ZO Optimization for LLM Fine-Tuning

Grigoriy Evseev, Egor Petrov,
Veprikov Andrey, Aleksandr Beznosikov

MIPT

2025

Motivation

- Large Language Models (LLMs) require significant memory for training.
- Backpropagation in first-order methods (SGD, Adam) causes high memory overhead.
- Zeroth-Order (ZO) methods only need forward passes, reducing memory usage.

Problem statement

Given a scalar-valued function $f(x)$ where $x \in \mathbb{R}^d$, the RGE (referred to as $\widehat{\nabla} f(x)$ is expressed using difference

$$\widehat{\nabla} f(x) = \frac{1}{q} \sum_{i=1}^{q} \left[\frac{f(x + \mu u_i) - f(x - \mu u_i)}{2\mu} u_i \right]$$

where u_i is a random direction vector typically drawn from the standard Gaussian distribution $\mathcal{N}(0, 1)$, q is the number of function queries, and $\mu > 0$ is a small perturbation stepsize (also known as smoothing parameter).

Our Contribution

- We propose a new ZO optimization method for LLM fine-tuning.
- Compared with existing methods like ZO-SGD and ZO-Adam.

Fine-tuning task

As LLM fine-tuning tasks we focus on four tasks:

- **Binary Classification:** Stanford Sentiment Treebank v2 (SST2) Socher 2013
- **Question Answering:** Choice Of Plausible Alternatives (COPA) Roemmele 2011
- **Commonsense Reasoning:** WinoGrande Sakaguchi 2021
- **Multi-Sentence Reading Comprehension:** MultiRC (for efficiency evaluation) Khashabi 2018

Language Models

- **Roberta-Large** (Liu 2019)
- **OPT** (Zhang 2022)
- **LLaMA2** (Touvron 2023)
- **Vicuna** (Zheng 2023)
- **Mistral** (Jiang 2023)

Fine-tuning schemes

- **Full-Tuning (FT):** Fine-tunes the entire pre-trained model.
- **Low-Rank Adaptation (LoRA):** Imposes low-rank weight perturbations Hu 2021.
- **Prefix-Tuning (Prefix):** Appends learnable parameters to token embeddings Li и Liang 2021.

Related Work

- Sign Operator for Coping with Heavy-Tailed Noise (2025) *(arXiv)*
- An Accelerated Directional Derivative Method for Smooth Optimization (2020) *(arXiv)*
- Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning (2024) *(arXiv)*
- Fine-Tuning Language Models with Just Forward Passes (2024) *(arXiv)*
- Simultaneous Computation and Memory Efficient Zeroth-Order Optimizer (2024) *(arXiv)*