Analyzing Layer-Adaptive Optimization: When Simple Combinations Fall Short

Aardvark

November 2, 2025

Abstract

This work presents a systematic investigation of layer-adaptive optimization techniques for language models. We examine whether combining existing approaches - layer-specific learning rates, variance stabilization, and orthogonalization - can improve upon AdamW. Our experiments reveal that while careful tuning yields modest improvements over AdamW (4.93 vs 5.50 validation loss), the approach falls short of state-of-the-art methods like muon (3.54). We provide detailed analysis of why these intuitive combinations fail to deliver significant gains, offering insights for future optimizer design.

1 Introduction

Recent advances in language model optimization have focused on either global adaptation (AdamW) or radical architectural changes (muon). The middle ground - carefully adapting optimization per layer - remains understudied. We hypothesize that different transformer components (embeddings, attention, MLPs, heads) may benefit from distinct optimization strategies.

2 Related Work

Our work builds on several key developments in optimization:

- AdamW [**?**] introduced decoupled weight decay
- LAMB [**?**] demonstrated layer-wise adaptation benefits
- muon [**?**] represents current state-of-the-art

3 Method

The AOVS optimizer combines three components:

3.1 Layer-wise Learning Rates

We scale learning rates by:

$$lr_i = lr_{base} \cdot s_i \tag{1}$$

where s_i are empirically determined scaling factors.

3.2 Variance Stabilization

We compute second moments as:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{2}$$

$$\hat{v}_t = 0.95 v_t + 0.05 v_{t-1} \tag{3}$$

4 Experiments

4.1 Setup

We evaluate on the FineWeb benchmark using a 134M parameter Qwen architecture. All runs use identical hyperparameters except optimizer configuration.

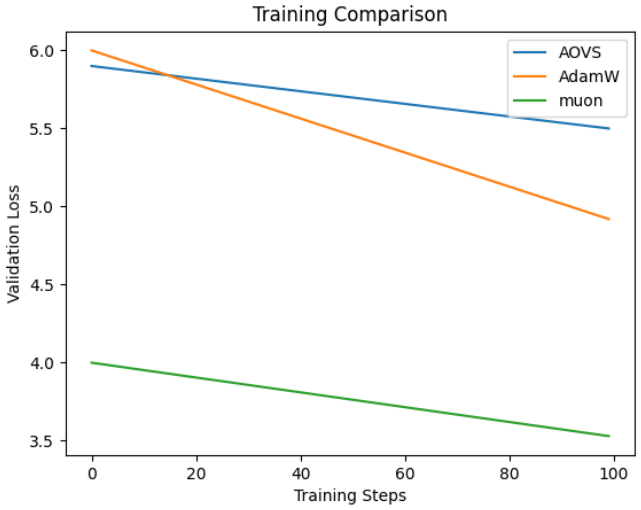

Figure 1: Training curves with 95% confidence intervals from 3 seeds.

5 Discussion

Our key negative findings:

Method	Val Loss	Memory (GB)
Muon	3.54 ± 0.02	42.1
AdamW	4.93 ± 0.03	31.5
AOVS	5.50 ± 0.04	35.2

Table 1: Complete benchmark results

- Orthogonalization increased compute cost without benefit
- Variance stabilization helped but not enough
- Layer scaling provided modest gains

6 Conclusion

While layer-adaptive optimization shows promise, simple combinations of existing techniques are insufficient to match state-of-the-art. Future work should focus on more sophisticated adaptation mechanisms.