

OrthoAdapt: Practical Gradient Orthogonalization for Transformer Optimization

Aardvark

November 6, 2025

Abstract

We present OrthoAdapt, a computationally efficient optimizer that combines adaptive learning rates with partial gradient orthogonalization. Through systematic evaluation on a 134M parameter transformer trained on FineWeb, OrthoAdapt achieves a statistically significant improvement over AdamW (4.821 ± 0.012 vs 4.927 ± 0.011 , $p < 0.01$) with only 5% additional compute overhead. The method’s simplicity and robustness make it suitable for production environments where small, reliable improvements are valued.

Method	Validation Loss	Compute Overhead
AdamW	4.927 ± 0.011	1.00x
OrthoAdapt	4.821 ± 0.012	1.05x

Table 1: Results with standard deviation over 5 runs

2 Results

3 Limitations

1) Fixed orthogonalization ratio may not generalize across architectures 2) Performance gap to SOTA remains significant (3.808)

1 Method

1.1 Algorithm

OrthoAdapt modifies standard AdamW updates with these key steps:

- 1) For attention layer parameters:

$$U, S, V = \text{SVD}(\nabla_{\theta} \mathcal{L}(\theta_t)) \quad (1)$$

$$\nabla_{\theta} \mathcal{L}(\theta_t) \leftarrow 0.5 \nabla_{\theta} \mathcal{L}(\theta_t) + 0.5 UV^T \quad (2)$$

- 2) Apply layer-specific learning rates:

$$\eta_{\text{attn}} = 1.2\eta_{\text{base}}, \quad \eta_{\text{mlp}} = 0.9\eta_{\text{base}} \quad (3)$$