Exploiting Weight-Space Symmetries for Approximating Curvature

Artem Artemev, Rui Xia, Benjamin M. Boyd, Youjing Yu, Felix Dangel,
Guillaume Hennequin, Alberto Bernacchia

June 2026

Curvature is useful but expensive

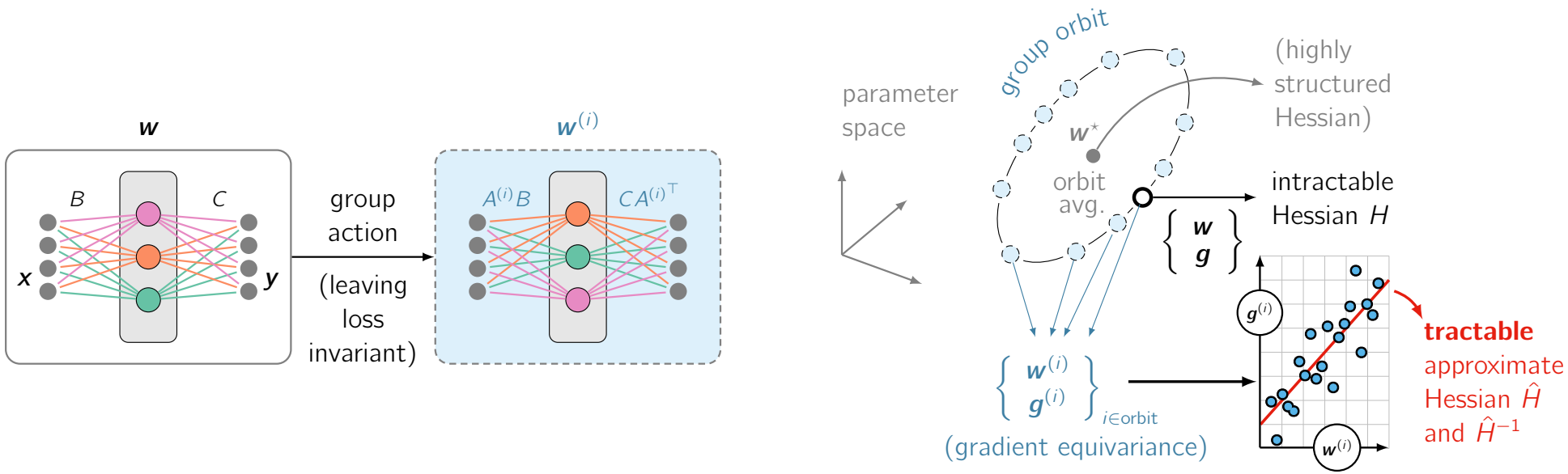
Curvature matters for:

- Second-order optimization
- Bayesian inference
- Continual learning
- Pruning / compression

The problem: Full Hessian is $O(D^2)$

Our focus is on second-order optimization where KFAC, Shampoo/Muon work well.
We introduce a framework based on weight-space symmetries.

Loss invariance → gradient equivariance → curvature



A single gradient reveals curvature along the entire orbit.

Method: orbit-averaged Hessian

Taylor expand around orbit average $\boldsymbol{w}^\star$:

$$\boldsymbol{g} - \boldsymbol{g}^\star \approx H^\star(\boldsymbol{w} - \boldsymbol{w}^\star)$$

Average the secant equation over the orbit:

$$S_{\boldsymbol{g}} \approx H^\star S_{\boldsymbol{w}} H^\star$$

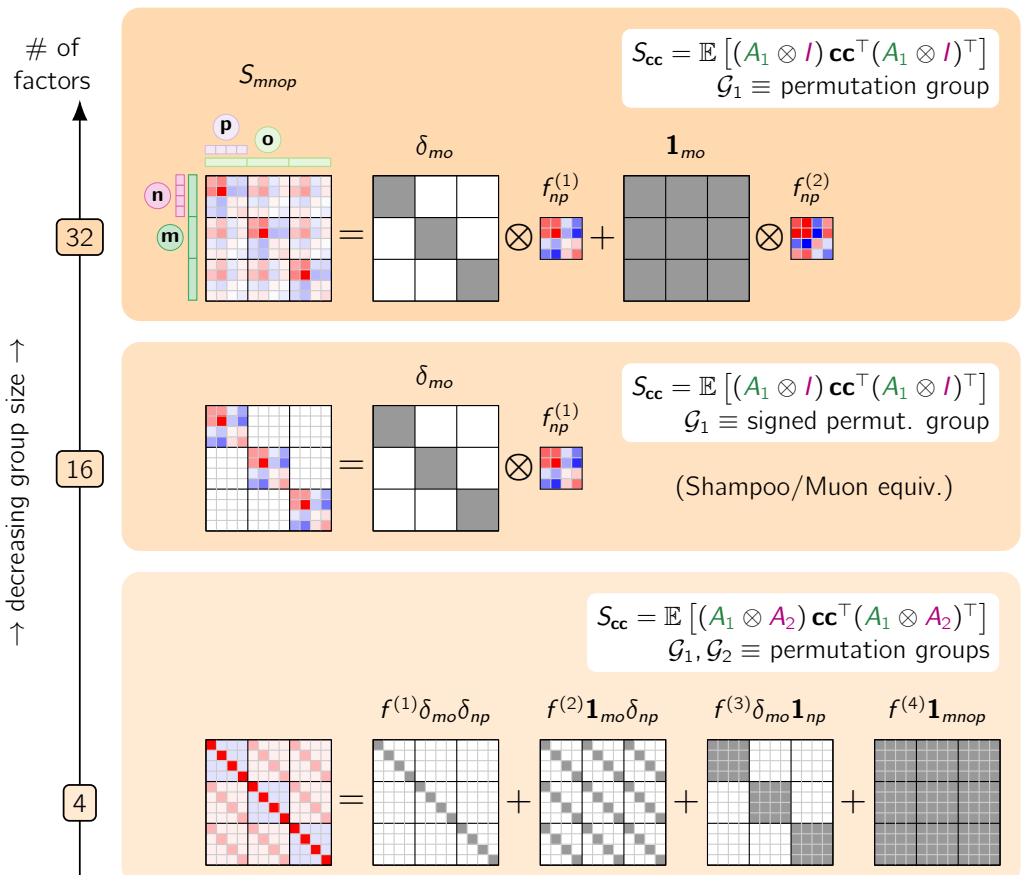
$$\begin{cases} \mathcal{G} \text{ is large} \rightarrow \text{cheap} \\ \mathcal{G} \text{ is small} \rightarrow \text{accurate} \end{cases}$$

Solve for the positive semi-definite curvature:

$$H^\star_{\mathrm{PD}} = S_{\boldsymbol{w}}^{-\frac{1}{2}} \left(S_{\boldsymbol{w}}^{\frac{1}{2}} S_{\boldsymbol{g}} S_{\boldsymbol{w}}^{\frac{1}{2}} \right)^{\frac{1}{2}} S_{\boldsymbol{w}}^{-\frac{1}{2}}$$

$$\hat{H}_{\boldsymbol{g}} = S_{\boldsymbol{g}}^{1/2}$$

Orbit average structure depends on group size



Larger group → fewer basis elements → cheaper.

Shampoo/Muon sits in the middle.

Symo recovers Shampoo/Muon as a Special case

When transformation groups alternate with identity maps, e.g. for an MLP:

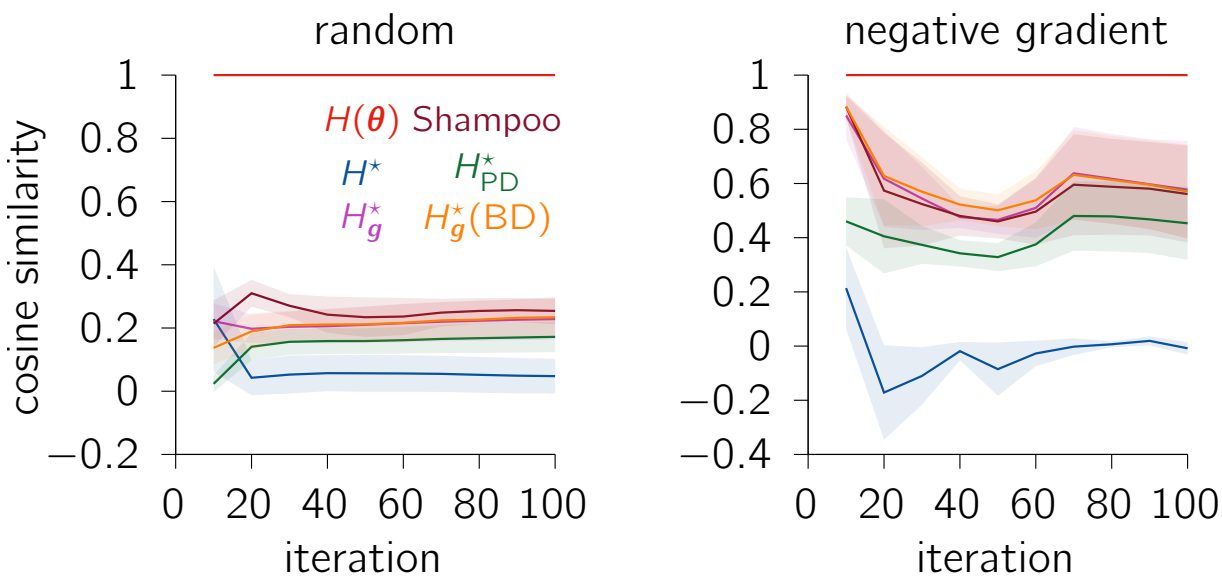
$$\mathcal{G} \overset{\text{def}}{=} I \times \mathcal{B}_1 \times I \times \mathcal{B}_2 \times \cdots$$

the Symo update coincides with Muon/Shampoo:

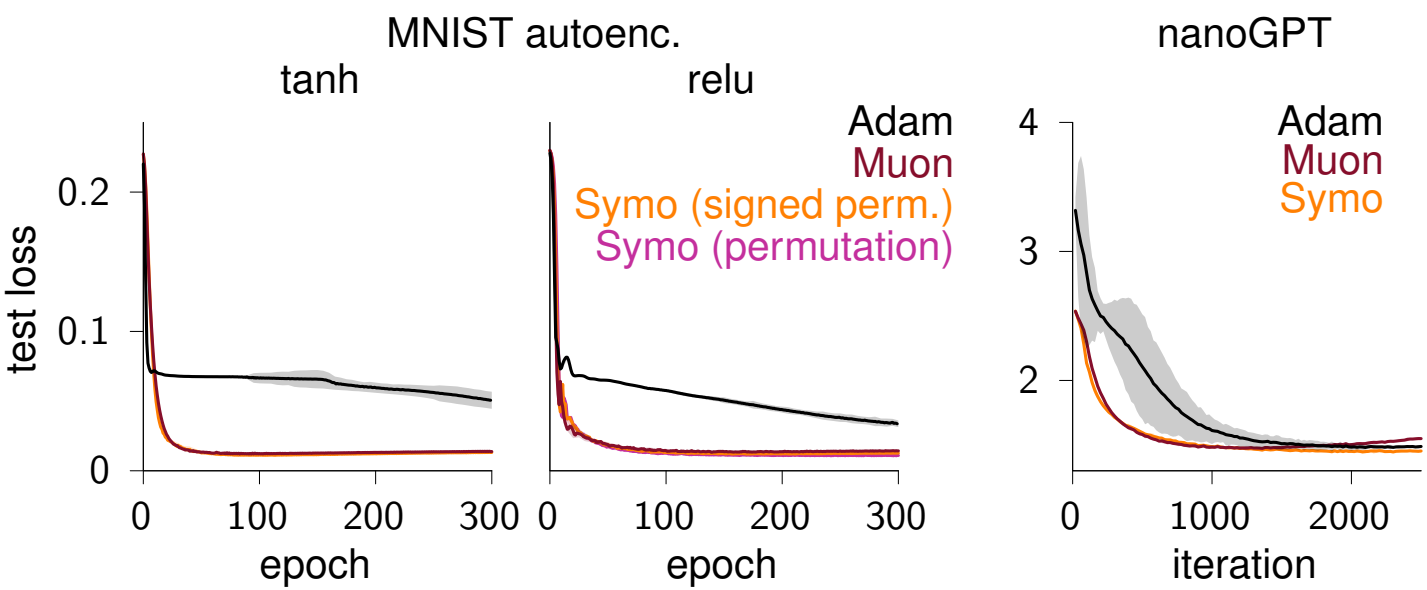
$$\Delta G = \sqrt{d}\, G\left(G^\top G\right)^{-1/2}$$

Hessian approximation quality

- Along *gradient direction*, $\hat{H}_{\boldsymbol{g}}$ achieves $0.4 - 0.8$ cosine similarity
- $\hat{H}_{\boldsymbol{g}}^{\mathrm{BD}}$ overlaps with Shampoo ✓
- $\hat{H}_{\boldsymbol{g}}$ beats $\hat{H}_{\mathrm{PD}}$ (rank deficiency of $S_{\boldsymbol{w}}$)

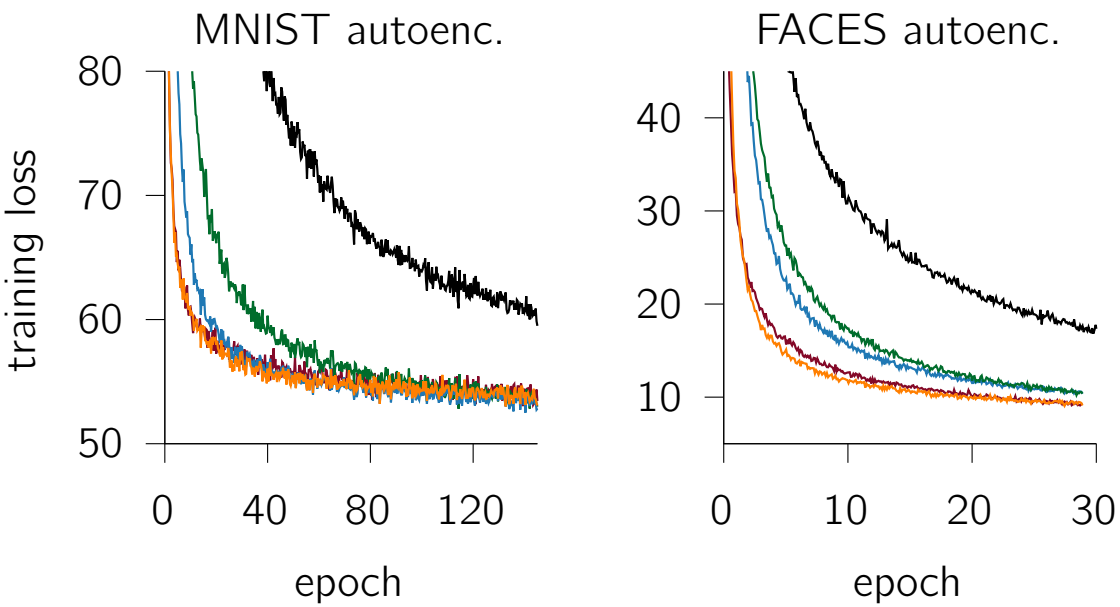


Empirical equivalence



Exploring different symmetry groups

Adam Shampoo Symo
Symo (larger group) Symo (smaller group, block-diag)



- *Larger group*: outperforms Adam (only 64 free parameters!)
- *Smaller group, block-diag*: matches Shampoo
- *Symo*: between the two

The group choice systematically controls the accuracy–cost trade-off.

Summary

Contributions

- Hessian approximation from *a single gradient* via orbit averaging.
- Complexity controlled by *group size*.
- Shampoo and Muon recovered as *special cases*.
- Validated on MLPs and Transformers

Future directions

- Bayesian inference, continual learning, pruning
- Engineering:
  - Automated model symmetry-to-optimizer compiler
  - Scaling to larger models

`github.com/mtkresearch/symm_opt`