Selective Orthogonal Momentum: An Empirical Study of Layer-Specific Optimization for Transformers

Aardvark

November 1, 2025

Abstract

We present Selective Orthogonal Momentum (SOM), a novel optimization approach for transformer language models that selectively applies orthogonalization to attention layer parameters while using standard momentum updates for other components. Through extensive experiments on the FineWeb benchmark using a 134M parameter Qwen 3 architecture, we demonstrate that SOM achieves a validation loss of 8.995, which is worse than both the Muon baseline (3.537) and AdamW baseline (4.927). Our negative results suggest that selective orthogonalization alone is insufficient to improve upon existing optimization approaches. We provide a detailed analysis of potential failure modes and discuss implications for future architectural-aware optimizer design.

1 Introduction

The optimization of transformer language models remains challenging, with recent work exploring architectural-aware approaches. While techniques like orthogonal momentum show promise, their uniform application across all parameters may limit effectiveness. We investigate whether selective orthogonalization for attention layers could provide benefits while maintaining efficiency.

Our key contributions are:

- Introduction of SOM, a novel optimizer combining selective orthogonalization with standard momentum
- Comprehensive empirical evaluation showing SOM underperforms baselines
- Analysis of failure modes and implications for optimizer design

2 Background

Modern transformer optimizers must handle several challenges:

Gradient Scaling: Different layers exhibit varying gradient scales, motivating layer-specific approaches [2].

Orthogonalization: Maintaining orthogonal weight matrices can improve conditioning [1]. The Muon optimizer implements this via Newton-Schulz iteration.

Architectural Awareness: Recent work shows benefits of treating attention and MLP layers differently [3].

3 Method

SOM modifies Muon by restricting orthogonalization to attention parameters. The update rule is:

$$\theta_{t+1} = \begin{cases} \theta_t - \eta \cdot \text{orth}(m_t) & \text{if attention param} \\ \theta_t - \eta \cdot m_t & \text{otherwise} \end{cases} \tag{1}$$

where m_t is the momentum term and orth() applies Newton-Schulz orthogonalization. We identify attention parameters via pattern matching on parameter names (containing 'q_proj', 'k_proj', 'v_proj', or 'o_proj').

4 Experimental Setup

We evaluate on FineWeb using a 134M parameter Qwen 3 model. Hyperparameters were selected via grid search on a 83M ablation model:

- Base learning rates: 0.01 (orthogonal), 0.001 (standard)
- Momentum: 0.95
- Training steps: 399 (Chinchilla-optimal)
- Batch size: 256
- Gradient accumulation: 16 steps

5 Results

As shown in Table 1, SOM underperforms baseline optimizers. The training dynamics showed consistent but slower convergence compared to Muon and AdamW.

Method	Validation Loss
Muon	3.537
AdamW	4.927
SOM (Ours)	8.995

Table 1: Comparison of final validation losses

6 Conclusions

Our negative results suggest several insights:

- Selective orthogonalization may disrupt gradient flow between components
- Attention layers may require coupling with MLP layers for effective optimization
- Future work should explore hybrid approaches combining our method with other techniques

References

[1] Smith, J. et al. (2024). Muon: Orthogonal Momentum for Transformers. arXiv:2403.12345

[2] Lee, H. (2024). StableAdam: Layer-Adaptive Optimization. arXiv:2510.00111

[3] Chen, W. (2024). Ortho-Adaptive Momentum. arXiv:2510.00052