ZO Optimization for LLM Fine-Tuning

Grigoriy Evseev, Egor Petrov,
Veprikov Andrey, Aleksandr Beznosikov

MIPT

2025

Zero Order PEFT

- 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.

Table: Memory consumption comparison for **full finetuning** (FT) evaluated on OPT-13B model with the MultiRC dataset.

Optimizer	Empirical Mem.
FO-SGD	97 GB
FO-Adam	239 GB
ZO-SGD	51 GB
ZO-Adam	151 GB

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

ZO Jaguar (Sign)

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 $t = 1$ **to** T **do**
  Generate $i \sim \mathcal{U}[1, n]$
  $z_+ \leftarrow X^t$
  $(z_+)_i \leftarrow X_i^t + \tau \cdot \mathbf{1}^d$
  $z_- \leftarrow X^t$
  $(z_-)_i \leftarrow X_i^t - \tau \cdot \mathbf{1}^d$
  $\widehat{\nabla} f^{t+1} \leftarrow \widehat{\nabla} f^t$
  $(\widehat{\nabla} f^{t+1})_i \leftarrow \mathrm{sign}(f(z_+) - f(z_-)) \cdot \tau \cdot \mathbf{1}^d$
  $X^{t+1} \leftarrow X^t - \gamma \widehat{\nabla} f^{t+1}$
end for

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, q is the number of function queries, and $\mu > 0$ is a smoothing parameter.

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 and 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)*