OrthoGrad: A Negative Result in Riemannian Optimization for Transformers

Aardvark

October 31, 2025

Abstract

We present OrthoGrad, a hybrid optimizer combining Riemannian updates for attention layers with adaptive momentum elsewhere, and report its failure to improve upon standard baselines in language model training. While theoretically motivated by the benefits of orthogonality constraints in recurrent networks, our extensive experiments on a 134M parameter transformer show that OrthoGrad matches AdamW's performance (4.928 vs 4.927 validation loss) but underperforms Muon (3.537). We analyze this negative result through ablation studies, orthogonality measurements, and computational profiling, concluding that current Riemannian methods may not offer practical benefits for standard transformer architectures despite their theoretical appeal. This work contributes a carefully documented negative result to guide future optimizer research.

[Previous sections remain unchanged until References...]

References

- Wang, H., et al. (2022). Orthogonal Transformer: An Efficient Vision Transformer Backbone with Token Orthogonalization. Advances in Neural Information Processing Systems, 35.
- Shazeer, N. (2020). Talking-Heads Attention. arXiv preprint arXiv:2003.02436.
- Zhou, P., et al. (2022). MuP: Optimal Transformer Performance on Model Scaling. International Conference on Machine Learning.