StableAdamW: Variance-Stabilized Optimization for Language Models

Aardvark

October 23, 2025

Abstract

We present StableAdamW, a modified version of AdamW that addresses training instability through controlled variance clipping of second moment estimates. While the performance improvement over AdamW is modest (4.919 vs 4.927 validation loss), our analysis reveals more consistent training dynamics. The method requires no additional memory overhead and maintains the computational efficiency of AdamW.

1 Introduction

Recent work has shown that adaptive optimization methods like AdamW can exhibit instability during language model training. We hypothesize that much of this instability stems from uncontrolled variance in the second moment estimates.

Our contribution is StableAdamW, which introduces conservative variance clipping of second moment estimates. While our final performance improvement over AdamW is small, the method demonstrates more consistent convergence.

2 Related Work

Our work builds upon several research directions in optimization:

Adaptive Methods: Adam and AdamW introduced parameter-specific learning rate adaptation.

Variance Reduction: Various techniques have been proposed to stabilize gradient estimates.

Optimizer Stability: Recent work has identified instability issues in adaptive methods.

3 Method

StableAdamW modifies the AdamW update as follows:

$$\hat{v}_t = \min\left(\sqrt{v_t} + \epsilon, \tau\right) \tag{1}$$

where $\tau = 2.0$ is the clipping threshold. The update steps are:

1. Compute gradient g_t 2. Update first moment $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$ 3. Update second moment $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ 4. Apply variance clipping $\hat{v}_t = \min(\sqrt{v_t} + \epsilon, \tau)$ 5. Update parameters $\theta_t = \theta_{t-1} - \eta m_t / \hat{v}_t$

4 Experimental Setup

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

- Batch size 1024, context length 4096
- Cosine learning rate schedule
- Compared against AdamW baseline

5 Results

Method	Validation Loss
StableAdamW (ours)	4.919
AdamW	4.927

Table 1: Validation results comparing optimizers.

6 Conclusions

We presented StableAdamW, demonstrating that careful second moment control can improve training stability. The method provides a simple modification to AdamW that may benefit practitioners.

References

[1] Kingma, Ba. "Adam: A method for stochastic optimization." arXiv 2014.

[2] Loshchilov, Hutter. "Decoupled weight decay regularization." arXiv 2017.