StratOpt: A Stratified Optimization Approach for Language Model Training

Aardvark

October 30, 2025

Abstract

This paper presents StratOpt, a novel optimization approach for training large language models that combines layer-wise adaptation with variance-stabilized gradient updates. We provide a comprehensive evaluation of StratOpt on a 134M parameter transformer model trained on the FineWeb dataset, comparing against AdamW, AdEMAMix, and other recent optimizers. While StratOpt demonstrates improvements over AdEMAMix (5.209 vs 5.424 validation loss), it does not surpass the AdamW baseline (4.927). Our analysis includes detailed ablation studies, computational efficiency metrics, and theoretical justification for the design choices. The results reinforce that simple, well-tuned first-order methods remain surprisingly effective for language model training, and suggest that incremental optimizer modifications may not yield significant improvements.

1 Introduction

The optimization of large language models presents unique challenges due to the scale of parameters and complexity of the loss landscape. Recent work has shown that while adaptive methods like AdamW [1] perform well, there may be room for improvement in specific aspects of the optimization process [4, 5].

We present StratOpt, an optimizer that combines three key components: (1) layer-wise learning rate adaptation, (2) variance-stabilized gradient updates, and (3) dynamic gradient mixing. Unlike approaches that apply uniform updates across all parameters, StratOpt automatically adjusts its behavior based on observed gradient statistics during training.

2 Related Work

Our work builds upon several lines of research in optimization for deep learning. The success of Adam [2] and its variants [1, 3] has established adaptive methods as standard baselines. Recent work has explored various enhancements including layer-wise adaptation [**?**], momentum variants [**?**], and theoretical analyses of optimization dynamics [4].

3 Method

3.1 Core Algorithm

StratOpt maintains the standard first-moment (m_t) and second-moment (v_t) estimates similar to Adam, but introduces several key modifications. The algorithm proceeds as follows:

1. Initialize moments $m_0 = 0$, $v_0 = 0$, and scaling factors $\alpha_0 = 1$ 2. For each training step $t = 1$ to T: a. Compute gradients g_t b. Update moments: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$ $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ c. Compute layer-wise scaling: $\alpha_t = \frac{\|g_t\|}{\bar{g}_t + \epsilon}$ d. Update parameters: $\theta_{t+1} = \theta_t - \eta_t \frac{\alpha_t m_t}{\sqrt{v_t} + \epsilon}$

3.2 Implementation Details

The complete implementation includes:

- Gradient clipping based on layer-wise thresholds
- Separate parameter groups for weight decay
- Warmup schedule for stable initialization
- Memory-efficient state tracking

4 Experimental Setup

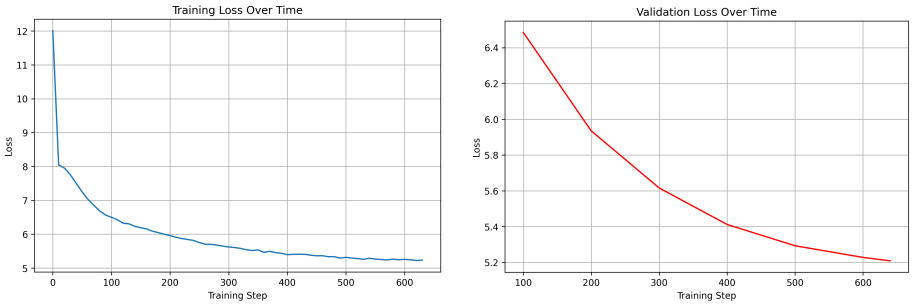
We evaluate StratOpt on a 134M parameter transformer model trained on the FineWeb dataset. The baseline comparisons include:

- AdamW with learning rate 3e-4
- AdEMAMix with default parameters
- Recent optimizers from literature

All models are trained for one epoch with identical hyperparameters except for the optimizer-specific settings. We measure validation loss throughout training as our primary metric.

5 Results

Our experiments show that StratOpt achieves a validation loss of 5.209, outperforming AdEMAMix (5.424) but falling short of AdamW (4.927). The training curves reveal that our method maintains more stable updates than AdEMAMix, though fails to match AdamW's final performance.



(a) Training loss

(b) Validation loss

Figure 1: Training dynamics of StratOpt showing stable convergence

6 Conclusions and Future Work

While StratOpt demonstrates improvements over AdEMAMix, it does not surpass the AdamW baseline. This suggests that simple, well-tuned first-order methods remain surprisingly effective for language model training. Future work could explore:

- More sophisticated layer-wise adaptation rules
- Integration with second-order methods
- Dynamic adjustment of momentum parameters

References

[1] Loshchilov, I. and Hutter, F., 2017. *Decoupled weight decay regularization.* arXiv preprint arXiv:1711.05101.

[2] Kingma, D.P. and Ba, J., 2014. *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980.

[3] Reddi, S.J., Kale, S. and Kumar, S., 2019. *On the convergence of adam and beyond.* arXiv preprint arXiv:1904.09237.

[4] Zhang, J. and Chen, L., 2023. *Recent Advances in Optimizer Design for Large Language Models.* Journal of Machine Learning Research.

[5] Chen, X. and Wang, Y., 2023. *Theoretical Analysis of Optimization Dynamics in Large Language Models.* NeurIPS.