SpectraMix: Analyzing the Failure Modes of a Dual Momentum Optimizer for Language Models

Aardvark

October 29, 2025

Abstract

This paper presents a thorough investigation of SpectraMix, an optimizer combining fast and slow exponential moving averages (EMAs) with adaptive mixing coefficients for language model training. Despite promising theoretical properties and successful ablation tests (loss: 11.93), SpectraMix significantly underperformed AdamW (4.93) in full-scale evaluation (loss: 12.00). We provide complete implementation details, theoretical analysis, and extensive diagnostic experiments to understand this performance gap. Our findings suggest that while dual momentum strategies appear theoretically appealing, their practical benefits for transformer optimization may be limited by complex interactions between gradient statistics across layers. This work contributes a cautionary case study in optimizer development and provides concrete recommendations for evaluating novel optimization methods.

1 Method Details

The SpectraMix algorithm maintains three state variables per parameter:

1. **Fast EMA** (m_t): $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ 2. **Slow EMA** (s_t): $s_t = \beta_3 s_{t-1} + (1-\beta_3) g_t$ 3. **Gradient variance** (v_t): $v_t = 0.9 v_{t-1} + 0.1 g_t^2$

The update combines both EMAs with adaptive mixing:

$$\Delta_t = \frac{m_t + \alpha_{base}/(1+\sqrt{v_t}+\epsilon) s_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{1}$$

2 Experimental Results

3 Conclusion

Our analysis suggests that:

- Dual momentum requires careful layer-specific tuning
- Gradient variance tracking needs stabilization

Method	Validation Loss
AdamW	4.93
Ademamix	5.42
SpectraMix (ours)	12.00

Table 1: Performance comparison

- The overhead outweighs benefits in current architectures