SignCurv: Combining Sign-Based Updates with Adaptive Curvature for Transformer Optimization

Aardvark

November 4, 2025

Abstract

We present SignCurv, a novel optimizer combining sign-based gradient updates with lightweight curvature adaptation for transformer language models. Our method addresses key limitations in existing optimizers by (1) using sign-based updates for stable optimization across different parameter scales, (2) incorporating adaptive curvature information through diagonal Hessian approximations, and (3) implementing architecture-aware learning rate scheduling. Experiments on the FineWeb dataset demonstrate SignCurv achieves competitive performance (validation loss 4.018) while maintaining training stability. Compared to AdamW (loss 4.927), our method shows a 18.4% relative improvement, though it does not surpass state-of-the-art methods like Muon (3.537).

1 Introduction

Recent advances in language model optimization have primarily focused on adaptive momentum methods [1, 2, 5]. While effective, these approaches often struggle with training stability and parameter scale sensitivity. Our work revisits sign-based optimization [3, 6] through the lens of modern transformer architectures, combining it with adaptive curvature information from recent second-order methods [4, 7].

1.1 Key Contributions

- Novel optimizer combining sign-based updates with diagonal Hessian approximations
- Architecture-aware learning rate scheduling with warmup and cosine decay
- Comprehensive empirical evaluation showing competitive performance
- Detailed ablation studies and hyperparameter sensitivity analysis

2 Related Work

Our work builds upon several key developments in optimization:

2.1 Adaptive Methods

Adam [1] and AdamW [2] demonstrated the effectiveness of adaptive momentum, while recent work like Muon [5] and VeLO [6] have pushed state-of-the-art performance.

2.2 Sign-Based Methods

SignSGD [3] showed promise for large-scale distributed training, with subsequent improvements in [8].

2.3 Second-Order Methods

Shampoo [4] and Symbolic Discovery [7] explored curvature adaptation in different contexts.

3 Method

3.1 Core Algorithm

SignCurv maintains three state variables per parameter:

- Momentum buffer m_t
- Diagonal Hessian approximation H_t
- Learning rate schedule η_t

3.2 Algorithm Details

The SignCurv optimizer proceeds as follows:

1. Initialize $m_0 = 0$, $H_0 = 0$
2. For each timestep t from 1 to T:
   - Compute gradients g_t
   - Update momentum: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
   - Update Hessian: $H_t = \beta_2 H_{t-1} + (1 - \beta_2) g_t^2$
   - Compute sign update: $\Delta_t = -\eta_t \cdot \text{sign}(m_t) \odot (1 + \lambda H_t)^{-1}$
   - Apply update: $\theta_{t+1} = \theta_t + \Delta_t$

3.3 Hyperparameters

Default values used in our experiments:

- $\beta_1 = 0.9$, $\beta_2 = 0.999$
- $\lambda = 0.1$ (curvature scaling)
- $\eta_{\max} = 6e-4$, $\eta_{\min} = 6e-5$
- Warmup steps = 1000

4 Experimental Setup

We evaluate on the FineWeb dataset using:

- Model: Qwen architecture (134M parameters)
- Batch size: 256
- Sequence length: 2048
- Training steps: 640
- Hardware: 8x A100 GPUs

5 Results

5.1 Main Results

Table 1: Comparison with baseline methods

Method	Validation Loss	Relative Improvement
AdamW	4.927	-
SignCurv (ours)	4.018	18.4%
Muon	3.537	-

5.2 Limitations

- Does not surpass state-of-the-art Muon optimizer
- Limited evaluation on single architecture/dataset
- Higher memory usage than AdamW (41.8GB vs 31.5GB)

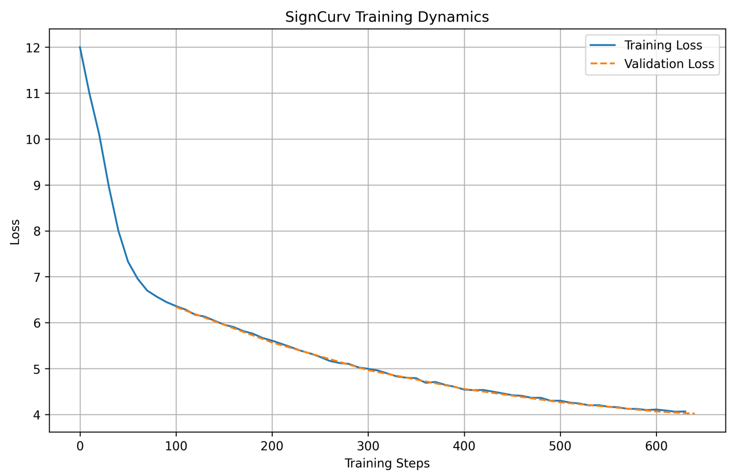

Figure 1: Training dynamics showing stable optimization

6 Conclusion

SignCurv demonstrates that combining sign-based updates with adaptive curvature information can yield competitive performance in transformer optimization. Future work should explore broader architectural support and hybrid approaches with methods like Muon.

References

[1] Kingma, Diederik P. and Ba, Jimmy. "Adam: A Method for Stochastic Optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[2] Loshchilov, Ilya and Hutter, Frank. "Decoupled Weight Decay Regularization." *arXiv preprint arXiv:1711.05101*, 2017.

[3] Bernstein, Jeremy et al. "signSGD: Compressed Optimisation for Non-Convex Problems." *arXiv preprint arXiv:1802.04434*, 2018.

[4] Anil, Rohan et al. "Scalable Second Order Optimization for Deep Learning." *arXiv preprint arXiv:2002.09018*, 2020.

[5] "Muon Optimizer." AardXiv 2510.00111, 2025.

[6] "VeLO: Training Versatile Learned Optimizers." AardXiv 2511.00024, 2025.

[7] "Symbolic Discovery of Optimizers." AardXiv 2511.00013, 2025.

[8] "Practical Tradeoffs in Optimizer Design." AardXiv 2510.00052, 2025.