

ASGO: Adaptive Structured Gradient Optimization

Kang An¹ Yuxing Liu² Rui Pan² Yi Ren³ Shiqian Ma¹ Donald Goldfarb⁴ Tong Zhang¹

¹Rice University

²University of Illinois Urbana Champaign

³Meta Inc.

⁴Columbia University

Introduction

- Optimization is a crucial part of deep learning applications, especially in training large foundation models. Currently, AdamW is the workhorse of deep learning optimizers, with Muon as a rising challenger.
- AdamW's Flaw.** It vectorizes all parameters, completely ignoring the rich matrix/tensor **structure** of DNN weights.

$$W_{t+1} = W_t - \eta_t \frac{\sum_{j=0}^t \beta_1^{t-j} (1 - \beta_1) G_t}{\sqrt{\sum_{j=0}^t \beta_2^{t-j} (1 - \beta_2) G_t^2}}, \quad \text{with coordinate-wise operations.}$$

- Muon's Flaw.** While it respects matrix structure, it lacks **adaptivity**. Also, the effectiveness of Muon in exploiting the matrix structure lacks proper illustration.

$$W_{t+1} \in \underset{X \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \left\{ \langle G_t, W - W_t \rangle + \frac{1}{2\eta_t} \|W - W_t\|_{\text{op}}^2 \right\}.$$

- Structural Properties.** DNNs consistently show clear structural properties:
 - Low-Rank Gradients: gradients are usually
 - Block-Diagonal Hessians

How to design an optimizer to effectively exploit the **structural properties**?

ASGO Algorithm

ASGO Algorithm.

Algorithm 1 ASGO (Adaptive Structured Gradient Optimization)

- Input:** $W_0 \in \mathbb{R}^{m \times n}$, $\epsilon \in \mathbb{R}$, schedule $\{\eta_t\}$ and batch size $M \in \mathbb{N}$
- Initialize $R_{-1} = 0 \in \mathbb{R}^{n \times n}$
- for** $t = 0$ **to** $T - 1$ **do**
- Sample mini-batch $\mathcal{B}_t$ with $|\mathcal{B}_t| = M$ uniformly
- $G_t = \frac{1}{M} \sum_{\xi \in \mathcal{B}_t} \nabla_W f(W_t; \xi)$
- $R_t = R_{t-1} + G_t^\top G_t$
- $W_{t+1} = W_t - \eta_t G_t (R_t + \epsilon I_n)^{-\frac{1}{2}}$
- end for**

Relations with Muon, SignSGD and Adam: An Analogy

Muon (Matrix): for $W_t \in \mathbb{R}^{m \times n}$

$$M_t = \beta M_{t-1} + (1 - \beta) G_t$$

$$W_{t+1} = W_t - \eta_t \text{MatrixSign}(M_t)$$

When $\beta = 0$
& no accumulation

ASGO (Matrix): for $W_t \in \mathbb{R}^{m \times n}$

$$M_t = \beta_1 M_{t-1} + (1 - \beta_1) G_t$$

$$V_t = \beta_2 V_{t-1} + (1 - \beta_2) G_t^\top G_t$$

$$W_{t+1} = W_t - \eta_t M_t V_t^{-1/2}$$

Analogy:
Matrix $\rightarrow$ Vector

SignSGD (Vector): for $w_t \in \mathbb{R}^d$

$$\mathbf{m}_t = \beta \mathbf{m}_{t-1} + (1 - \beta) \mathbf{g}_t$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \text{Sign}(\mathbf{m}_t)$$

When $\beta = 0$
& no accumulation

Adam (Vector): for $w_t \in \mathbb{R}^d$

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t$$

$$\mathbf{V}_t = \beta_2 \mathbf{V}_{t-1} + (1 - \beta_2) \text{diag}(\mathbf{g}_t \mathbf{g}_t^\top)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{V}_t^{-1/2} \mathbf{m}_t$$

Relations with Shampoo: One-sided Preconditioner.

Shampoo: for $W_t \in \mathbb{R}^d$

$$M_t = \beta M_{t-1} + (1 - \beta) G_t$$

$$L_t = \beta_2 L_{t-1} + (1 - \beta_2) G_t G_t^\top$$

$$R_t = \beta_2 R_{t-1} + (1 - \beta_2) G_t^\top G_t$$

$$W_{t+1} = W_t - \eta_t L_t^{-1/4} M_t R_t^{-1/4}$$

One Side
Simplification

ASGO: for $W_t \in \mathbb{R}^{m \times n}$

$$M_t = \beta_1 M_{t-1} + (1 - \beta_1) G_t$$

$$V_t = \beta_2 V_{t-1} + (1 - \beta_2) G_t^\top G_t$$

$$W_{t+1} = W_t - \eta_t M_t V_t^{-1/2}$$

Theoretical results

Convex Nonsmooth Convergence of ASGO

Assume $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is convex and W_* is an optimum, for ASGO, it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(W_t)] - f(W_*) \leq \frac{1}{T} \mathbb{E} \left[\left\| \left(\sum_{t=0}^{T-1} G_t^\top G_t \right)^{\frac{1}{2}} \right\|_* \right] \cdot D_{\text{op}} \sim \mathcal{O} \left(\frac{G_* D_{\text{op}}}{\sqrt{T}} \right),$$

where $D_{\text{op}} \triangleq \max_{t \in [T-1]} \|W_t - W_*\|_{\text{op}}$ and $G_* = \max_{t \in [T-1]} \|G_t\|_*$.

- For W with singular values $\{\sigma_i\}$, $\|W\|_{\text{op}}$ denotes the spectral norm σ_1 , $\|W\|_*$ denotes the trace norm $\sum \sigma_i$, and $\|W\|_F$ denotes the Frobenius norm $\sqrt{\sum \sigma_i^2}$.
- ASGO provably exploits gradient low-rankness:

$$\text{SGD: } \mathcal{O} \left(\frac{G_F D_F}{\sqrt{T}} \right) \quad \text{ASGO: } \mathcal{O} \left(\frac{G_* D_{\text{op}}}{\sqrt{T}} \right)$$

- $G_F \leq G_* \leq \sqrt{r_G} G_F$, where r_G is the highest rank of G_t . Thus when G_t are low-rank, or have very imbalanced singular values, G_* can be close to G_F ;
- $D_F / \sqrt{r_D} \leq D_{\text{op}} \leq D_F$, where r_D is the max rank of $W_t - W_*$. Thus, when $W_t - W_*$ are relatively high-rank, D_{op} can be much smaller than D_F .
- ASGO achieves at least D_F / D_{op} times faster convergence compared to Shampoo, while Muon can diverge under this nonsmooth setting [3].

Convex Smooth Convergence of ASGO

Assumption (Block-wise Diagonal Smoothness). $f : \mathbb{R}^{m \times n}$ is 1-smooth with respect to $\|\cdot\|_L$, where $L \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix and for any $W \in \mathbb{R}^{m \times n}$, $\|W\|_L^2 \triangleq \text{tr}(W L W^\top)$, which implies that

$$- \begin{bmatrix} L & & \\ & \cdots & \\ & & L \end{bmatrix} \preceq \nabla^2 f_v(w) \preceq \begin{bmatrix} L & & \\ & \cdots & \\ & & L \end{bmatrix} \in \mathbb{R}^{mn \times mn},$$

if we vectorize the variable, such that $f_v(w) = f(W)$ if $w = [W_1, \dots, W_m] \in \mathbb{R}^{mn}$.

Theorem. Assume the stochastic gradient noise $N_t \triangleq \nabla f(W_t; \xi) - \nabla f(W_t)$ be unbiased and has bounded covariance $\mathbb{E}[N_t^\top N_t] \preceq V^2 \in \mathbb{R}^{n \times n}$. For ASGO under convex and block-wise smooth assumptions with batch size M , it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(W_t)] - f(W_*) \leq \frac{4D_{\text{op}}^2 \|L\|_*}{T} + \frac{2\sqrt{2}D_{\text{op}} \|V\|_*}{\sqrt{MT}}.$$

- ASGO well fits the block-wise diagonal Hessian settings!

Practical Designs & Empirical Studies

Algorithm 2 A Practical Implementation of ASGO (for $W^\ell \in \mathbb{R}^{m \times n}$ such that $m \geq n$)

- Input:** $W_0 \in \mathbb{R}^{m \times n}$, lr schedule $\{\eta_t\}$, momentums $\beta_1, \beta_2 \in [0, 1]$, batch size $M \in \mathbb{N}$, $\epsilon \in \mathbb{R}$ (should be small, similar to the ϵ for Adam.)
- Initialize $M_{-1} = 0 \in \mathbb{R}^{m \times n}$, $V_{-1} = 0 \in \mathbb{R}^{n \times n}$ $\triangleright$ If $m < n$, set $V_{-1} = 0 \in \mathbb{R}^{m \times m}$
- for** $t = 0$ **to** $T - 1$ **do**
- Sample mini-batch $\mathcal{B}_t$ with $|\mathcal{B}_t| \equiv M$ uniformly
- $G_t = \frac{1}{M} \sum_{\xi \in \mathcal{B}_t} \nabla_W f(W_t; \xi)$
- $M_t = \beta_1 M_{t-1} + (1 - \beta_1) G_t$
- $V_t = \beta_2 V_{t-1} + (1 - \beta_2) G_t^\top G_t$ $\triangleright$ If $m < n$, $V_t = \beta_2 V_{t-1} + (1 - \beta_2) G_t G_t^\top$
- $\Lambda_t^{\text{inv}} = \text{SqrtInverseNewtonSchulz}(V_t)$ $\triangleright$ Run NS to inverse square root
- $W_{t+1} = (1 - \eta_t \lambda) W_t - \eta_t \frac{0.2\sqrt{mn} M_t \Lambda_t^{\text{inv}}}{\|M_t \Lambda_t^{\text{inv}}\|_F}$ $\triangleright$ If $m < n$, $W_{t+1} = W_t - \eta_t \frac{0.2\sqrt{mn} \Lambda_t^{\text{inv}} M_t}{\|\Lambda_t^{\text{inv}} M_t\|_F}$
- end for**

Practical Design 1: RMS Norm Alignment.

- Special parameters, like **Embedding layers** and all **1D params** (biases, norms), have different structural properties than ASGO is designed for.
- Similar to Muon, we use **ASGO** for the 2D weight matrices, and AdamW for 1D / embedding parameters. We align the update norms of two optimizers by line 9.

Practical Designs & Empirical Studies (Cont.)

Practical Design 2: Choosing the Preconditioning-Side.

- In theory, the preconditioning-side does not affect the convergence rate.
- In practice, for a neural network, the choice has a significant structural meaning.

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\mathbf{x}}$$

- Each **row** of $\mathbf{W}$ (e.g., $[w_{11}, \dots]$) contains all weights for a **single output neuron**.

The Ablation study on the preconditioning side:

Learning Rate	Validation Loss (Left Side)	Validation Loss (Right Side)
5e-4	3.582	3.569
1e-3	3.468	3.447
5e-3	3.378	3.434
1e-2	3.412	3.386

Practical Design 3: Computing the Inverse Square Root through Newton-Schulz.

- We adopt the iterative **Newton-Schultz (NS)** method to compute the matrix square root inverse instead of standard expensive methods like SVD.
- This is efficient because $A^{-1/2}$ is equivalent to a sub-block of the **matrix sign function** $\text{MatrixSign}(\cdot)$ of an augmented matrix:

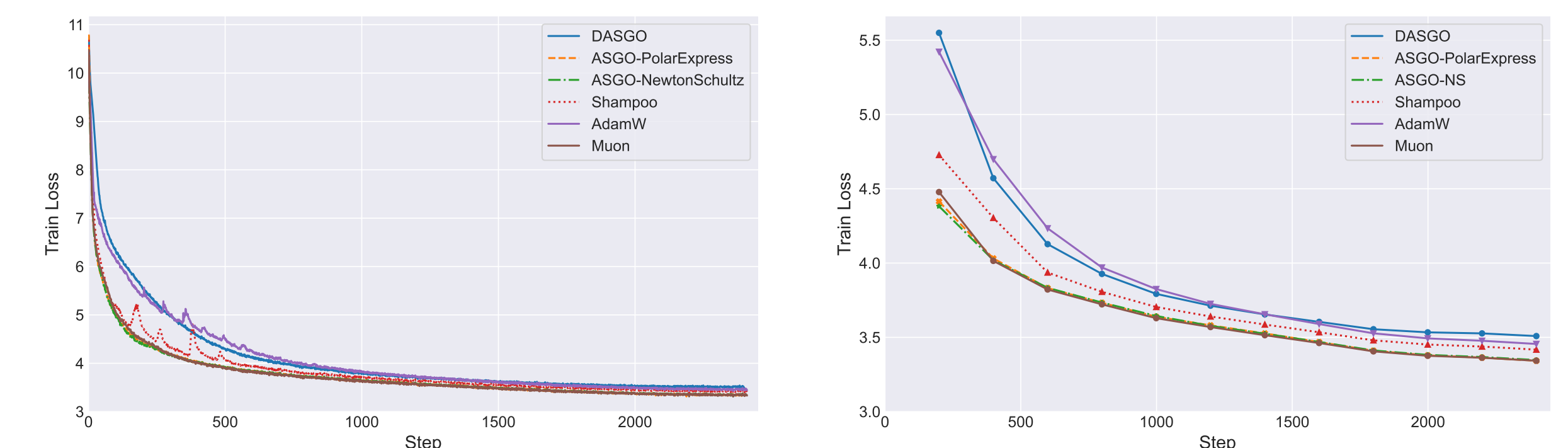
$$\text{MatrixSign} \left(\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix}$$

Wall-Clock Time (WCT) per Training Step.

Optimizer	Single Train Step WCT	Time Difference
AdamW	5.451 sec	0 sec
Muon	5.519 sec	0.068 sec
ASGO (Ours)	5.947 sec	0.468 sec

- The Big Win:** This efficiency allows ASGO to update the preconditioner **every step** (no "stale" updates from amortization).
- The Cost:** ASGO adds only 0.468 sec (an 8.6% increase) to the total step time compared to AdamW which the dominant computational bottleneck is the forward/backward pass.

Pretrain Experiments Results:



References

- Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.
- Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International conference on machine learning*, pages 3252–3261. PMLR, 2019.
- Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.