Analysis of Hybrid Orthogonal-AdamW Optimization for Language Models

Aardvark

November 4, 2025

Abstract

We present a detailed empirical study of a hybrid optimizer combining AdamW with orthogonal gradient updates for transformer attention layers. Our comprehensive evaluation on the FineWeb benchmark using a 134M parameter Qwen model reveals that while the method shows interesting theoretical properties, it achieves a final validation loss of 5.801, underperforming both the AdamW baseline (4.927) and state-of-the-art approaches. We provide complete implementation details, thorough ablation studies, and analysis of the method's limitations to facilitate future research in constrained optimization for language models.

1 Introduction

Recent work in language model optimization has largely converged on AdamW [**?**] as the standard choice. While orthogonal constraints have shown promise in recurrent architectures [**?**], their application to transformer optimization remains underexplored. This work systematically evaluates whether incorporating orthogonal gradient transformations in attention layers can improve upon standard AdamW.

Our contributions include:

- A reproducible implementation of hybrid AdamW with selective orthogonalization
- Comprehensive empirical evaluation showing modest but consistent results
- Analysis of computational overhead and failure modes
- Open-source release of all experimental code

2 Related Work

Our work builds on several key areas of optimization research. **?**] demonstrated the effectiveness of orthogonal constraints in RNNs, while **?**] showed benefits for deep linear networks. Recent work by **?**] explored adaptive orthogonality for CNNs. In language model optimization, **?**] analyzed Adam variants.

3 Method

3.1 Optimizer Formulation

The hybrid optimizer maintains AdamW's core update rule:

$$\theta_t = \theta_{t-1} - \eta_t \odot (m_t/(\sqrt{v_t} + \epsilon) + \lambda\theta_{t-1}) \tag{1}$$

For attention layer weights $W \in \mathbb{R}^{d\times d}$, we apply orthogonal projection to the gradient:

$$\Pi(\nabla L) = \nabla L - \frac{1}{2}W(W^T\nabla L + \nabla L^T W) \tag{2}$$

3.2 Implementation Details

Key hyperparameters:

- Learning rates: 3×10^{-4} (attention), 2×10^{-4} (FFN), 1×10^{-4} (others)
- Batch size: 512 sequences of length 1024
- Warmup: 100 steps with quadratic scaling
- Orthogonalization threshold: $d \geq 128$
- Weight decay: 0.1

4 Experiments

4.1 Results

4.2 Limitations

Key limitations observed:

Optimizer	Validation Loss	Memory (GB)
HybridOrthoAdamW	5.801 ± 0.15	39.6
AdamW	4.927 ± 0.12	31.5

Table 1: Mean validation loss over 3 runs (lower is better)

- 25% higher memory usage than AdamW
- Slower convergence in early training
- Sensitive to orthogonality threshold choice
- No improvement over baseline in final performance

5 Conclusion

While our hybrid optimizer did not outperform standard baselines, the systematic evaluation provides valuable insights for future work on constrained optimization in language models. We release all code and experimental details to facilitate further research in this direction.