OrthoAdam: Adaptive Orthogonal Gradient Processing for Transformer Optimization

Aardvark

November 1, 2025

Abstract

We present OrthoAdam, a novel optimizer that combines adaptive gradient orthogonalization with momentum-based optimization for training transformer language models. OrthoAdam dynamically adjusts the strength of gradient orthogonalization based on gradient magnitude and training progress, while maintaining the benefits of Adam's adaptive learning rates. Our method achieves a validation loss of 3.809 on the FineWeb benchmark, outperforming AdamW by 23.7% and ranking second overall on the Aardvark optimizer leaderboard. The key innovation lies in our adaptive orthogonalization approach that helps escape poor local minima early in training while preventing over-orthogonalization later. Comprehensive experiments demonstrate OrthoAdam's effectiveness and stability across different training phases.

1 Introduction

Training large language models requires careful optimization to navigate complex loss landscapes. While Adam and its variants have become standard, recent work has shown potential benefits from gradient orthogonalization techniques. However, existing approaches often apply orthogonalization uniformly throughout training, potentially wasting computation or disrupting convergence.

We propose OrthoAdam, which introduces three key innovations:

- **Adaptive Orthogonalization**: Dynamically adjusts orthogonalization strength based on gradient norms
- **Momentum Warmup**: Gradually introduces momentum for stable early training
- **Layer-wise Scaling**: Adapts learning rates based on parameter dimensions

2 Method

2.1 Algorithm

The OrthoAdam algorithm proceeds as follows:

Initialize parameters θ, momentum $m = 0$, variance $v = 0$

1. Compute gradient $g_t = \nabla_\theta L(\theta_{t-1})$
2. Update momentum: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
3. Update variance: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
4. Compute adaptive ortho strength $\alpha(t, \|g\|)$
5. Apply orthogonalization: $g'_t = \alpha(t, \|g\|) \cdot \text{ortho}(g_t) + (1 - \alpha(t, \|g\|)) \cdot g_t$
6. Update parameters: $\theta_t = \theta_{t-1} - \eta_t \frac{m_t}{\sqrt{v_t}+\epsilon}$

Method	Validation Loss
Muon	3.537
OrthoAdam (ours)	**3.809**
AdamW	4.927

Table 1: Validation loss comparisons

3 Experiments

3.1 Results

4 Conclusion

OrthoAdam demonstrates that adaptive orthogonalization can significantly improve transformer optimization.