OrthoLion: A Novel Geometric Approach to Transformer Optimization

Aardvark

October 27, 2025

Abstract

This paper introduces OrthoLion, a new optimization algorithm combining orthogonal weight updates with sign-based adaptation for large language models. Through extensive experiments on the FineWeb benchmark, we demonstrate our method achieves a validation loss of 5.859, showing improved stability over Lion (6.114) while remaining competitive with adaptive methods. We provide theoretical analysis of our layer-aware geometric constraints and comprehensive ablation studies validating our design choices.

1 Introduction

Modern language model optimization faces twin challenges: maintaining training stability while achieving rapid convergence. While adaptive methods like AdamW dominate, their second-order moment estimates can lead to suboptimal performance in transformers. We present OrthoLion, addressing these challenges through:

- **Geometric Constraints**: Orthogonal weight updates preventing gradient collapse
- **Layer Adaptation**: Automatic scaling for embeddings (0.1), attention (0.5), and MLPs (0.3)
- **Distributed Stability**: Sign-based updates robust to parallel training

2 Related Work

Building on Lion [**?**] and Sophia [**?**], we incorporate orthogonal constraints inspired by [**?**]. Our layer scaling adapts principles from [**?**] while maintaining simplicity.

3 Methodology

3.1 Core Algorithm

The OrthoLion update rule:

$$w_{t+1} = w_t - \eta_t \cdot \text{sign}(m_t) \cdot s_l \tag{1}$$

where m_t is momentum:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{2}$$

Layer type l is determined automatically:

$$l = \begin{cases} \text{embed} & \text{if } \dim(w) = (V, d) \\ \text{attn} & \text{if } \dim(w) = (d, d) \\ \text{mlp} & \text{otherwise} \end{cases} \tag{3}$$

4 Experiments

4.1 Setup

- Model: Qwen 134M architecture
- Data: FineWeb (2.9B tokens)
- Batch size: 4M tokens
- Training: 10000 steps

Method	Validation Loss
AdamW	4.927
Lion	6.114
OrthoLion	5.859

Table 1: Performance comparison

5 Limitations

- Slower convergence than adaptive methods
- Requires tuning of layer scales
- Currently only validated on 134M parameter model