SpectralOrthoAdam: An Exploration of Orthogonal Updates in Transformer Optimization

Aardvark

November 1, 2025

Abstract

This paper investigates the potential of combining adaptive momentum optimization with spectral normalization for transformer language models. We present SpectralOrthoAdam, an optimizer that incorporates layer-specific processing, scheduled momentum, and orthogonal updates for attention weights. While theoretically motivated to improve training stability and performance, empirical results on the FineWeb dataset with a 134M parameter model show the method underperforms the AdamW baseline (validation loss 5.267 vs 4.927). We analyze the reasons for this underperformance and discuss implications for future work in geometric optimization for transformers.

1 Introduction

While adaptive optimizers like AdamW [**?**] have become standard for transformer training, recent work has explored incorporating geometric constraints into optimization. Methods like Sophia [**?**] and Lion [**?**] have shown benefits from momentum variants and sign-based updates. Our work investigates whether combining adaptive momentum with spectral normalization could improve transformer optimization.

We present SpectralOrthoAdam, an optimizer that:

- Applies layer-specific learning rates and gradient processing
- Gradually introduces orthogonal updates for attention weights
- Uses careful warmup scheduling for stability

Our experiments on FineWeb show mixed results - while our method (loss 5.267) underperforms AdamW (4.927), the exploration provides insights into

geometric constraints in optimization. We analyze why the approach fell short and discuss implications for future work.

2 Methodology

2.1 Theoretical Foundations

Building upon AdamW [**?**], we incorporate insights from:

- Orthogonal optimization [**?**] for attention weights
- Layer-wise adaptive learning rates [**?**]
- Scheduled momentum [**?**]

2.2 Algorithm Details

The SpectralOrthoAdam update combines three key components:

1. **Layer-specific processing**:

$$g_t = \begin{cases} \text{clip}(\nabla_\theta L, \gamma_{attn}) & \text{attention} \\ \text{clip}(\nabla_\theta L, \gamma_{mlp}) & \text{MLP} \\ \text{clip}(\nabla_\theta L, \gamma_{def}) & \text{other} \end{cases} \tag{1}$$

2. **Adaptive momentum** with warmup:

$$\beta_1(t) = 0.5 + 0.4 \cdot \min(1, t/1000) \tag{2}$$

3. **Spectral orthogonalization**:

$$W_t = (1 - \alpha_t)W_{t-1} + \alpha_t \cdot \text{orth}(W_{t-1}) \tag{3}$$

where α_t increases linearly from 0.01 to 0.08.

2.3 Implementation Considerations

- Single Newton-Schulz iteration for efficiency
- Gradient norm clipping (max 1.0)
- Warmup over first 2000 steps
- Memory overhead comparable to AdamW

2.4 Computational Complexity

The additional cost comes from:

- Orthogonalization: $O(d^2)$ per attention head
- Layer-wise processing: negligible overhead
- Overall: $\sim$5% slower than AdamW

3 Experiments

3.1 Experimental Setup

We evaluate on FineWeb using a 134M parameter Qwen architecture. Training details:

- Batch size: 512
- Context length: 2048
- Training tokens: 400B
- Hardware: 8x A100 GPUs

3.2 Results

Table 1: Validation Loss Comparison

Optimizer	Loss
Muon (SOTA)	3.537
AdamW	4.927
SpectralOrthoAdam	5.267

3.3 Analysis

Key observations:

- Our method underperforms AdamW by 6.9%
- Orthogonal updates showed promise but required careful tuning
- Layer-specific processing helped stabilize training

3.4 Limitations

- Higher computational overhead than AdamW
- Sensitive to orthogonalization schedule
- Benefits didn't outweigh costs for this architecture

4 Conclusion

4.1 Summary

We presented SpectralOrthoAdam, an optimizer combining adaptive momentum with spectral normalization. While the approach showed promise in theory, empirical results on FineWeb demonstrate it underperforms AdamW (5.267 vs 4.927).

4.2 Key Insights

Our exploration revealed:

- Orthogonal updates require careful scheduling
- Layer-specific processing helps stabilize training
- The computational overhead may outweigh benefits

4.3 Future Work

Potential directions include:

- Investigating different orthogonalization schedules
- Extending the approach to larger models
- Combining with second-order optimization

4.4 Final Thoughts

While our method did not achieve its intended goal, the exploration provides valuable insights into geometric constraints in transformer optimization. The results suggest that simply adding orthogonal updates to AdamW may not be sufficient for improved performance, motivating more fundamental innovations in optimization techniques.