NEURAL INFORMATION PROCESSING SYSTEMS

Understanding Adam Requires Better Rotation Dependent Assumptions

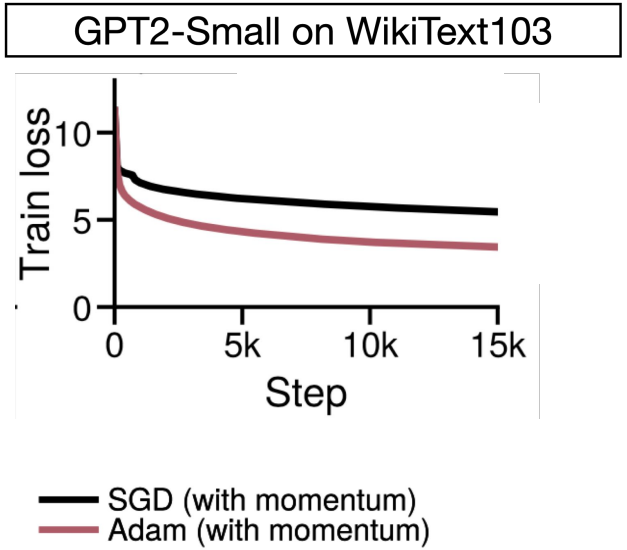
NeurIPS 2025

Tianyue H. Zhang*, Lucas Maes*, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, Charles Guille-Escuret

Mila Université de Montréal UBC SAMSUNG

TL;DR

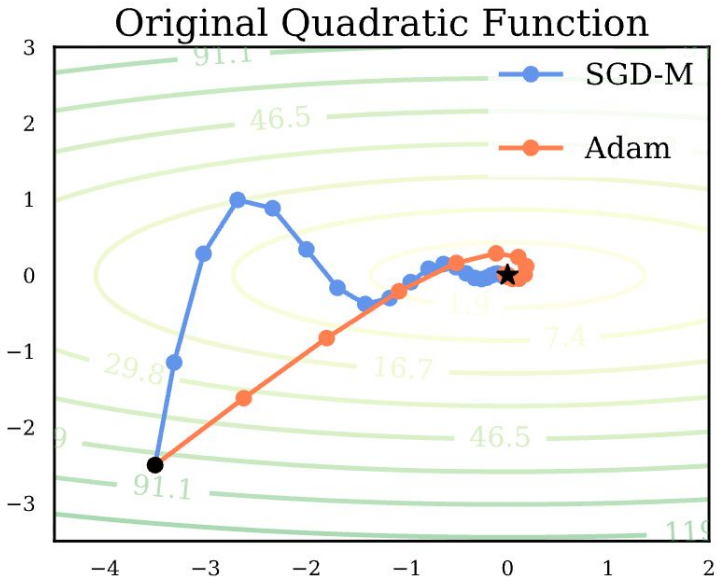
- We quantify how rotation on network parameters impact Adam's performance;
- We show that most theoretical assumptions do not capture this dependency well.

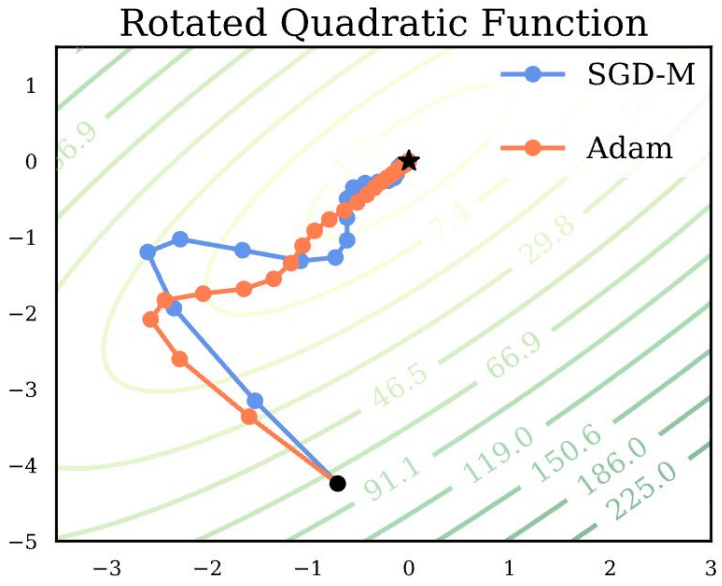


[Kunstner et al., 2025]

Rotation (non-)Equivariance

- SGD(M) maintains the same trajectory up to rotation (f(Rw) = Rf(w))
- Adam is not rotation equivariant, due to its element-wise division

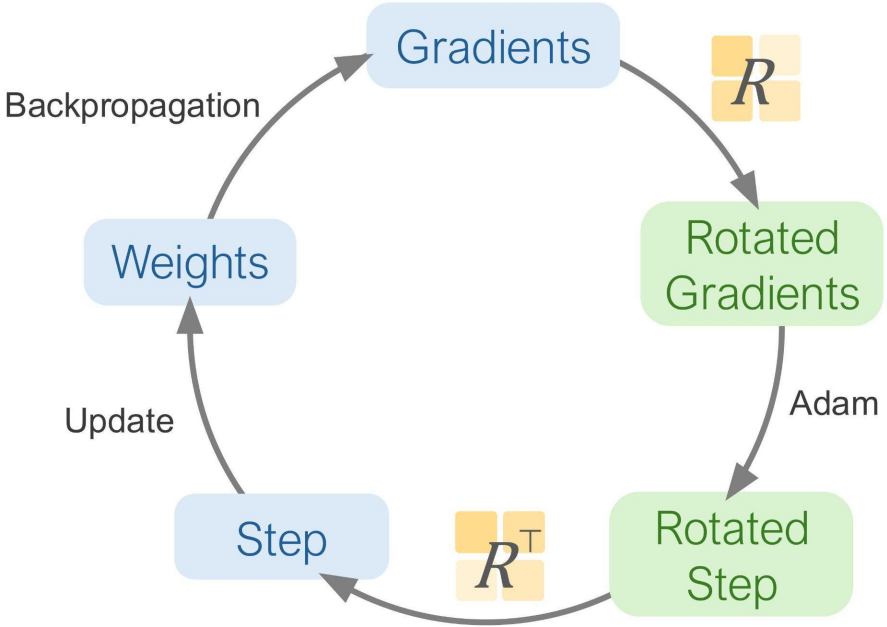




Q1. How do various types of rotations influence Adam’s performance in practice?

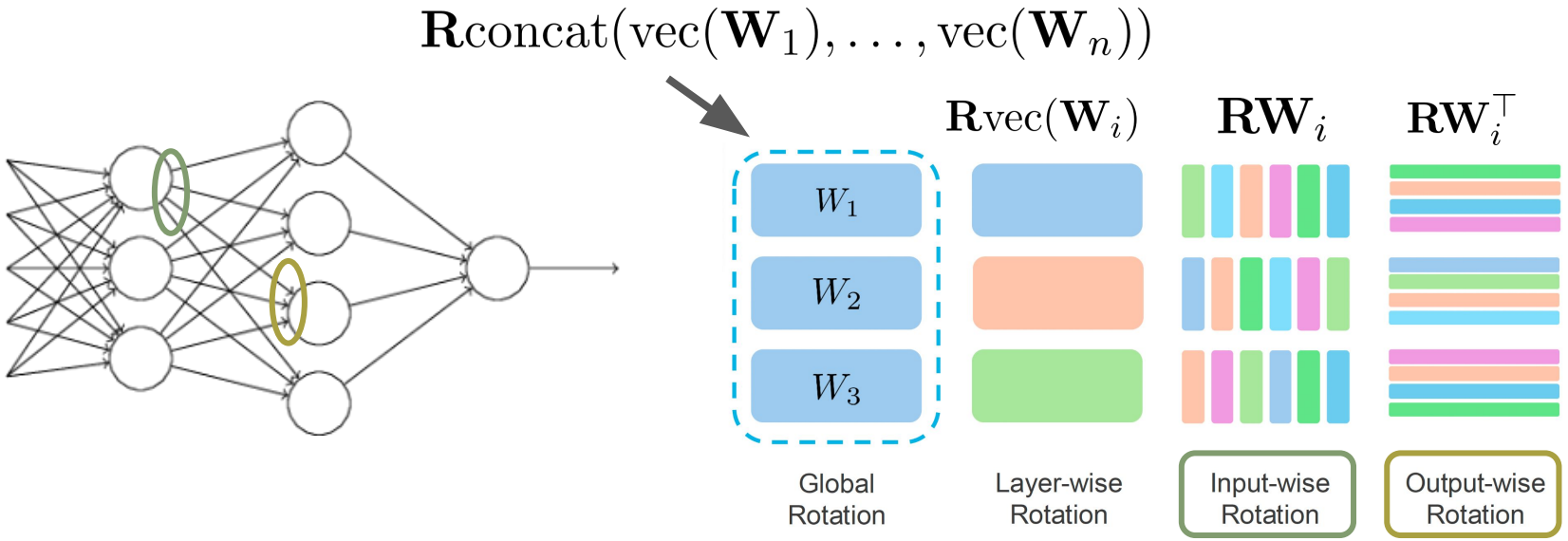
Q2. What rotation-dependent assumptions adequately capture Adam’s behavior under rotations?

Training in Rotated Parameter Spaces

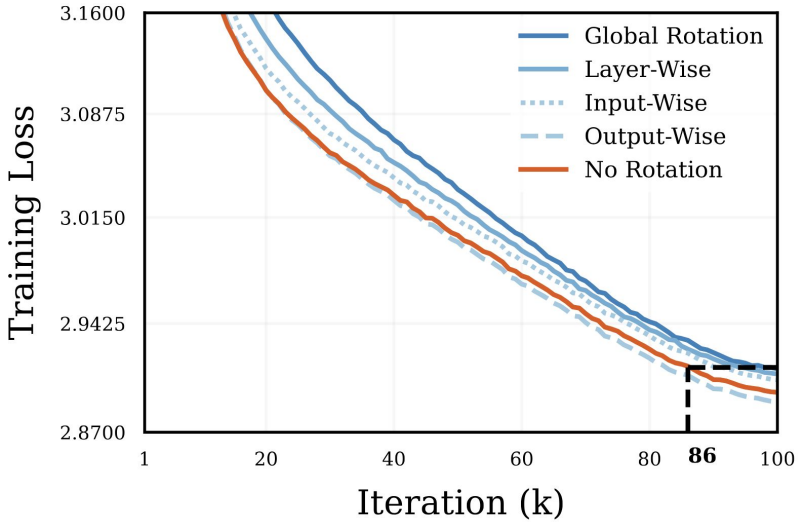


Rotation Scopes

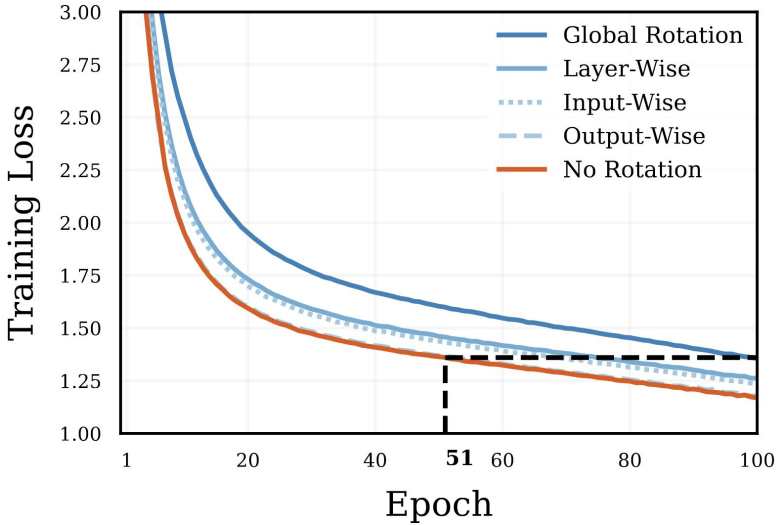
We want to examine the impact of various scopes of rotations on Adam's performance



Random Rotation Degrades Performance



(a) GPT2 (124M)



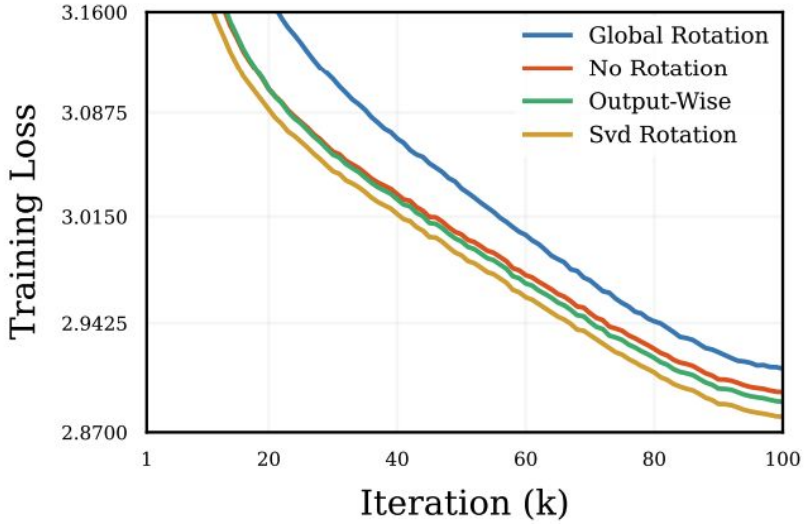
(b) ViT/S (22M)

Structured Rotation Improves Performance

We can also use a structured rotation, such as singular vectors

Rotate the gradient with its SVD

$$
\begin{aligned}
&\textbf{if } t \mod F = 0 \textbf{ then} \\
&\quad \mathbf{U}, \mathbf{S}, \mathbf{V}^\top \leftarrow \mathrm{SVD}(\mathbf{g}_t) \\
&\textbf{end if} \\
&\tilde{\mathbf{g}}_t = \mathbf{U}^\top \mathbf{g}_t \mathbf{V}
\end{aligned}
$$



Rotate update back

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha \mathbf{U}(\hat{\mathbf{m}}_t / (\sqrt{\hat{\mathbf{v}}_t} + \epsilon))\mathbf{V}^\top - \alpha\lambda\boldsymbol{\theta}_{t-1}$$

Q1. How do various types of rotations influence Adam’s performance in practice?

Q2. What rotation-dependent assumptions adequately capture Adam’s behavior under rotations?

Assumptions in Theory

Theoretical analysis uses assumptions to describe characteristics of functions, which can be used to prove performance guarantees of algorithms.

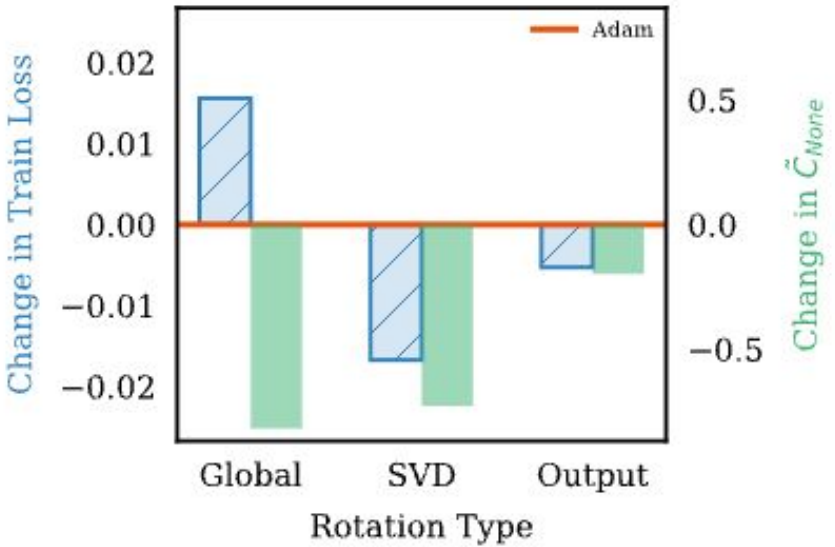
Theorem 3.1. *Suppose* G *is* L*-smooth and that* $\eta = \frac{1}{L}$*, then*

$$G(w_t) - G(w_*) \leq \frac{L\|w_0 - w_*\|^2}{t}$$

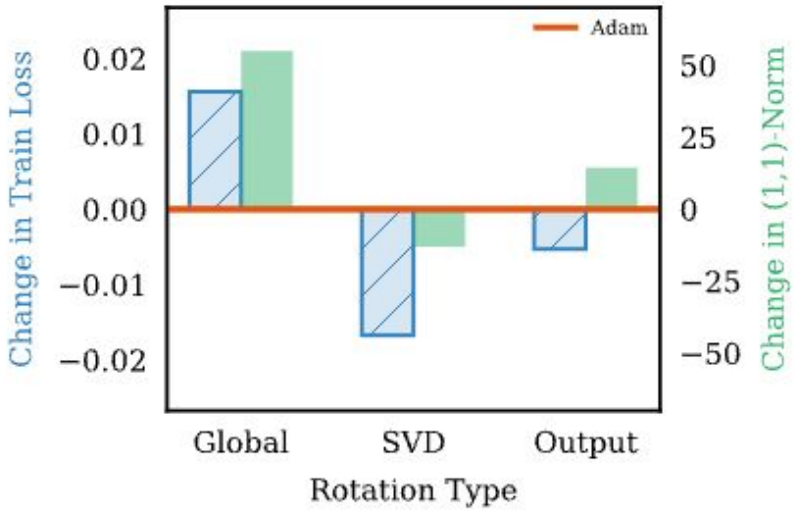
- Many common assumptions, like Lipschitz smoothness in euclidean norm, are rotation invariant
- We examine a few rotation dependent assumptions

L_∞ Bounded Gradient; L_∞ Smoothness

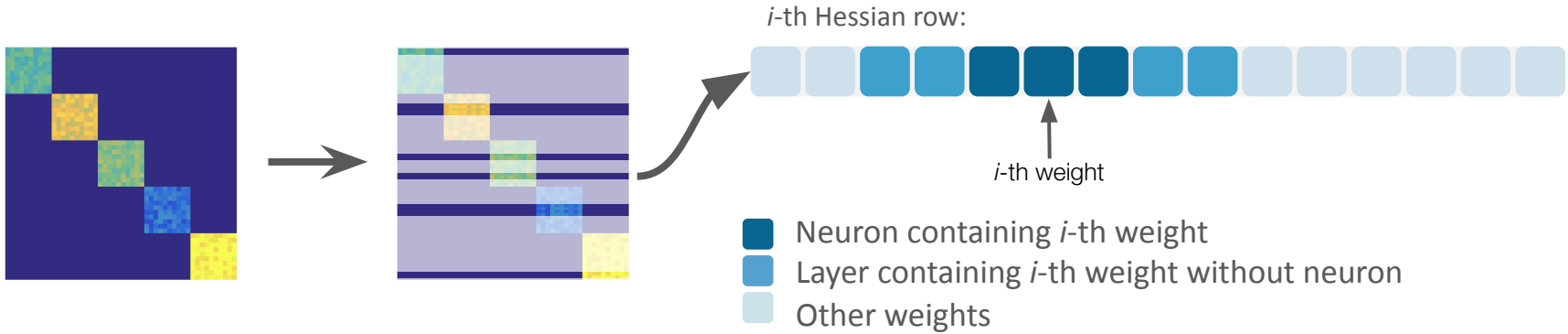
$$\|\nabla f_B(\mathbf{w})\|_\infty \leq C$$



$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_1 \leq C\|\mathbf{x} - \mathbf{y}\|_\infty$$



Block Diagonal Hessian



We found that Hessian values outside the block diagonal are the primary drivers of gradient evolution, despite their smaller magnitude

Update (Semi-)orthogonality

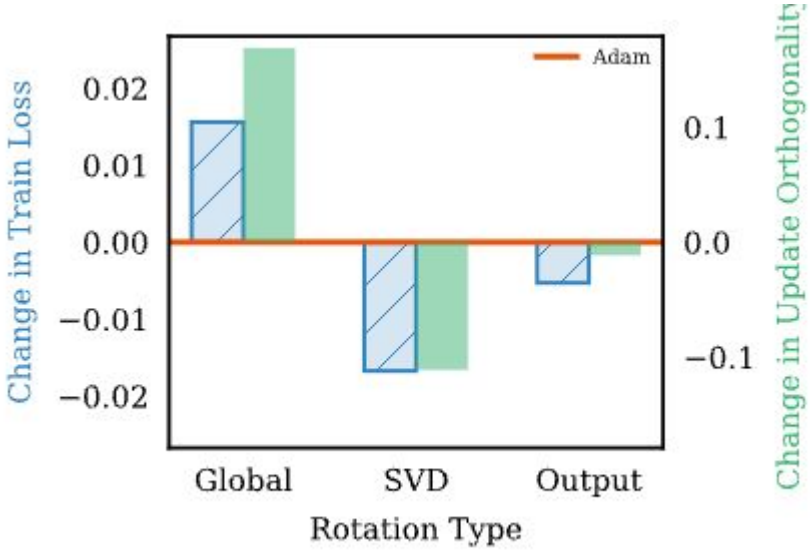
Recent optimizers (Muon, Shampoo), “orthogonalize” update matrices.

Compute the Adam Update (ignoring weight decay)

$$\frac{\mathbf{W}_{t+1} - (1 - \alpha_t \lambda)\mathbf{W}_t}{-\alpha_t} = \mathbf{R}^\top \frac{\mathbf{M}_t}{\sqrt{\mathbf{V}_t} + \epsilon}$$

Measure the singular value variation

$$\mathrm{CV}(s_i) = \frac{\sigma_s}{\mu_s} :$$



Summary

- Adam is sensitive to parameter space rotations in practical training:
  - Broader scope of random rotation leads to greater performance degradation
  - Structured SVD-based rotation improves performance
- Existing theoretical assumptions are not properly equipped to understand the beneficial properties of Adam;
- (Semi-)orthogonality of layer updates is a strong predictor of Adam's performance