Stiefel-AdamW: Geometry-Aware AdamW for Linear Factorization Blocks

Marco Sutti

Postdoctoral fellow at GSSI

Joint work with Emanuele Zangrando and Francesco Tudisco

SIAM Conference on Optimization (OP26)

June 3, 2026

Overview

Talk based on:

- Stiefel-AdamW: Geometry-Aware AdamW for Linear Factorization Blocks, by E. Zangrando, M. S., and F. Tudisco, Tech. report, submitted.

Main contributions:

(i) Stiefel-AdamW, a variant of AdamW that carries geometric information.

(ii) Relaxation of the full $\mathrm{GL}(\mathbb{R}^r)$ quotient invariance to orthogonal invariance.

(iii) All adaptive moment estimation is performed in ambient Euclidean space, with geometric corrections applied only at the update step.

(iv) Theoretical guarantees (boundedness of gradients, regret analysis).

(v) Empirical validation of Stiefel-AdamW across a range of representative tasks: LoRA-style fine-tuning, full LLM pre-training.

I. Motivation & Proposed Method

A pervasive structure in DL: linear factorization blocks/1

Linear factorization blocks: trainable submodules of the form $W = BA$, with $B \in \mathbb{R}^{m \times r}$ and $A \in \mathbb{R}^{r \times n}$, and the rank $r \ll \min(m, n)$.

Examples:

- LoRA-style parameter-efficient fine-tuning (PEFT):

$$\mathcal{L}_t(\ W_0 + \Delta W\ ).$$

pre-trained weight matrix

weight correctors, $\Delta W = BA$

- Self-attention layers: the score matrix factorizes as $W_Q W_K^\top$.
- Analogous matrix-matrix factorizations appear in the recurrence kernels of several linear state-space models (SSMs).

LoRA-style PEFT: [Hu et al. 2022, Schotthöfer et al. 2025, Hayou et al. 2024], Low-rank pre-training: [Khodak et al. 2021], Network compression [Vogels/Karimireddy/Jaggi 2019, Saha/Srivastava/Pilanci 2023], Self-attention: [Vaswani et al. 2017]

A pervasive structure in DL: linear factorization blocks/2

- Gauge Symmetry: The map $\Phi(A, B) = BA$ is highly non-unique:

$$\forall M \in \mathrm{GL}(\mathbb{R}^r), \qquad \Phi(A, B) = \Phi(M^{-1}A, BM).$$

- Numerical Instability: e.g., $(M_n^{-1}A_n, B_nM_n)$ with $M_n \to 0$.
- Standard practice applies Euclidean optimizers (e.g., AdamW) directly to (A, B), entirely ignoring the underlying geometry.
- The Quotient Remedy: Optimizing on $\mathcal{M}_r = \mathbb{R}_*^{r\times n} \times \mathbb{R}_*^{m\times r}/\mathrm{GL}(\mathbb{R}^r)$ resolves instabilities but introduces a core "tension":
  - Full $\mathrm{GL}(\mathbb{R}^r)$ invariance conflicts with Adam's coordinate-wise diagonal preconditioning.

⚠ Existing methods must either sacrifice coordinate-wise adaptivity or heavily modify the gradient structure [Bécigneul/Ganea 2019, Bian et al. 2025].

Our Approach: Relaxed Quotient Invariance

- A New Trade-off: Instead of enforcing full quotient invariance, we relax it to orthogonal symmetry by restricting to the product manifold

$$\mathrm{St}(n,r) \times \mathbb{R}_*^{m\times r},$$

  i.e., $A^\top \in \mathrm{St}(n,r)$ ($AA^\top = I_r$), while $B \in \mathbb{R}_*^{m\times r}$ remains unconstrained.

$\rightsquigarrow$ Φ is invariant only under $Q \in \mathrm{O}(r)$:

$$\Phi(A,B) = \Phi(Q^\top A, BQ).$$

- Benefits:
  - Constraining one factor to $\mathrm{St}(n,r)$ makes the fibers of Φ compact, ruling out factor blow-up and stabilizing training.
  - Moment accumulation for both (A,B) factors happens in the ambient space.
  - Projecting onto the tangent space (the A factor) right before the update retains AdamW's coordinate-wise diagonal preconditioning.

Stiefel-AdamW

Algorithm 1: Single iteration of Stiefel-AdamW.

Require: A_t with $A_t A_t^\top = I_r$, B_t, M_{t-1}^A, M_{t-1}^B, V_{t-1}^A, V_{t-1}^B, η_t, β_1, β_2, ε.

$G_t^B \leftarrow \nabla_B \mathcal{L}(B_t A_t) A_t^\top$ ▷ Euclidean gradient w.r.t. B

$G_t^A \leftarrow B_t^\top \nabla_A \mathcal{L}(B_t A_t)$ ▷ Euclidean gradient w.r.t. A

$M_t^B \leftarrow \beta_1 M_{t-1}^B + (1-\beta_1) G_t^B$ ▷ First moment, B

$M_t^A \leftarrow \beta_1 M_{t-1}^A + (1-\beta_1) G_t^A$ ▷ First moment, A

$V_t^B \leftarrow \beta_2 V_{t-1}^B + (1-\beta_2)(G_t^B)^{\circ 2}$ ▷ Second moment, B

$V_t^A \leftarrow \beta_2 V_{t-1}^A + (1-\beta_2)(G_t^A)^{\circ 2}$ ▷ Second moment, A

$B_{t+1} \leftarrow B_t - \eta_t M_t^B / \left(\sqrt{V_t^B + \varepsilon}\right)$ ▷ Standard AdamW step on B

$X_t \leftarrow A_t^\top$ ▷ Column convention, $X_t \in \mathrm{St}(n, r)$

$D_t \leftarrow M_t^A / (\sqrt{V_t^A + \varepsilon})$ ▷ Preconditioned direction

$\xi_t \leftarrow -\eta_t \, \mathrm{P}_{X_t}(D_t)$ ▷ Project onto $\mathrm{T}_{X_t}\mathrm{St}(n, r)$

$X_{t+1} \leftarrow \mathrm{R}_{X_t}(\xi_t), \quad A_{t+1} \leftarrow X_{t+1}^\top$ ▷ Retract to $\mathrm{St}(n, r)$

II. Theoretical Guarantees

Regret analysis/1

Regret analysis: standard tool in online optimization.

- Quantifies how much a dynamic trajectory of iterates $\{W_t\}_{t=1}^T$ is suboptimal with respect to the best static choice in hindsight:

$$R(T) := \sum_{t=1}^{T} \mathcal{L}_t(W_t) - \min_{W} \sum_{t=1}^{T} \mathcal{L}_t(W).$$

- Stiefel-AdamW achieves zero average regret ($R(T)/T \to 0$ as $T \to +\infty$) up to a controlled learning rate and retraction error.

Core Theoretical Assumptions:

- Setup: Convex extensions $\mathcal{L}_t$ over the ambient space; standard alignment condition $\langle \xi_t, A^* - A_t \rangle \geqslant 0$ toward the global minimizer.
- Moments: Geometrically decaying first momentum ($\beta_{1,t} \to 0$) alongside an AMSGrad-style monotonic second momentum update.
- Boundedness: Iterates B_t and Euclidean gradients $\nabla \mathcal{L}_t$ remain bounded in the max-norm across iterations.

These assumptions adapt standard machinery used for proving Euclidean Adam convergence [Reddi/Kale/Kumar 2018].

Regret analysis/2

Theorem (Regret Bound)

Under our core assumptions with $\eta_t = \eta/\sqrt{t}$, the regret $R(T)$ satisfies a sub-linear bound:

$$R(T) \leqslant C_1 + C_2\sqrt{T} + C_3\sqrt{1+\log(T)} + C_4\log(T) + C_5 T^{-1/2},$$

where C_i are constants independent of T. In particular, $\lim_{T\to+\infty} \frac{R(T)}{T} = 0$.

Bridging Theory and Practice:

- Approximate Retractions: Our practical implementation use a fixed-point iteration method with an approximation error $\delta > 0$.
  - The proof naturally extends by adding a manageable $C_6\delta$ term to the bound without affecting the fundamental convergence result.
- The Max Update: While the proof requires an AMSGrad-like update, empirical testing shows that Stiefel-AdamW converges even without it.

Gradient Boundedness & Learning Rate Stability

- The Redundancy Issue: In unconstrained $W = BA$, the learning rate bound depends on the parameterization (singular values), not just the loss landscape.
- Example (Rank-r Recovery): Minimize $\mathcal{L} = \frac{1}{2}\|BA - \alpha I\|_{\mathrm{F}}^2$ with $B_0 = 0$.
  - Euclidean GD: Decouples to scalar recursion $\sigma_{i,k+1} = \sigma_{i,k}\left(2 - \eta\,\sigma_{i,k}^2\right)$. Diverges if $\sigma_{i,k} > \sqrt{3/\eta}$. Stable step-size depends on iterate scale.
  - Stiefel Constraint ($AA^\top = I_r$): Updates for B collapse to the affine recursion $B_{k+1} = (1-\eta)B_k + \eta\alpha A_k^\top$. Bounded for all $0 < \eta < 2$ since $\|A_k\| = 1$.

Proposition (Gradient Boundedness on Fibers)

Let $W = BA$ be a fixed matrix with $\operatorname{rank}(W) \leqslant r$ and $\nabla\mathcal{L}(W) \neq 0$. For the full unconstrained fiber $\mathcal{F} = \Phi^{-1}(W)$ and the **Stiefel-restricted fiber** $\widetilde{\mathcal{F}} = \widetilde{\Phi}^{-1}(W)$:

$$\|\nabla(\mathcal{L} \circ \Phi)\|_{L^\infty(\mathcal{F})} = +\infty, \qquad \|\nabla(\mathcal{L} \circ \widetilde{\Phi})\|_{L^\infty(\widetilde{\mathcal{F}})} < +\infty.$$

- **Takeaway**: Bounded gradients translate to wider stable learning rates.

III. Numerical experiments

Fine-tuning performance with low-rank adapters/1

GPT2 on the E2E NLU

Stiefel-AdamW for fine-tuning GPT2 on the E2E Natural Language Generation (NLU) challenge, with LoRA of rank = 4.

Method	**BLEU**	**NIST**	**MET**	**ROUGE-L**	**CIDEr**
AdamW [Loshchilov/Hutter, 2019]	68.90	8.69	46.50	71.30	2.51
Scaled AdamW [Zhang/Pilanci 2024]	**69.60**	8.77	46.60	**71.80**	2.52
Stiefel-AdamW	**69.60**	**8.79**	**46.70**	71.70	**2.53**
GeoLoRA [Schotthöfer et al. 2025]	67.70	8.50	46.20	70.80	2.40
LoRA-RITE [Yen et al. 2025]	67.80	8.55	46.20	70.80	2.40
LoRA-Pro [Wang et al., 2025]	67.75	8.54	46.23	70.77	2.43
Cayley Adam [Li/Li/Todorovic 2020]	67.75	8.55	46.23	70.77	2.40

The models have been trained for 5 epochs and a batch size of 8.
Stiefel-AdamW outperforms all baselines across all tasks except ROUGE-L.

Results for AdamW and Scaled AdamW are reported from [Zhang/Pilanci 2024].

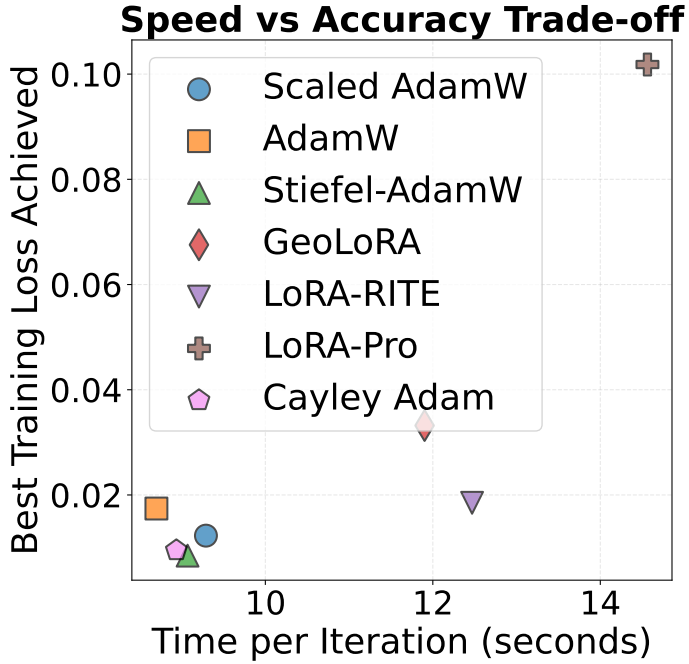
Fine-tuning performance with low-rank adapters/2

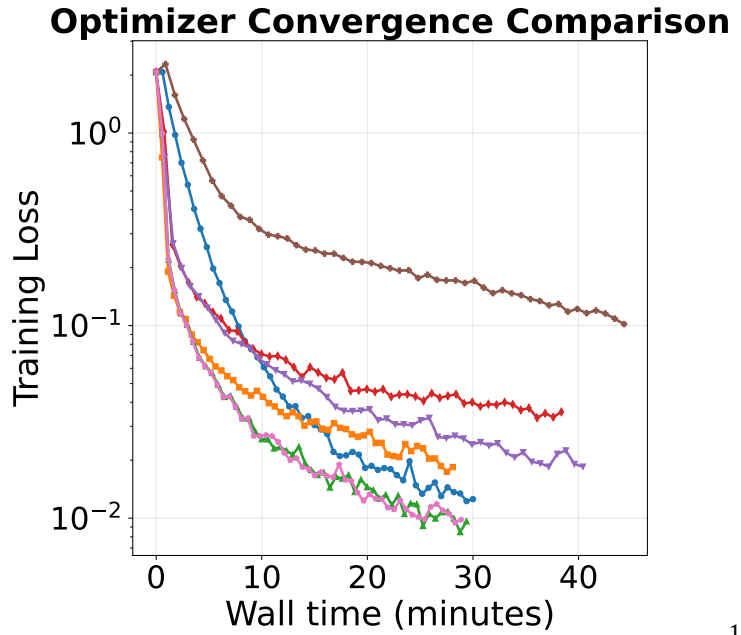
ViT Base on CIFAR-10 LoRA fine-tuning

F.-t. the base ViT of [Dosovitskiy et al., 2021] on CIFAR-10 [Krizhevsky/Hinton, 2009].

	Rank		
Method	**32**	**64**	**128**
AdamW [Loshchilov/Hutter, 2019]	95.78	95.94	96.10
Scaled AdamW [Zhang/Pilanci 2024]	92.25	94.71	95.60
Stiefel-AdamW	**96.19**	**96.25**	**96.45**
GeoLoRA [Schotthöfer et al. 2025]	96.02	95.86	95.87
LoRA-RITE [Yen et al. 2025]	94.95	95.06	95.15
LoRA-Pro [Wang et al., 2025]	95.91	95.76	95.57
Cayley Adam [Li/Li/Todorovic 2020]	96.01	96.13	96.39

Stiefel-AdamW outperforms all baselines, with convergence speed comparable to that of AdamW [Loshchilov/Hutter, 2019] and Scaled AdamW [Zhang/Pilanci 2024].





Fine-tuning performance with low-rank adapters/3

Mistral 7B on GLUE

Scores for rank-16 LoRA f.-t. of 4-bit quantized Mistral 7B LLM [Jiang et al., 2023] on the GLUE benchmark [Wang et al., 2019] for natural language understanding challenges with different optimizers. In parentheses, percentage deviation from the best performer.

Method	**GLUE**									
	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	WNLI	Avg.
SGD	88.15 (-4.14%)	96.10 (-1.18%)	70.10 (-22.07%)	55.89 (-22.23%)	94.22 (-1.21%)	88.59 (-3.94%)	50.90 (-44.27%)	47.64 (-48.37%)	49.30 (-43.54%)	71.21
Scaled GD [Zhang/Pilanci 2024]	90.21 (-1.89%)	96.90 (-0.36%)	81.62 (-9.26%)	68.17 (-5.14%)	94.40 (-1.01%)	91.15 (-1.16%)	54.15 (-40.72%)	90.31 (-2.13%)	56.34 (-35.48%)	80.36
AdamW	91.64 (-0.34%)	**97.25**	87.01 (-3.27%)	**71.87**	94.79 (-0.61%)	91.81 (-0.44%)	90.25 (-1.19%)	90.51 (-1.92%)	85.91 (-1.61%)	89.00
Scaled AdamW [Zhang/Pilanci 2024]	90.68 (-1.38%)	**97.25**	89.46 (-0.55%)	71.30 (-0.79%)	94.67 (-0.73%)	**92.22**	**91.34**	91.10 (-1.28%)	83.10 (-4.83%)	89.01
Stiefel-AdamW	**91.95**	96.79 (-0.47%)	**89.95**	70.61 (-1.75%)	94.78 (-0.62%)	91.83 (-0.42%)	90.61 (-0.80%)	**92.28**	**87.32**	**89.57**
GeoLoRA [Schotthöfer et al. 2025]	91.30 (-0.71%)	94.61 (-2.71%)	87.26 (-2.99%)	69.78 (-2.91%)	**95.37**	90.80 (-1.54%)	88.81 (-2.77%)	91.45 (-0.90%)	**87.32**	88.52

SGD, Scaled GD, and Scaled AdamW results reported from Table 2 in [Zhang/Pilanci 2024].

GPT2 Pre-training on OpenWebText

Beyond Fine-Tuning:

- Stiefel-AdamW can directly target low-rank structures arising natively during **pre-training**.

Self-Attention Parametrization:

- The map $(W_Q, W_K) \mapsto W_Q W_K^\top$ is highly non-injective.
- We restrict optimization to $\mathcal{M}_r \cong \mathrm{St}(n,r) \times \mathbb{R}_*^{n\times r}/\mathrm{O}(r)$.
- Biases and un-structured parameters are left to standard AdamW.

Experimental Setup:

- Pre-trained GPT-2 on OpenWebText.
- Both models trained for 6000 steps.

Results for p.-t. GPT2 [Radford et al. 2019] on OpenWebText [Gokaslan/Cohen, 2019] using Karpathy's reproduction.

Method	**Test Loss**	**Peak Memory (GB)**
AdamW	3.32	13.8
Stiefel-AdamW	**3.31**	**13.8**

- Stiefel-AdamW achieves comparable or slightly superior loss without any hyperparameter tuning (copied directly from AdamW configurations).
- Maintains the exact same peak GPU memory footprint.

Karpathy's reproduction: https://github.com/karpathy/nanoGPT.

Conclusions

Main contributions:

- **Stiefel-AdamW**: A geometry-aware variant of AdamW carrying natural convergence guarantees and preventing numerical instabilities.
- **Efficiency & Scalability**: Simple to implement with minimal Riemannian machinery; empirically validated from LLM fine-tuning to pre-training.
- **Limitation**: Entrywise nature leaves a retained orthogonal invariance; fixing this via gauge conditions on quotient space $\mathcal{M}_r$ breaks full parallelization.

Thank you for your attention!

Questions?

Bonus material

Stiefel-AdamW: Implementation & Versatility

- Structural Modification: A near drop-in replacement for AdamW:
  - *Euclidean factor*: Updated via standard AdamW.
  - *Stiefel factor*: Moments computed in ambient space, projected onto the tangent space, and retracted (via QR, Cayley, Polar, Newton–Schulz).
  - Stiefel factor:
    - First and second moments are computed in the ambient space
    - The preconditioned direction is projected onto the tangent space
    - Retract back to the manifold. The retraction step is treated as a modular component: any efficient retraction on the Stiefel manifold can be used, including the Cayley transform, QR-based retraction, polar decomposition, or Newton–Schulz iteration
- Key Advantages: Inherits the core benefits of geometric optimization for free: numerical stability, memory efficiency, and provable convergence.
- Universal Application: Not limited to LoRA. It applies to *any* matrix factorization block $W = BA$ without an intervening nonlinearity.
- Use Cases: Successfully handles LoRA adapters, low-rank compressed layers, and query/key attention projections.

$\rightsquigarrow$ Principal benefits of geometric optimization: numerical stability, parameters invariance to orthogonal reparametrization, low memory footprint, and provable convergence under standard assumptions.

A Closer Look to Stiefel-AdamW

- Ambient Moment Accumulation: Compute Euclidean gradients G_t^B, G_t^A and update first (M_t) and second (V_t) moments identically for both factors in the ambient space:

$$\begin{cases} M_t^{A,B} = \beta_1 M_{t-1}^{A,B} + (1-\beta_1)\, G_t^{A,B}, \\ V_t^{A,B} = \beta_2 V_{t-1}^{A,B} + (1-\beta_2)\,(G_t^{A,B})^{\circ 2}. \end{cases}$$

- Unconstrained Factor Update (B_t): Performed via standard, purely Euclidean AdamW steps:

$$B_{t+1} = B_t - \eta_t \cdot M_t^B / \left(\sqrt{V_t^B} + \varepsilon \right).$$

- Orthonormal Factor Update (A_t): Form the Euclidean adaptive direction $D_t = M_t^A / (\sqrt{V_t^A} + \varepsilon)$, then map it to the manifold:
  - $\rightsquigarrow$ Tangent Projection: Project D_t onto $\mathrm{T}_{X_t}\mathrm{St}(n,r)$ to get the Riemannian direction:

$$\xi_t = -\eta_t \, \mathrm{P}_{X_t}(D_t).$$

  - $\rightsquigarrow$ Retraction: Apply a retraction to map ξ_t back onto $\mathrm{St}(n,r)$.

Orthogonal Projection onto $\mathrm{T}_X\mathrm{St}(n,r)$

The Projection Formula

For any ambient vector $\xi \in \mathbb{R}^{n\times r}$, its orthogonal projection onto the tangent space $\mathrm{T}_X\mathrm{St}(n,r)$ with respect to the Euclidean inner product is defined as:

$$\mathrm{P}_X(\xi) = \underbrace{X\,\mathrm{skew}(X^\top \xi)}_{\text{Tangential component}} + \underbrace{(I_n - XX^\top)\,\xi}_{\text{Normal complement}},$$

$\mathrm{skew}(M) := \frac{1}{2}(M - M^\top)$ being the skew-symmetric part of a square matrix M.

Geometry Behind the Math:

- Rotational/Tangential Component ($X\,\mathrm{skew}(X^\top\xi)$): Accounts for movements *along* the manifold fiber. It takes the interaction matrix $X^\top\xi$ and projects it directly onto the Lie algebra of the orthogonal group.
- Normal Subspace Complement ($(I_n - XX^\top)\,\xi$): Acts as a standard orthogonal projection onto the subspace perpendicular to the column span of X, keeping updates strictly tied to valid tangential directions.

↩ Back to algorithm pseudocode.

Choice of Retraction

- Definition: Maps a tangent vector back onto the manifold:

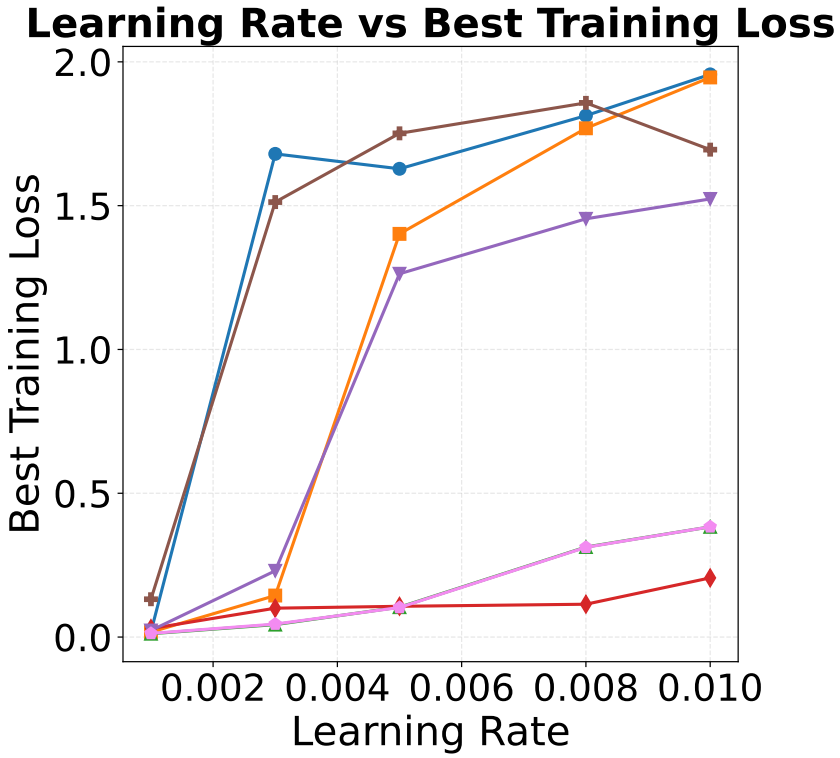
$$X_{t+1} = \mathrm{R}_{X_t}(\xi_t).$$

  Replaces the computationally prohibitive exponential map with a (typically first-order) approximation.
- Efficiency: For $\mathrm{St}(n, r)$, efficient choices cost only $\mathcal{O}(nr^2 + r^3)$, making them highly scalable when $r \ll n$:
  - *QR Decomposition*: Extracts Q from a standard QR factorization.
  - *Polar Decomposition*: Finds the closest orthogonal matrix via Newton–Schulz.
  - *Cayley Transform*: Avoids large inversions using the Sherman–Morrison–Woodbury (SMW) identity or fixed-point iterations.
- Our Default: Cayley retraction approximated via fixed-point iteration, selected for its implementation simplicity and strong empirical results.
- Robustness to Approximation: Our theorem proves that a controlled Frobenius error δ only adds a linear term to the regret bound.

↩ Back to algorithm pseudocode.

Stepsize stability

Motivated by the proposition ↩ gradient boundedness on fibers, we numerically demonstrate the stability of Stiefel-AdamW with respect to the learning rate.

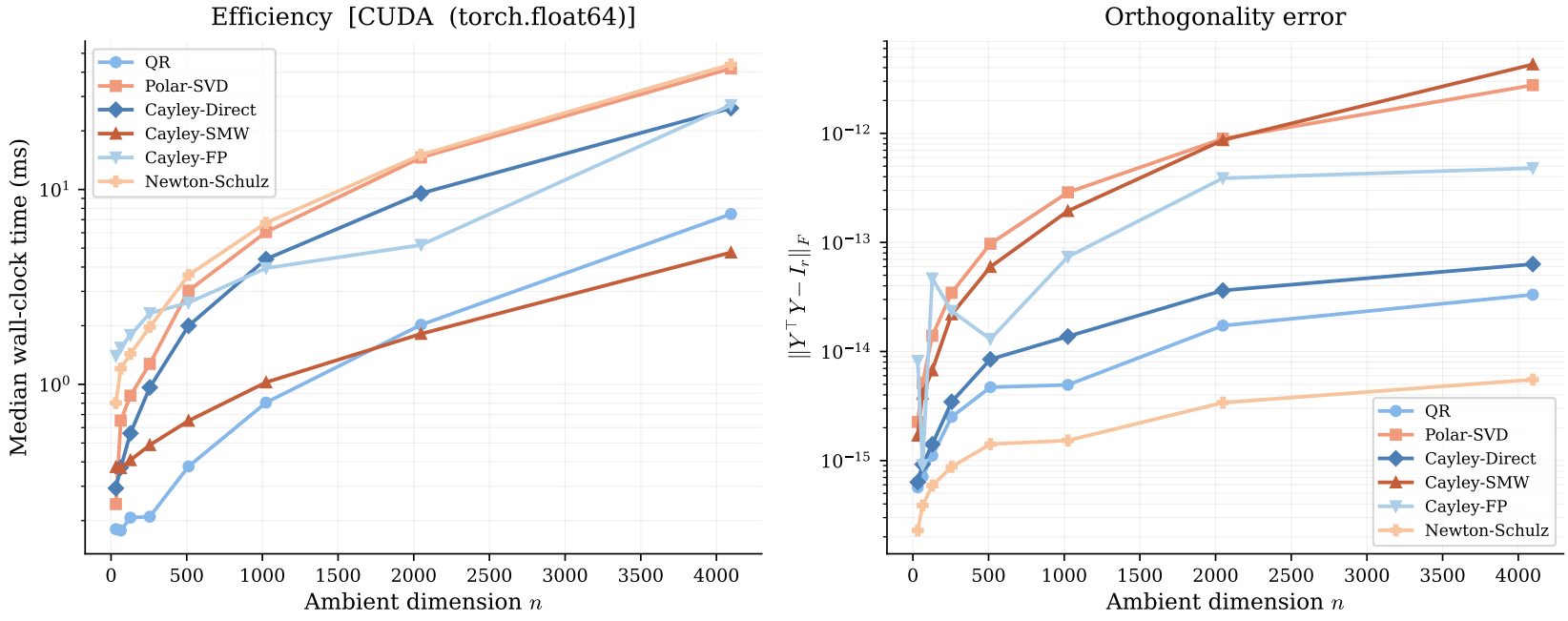


"Pure" Riemannian methods such as GeoLoRA [Schotthöfer et al. 2025] and the proposed Stiefel-AdamW are more stable with respect to the learning rate magnitude.

Methods in which the fiber of Φ is not compact (such as AdamW and Scaled AdamW, which are defined in $\mathbb{R}^{r\times n} \times \mathbb{R}^{m\times r}$), appear to be less stable with respect to larger learning rates, despite the preconditioning employed in Scaled AdamW.

Ablation over different retractions

Comparison of several possible choices of retraction on the Stiefel manifold. Matrix size against GPU wall-clock time and orthogonality error.



Retraction	**Test acc.** ($r = 32$)	**Test acc.** ($r = 64$)	**Test acc.** ($r = 128$)
Cayley FP	96.19	96.25	96.45
Cayley SMW	96.11	96.09	96.52
Cayley direct	96.11	96.22	96.61
QR	95.41	94.83	93.71
Polar	96.11	96.22	96.61
Newton–Schulz	96.11	96.2	96.52