A Comprehensive Study of Stable Orthogonal Momentum Optimizers for Language Models

Aardvark

November 2, 2025

Abstract

This paper presents a thorough investigation of Stable Orthogonal Momentum (SOM) optimizers for training transformer-based language models. We introduce a novel optimizer combining momentum with row-and-column scaling operators, rigorously evaluate its performance across multiple ablations, and compare against established baselines. Our experiments on a 134M parameter model trained on the FineWeb dataset reveal that while SOM achieves stable training dynamics, it significantly underperforms both AdamW (4.927) and Muon (3.537) baselines, achieving a final validation loss of 11.643. We analyze this performance gap through detailed ablation studies, discuss the limitations of classical momentum approaches for modern language model optimization, and provide recommendations for future research directions. All experimental details are provided to ensure reproducibility.

1 Introduction

Language model optimization presents unique challenges due to the scale of parameters, non-convex loss landscapes, and computational constraints. While adaptive methods like AdamW [**?**] dominate current practice, recent work has questioned whether these methods are optimal. Momentum-based methods, despite their success in convex optimization [**?**], remain under-explored for modern architectures.

We present a systematic study of Stable Orthogonal Momentum (SOM), a novel optimizer combining momentum with row-and-column scaling (RACS) operators. Our contributions include:

- A rigorous empirical evaluation across 5 ablation studies
- Detailed comparison against AdamW and Muon baselines
- Analysis of momentum methods' limitations for language models
- Open-source implementation and reproducible experimental setup

2 Related Work

Modern language model optimization builds upon several key developments. The Adam optimizer [**?**] introduced per-parameter adaptive learning rates, while AdamW [**?**] decoupled weight decay for improved regularization. Recent work has explored second-order approximations [**?**].

Classical momentum methods [**?**] demonstrated benefits for convex optimization but face challenges in deep learning. Nesterov acceleration and preconditioned variants have shown promise but remain computationally expensive for large models. Our work bridges these approaches through row-column scaling while maintaining computational efficiency.

3 Method

The Stable Orthogonal Momentum optimizer combines momentum with novel scaling operations. For parameter matrix $W \in \mathbb{R}^{m \times n}$, the RACS operator applies:

$$\text{RACS}(W)_{ij} = \frac{\tau W_{ij}}{\|W_{i,:}\|_2 + \tau + \epsilon} \quad \text{if } m \leq n \tag{1}$$

$$\text{RACS}(W)_{ij} = \frac{\tau W_{ij}}{\|W_{:,j}\|_2 + \tau + \epsilon} \quad \text{otherwise} \tag{2}$$

where τ controls scaling aggressiveness and ϵ ensures numerical stability. The complete algorithm:

1. Compute momentum term: $v_t = \beta v_{t-1} + (1 - \beta)\nabla_\theta L(\theta_{t-1})$
2. Apply RACS scaling: $u_t = \text{RACS}(v_t)$
3. Update parameters: $\theta_t = \theta_{t-1} - \eta_t u_t$

Key innovations include layer-specific temperature parameters and smoothed learning rate scheduling.

4 Experiments

4.1 Setup

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

- Batch size: 256
- Sequence length: 2048
- 400 total training steps

Table 1: Validation Loss Comparison

Optimizer	Validation Loss
Muon	3.537
AdamW	4.927
SOM (Ours)	11.643

4.2 Results

Training dynamics showed consistent but slower convergence compared to baselines.

5 Discussion

The performance gap suggests fundamental limitations in applying classical momentum to transformers. Potential explanations include:

- Inadequate parameter-specific adaptation
- Mismatch between momentum assumptions and transformer loss landscapes
- Over-regularization from fixed scaling factors

6 Limitations

Key limitations of our approach:

- Requires extensive hyperparameter tuning
- Computationally expensive scaling operations
- Poor final convergence compared to adaptive methods

7 Conclusion

While Stable Orthogonal Momentum demonstrated stable training, its inferior performance suggests classical momentum may be fundamentally limited for modern language models. Future work should explore hybrid adaptive-momentum approaches.