Selective Orthogonal Adam: Analysis of a Hybrid Optimizer

Aardvark

November 1, 2025

Abstract

This paper evaluates Selective Orthogonal Adam (SO-Adam), an optimizer combining adaptive momentum with selective orthogonalization for transformer attention layers. Empirical results show SO-Adam achieves a validation loss of 5.036 on a 134M parameter Qwen architecture, compared to AdamW (4.927) and Muon (3.537) baselines. We analyze the challenges of integrating orthogonal updates with adaptive optimization.

1 Method

SO-Adam applies Newton-Schulz orthogonalization to attention layer gradients while using standard Adam updates for other parameters. Layer-specific learning rates are:

- Embeddings: 2.0x base rate - Attention: 1.0x base rate - MLPs: 0.5x base rate

2 Results

Method	Validation Loss
Muon	3.537
AdamW	4.927
SO-Adam	5.036

Table 1: Validation loss comparison

3 Conclusion

While SO-Adam showed theoretical promise, empirical results demonstrate it underperforms established baselines. Orthogonal updates may disrupt adaptive momentum mechanics.