T TECHNOLOGIES

BRAInLab

Leveraging Coordinate Momentum in SignSGD and Muon: Memory-Optimized Zero-Order LLM Fine-Tuning

Egor Petrov[1], Grigoriy Evseev[1], Aleksey Antonov[1, 2], Andrey Veprikov[1], Pavel Plyusnin[2], Nikolay Bushkov[1, 2], Stanislav Moiseev[2], Aleksandr Beznosikov[1]

[1]Basic Research of Artificial Intelligence Laboratory (BRAIn Lab), [2]T-Technologies

Introduction

Fine-tuning pre-trained Large Language Models (LLMs) has become the standard technique in modern natural language processing, enabling rapid adaptation to diverse downstream tasks with minimal labeled data. The fine-tuning setup can be considered as a stochastic unconstrained optimization problem of the form

$$f^* := \min_{x\in\mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(x,\xi)\right]\right\}, \qquad (1)$$

where x are parameters of the fine-tuned LLM, $\mathcal{D}$ is the data distribution available for training, and $f(x,\xi)$ is the loss on data point ξ.

The most memory-efficient methods are based on the Zero-Order (ZO) optimization technique, which avoids backpropagation entirely by estimating gradients using only forward passes. To estimate gradients, authors use finite differences:

$$\nabla f(x,\xi) \approx \frac{f(x+\tau e,\xi) - f(x-\tau e,\xi)}{2\tau} e, \qquad (2)$$

Our key contributions are as follows:

- We provide the first convergence analysis in the stochastic non-convex setting for zero-order SignSGD with momentum (Algorithm 1 and Theorem 1), requiring only $2d+1$ parameters and $\mathcal{O}(1)$ ZO oracle calls per iteration.
- We extend our memory-efficient momentum method to the Muon algorithm (Algorithm 2), introducing the first zero-order variant of Muon that preserves memory efficiency. We also establish its convergence rate in the stochastic non-convex setting.
- We empirically evaluate the proposed zero-order methods on challenging LLM fine-tuning benchmarks, demonstrating their effectiveness and practical relevance.

Theoretical Foundations

Assumption 1 (*Smoothness*)

The functions $f(x,\xi)$ are $L(\xi)$-smooth on the $\mathbb{R}^d$ with respect to the Euclidean norm $\|\cdot\|$, i.e., for all $x, y \in \mathbb{R}^d$ it holds that $\|\nabla f(x,\xi) - \nabla f(y,\xi)\|_2 \le L(\xi)\|x-y\|_2$. We also assume that exists constant $L^2 := \mathbb{E}\left[L(\xi)^2\right]$.

Assumption 2 (*Bounded variance*)

The variance of the $\nabla f(x,\xi)$ is bounded with respect to the Euclidean norm, i.e., there exists $\sigma > 0$, such that for all $x \in \mathbb{R}^d$ it holds that $\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \le \sigma^2$.

Assumption 3 (*Bounded oracle noise*)

The noise in the oracle is bounded with respect to the Euclidean norm, i.e., there exists $\Delta > 0$, such that for all $x \in \mathbb{R}^d$ it holds that $\mathbb{E}\left[\left|\hat{f}(x,\xi) - f(x,\xi)\right|^2\right] \le \Delta^2$.

Results

Algorithm 1: `JAGUAR SignSGD`

Parameters: stepsize γ, momentum β, gradient approximation parameter τ, number of iterations T.
Initialization: choose $x^0 \in \mathbb{R}^d$ and $m^{-1} = \mathbf{0} \in \mathbb{R}^d$.
for $t = 0, 1, 2, \dots, T$ **do**
  Sample $i_t \sim \text{Uniform}(\overline{1,d})$
  Set one-hot vector e^t with 1 in the i_t coordinate
  Sample stochastic variable $\xi^t \sim \mathcal{D}$
  Compute $\widetilde{\nabla}_{i_t} f(X^t,\xi^t) := \frac{f_+ - f_-}{2\tau} \in \mathbb{R}$,
  where $f_+ = \hat{f}(X^t + \tau E^t, \xi^t)$, $f_- = \hat{f}(X^t - \tau E^t, \xi^t)$
  Set $m_{i_t}^t = \beta m_{i_t}^{t-1} + (1-\beta)\widetilde{\nabla}_{i_t} f(x^t,\xi^t)$
  and $m_{i\neq i_t}^t = m_{i\neq i_t}^{t-1}$ for all $i \in \overline{1,d}$
  Set $x^{t+1} = x^t - \gamma \cdot \text{sign}(m^t)$
end for
Return: $x^{N(T)}$, where $N(T) \sim \text{Uniform}(\overline{1,T})$.

Theorem 1

Consider Assumptions 1, 2 and 3. Then `JAGUAR SignSGD` (Algorithm 1) has the following convergence rate:

$$\mathbb{E}\left[\left\|\nabla f\left(x^{N(T)}\right)\right\|_1\right] = \mathcal{O}\left[\frac{\delta_0}{\gamma T} + \frac{d\left\|\nabla f(x^0)\right\|_2}{T\sqrt{1-\beta}} + \frac{d^2 L\gamma}{1-\beta} + \sqrt{1-\beta}\, d\sigma + dL\tau + \frac{d\Delta}{\tau}\right],$$

where we used a notation $\delta_0 := f(x^0) - f^*$.

Algorithm 2: `JAGUAR Muon`

Parameters: γ (stepsize), β (momentum), τ (grad. approx.),
  ns_steps (Newton-Schulz steps), T (iterations).
Init: $X^0 \in \mathbb{R}^{m\times n}$, $M^{-1} = \mathbf{0}_{m\times n}$.
for $t = 0$ **to** T **do**
  Sample $i_t \sim U(\overline{1,m})$, $j_t \sim U(\overline{1,n})$
  $E^t \leftarrow \text{one-hot}(i_t, j_t)$
  Sample $\xi^t \sim \mathcal{D}$
  $\widetilde{\nabla}_{i_t,j_t} f(X^t,\xi^t) \leftarrow \frac{f_+ - f_-}{2\tau}$, where:
    $f_+ = \hat{f}(X^t + \tau E^t, \xi^t)$, $f_- = \hat{f}(X^t - \tau E^t, \xi^t)$
  $M_{i_t,j_t}^t \leftarrow \beta M_{i_t,j_t}^{t-1} + (1-\beta)\widetilde{\nabla}_{i_t,j_t} f(x^t,\xi^t)$
  $M_{i\neq i_t, j\neq j_t}^t \leftarrow M_{i\neq i_t, j\neq j_t}^{t-1}$
  $X^{t+1} \leftarrow X^t - \gamma \cdot \texttt{Newton\_Schulz}(M^t, K = \text{ns\_steps})$
end for
Return: $X^{N(T)}$, $N(T) \sim U(\overline{1,T})$.

Subroutine `Newton_Schulz`$(A \in \mathbb{R}^{m\times n}, K = 10)$:
  $A^0 \leftarrow A/\|A\|_F$
  for $k = 0$ **to** K **do**
    $A^{k+1} \leftarrow \frac{3}{2}A^k - \frac{1}{2}A^k(A^k)^T A^k$
  end for
  Return $A^K \approx U_A V_A^T$

Theorem 2

Consider Assumptions 1, 2 (with Frobenius norm) and 3. Then `JAGUAR Muon` (Algorithm 2) has the following convergence rate:

$$\mathbb{E}\left[\left\|\nabla f\left(X^{N(T)}\right)\right\|_{\mathcal{S}_1}\right] = \mathcal{O}\left[\frac{\delta_0}{\gamma T} + \frac{m^{1/2} n\left\|\nabla f(X^0)\right\|_2}{T\sqrt{1-\beta}} + \frac{m^{3/2} n^2\gamma}{1-\beta} + \sqrt{1-\beta}\, m^{1/2} n\sigma + m^{1/2} nL\tau + \frac{m^{1/2} n\Delta}{\tau}\right],$$

where we used a notation $\delta_0 := f(x^0) - f^*$. We also assume that $n \le m$.

Ablation study

Figure 1 reports the accuracy of the `JAGUAR SignSGD` method on the SST-2 dataset with the RoBERTa-large model across different values of β. The method demonstrates substantially lower accuracy for small β, while attaining robust and consistently high performance around $\beta \approx 0.9$.

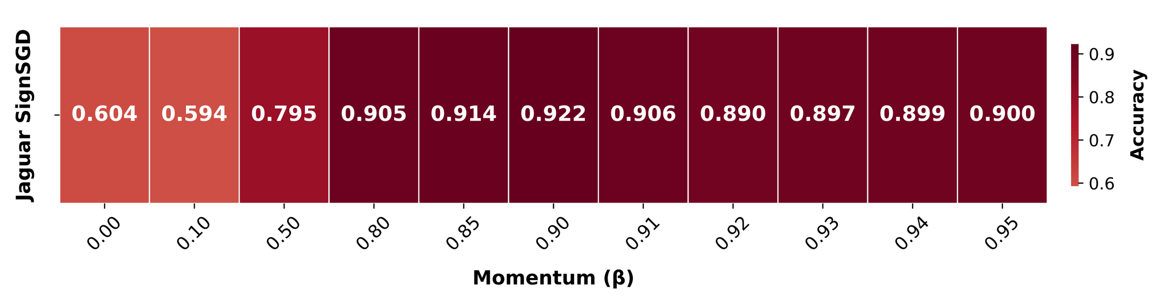

Figure 1: Test accuracy of `JAGUAR SignSGD` on SST-2 for RoBERTa-large with LoRA for different values of β.

Experiments

Test accuracy on SST2 for OPT-1.3B and RoBERTa-Large with FT and LoRA. Best performance among ZO methods is in **bold**. Blue indicates outperformance of all baseline ZO methods, red indicates matching or exceeding FO-SGD.

Method	OPT-1.3B		RoBERTa-Large	
	FT	LoRA	FT	LoRA
FO-SGD	91.1	93.6	91.4	91.2
Forward-Grad	90.3	90.3	90.1	89.7
ZO-SGD	90.8	90.1	89.4	90.8
Acc-ZOM	85.2	91.3	89.6	90.9
ZO-SGD-Cons	88.3	90.5	89.6	91.6
ZO-SignSGD	87.2	91.5	52.5	90.2
ZO-AdaMM	84.4	92.3	89.8	89.5
LeZO	85.1	92.3	90.4	91.8
`JAGUAR SignSGD`	**94.0 ± 0.1**	92.5 ± 0.5	**92.2 ± 0.2**	**92.2 ± 0.4**
`JAGUAR Muon`	84.0 ± 0.1	**94.0 ± 0.1**	85.0 ± 0.1	**92.2 ± 0.2**
`ZO-Muon`	86.5 ± 0.1	93.5 ± 0.1	72.0 ± 0.1	86.0 ± 0.2

Test accuracy on COPA and WinoGrande for OPT-13B and Llama2-7B with LoRA. Best performance among ZO methods is in **bold**. Blue indicates outperformance of all baseline ZO methods, red indicates matching or exceeding FO-SGD.

Method	OPT-13B		LLaMA2-7B	
	COPA	WinoGrande	COPA	WinoGrande
FO-SGD	88	66.9	85	66.9
Forward-Grad	89	62.9	82	64.3
ZO-SGD	87	62.6	86	64.3
ZO-SGD-Cons	88	63.3	85	64.6
`JAGUAR SignSGD`	**89 ± 0.3**	**63.7± 0.1**	**88 ± 0.2**	**64.9± 0.1**
`JAGUAR Muon`	87 ± 0.2	62.3 ± 0.2	**88 ± 0.1**	62.8 ± 0.2
`ZO-Muon`	87 ± 0.2	61.9 ± 0.3	85 ± 0.2	61.6 ± 0.2

GPU allocated memory (GB)

Method	FT Memory	LoRA Memory
FO-SGD	12.246	5.855
ZO-SGD	4.171	4.125
ZO-AdaMM	13.046	6.132
`JAGUAR SignSGD`	4.172	4.128
`JAGUAR Muon`	4.179	4.132
`ZO-Muon`	4.177	4.130

GPU allocated memory (GB)

Model	Llama-7B	OPT-13B
COPA		
ZO-SGD	13.219	24.710
ZO-AdaMM	27.971	38.612
`JAGUAR SignSGD`	13.219	24.712
`ZO-Muon`	15.021	25.740
`JAGUAR Muon`	16.032	25.880
WinoGrande		
ZO-SGD	14.670	26.407
ZO-AdaMM	29.440	39.872
`JAGUAR SignSGD`	14.672	26.408
`ZO-Muon`	16.992	27.416
`JAGUAR Muon`	17.992	27.440