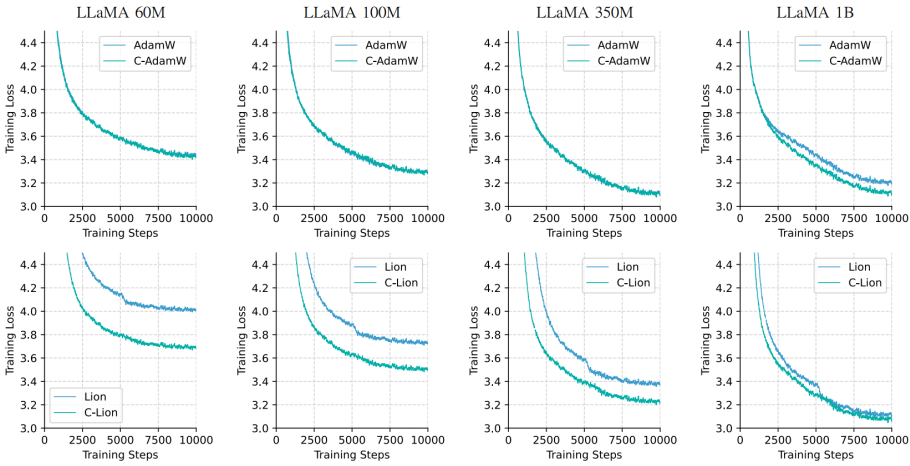


Figure 5: Training loss curves for AdamW, C-AdamW, Lion, C-Lion on LLaMA with 60M, 100M, 350M, and 1B parameters.

3. Adafactor & Shampoo: Structure-aware Preconditioner

Adafactor: Adaptive learning rates with sublinear memory cost [ICML 2018]

When storing the full momentum is infeasible, one choice for rank-1 approximation is to have $V = (V1_m)(\frac{1_n^T V}{1_n^T V 1_m})$ which is optimal w.r.t generalized KL divergence:

- $V1_m$ and $1_n^T V$ are the per-row and per-column sums of V.
- By maintaining only the moving average of per-row and per-column sums of squared gradients, the per-parameter second moments can be estimated [10]:

$$
\begin{aligned}
R_t &\leftarrow \beta_2 R_{t-1} + (1-\beta_2) g_t^2 1_m \\
C_t &\leftarrow \beta_2 C_{t-1} + (1-\beta_2) 1_n^T g_t^2 \\
\hat{V}_t &\leftarrow \frac{R_t C_t}{1_n^T R_t}, \quad \theta_t \leftarrow \theta_{t-1} - \eta_t \frac{g_t}{\sqrt{\hat{V}_t} + \epsilon}
\end{aligned}
$$

What if we use a full-rank approximation to the second-order momentum?

Shampoo: Preconditioned stochastic tensor optimization [ICML 2018]

Shampoo maintains a separate preconditioning matrix H_t^i corresponding to for each dimension $i \in [k]$ of the gradient [3].

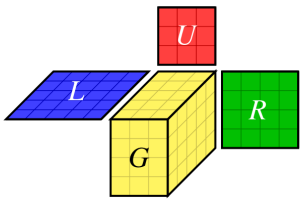


Figure 1. Illustration of Shampoo for a 3-dim tensor $G \in \mathbb{R}^{3\times4\times5}$.

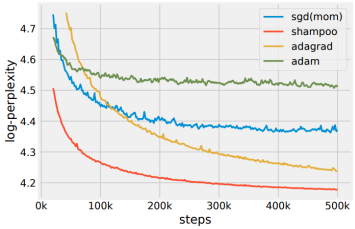


Figure 4. Convergence of test loss for the Transformer model for machine translation (Vaswani et al., 2017) on LM1B.

- The set of preconditioners is updated in an online fashion with the second-order statistics of the accumulated gradients.
- Adaptation to the high-order tensor case is non-trivial, e.g., 4d tensors in CNNs.

Shampoo: Preconditioned stochastic tensor optimization [ICML 2018]

- Shampoo, matrix case:

$$L_t \leftarrow L_{t-1} + g_t g_t^T$$
$$R_t \leftarrow R_{t-1} + g_t^T g_t$$
$$\theta_t \leftarrow \theta_{t-1} - \eta \cdot L_t^{-1/4} g_t R_t^{-1/4}$$

- Space Complexity: $O(\sum_{i=1}^{k} d_i^2)$.
- Compute Complexity: $O(\sum_{i=1}^{k} d_i^3)$.

- Shampoo, general tensor case:

for $i = 1, \ldots, k$ do:

$$H_t^i \leftarrow H_{t-1}^i + g_t^{(i)}$$
$$\theta_t \leftarrow \theta_{t-1} - \eta \cdot g_t \times_i (H_t^i)^{(-1/2k)},$$

- Step-size decay rate: $O(1/\sqrt{k})$.
- Overall regret bound: $O(\sqrt{T})$.

The gradient outer product approximates the Hessian: [1]

$$\mathbb{E}[g_t g_t^T] = \mathbb{E}[\nabla \log p \nabla \log p^T] \approx \mathbb{E}[\nabla_\theta^2 \log p] = \mathbb{E}[H_t(\mathcal{L})]$$

[1] Derived from Fisher information matrix under certain regularity conditions.

Stepping from Shampoo to Muon

What if we remove preconditioner accumulation in Shampoo?

$$\begin{aligned}
\theta_t &\leftarrow \theta_{t-1} - \eta (g_t g_t^T)^{-1/4} g_t (g_t^T g_t)^{-1/4} \\
&= \theta_{t-1} - \eta (US^2U^T)^{-1/4} (USV^T)(VS^2V^T)^{-1/4} \\
&= \theta_{t-1} - \eta (US^{-1/2}U^T)(USV^T)(VS^{-1/2}V^T) \\
&= \theta_{t-1} - \eta US^{-1/2} SS^{-1/2} V^T) \\
&= \theta_{t-1} - \eta UV^T
\end{aligned}$$

Given matrix-sign function $\text{sign}_M(A) = \text{sign}_M(U_A \Sigma_A V_A^T) = U_A V_A^T$ $(\Sigma_A^* = I)$,

$$\theta_t \leftarrow -\eta \underset{M}{\text{sign}}(g_t)$$

We have derived the Muon Optimizer (with no momentum)!

4. Muon: All Hail the New King?

Muon: MomentUm Orthogonalized by Newton-Schulz

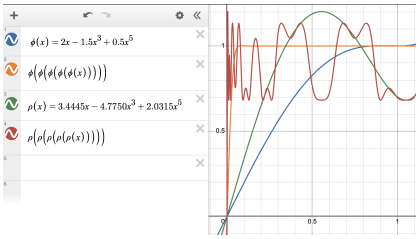
Given the Muon intuition derived from Shampoo with sgd-momentum, we have:

$$m_t \leftarrow \beta m_{t-1} + g_t$$
$$\theta_t \leftarrow \theta_{t-1} - \eta \underset{M}{\text{sign}}(m_t) = \theta_{t-1} - \eta(U_{m_t} V_{m_t}^T)$$

However, performing SVD decomposition on the momentum is costly. We need an effective derivation of orthogonal eigenvectors with eigenvalues close to 1 $\Sigma^* = I$.

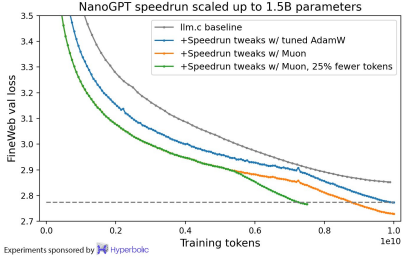
Newton-Schulz iteration: approximates the nearest semi-orthogonal matrix:

$$X_{t+1} = aX_t + bX_t(X_t^T X_t) + cX_t(X_t^T X_t)^2$$
$$\Rightarrow a = 3.4445, b = 4.7750, c = 2.0315$$



Muon: MomentUm Orthogonalized by Newton-Schulz

Orthogonalization mat effectively increases the scale of other "rare directions" which have small magnitude in the update but are nevertheless important for learning [4].



Empirical considerations:

- Muon only applies to 2D parameters; remaining scalar and vector parameters must be optimized using, e.g., AdamW.
- Empirically, input and output parameters optimized using AdamW attained best performance (e.g., embedding and lm_head).

With so many optimizers claimed to beat AdamW, now 'dead', can Muon be really adopted to the deep learning community this time?

Muon is Scalable for LLM Training [Moonshot AI] / [Essential AI]

Moonshot AI and Essential AI have declared to embrace Muon in 2025 [7, 9].

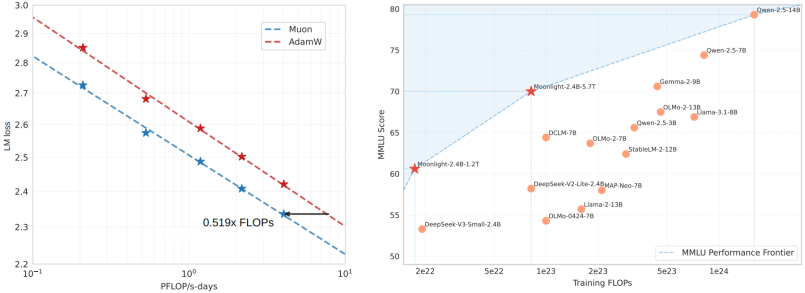
- Performance gains vs. training FLOPs on 3B/16B-MoE, 5.7T tokens (Moonshot AI) and 3.7B, 160B tokens (Essential AI).

Why focus on Muon?

1. Simplest derivation and implementation among second-order methods.
2. Lightest memory footprint by only maintaining the first momentum.
3. Moonshot AI proposed to rescale Muon's update RMS to align with AdamW, so learning rate and weight decay parameters can be reused:

$$
\begin{aligned}
m_t &\leftarrow \beta m_{t-1} + g_t \\
O_t &\leftarrow \textbf{Newton-Schulz}(m_t) \\
\theta_t &\leftarrow \theta_{t-1} - \eta(0.2\sqrt{\max(A, B)})O_t + \lambda\theta_{t-1})
\end{aligned}
$$

The End of History and the Last Optimizer?



(a) (b)

Figure 1: Scaling up with Muon. **(a)** Scaling law experiments comparing Muon and Adam. Muon is $\sim 2\times$ more computational efficient than Adam with compute optimal training. **(b)** The MMLU performance of our Moonlight model optimized with Muon and other comparable models. Moonlight advances the Pareto frontier of performance vs training FLOPs.

Thank you for your Attention!

References (1/2)

[1] Bremen. *Neural Networks (Maybe) Evolved to Make Adam the Best Optimizer*. 2020. URL: https://parameterfree.com/2020/12/06/neural-network-maybe-evolved-to-make-adam-the-best-optimizer/.

[2] Xiangning Chen et al. "Symbolic discovery of optimization algorithms". In: *Advances in neural information processing systems* 36 (2023), pp. 49205–49233.

[3] Diederik P Kingma. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[4] Kaizhao Liang et al. "Cautious optimizers: Improving training with one line of code". In: *arXiv preprint arXiv:2411.16085* (2024).

[5] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[6] Jianlin Shu. *Tiger: A Tight-first Optimizer*. Mar. 2023. URL: https://spaces.ac.cn/archives/9512.

References (2/2)

[1] Vineet Gupta, Tomer Koren, and Yoram Singer. "Shampoo: Preconditioned stochastic tensor optimization". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1842–1850.

[2] Keller Jordan et al. *Muon: An optimizer for hidden layers in neural networks*. 2024. URL: https://kellerjordan.github.io/posts/muon/.

[3] Jingyuan Liu et al. "Muon is scalable for LLM training". In: *arXiv preprint arXiv:2502.16982* (2025).

[4] Ishaan Shah et al. "Practical efficiency of muon for pretraining". In: *arXiv preprint arXiv:2505.02222* (2025).

[5] Noam Shazeer and Mitchell Stern. "Adafactor: Adaptive learning rates with sublinear memory cost". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4596–4604.

[6] Nikhil Vyas et al. "Soap: Improving and stabilizing shampoo using adam". In: *arXiv preprint arXiv:2409.11321* (2024).