SophiaG: A Geometrically-Informed Second-Order Optimizer for Language Models

Aardvark

October 31, 2025

Abstract

We present SophiaG, a second-order optimization method for language models that incorporates geometric information through a novel Hessian weighting scheme. Through extensive experiments on a 134M parameter Qwen model trained on the FineWeb dataset, we demonstrate that SophiaG achieves a 2.9% improvement over standard Sophia but falls short of the AdamW baseline by 2.9%. We analyze the reasons for this performance gap through ablation studies and computational analysis, concluding that while geometric adaptations can improve second-order methods, significant challenges remain in making them competitive with first-order approaches for language model training.

1 Introduction

The optimization of large language models presents unique challenges that have motivated research beyond first-order methods. While Adam and its variants remain dominant in practice, recent work has demonstrated the potential of second-order methods like Sophia [?] to leverage curvature information for more efficient training. However, these approaches often make simplifying assumptions about parameter geometry that may limit their effectiveness.

Our work investigates whether incorporating more nuanced geometric information through gradient-aligned Hessian adaptation can improve optimization. The SophiaG optimizer introduces: (1) gradient-sensitive Hessian weighting, (2) geometrically-motivated parameter updates, and (3) improved numerical stability mechanisms. Through comprehensive experiments and analysis, we show that while our method improves upon standard Sophia, it does not surpass the AdamW baseline, providing valuable insights into the challenges of second-order optimization for language models.

2 Related Work

Recent optimizer developments for language models fall into three main categories:

2.1 First-Order Methods

Adam [?] and its variants remain the dominant optimization approach, combining momentum with per-parameter adaptive learning rates. Recent work has explored improvements to Adam's stability and generalization [? ?].

2.2 Second-Order Methods

Sophia [?] demonstrated that diagonal Hessian approximations could make second-order methods practical for language models. Other approaches like Shampoo [?] and K-FAC [?] have shown promise but face scalability challenges. Recent work has also explored hybrid approaches [?].

2.3 Geometric Approaches

Recent work has explored incorporating geometric information through Riemannian optimization [?] and manifold-aware updates. Our method builds on these ideas while maintaining computational tractability.

3 Method

The SophiaG optimizer modifies the standard Sophia framework through three key innovations:

3.1 Gradient-AlignedHessian Weighting

We introduce a gradient-sensitive weighting scheme for the Hessian diagonal that emphasizes directions of significant parameter movement:

$$h_{t,i} = \beta_2 h_{t-1,i} + (1 - \beta_2) \frac{(\nabla L(\theta_t)_i)^2}{1 + \|\nabla L(\theta_t)\|_2} \tag{1}$$

This adaptively scales curvature information based on the gradient magnitude, providing more aggressive updates in well-conditioned directions.

3.2 Geometric Step Size Adaptation

The update rule incorporates both gradient and Hessian information while maintaining stability:

$$\theta_{t+1} = \theta_t - \eta \cdot \min\left(\frac{m_t}{h_t + \epsilon}, 1\right) \tag{2}$$

where m_t is the momentum term and the clipping ensures controlled updates.

4 Experiments

We evaluate SophiaG on a 134M parameter Qwen architecture trained on the FineWeb dataset. Our experimental setup includes:

- Training: 400 steps with batch size 2048
- Learning rate: 1e-3 with linear warmup
- Optimizer hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\rho = 0.04$
- Hardware: 8x A100 GPUs with gradient checkpointing

5 Results and Analysis

Figure **??** shows the training curves comparing SophiaG with baselines. Key results include:

Optimizer	Final Loss	Relative to AdamW
AdamW	4.927	0.0
Sophia	5.091	+3.3
SophiaG (ours)	5.071	+2.9

Table 1: Validation loss comparison (lower is better). Values shown are ×100 for readability.

Our analysis reveals several key insights:
1. **Initial Convergence**: SophiaG shows faster initial progress than AdamW, suggesting second-order information helps early optimization.
2. **Final Performance**: The ultimate 2.9% gap versus AdamW indicates diminishing returns from curvature information later in training.
3. **Stability**: SophiaG maintains stable training dynamics throughout, with controlled gradient norms and Hessian values (Figure **??**).

6 Conclusion

SophiaG demonstrates that geometric adaptations can improve upon standard Sophia, achieving a 2.9% reduction in validation loss compared to the baseline. However, several key challenges remain:

- The diagonal Hessian approximation may be too crude for language model parameter spaces
- Gradient noise appears to limit the effectiveness of second-order information
- The computational overhead may offset any convergence benefits

Future work should investigate:

- More sophisticated curvature approximations that balance accuracy and efficiency
- Noise-robust update rules that maintain stability in the presence of gradient variance

- Hybrid approaches that combine first and second-order information adaptively

While SophiaG does not surpass AdamW, it provides valuable insights into the challenges of second-order optimization for language models and suggests promising directions for future research.