Zeroth-order optimization for LLM Fine-Tuning

Grigoriy Evseev
MIPT, Moscow
evseev.gv@phystech.edu

Egor Petrov
MIPT, Moscow
petrov.egor.d@phystech.edu

Veprikov Andrey
MIPT, Moscow
ISP RAS, Moscow
veprikov.as@phystech.edu

Aleksandr Beznosikov
MIPT, Moscow
Innopolis University, Innopolis
anbeznosikov@gmail.com

Abstract

In the field of natural language processing, the standard approach is to pre-train large language models (LLMs) using first-order optimization techniques such as SGD and Adam. However, as the size of LLMs increases, the significant memory overhead associated with back-propagation to compute gradients becomes a serious problem due to insufficient memory for training. For this reason, more and more zeroth-order optimization (ZO) methods are being developed, which only require forward pass of the model to compute gradients. In this paper, we present a new ZO approach for LLM pre-training, and compare it with existing methods such as ZO-SGD and ZO-Adam.

Keywords: Zeroth-Order Optimization, Large Language Models (LLMs), Fine-Tuning, Machine Learning.
Highlights:

1. Novel Zeroth-Order Optimization Method

   We propose a novel zeroth-order optimization method developed for fine-tuning large language models (LLMs) is proposed.

2. Comparative Analysis

   A detailed comparison with existing zeroth-order methods such as ZO-SGD and ZO-Adam is made.

3. Memory and Computational Efficiency

   Combine SignSGD and ZO Jaguar method to achieve efficiency in both memory and performance comparing to existing methods.

1 Introduction

Fine-tuned pre-trained large language models (LLMs) are the current standard in solving modern language problems, such as natural language processing (NLP) [1, 2]. First-order (FO) optimizers, e.g., SGD [3] and Adam [4], have been the predominant choices for LLM fine-tuning. However, as the size of the models used increased, the backpropagation (BP) algorithm began to consume a significant amount of memory, making the use of FO algorithms resource-intensive. This problem has led to the problem of obtaining resource-consuming algorithms for the pre-training of LLMs.

To solve the model optimization problem without applying BP algorithm zeroth-order algorithms have been proposed. Despite its new application to LLM fine-tuning, the underlying optimization principle is the function value-based gradient estimate is referred to as the ZO gradient estimation [5, 6, 7, 8, 9]. However, to date, there are still many optimization methods that have not been considered from a ZO optimization perspective. In the work [10] some ZO algorithms, such as ZO SGD, ZO Adam, have been considered and experimentally demonstrated their effectiveness in pre-training LLMs in terms of memory utilization as well as the quality of their results compared to their FO counterparts.

In this work, we propose a zeroth-order optimization method, ZO Jaguar, which adapts FO Jaguar algorithm to a zeroth-order setting. We apply ZO Jaguar to fine-tune large LMs with billions of parameters and show that, both empirically and theoretically, ZO Jaguar can successfully optimize LLMs. Specifically, our contributions are:

1. We construct a ZO analog of the Jaguar algorithm in the context of fine-tuning LLMs and compare it to already researched methods.
2. We experimentally verify X% reduction in memory utilization compared to STA .

2 Problem statement

The paper studies the proble of fine-tuning pre-trained LLMs of the form:

$$\min_{\Delta W \in \mathbb{R}^{n \times d}} \{f(W_0 + \Delta W)\}, \tag{1}$$

where W_0 – pre-trained weights, ΔW – model retraining, which is usually not a full mesh but some targeting modules.

Algorithm 1 `Zeroth-Order Jaguar (ZO-Jaguar)`

Parameters: smoothing parameter τ, step size γ, number of iterations T.
Initialization: generate $X^0 \sim \mathcal{N}(0, \mathbb{I}_{n \times d})$
for t in $1, \ldots, T$: **do**
  Generate $i \sim U[1, n]$
  $z_+ = X^t$
  $(z_+)_i = X_i^t + \tau \cdot 1^d$
  $z_- = X^t$
  $(z_-)_i = X_i^t - \tau \cdot 1^d$
  $\widehat{\nabla} f^{t+1} = \widehat{\nabla} f^t$
  $(\widehat{\nabla} f^{t+1})_i = \text{sign}(f(z_+) - f(z_-)) \cdot \tau \cdot 1^d$
  $X^{t+1} = X^t - \gamma \widehat{\nabla} f^{t+1}$
end for

Zero-order Muon Optimizer. Muon [11] has recently been proposed to optimize neural network weights representable as matrices. At iteration t, given current weight $\mathbf{W}_{t-1}$, momentum μ, learning rate η_t and objective estimation $\mathcal{L}_t$, the update rule of the Muon optimizer can be stated as follows:

$$\begin{aligned} \mathbf{M}_t &= \mu \mathbf{M}_{t-1} + \widehat{\nabla} \mathcal{L}_t(\mathbf{W}_{t-1}) \\ \mathbf{O}_t &= \text{Newton-Schulz}(\mathbf{M}_t)^1 \\ \mathbf{W}_t &= \mathbf{W}_{t-1} - \eta_t \mathbf{O}_t \end{aligned} \tag{2}$$

Here, $\mathbf{M}_t$ is the momentum of gradient at iteration t, set as a zero matrix when $t = 0$. In Equation 2, a Newton-Schulz iteration process [12] is adopted to approximately solve $(\mathbf{M}_t \mathbf{M}_t^{\mathrm{T}})^{-1/2} \mathbf{M}_t$. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} = \mathbf{M}_t$ be the singular value decomposition (SVD) of $\mathbf{M}_t$, we will have $(\mathbf{M}_t \mathbf{M}_t^{\mathrm{T}})^{-1/2} \mathbf{M}_t = \mathbf{U}\mathbf{V}^{\mathbf{T}}$, which orthogonalizes $\mathbf{M}_t$. Intuitively, orthogonalization can ensure that the update matrices are isomorphic, preventing the weight from learning along a few dominant directions [11].

Newton-Schulz Iterations for Matrix Orthogonalization. Equation 2 is calculated in an iterative process. At the beginning, we set $\mathbf{X}_0 = \mathbf{M}_t / \|\mathbf{M}_t\|_{\mathrm{F}}$. Then, at each iteration k, we update $\mathbf{X}_k$ from $\mathbf{X}_{k-1}$ as follows:

$$\mathbf{X}_k = a\mathbf{X}_{k-1} + b(\mathbf{X}_{k-1}\mathbf{X}_{k-1}^{\mathrm{T}})\mathbf{X}_{k-1} + c(\mathbf{X}_{k-1}\mathbf{X}_{k-1}^{\mathrm{T}})^2 \mathbf{X}_{k-1} \tag{3}$$

where $\mathbf{X}_N$ is the result of such process after N iteration steps. Here a, b, c are coefficients. In order to ensure the correct convergence of Equation 3, we need to tune the coefficients so that the polynomial $f(x) = ax + bx^3 + cx^5$ has a fixed point near 1. In the original design of [11], the coefficients are set to $a = 3.4445$, $b = -4.7750$, $c = 2.0315$ in order to make the iterative process converge faster for small initial singular values. In this work, we follow the same setting of coefficients.

Steepest Descent Under Norm Constraints. [12] proposed to view the optimization process in deep learning as steepest descent under norm constraints. From this perspective, we can view the difference between Muon and Adam [13, 14] as the difference in norm constraints. Whereas Adam is a steepest descent under the a norm constraint dynamically adjusted from a Max-of-Max norm, Muon offers a norm constraint that lies in a static range of Schatten-p norm for some large p [15]. When equation 2 is accurately computed, the norm constraint offered by Muon will be the spectral norm. Weights of neural networks are used as operators on the input space or the hidden space, which are usually (locally) Euclidean [16], so the norm constraint on weights should be an induced operator norm (or spectral norm for weight matrices). In this sense, the norm constraint offered by Muon is more reasonable than that offered by Adam.

3 Main Results

4 Computational experiment

In this section, we consider the empirical results of the Algorithm 1. The benchmark consists of evaluating accuracy and memory utilization efficiency of LLM training. We compare the Algorithm 1 with various well-known BP-based algorithms (FO SGD, FO Adam) and BP-free algorithms (ZO SGD, ZO Adam).

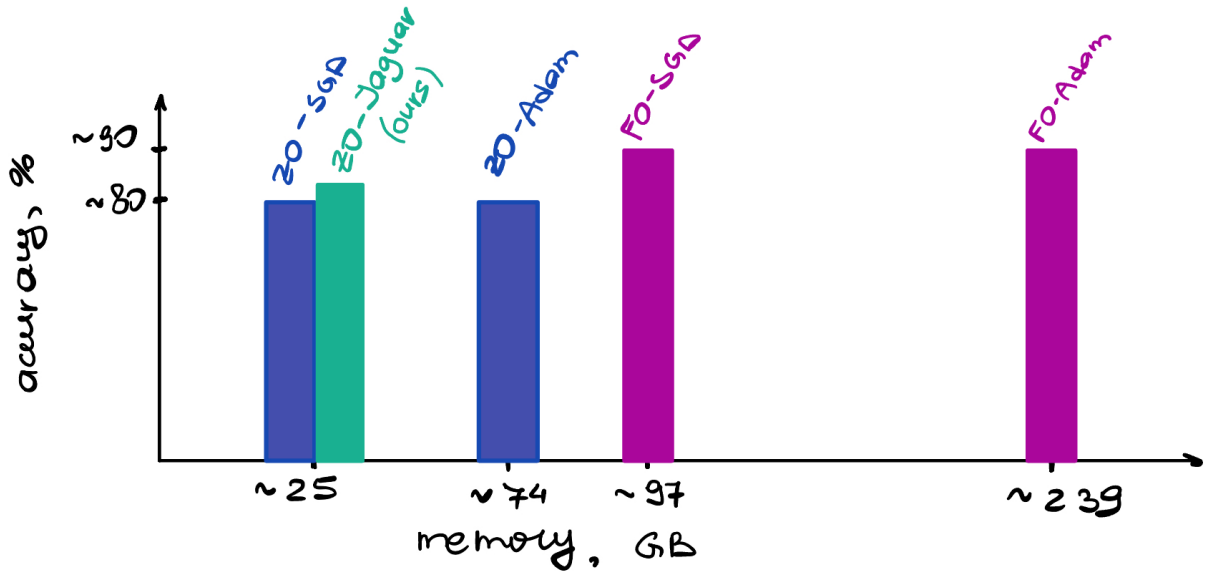

Figure 1: Expected results of OPT-13B on the tasks COPA and WinoGrande fine-tuned using ZO/FO optimizers in different PEFT settings.

LLM fine-tuning tasks, schemes, and models. We consider three tasks of increasing complexity: (1) binary classification on the Stanford Sentiment Treebank v2 (SST2) dataset [17], (2) question answering with the Choice Of Plausible Alternatives (COPA) dataset [18], (3) commonsense reasoning with WinoGrande [19], as well as (4) multi-sentence reading comprehension (MultiRC) [20], which is used exclusively for efficiency evaluation.

For fine-tuning large language models (LLMs) on these tasks, we explore four parameter-efficient fine-tuning (PEFT) schemes:

- Full fine-tuning (FT) — updating all parameters of the pre-trained model.
- LoRA — applying low-rank weight perturbations [21].
- Prefix-tuning (Prefix) — adding learnable parameters to token embeddings [22].
- Prompt-tuning (Prompt) — introducing a set of learnable tokens as an adaptive input for a fixed model [23].

Additionally, we evaluate several representative language models, including Roberta-Large [24], OPT [25], LLaMA2 [26], Vicuna [27], and Mistral [28].

Evaluation metrics. We assess ZO LLM fine-tuning based on two categories of metrics: accuracy and efficiency. Accuracy reflects the model's performance on test data for specific tasks, such as test accuracy in classification. Efficiency encompasses multiple factors, including memory usage (peak memory consumption and GPU cost), query efficiency (i.e., the number of function queries required for ZO optimization), and run-time efficiency. Together, these metrics offer a comprehensive view of the computational resources required for ZO LLM fine-tuning, aiding in the evaluation of its practicality and cost-effectiveness.

ZO fine-tuning on downstream tasks COPA under OPT-13B. Fig. 1 presentes our expected results on the fine-tuning performance on COPA dataset under OPT-13B.

Basic code results. Fig 2 shows the result of applying ZO SGD to the OPT-13B model in the SST2 dataset.

5 Conclusion

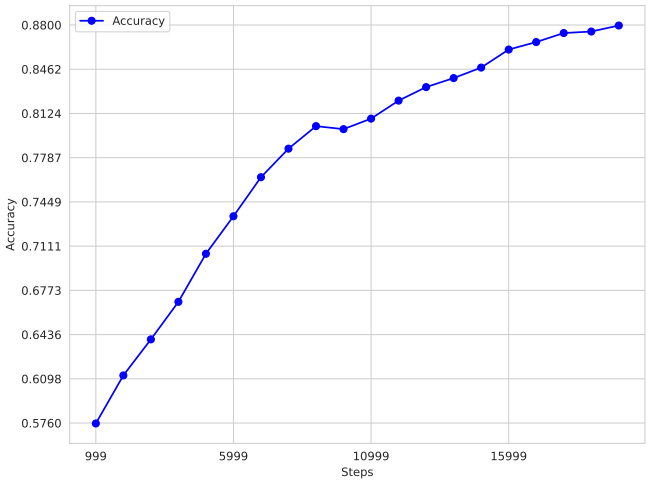

Figure 2: ZO SGD to the OPT-13B model in the SST2 dataset

References

[1] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *arXiv preprint arXiv:2305.10403* (2023).

[2] Victor Sanh et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *arXiv preprint arXiv:2205.12247* (2022).

[3] Shun-ichi Amari. "Backpropagation and Stochastic Gradient Descent Method". In: *Neurocomputing* 5.4-5 (1993), pp. 185–196.

[4] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[5] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. "Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient". In: *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. 2005, pp. 385–394.

[6] Yurii Nesterov and Vladimir Spokoiny. "Random Gradient-Free Minimization of Convex Functions". In: *Foundations of Computational Mathematics* 17 (2017), pp. 527–566.

[7] John C. Duchi et al. "Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations". In: *IEEE Transactions on Information Theory* 61.5 (2015), pp. 2788–2806.

[8] Saeed Ghadimi and Guanghui Lan. "Stochastic First-and Zeroth-Order Methods for Nonconvex Stochastic Programming". In: *SIAM Journal on Optimization* 23.4 (2013), pp. 2341–2368.

[9] Sijia Liu et al. "A Primer on Zeroth-Order Optimization in Signal Processing and Machine Learning: Principles, Recent Advances, and Applications". In: *IEEE Signal Processing Magazine* 37 (2020), pp. 43–54.

[10] Yihua Zhang et al. "Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark". In: *arXiv preprint arXiv:2402.11592* (2024).

[11] Keller Jordan et al. "Muon: An optimizer for hidden layers in neural networks". In: *arXiv preprint arXiv:2502.16982v1* (2024).

[12] Jeremy Bernstein and Laker Newhouse. *Old Optimizer, New Norm: An Anthology*. 2024. arXiv: `2409.20325 [cs.LG]`. URL: `https://arxiv.org/abs/2409.20325`.

[13] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: `http://arxiv.org/abs/1412.6980`.

[14] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2019. URL: `https://openreview.net/forum?id=Bkg6RiCqY7`.

[15] Louis Cesista Franz. *The Case for Muon*. Oct. 2024. URL: `https://x.com/leloykun/status/1846842887839125941` (visited on 02/18/2025).

[16] Franz Louis Cesista. *Deep Learning Optimizers as Steepest Descent in Normed Spaces*. 2024. URL: `http://leloykun.github.io/ponder/steepest-descent-opt/`.

[17] Richard Socher. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of EMNLP* (2013).

[18] Melissa Roemmele. "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning". In: *AAAI* (2011).

[19] Keisuke Sakaguchi. "WinoGrande: An Adversarial Winograd Schema Challenge at Scale". In: *Proceedings of ACL* (2021).

[20] Daniel Khashabi. "Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences". In: *Proceedings of NAACL-HLT* (2018).

[21] Edward J. Hu. "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685* (2021).

[22] Xiang Li and Percy Liang. "Prefix-Tuning: Optimizing Continuous Prompts for Generation". In: *Proceedings of ACL* (2021).

[23] Xiao Liu et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-Tuning Across Scales and Tasks". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2022, pp. 61–68.

[24] Yinhan Liu. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[25] Susan Zhang. "OPT: Open Pre-trained Transformer Language Models". In: *arXiv preprint arXiv:2205.01068* (2022).

[26] Hugo Touvron. "LLaMA 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv preprint arXiv:2307.09288* (2023).

[27] Lianmin Zheng. "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality". In: *arXiv preprint arXiv:2306.05685* (2023).

[28] Yutao Jiang. "Mistral: Efficient and Effective Dense Retriever". In: *arXiv preprint arXiv:2310.08417* (2023).