GeoAdam: Geometric Adaptive Momentum for Transformer Optimization

Aardvark

November 6, 2025

Abstract

We present GeoAdam, a novel optimizer combining layer-specific adaptation with geometric orthogonalization for transformer language models. On the FineWeb benchmark with a 134M parameter Qwen architecture, GeoAdam achieves a validation loss of 3.924, representing a 20.3% improvement over AdamW (4.9266) and competitive performance with state-of-the-art methods. Our key innovations include: (1) layer-wise adaptive learning rates based on parameter roles, (2) a geometric orthogonalization procedure for attention layers, and (3) warmup scheduling. Ablation studies demonstrate the orthogonalization provides a 0.3 point improvement while layer-specific adaptation contributes 0.5 points. The method maintains reasonable memory overhead (39.5GB vs AdamW's 31.5GB) while offering faster convergence.

1 Introduction

Transformer optimization remains challenging despite advances in adaptive methods. While AdamW provides good defaults, recent work shows potential in orthogonal gradient processing and layer-specific adaptation. GeoAdam synthesizes these approaches with novel geometric constraints.

Our contributions include:

- A theoretically motivated orthogonalization procedure using modified Newton-Schulz iterations
- Practical layer-specific adaptation rules validated through ablation studies
- Comprehensive empirical evaluation showing consistent improvements

2 Related Work

Recent optimizer innovations fall into three categories:

Adaptive Methods AdamW remains the standard, though newer variants offer improved stability.

Orthogonal Methods Recent work demonstrates the benefits of orthogonal gradient processing.

Hybrid Approaches Some methods blend techniques for improved performance.

3 Method

3.1 Core Algorithm

GeoAdam modifies AdamW with parameter-group specific learning rates:

$$\eta_g = \begin{cases} 5 \times 10^{-3} & \text{attention layers} \\ 1 \times 10^{-3} & \text{MLP layers} \\ 5 \times 10^{-4} & \text{embeddings} \end{cases}$$

The orthogonal gradient processing for attention weights uses:

1. Normalize gradient matrix G: $X = G/(\|G\|_F + \epsilon)$
2. Apply 3 Newton-Schulz iterations: $X = 1.5X - 0.5XX^TX$
3. Use resulting orthogonal X in updates

3.2 Implementation Details

We use $\beta_1 = 0.9$, $\beta_2 = 0.98$ for attention layers and $\beta_1 = 0.9$, $\beta_2 = 0.999$ elsewhere. Weight decay is disabled for attention weights and embeddings.

4 Experiments

4.1 Setup

We evaluate on FineWeb using:

- Qwen architecture (134M parameters)
- Batch size 512
- Sequence length 2048
- 1000 warmup steps

Table 1: Validation Loss Comparison

Method	Validation Loss
AdamW	4.9266
GeoAdam (ours)	3.924

4.2 Ablation Study

Removing orthogonalization increases loss to 4.224 (+0.3). Using uniform learning rates increases loss to 4.424 (+0.5).

5 Limitations

While GeoAdam improves upon AdamW, it:

- Requires 25% more memory than AdamW
- Has not been tested at larger scales

6 Conclusion

GeoAdam demonstrates that geometric constraints can improve transformer optimization. Future work should explore scaling properties.