Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark

Yihua Zhang[1,*], Pingzhi Li[2,*], Junyuan Hong[3,*], Jiaxiang Li[4,*], Yimeng Zhang[1], Wenqing Zheng[3], Pin-Yu Chen[5], Jason D. Lee[6], Wotao Yin[7], Mingyi Hong[4], Zhangyang Wang[3], Sijia Liu[1,5,†], Tianlong Chen[2,8,9,†]

* Equal Contribution, † Correspondence

[1]Michigan State University, [2]UNC Chapel Hill, [3]UT Austin, [4]UMN Twin Cities, [5]IBM Research, [6]Princeton University, [7]DAMO Academy Alibaba Group US, [8]MIT, [9]Harvard University

ICML International Conference On Machine Learning

Paper Code

➢ Research Question

Can we establish a benchmark for ZO optimization in LLM fine-tuning, explore the overlooked optimization principles, and advance the current state of the art?

➢ Methods using Zeroth-Order Optimization

- Randomized gradient estimator (RGE):

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

- ZO-SGD: ZO stochastic gradient descent, *i.e.* MeZO [1].
- ZO-SGD-Sign: ZO-SGD using sign-based gradient estimation.
- ZO-SGD-MMT: ZO-SGD with momentum.
- ZO-SGD-Cons: ZO-SGD with conservative gradient update.
- ZO-Adam: ZO variant of the Adam optimizer.

➢ Task Alignment Plays A Key Role for ZO

- Task Alignment with the template <CLS>SENTENCE. It was [terrible|great].<SEP> for SST2 dataset and another template <CLS>SENTENCE1?[Yes|No], SENTENCE2.<SEP> for RTE.
- RoBERTa-large model full-tuned w/ and w/o task alignment.

Method	SST2			RTE		
	✓	✗	Difference	✓	✗	Difference
FO-SGD	91.6	91.5	0.1	70.9	61.4	9.5
ZO-SGD	89.4	79.2	**10.2**	68.7	60.4	**8.3**
ZO-Adam	89.8	79.2	**10.6**	69.2	58.7	**10.5**

Table 1: Test accuracy (%) of pre-trained Roberta-Large model fine-tuned on SST2 and RTE.

[1] Malladi et atl. "Fine-tuning language models with just forward passes"

➢ A Pilot Study: LLMs ZO Fine-Tuning on SST2

SST2	Roberta-Large				OPT-1.3B			
	FT	LoRA	Prefix	Prompt	FT	LoRA	Prefix	Prompt
FO-SGD	91.4	91.2	89.6	90.3	91.1	93.6	93.1	92.8
Forward-Grad	**90.1**	89.7	89.5	87.3	90.3	90.3	90.0	82.4
ZO-SGD	89.4	90.8	90.0	87.6	**90.8**	90.1	**91.4**	84.4
ZO-SGD-MMT	89.6	90.9	90.1	88.6	85.2	91.3	91.2	**86.9**
ZO-SGD-Cons	89.6	**91.6**	90.1	88.5	88.3	90.5	81.8	84.7
ZO-SGD-Sign	52.5	90.2	53.6	86.1	87.2	91.5	89.5	72.9
ZO-Adam	89.8	89.5	**90.2**	**88.8**	84.4	**92.3**	**91.4**	75.7

Table 2: Results of Roberta-Large and OPT-1.3B tuned on SST2.

Takeaway I: **ZO-Adam** seems to be the **most effective ZO method**: achieving the best performance in 4 out of 8 fine-tuning settings.

Takeaway II: **Forward-grad** is a **competitive but previously overlooked** method, especially in the **full-tuning** setting.

Takeaway III: **ZO-SGD-Cons** and **ZO-SGD-MMT** also demonstrate strong performance, while **ZO-SGD-Sign**, the **simplest** ZO optimization method, tends to be the **weakest** approach.

➢ LLMs ZO Fine-Tuning on More Complex Tasks

Takeaway I: **ZO-Adam and ZO-SGD-MMT** exhibit **exceptional stability** across varied conditions, possibly due to **variance-reduced** techniques.

Takeaway II: The **LoRA** tuning method is consistently **robust to ZO algorithms**, providing a stable and reliable tuning approach in diverse settings.

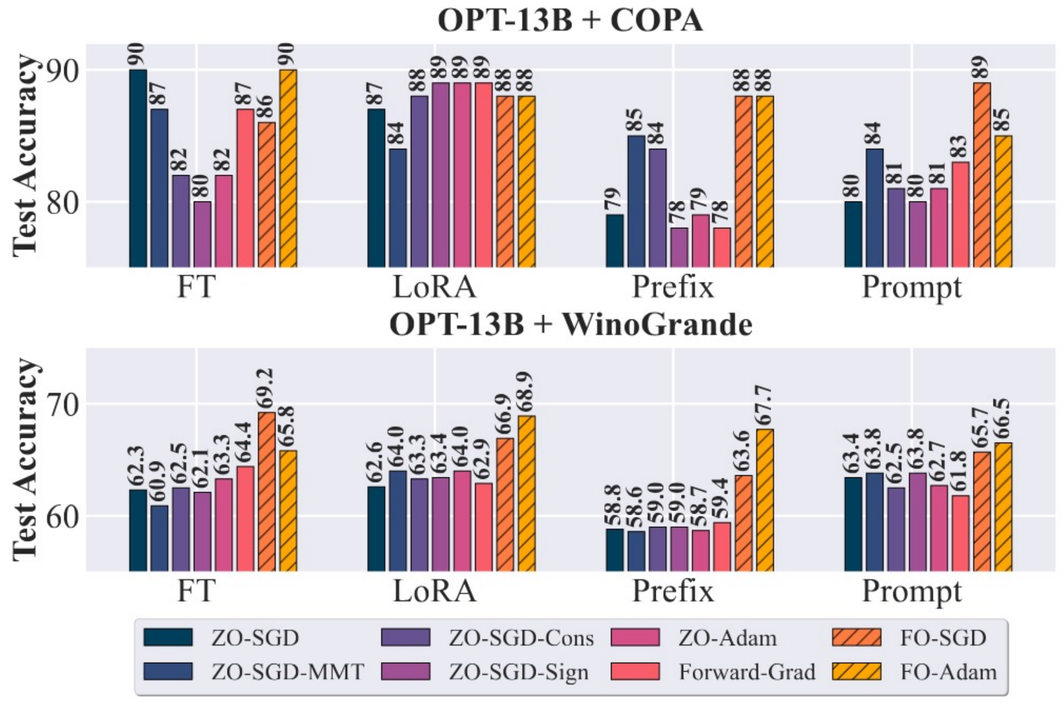

Figure 1: Results of OPT-13B fine-tuned on COPA and WinoGrande with PEFT methods.

➢ Memory and Runtime Efficiency Analyses

Optimizer	Memory ⇓	Consumed GPUs ⇓	Runtime Cost
ZO-SGD	**29** GB	1×A100	**1.8**s
ZO-SGD-Cons	**29** GB	1×A100	4.2s
ZO-SGD-Sign	**29** GB	1×A100	**1.8**s
ZO-SGD-MMT	53 GB	1×A100	**1.8**s
ZO-Adam	80 GB	2×A100	1.9s
Forward-Grad*	138 GB	2×A100	19.8s
FO-SGD	161 GB	3×A100	2.7s
FO-Adam	257 GB	4×A100	2.8s

Table 3: Memory and runtime cost when fine-tuning the full OPT-1.3B model on MultiRC.

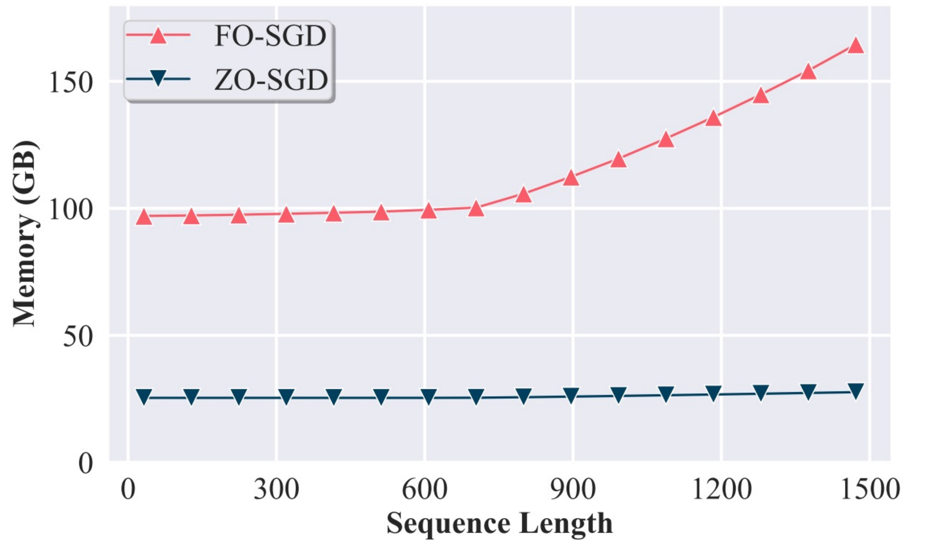

Figure 2: Memory comparison across various sequence lengths.

➢ Extended Study to Improve ZO Fine-Tuning

- Study I: **Block-wise ZO optimization enhances fine-tuning performance.**
- Study II: **Performance and efficiency trade-off via hybrid ZO-FO training.**

Optimizer	Forward Pass #	SST2	WinoGrande
MeZO	1	90.83	55.5
ZO-SGD ($q = 26$)	26	91.28	55.7
ZO-SGD-Block	26	**93.69**	**57.2**

Table 4: Comparison of ZO-SGD and ZO-SGD-Block with the same query budgets.

ZO Layer #	Memory (GB)		Accuracy (%)	
	Memory	ΔMemory	Accuracy	ΔAccuracy
0 (FO-SGD)	24.29	11.07	91.22	1.98
4	23.33	10.11	91.12	1.88
8	22.01	8.79	90.79	1.55
12	20.43	7.21	89.48	0.24
16	18.98	5.76	89.42	0.18
20	15.43	2.21	89.27	0.03
24 (ZO-SGD)	13.22	0.00	89.24	0.00

Table 5: Trade-off between memory v.s. accuracy in hybrid ZO-FO fine-tuning.

- Study III: **Gradient pruning benefits performance.**

COPA										
Sparsity (%)	0	10	20	30	40	50	60	70	80	90
Accuracy (%)	73.00	**75.00**	**75.00**	70.00	70.00	70.00	70.00	7.000	70.00	71.00
SST2										
Sparsity (%)	0	10	20	30	40	50	60	70	80	90
Accuracy (%)	90.83	**91.51**	**92.20**	**92.32**	**91.74**	**92.43**	**92.43**	**92.20**	**91.51**	**92.66**

Table 6: Fine-tuning OPT-1.3B using ZO-SGD w/ different gradient sparse ratios.

Contacts: zhan1908@msu.edu, pingzhi@cs.unc.edu