Analysis of Adaptive Muon-Adam: Lessons from a Failed Optimizer

Aardvark

November 3, 2025

Abstract

We analyze Adaptive Muon-Adam, an optimizer combining orthogonal gradient processing, adaptive momentum, and second-order information. Despite careful implementation, our method underperformed baselines (loss=10.854 vs Muon=3.537). We identify key failure modes and provide recommendations for future optimizer designs.

1 Method

1.1 Algorithm

The optimizer processes gradients as:

- For 2D weights:
  - Compute SVD: $G = U\Sigma V^T$
  - Clip singular values: $\Sigma_{ii} = \min(\sigma_i, \tau)$
  - Estimate diagonal Hessian: $h_i = u_i^T(\nabla^2 L)u_i$
  - Update: $\Delta = -\eta G/(|h| + \epsilon)$
- For other parameters: Standard Adam update

2 Experiments

2.1 Setup

- Model: Qwen 3 (134M params)
- Dataset: FineWeb (10B tokens)
- Learning rate: 10^{-4} with warmup
- Gradient clipping: 5.0

Table 1: Validation Loss

Method	Loss
Our Method	10.854
AdamW	4.927
Muon	3.537

2.2 Results

3 Analysis

Failure modes:

- Orthogonalization caused instability
- Second-order estimates were noisy
- Update rules conflicted

4 Conclusion

Key lessons:

- Need better orthogonalization scaling
- Second-order methods require stabilization
- Parameter grouping needs validation