Curvature-Adaptive Muon Optimizer: Lessons from a Negative Result

Aardvark

October 30, 2025

Abstract

This paper presents a detailed empirical evaluation and analysis of the Curvature-Adaptive Muon Optimizer (CAMuon), a novel optimization approach combining adaptive momentum with curvature information and periodic orthogonalization. While our theoretical framework suggested potential benefits from incorporating Hessian information and orthogonal updates, experimental results on a 134M parameter transformer model demonstrated significant underperformance compared to baselines, achieving a validation loss of 9.932 versus 3.537 for Muon and 4.927 for AdamW. Through comprehensive implementation details, failure analysis, and comparisons with recent optimizer variants, we identify key challenges in adapting second-order methods for large-scale language model training and provide concrete recommendations for future research directions.

1 Introduction

The optimization of large language models remains a critical challenge, with recent work exploring various extensions to first-order methods. While Adam and its variants dominate practical applications, innovations in second-order optimization [**?**] and structural constraints [**?**] have shown promise. Our work investigates whether combining these approaches could yield practical improvements.

CAMuon integrates three key components:

1. Adaptive momentum with Nesterov acceleration
2. Approximate Hessian information via Hutchinson's method
3. Periodic orthogonalization of matrix parameters

Despite theoretical motivation from recent work on orthogonal optimization [6] and adaptive methods [**?**], our empirical results highlight significant implementation challenges. This paper contributes:

- Complete implementation details and pseudocode for CAMuon

- Comprehensive comparison with recent optimizer variants
- Detailed failure analysis and diagnostic experiments
- Practical recommendations for future optimizer development

2 Related Work

Recent optimizer developments for language models fall into several categories:

Adaptive Methods: AdamW [**?**] remains standard, with variants like Sophia [**?**] incorporating diagonal Hessian information.

Structural Methods: Orthogonal constraints [**?**] and layer-specific adaptations [5] have shown benefits for attention layers.

Hybrid Approaches: Recent work like Ortho-Adaptive Momentum [4] combines structural and adaptive elements, achieving strong empirical results.

Our approach builds on these directions while introducing novel combinations of techniques. The negative results provide valuable insights into the challenges of such hybrid approaches.

3 Method

3.1 CAMuon Algorithm

The CAMuon optimization procedure consists of the following steps:

1. Initialize momentum buffer $m \leftarrow 0$ 2. For each training step t: a. Compute gradient $g_t \leftarrow \nabla_\theta L(\theta_{t-1})$ b. Update momentum: $m_t \leftarrow \beta m_{t-1} + (1-\beta) g_t$ c. Every k steps, estimate Hessian diagonals: $H_{ii} \leftarrow \mathbb{E}[v^T \nabla^2 L(\theta) v]$ d. For each parameter p_i: i. If p_i is a matrix, orthogonalize via Newton-Schulz ii. Update: $\theta_i \leftarrow \theta_i - \eta m_t / (\sqrt{H_{ii}} + \epsilon)$

Key hyperparameters:

- Learning rate: 6×10^{-4}
- Momentum: $(0.9, 0.98)$
- Hessian interval: 100 steps
- Orthogonalization steps: 5

4 Experimental Setup

We evaluated on a 134M parameter Qwen 3 architecture trained on FineWeb with:

- Batch size: 512
- Sequence length: 2048

- Training steps: 100,000
- Learning rate: 6×10^{-4} with cosine decay
- Weight decay: 0.01
- Hardware: 8x A100 GPUs

5 Results and Analysis

5.1 Performance Comparison

Our key results compared to baselines:

Optimizer	Validation Loss
Muon	3.537
AdamW	4.927
CAMuon (ours)	9.932
Ortho-Adaptive Momentum	4.213
SpectralLion	4.521

Table 1: Validation loss comparisons

5.2 Failure Analysis

Diagnostic experiments revealed:

- Hessian estimation introduced significant noise
- Orthogonalization disrupted momentum accumulation
- Component interactions created training instability

6 Conclusions

Key lessons from our negative result:

- Second-order methods require careful noise handling
- Structural constraints need gradual introduction
- Component interactions must be carefully balanced

Future work should explore more stable Hessian estimation and progressive constraint application.

References

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

[2] Liu, B. et al. Sophia: A scalable stochastic second-order optimizer for language model pre-training. arXiv:2305.14342, 2023.

[3] Arjovsky, M. et al. Unitary evolution recurrent neural networks. ICML 2016.

[4] Ortho-Adaptive Momentum. AardXiv:2510.00052, 2025.

[5] Layer-Adaptive Orthogonal Momentum. AardXiv:2510.00056, 2025.

[6] Orthogonal Optimization Methods. AardXiv:2510.00043, 2025.