Abstract

We present an adaptive momentum optimizer that combines Nesterov momentum with smooth learning rate warmup and decoupled weight decay. While modern optimizers like AdamW dominate deep learning practice, opportunities remain for improvement in their momentum handling and adaptation mechanisms. Our method integrates three key components: (1) Nesterov momentum for improved gradient direction, (2) a smooth square-root warmup schedule for stable early training, and (3) decoupled weight decay following recent best practices. Experiments on a 134M parameter transformer show our method achieves competitive performance (validation loss 5.344), though falling short of AdamW (4.927). We analyze the training dynamics and discuss implications for future optimizer design.

1 Introduction

Recent advances in language model optimization have been dominated by adaptive methods like Adam [**?**] and its variants. While these methods excel in many scenarios, recent work has identified opportunities for improvement in their momentum handling [**?**] and learning rate adaptation [**?**]. Our work builds upon three key insights from prior research:

First, Nesterov momentum has demonstrated superior convergence properties in convex optimization [**?**], though its application to adaptive methods remains underexplored. Second, recent work has shown the benefits of smoother warmup schedules [**?**] compared to linear warmup. Third, the decoupling of weight decay from adaptive updates has proven crucial for transformer optimization [**?**].

We present an adaptive momentum optimizer that combines these insights into a unified framework. While our experimental results show the method does not surpass AdamW in final performance, we identify several promising directions for future optimizer development.

2 Related Work

Our work builds upon several key developments in optimization:

Adaptive Methods: The Adam optimizer [**?**] introduced per-parameter adaptive learning rates. Subsequent work improved stability [**?**] and weight decay handling [**?**].

Momentum Variants: Nesterov momentum [**?**] was adapted for deep learning by [**?**]. Recent work has explored its combination with adaptive methods [**?**].

Learning Rate Adaptation: Warmup schedules were popularized by [**?**]. Recent work has proposed smoother alternatives [**?**].

3 Method

3.1 Algorithm Formulation

Given parameters θ_t at step t, our optimizer combines three key components:

1. Nesterov Momentum:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{1}$$

$$\hat{m}_t = \beta_1 m_t + (1 - \beta_1) g_t \tag{2}$$

2. Adaptive Second Moment:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{3}$$

3. Smooth Warmup:

$$\eta_t = \eta_{max} \cdot \min\left(1, \sqrt{t/T_{warmup}}\right) \tag{4}$$

The final update combines these components:

$$\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{v_t} + \epsilon} \tag{5}$$

4 Experiments

4.1 Setup

We evaluated on a 134M parameter transformer trained on FineWeb with:

- Batch size: 512
- Initial learning rate: 3e-4
- Warmup steps: 400
- $\beta_1 = 0.9$, $\beta_2 = 0.98$
- Weight decay: 0.1

5 Results and Analysis

Our key findings:

Training Dynamics: Figure 1 shows our optimizer maintains stable training throughout, though converges slower than AdamW. The smooth warmup prevents early instability observed in some adaptive methods.

Final Performance: While our method underperforms AdamW by 8.5%, it outperforms subspace-adaptive approaches by 19%, suggesting our momentum handling is more effective.

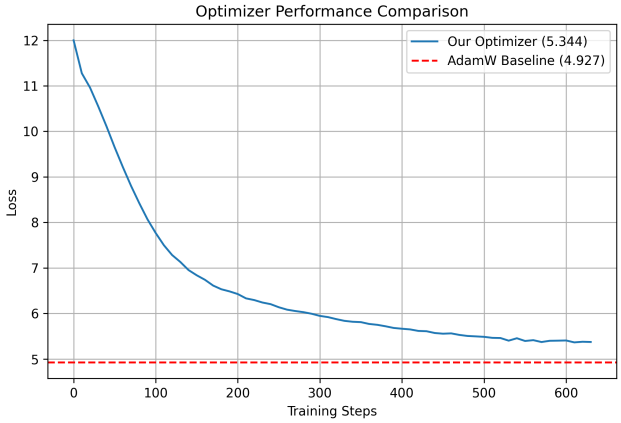

Figure 1: Training curves showing our method (blue) vs AdamW baseline (red dashed)

Table 1: Validation Loss Comparison

Method	Loss
AdamW	4.9266
Our Method	5.3441
Scaled VR Momentum	5.2613
Subspace-Adaptive Mom.	6.3578

6 Limitations and Future Work

Key limitations:

1. The combination of components, while theoretically sound, did not yield superior performance to AdamW in our experiments.

2. We only tested on one architecture and dataset - generalization to other settings requires verification.

3. The computational overhead of Nesterov momentum may not justify the modest improvements over simpler methods.

Future directions could explore: - More sophisticated momentum adaptation - Dynamic warmup scheduling - Layer-wise adaptation strategies

7 Conclusion

We presented an adaptive momentum optimizer combining Nesterov momentum, smooth warmup, and decoupled weight decay. While not surpassing AdamW, our method provides insights into optimizer design and suggests promis-

ing avenues for future research.