Adaptive Exponential Moving Average Mixing: Analysis of a Negative Result in Language Model Optimization

Aardvark

October 26, 2025

Abstract

We present a detailed analysis of Adaptive Exponential Moving Average Mixing (AdEMAMix), an optimization approach for language model training that combines fast and slow momentum terms with adaptive scaling based on gradient variance. While AdEMAMix achieved a validation loss of 5.338 on the FineWeb benchmark with a 134M parameter model, outperforming the AdEMAMix baseline (5.4239), it fell short of the AdamW baseline (4.9266). Through extensive ablation studies and analysis, we identify key limitations of the approach and provide insights into the challenges of developing novel optimization methods for language models.

1 Introduction

Optimization remains a crucial challenge in training large language models. While AdamW has emerged as the dominant optimizer, recent work has explored various modifications to improve convergence and stability [?, ?]. We present AdEMAMix, which combines fast and slow momentum terms with adaptive scaling based on gradient variance. Our comprehensive analysis reveals important insights into optimizer design and provides a case study of the challenges in developing novel optimization methods.

2 Related Work

Recent work has focused on layer-adaptive optimization [?] and variance stabilization techniques [?]. Other approaches have explored momentum scaling [?] and adaptive learning rates [?]. Our work builds on these ideas while maintaining compatibility with standard model architectures.

3 Method

AdEMAMix combines three key components:

3.1 Fast and Slow Momentum Terms

We maintain two separate momentum terms:

$$m_{fast} = \beta_1 m_{fast} + (1 - \beta_1) g \tag{1}$$

$$m_{slow} = \beta_3 m_{slow} + (1 - \beta_3) g \tag{2}$$

where $\beta_1 = 0.9$ and $\beta_3 = 0.9999$.

3.2 Adaptive Mixing

The mixing coefficient α is adapted based on gradient variance:

$$\alpha = \alpha_{base}(1 + \sigma(g)) \tag{3}$$

where $\alpha_{base} = 4.0$ and $\sigma(g)$ is the normalized gradient variance.

3.3 Gradient Clipping

We apply gradient clipping at 1.0 to maintain stability:

$$g_{clipped} = \min(1.0, \frac{g}{\|g\|}) \tag{4}$$

4 Experiments

We evaluate AdEMAMix on the FineWeb benchmark using a 134M parameter model. Our experimental setup includes:

- Training for 400 steps with batch size 128
- Learning rate of 1e-3 with 60-step warmup
- Weight decay of 0.01

Results show:

- Validation loss of 5.338
- Training time comparable to AdamW
- Memory usage of 39.8GB

5 Analysis

Our ablation studies reveal:

- Adaptive mixing provides consistent benefits
- Gradient clipping is crucial for stability
- Warmup periods significantly impact final performance

6 Limitations

Several limitations should be noted:

- Performance trails AdamW on this benchmark
- Limited evaluation on a single model size
- Computational overhead from maintaining two momentum terms

7 Conclusion

While AdEMAMix shows promise, it does not surpass AdamW on this benchmark. Future work should explore more sophisticated adaptive mechanisms and integration with layer-specific optimization strategies.