ICML
International Conference
On Machine Learning

UNIVERSITY of WASHINGTON

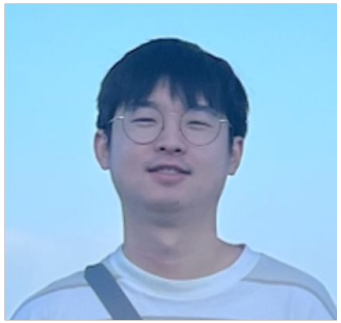
ETH zürich

amazon

DPZero: Private Fine-Tuning of Language Models without Backpropagation

Liang Zhang

(ETH Zurich & Max Planck Institute)



Bingcong Li

(ETH Zurich)

Kiran Thekumparampil

(Amazon)

Sewoong Oh

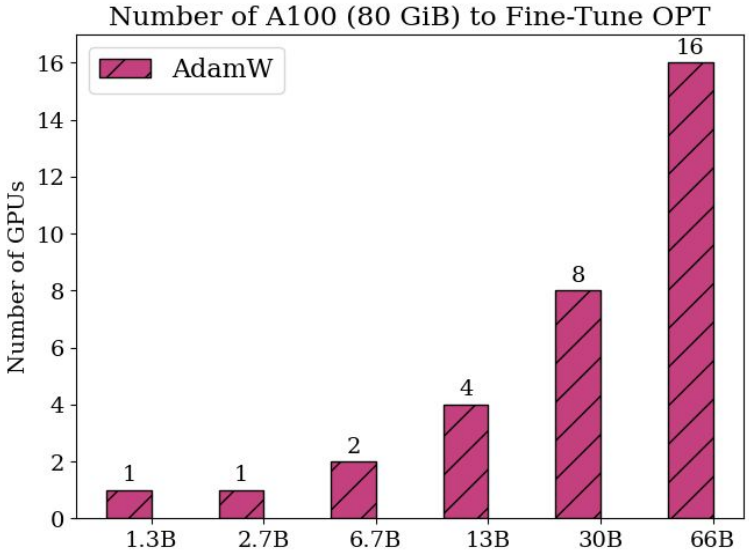
(University of Washington)

Niao He

(ETH Zurich)

Bottleneck in Fine-Tuning LLMs

- **Memory** becomes a bottleneck when fine-tuning billion-sized LLMs

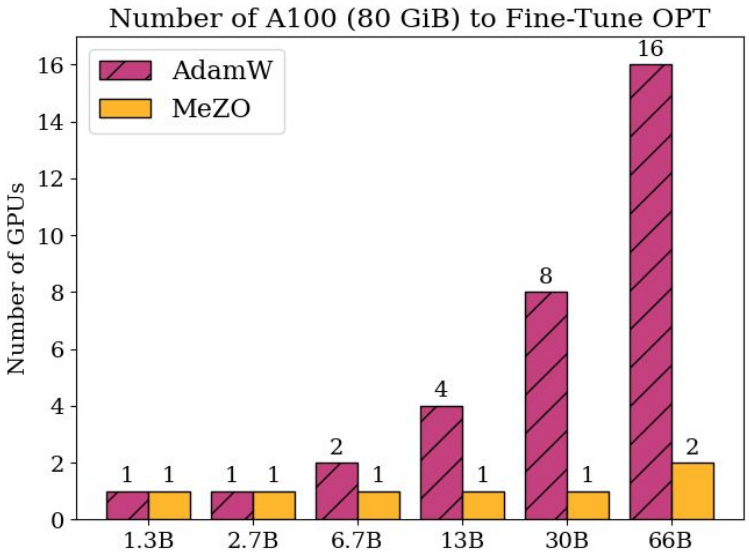


- Backpropagation heavy in memory

[1] Malladi et al. Fine-Tuning Language Models with Just Forward Passes. NeurIPS, 2023.

Bottleneck in Fine-Tuning LLMs

- **Memory** becomes a bottleneck when fine-tuning billion-sized LLMs

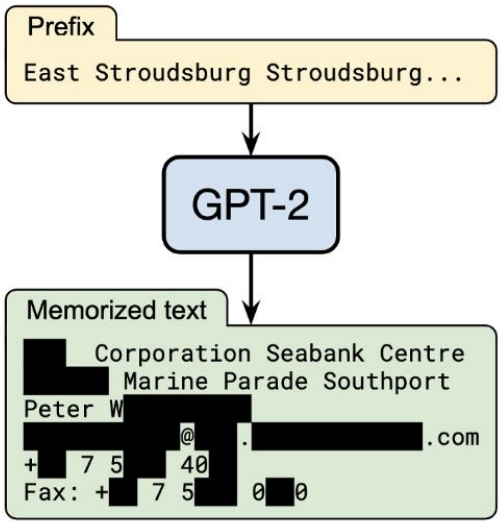


- Backpropagation heavy in memory
- MeZO [1]: **zeroth-order** methods with **only forward passes**

[1] Malladi et al. Fine-Tuning Language Models with Just Forward Passes. NeurIPS, 2023.

Bottleneck in Fine-Tuning LLMs

- High quality data are often **private**

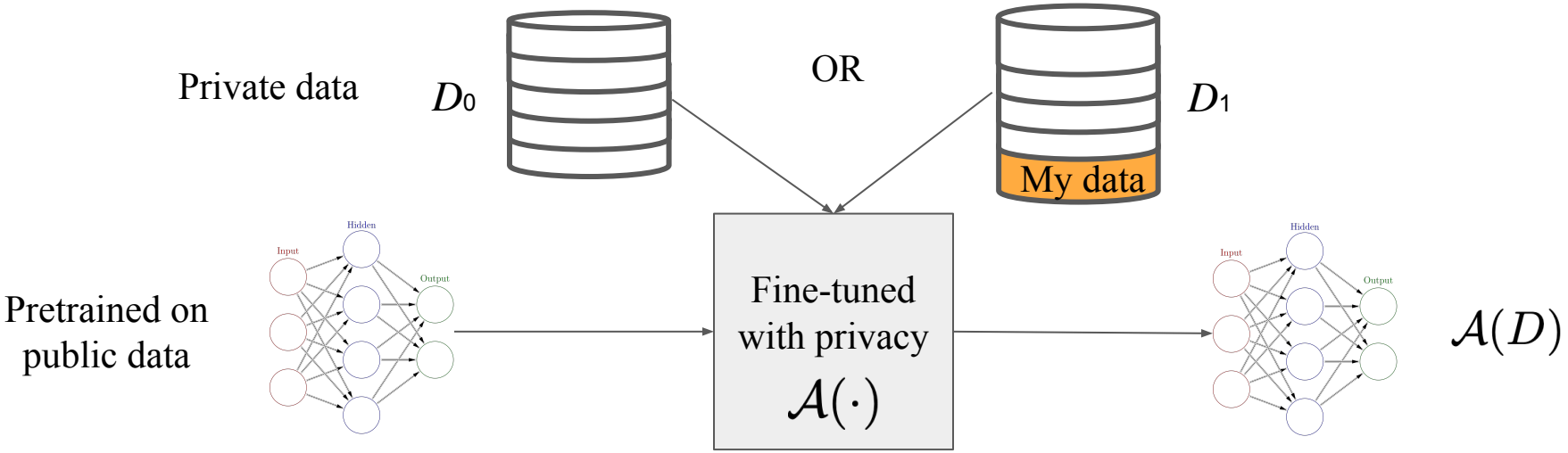


- **LLMs memorize and reveal sensitive data !!!**
- Fine-tuning LLMs with **differential privacy (DP)**

[2] Carlini et al. Extracting Training Data from Large Language Models. USENIX Security Symposium, 2021.

Bottleneck in Fine-Tuning LLMs

- High quality data are often **private** → fine-tuning with **(ε,δ)-Differential Privacy**



$$\mathbb{P}(\mathcal{A}(D_1) \in R) \leq e^{\varepsilon}\, \mathbb{P}(\mathcal{A}(D_0) \in R) + \delta$$

Private and Zeroth-Order Optimization for LLMs Fine-Tuning

Gap Between Theory and Practice

$$\text{minimize}_x \quad F_S(x) \quad := \frac{1}{n}\sum_{i=1}^{n} f(x;\xi_i)$$

Theory

Convergence rates of both

👉 **private** first-order optimization

👉 non-private **zeroth-order** optimization

depend on the **dimension**

Practice

They perform well on LLMs fine-tuning with **dimension** scaling to **billions**

Non-Private Zeroth-Order Optimization

$$x_{t+1} \leftarrow x_t - \alpha \Bigg(\underbrace{\frac{1}{n} \sum_{i=1}^{n} \frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t}_{\text{0-th order gradient estimate}} \Bigg)$$

Non-Private Zeroth-Order Optimization

$$x_{t+1} \leftarrow x_t - \alpha \Bigg(\underbrace{\frac{1}{n} \sum_{i=1}^{n} \frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t}_{\text{0-th order gradient estimate}} \Bigg)$$

Theory

- **Unbiased** when $\lambda \to 0$ & $\mathrm{E}[uu^{\mathrm{T}}] = \mathrm{I}_d$

$$\|\nabla F_S(x)\|^2 \lesssim \frac{d}{T}$$

Non-Private Zeroth-Order Optimization

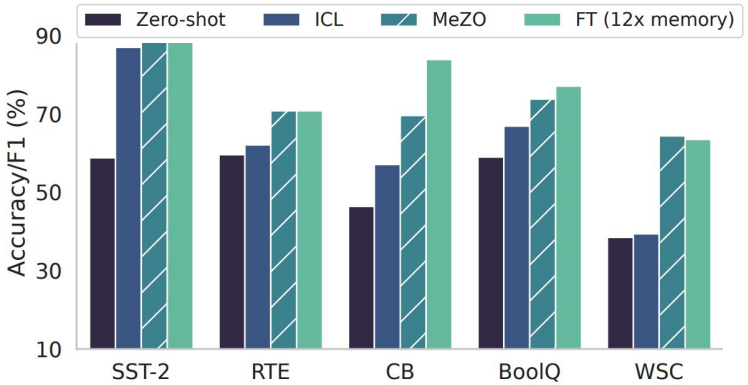
$$x_{t+1} \leftarrow x_t - \alpha \Big(\underbrace{\frac{1}{n} \sum_{i=1}^{n} \frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t}_{\text{0-th order gradient estimate}} \Big)$$

Theory

- **Unbiased** when $\lambda \to 0$ & $\mathrm{E}[uu^{\mathrm{T}}] = \mathrm{I}_d$

$$\|\nabla F_S(x)\|^2 \lesssim \frac{d}{T}$$

Practice



OPT-13B

[1] Malladi et al. Fine-Tuning Language Models with Just Forward Passes. NeurIPS, 2023.

Private First-Order Optimization: DP-GD

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{clip}_C\big(\nabla f(x_t; \xi_i) \big) + \frac{C}{n} z_t \right)$$

Private First-Order Optimization: DP-GD

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \text{clip}_C\big(\nabla f(x_t; \xi_i) \big) + \frac{C}{n} z_t \right)$$

Theory

- (ε,δ)-differential privacy with

$$z_t \sim \mathcal{N}(0, (4\sqrt{2T \log(1.25/\delta)}/\varepsilon)^2 \mathbf{I}_{d\times d})$$

$$\|\nabla F_S(x)\|^2 \lesssim \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon}$$

Private First-Order Optimization: DP-GD

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \text{clip}_C\big(\nabla f(x_t; \xi_i) \big) + \frac{C}{n} z_t \right)$$

Theory

- (ε,δ)-differential privacy with

$$z_t \sim \mathcal{N}(0, (4\sqrt{2T \log(1.25/\delta)}/\varepsilon)^2 \mathbf{I}_{d \times d})$$

$$\|\nabla F_S(x)\|^2 \lesssim \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon}$$

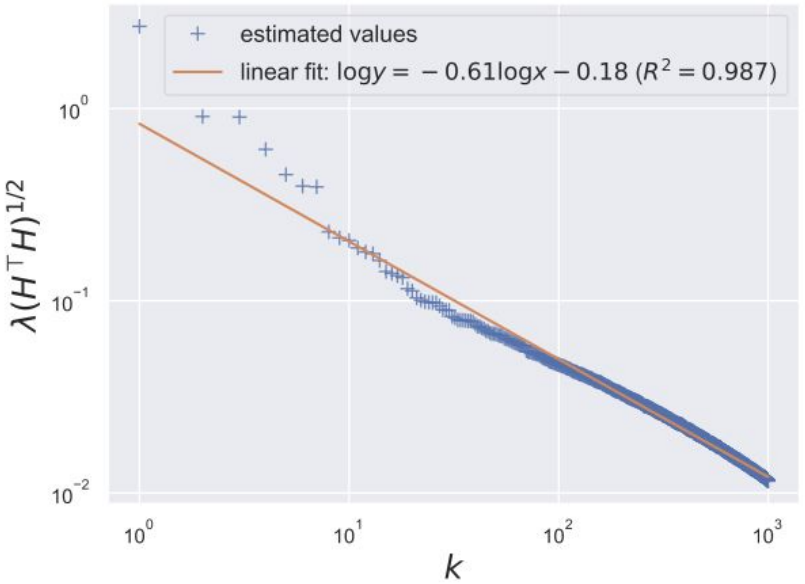
Practice

Model	BLEU	BLEU (DP)	Drop
GPT2-(345M)	47.1	42.0	5.1
GPT2-(774M)	47.5	43.1	4.4
GPT2-(1.5B)	48.1	43.8	4.3

$(\varepsilon = 6.8, \delta =$ 1e-5$)$ on DART

[3] Yu et al. Differentially Private Fine-Tuning of Language Models. ICLR, 2022.

Low-Dimensional Structure



Effective Rank

$$-H \preceq \nabla^2 F_S(x) \preceq H$$

$$\mathrm{Tr}(H) \leq r\|H\|_2$$

- Not necessarily low rank
- Recover smoothness when $r = d$

[4] Li et al. When Does Differentially Private Learning Not Suffer in High Dimensions. NeurIPS, 2022.

Dimension-Independent Private and Zeroth-Order Optimization under Low Effective Rank Structure

First Attempt: DPGD-0th

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \text{clip}_C \left(\underbrace{\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t}_{\text{0-th order gradient estimate}} \right) + \frac{C}{n} z_t \right)$$

First Attempt: DPGD-0th

$$x_{t+1} \leftarrow x_t - \alpha\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{clip}_C\Big(\underbrace{\frac{f(x_t+\lambda u_t;\xi_i)-f(x_t-\lambda u_t;\xi_i)}{2\lambda}u_t}_{\text{0-th order gradient estimate}}\Big)+\frac{C}{n}z_t\right)$$

$f(x;\xi_i)$ is L-Lipschitz and ℓ-smooth

- Without effective rank

$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2\right] \lesssim \left((F_S(x_0)-F_S^*)\ell+L^2\right)\frac{d\sqrt{d\log(1/\delta)}}{n\varepsilon}$$

- With effective rank

$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2\right] \lesssim \left((F_S(x_0)-F_S^*)\ell+L^2\right)\frac{d\sqrt{r\log(1/\delta)}}{n\varepsilon}$$

First Attempt: DPGD-0th

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{clip}_C \left(\underbrace{\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t}_{\text{0-th order gradient estimate}} \right) + \frac{C}{n} z_t \right)$$

$f(x; \xi_i)$ is L-Lipschitz and ℓ-smooth

$C \sim Ld$

- Without effective rank

$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2\right] \lesssim \left((F_S(x_0) - F_S^*)\ell + L^2\right) \frac{d\sqrt{d\log(1/\delta)}}{n\varepsilon}$$

- With effective rank

$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2\right] \lesssim \left((F_S(x_0) - F_S^*)\ell + L^2\right) \frac{d\sqrt{r\log(1/\delta)}}{n\varepsilon}$$

DPZero: Dimension-Independent

- No need to add noise to the update direction

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \underbrace{\text{clip}_C \left(\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} \right)}_{\text{(approx.) directional derivative}} + \underbrace{\frac{C}{n} z_t}_{\textbf{scalar} \text{ noise}} \right) u_t$$

DPZero: Dimension-Independent

- No need to add noise to the update direction

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \underbrace{\text{clip}_C \left(\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} \right)}_{\text{(approx.) directional derivative}} + \underbrace{\frac{C}{n} z_t}_{\textbf{scalar} \text{ noise}} \right) u_t$$

$$\simeq \langle \nabla f(x_t; \xi_i), u_t \rangle \simeq \begin{cases} \sqrt{d}\, L & \text{worst-case} \\ L & \text{w.h.p} \end{cases}$$

$$C \sim L$$

DPZero: Dimension-Independent

- No need to add noise to the update direction

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \underbrace{\text{clip}_C \left(\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} \right)}_{\text{(approx.) directional derivative}} + \underbrace{\frac{C}{n} z_t}_{\textbf{scalar} \text{ noise}} \right) u_t$$

$$\simeq \langle \nabla f(x_t; \xi_i), u_t \rangle \simeq \begin{cases} \sqrt{d}\, L & \text{worst-case} \\ L & \text{w.h.p} \end{cases}$$

$$C \sim L$$

- Under effective rank structure

$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2 \right] \lesssim \left((F_S(x_0) - F_S^*)\ell + L^2 \right) \frac{\sqrt{r \log(1/\delta)}}{n\varepsilon}$$

DPZero

Algorithm 3 DPZero

Input: Dataset $S = \{\xi_1, \cdots, \xi_n\}$, initialization $x_0 \in \mathbb{R}^d$, number of iterations T, stepsize $\alpha > 0$, smoothing parameter $\lambda > 0$, clipping threshold $C > 0$, privacy parameters $\varepsilon > 0, \delta \in (0,1)$.

1: **for** $t = 0, 1, \cdots, T-1$ **do**

2: Sample u_t uniformly at random from the Euclidean sphere $\sqrt{d}\,\mathbb{S}^{d-1}$.

3: Sample $z_t \sim \mathcal{N}(0, \sigma^2)$ with variance $\sigma = 4\sqrt{2T\log(e + (\varepsilon/\delta))}/\varepsilon$, and

$$x_{t+1} \leftarrow x_t - \alpha \left(\frac{1}{n} \sum_{i=1}^{n} \mathrm{clip}_C \left(\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} \right) + \frac{C}{n} z_t \right) u_t.$$

Output: x_τ for τ sampled uniformly at random from $\{0, 1, \cdots, T-1\}$.

Summary of Results

- **Effective rank:** $-H \preceq \nabla^2 F_S(x) \preceq H$

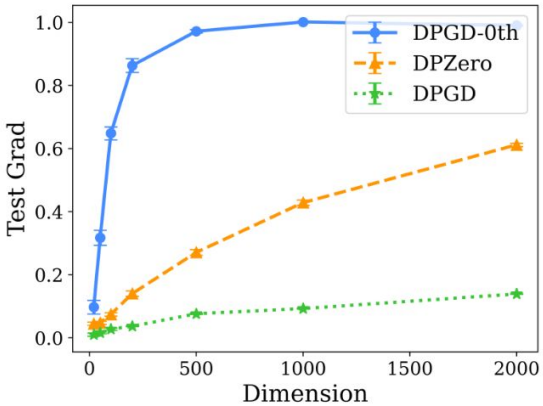
$$\mathrm{Tr}(H) \le r\|H\|_2$$

	without effective rank	with effective rank r
DPGD-0th	$d\sqrt{d}$	$d\sqrt{r}$
DPZero	$(\log d)\sqrt{d}$	$(\log d)\sqrt{r}$
DP-GD	$\sqrt{d}$	$\sqrt{r}$

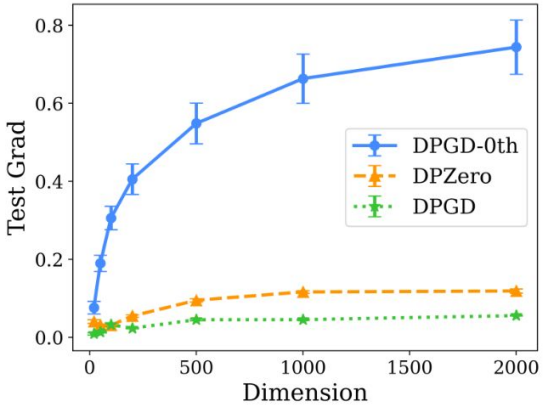
Experiments on Synthetic Examples

- $(\varepsilon = 2, \delta = 10^{-6})$ - DP on a quadratic

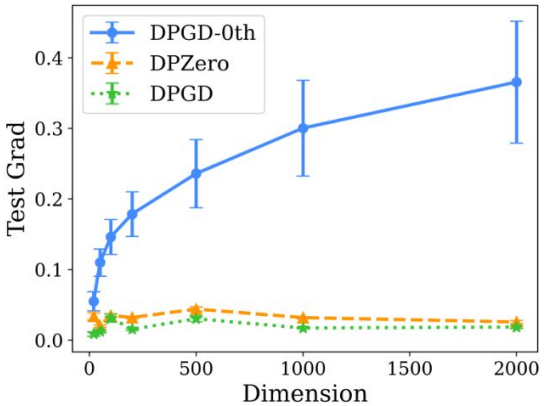
$$\min_{x \in \mathbb{R}^d} F_S(x) = \frac{1}{2n} \sum_{i=1}^{n} (x - x_i)^\top A (x - x_i)$$



(a) $\mathrm{Tr}(A) = \mathcal{O}(d)$.

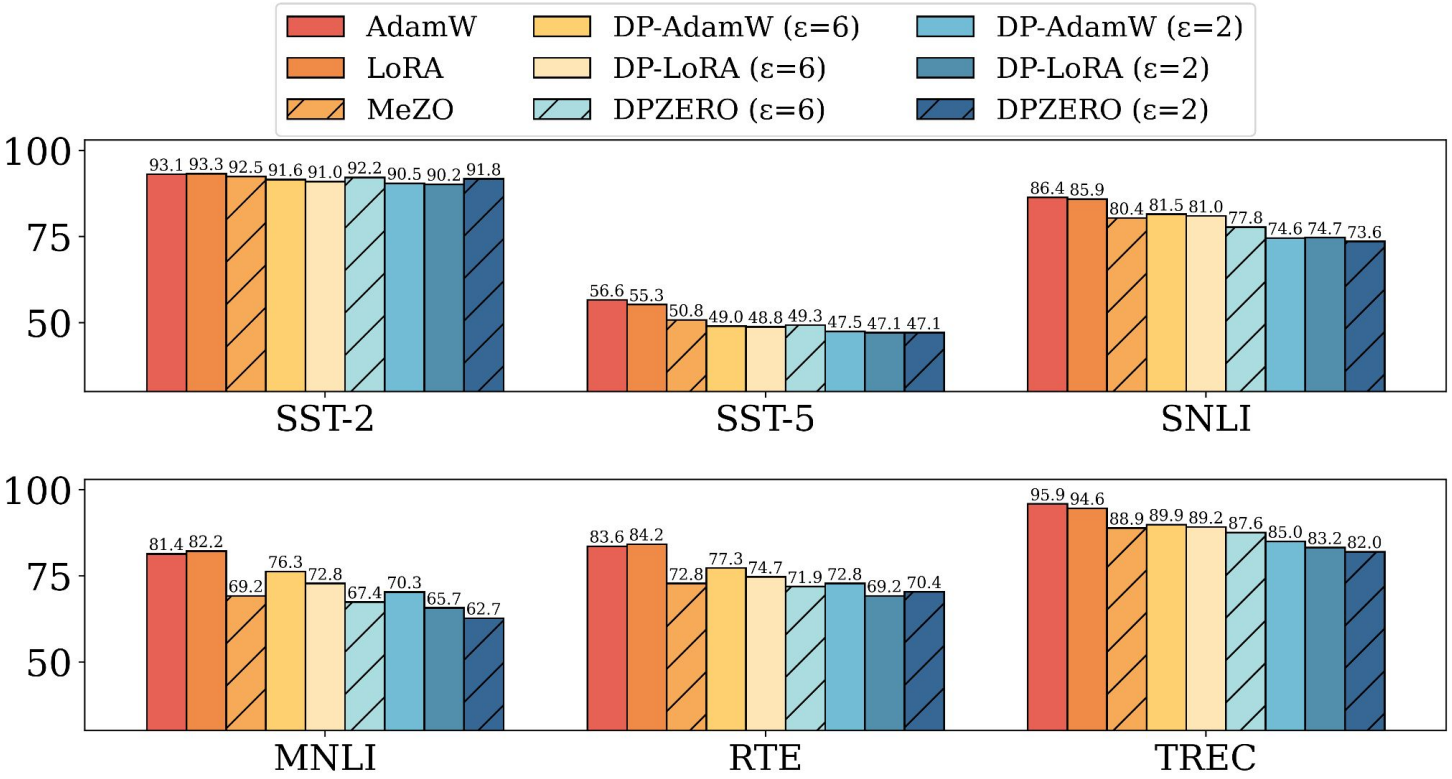


(b) $\mathrm{Tr}(A) = \mathcal{O}(\sqrt{d})$.

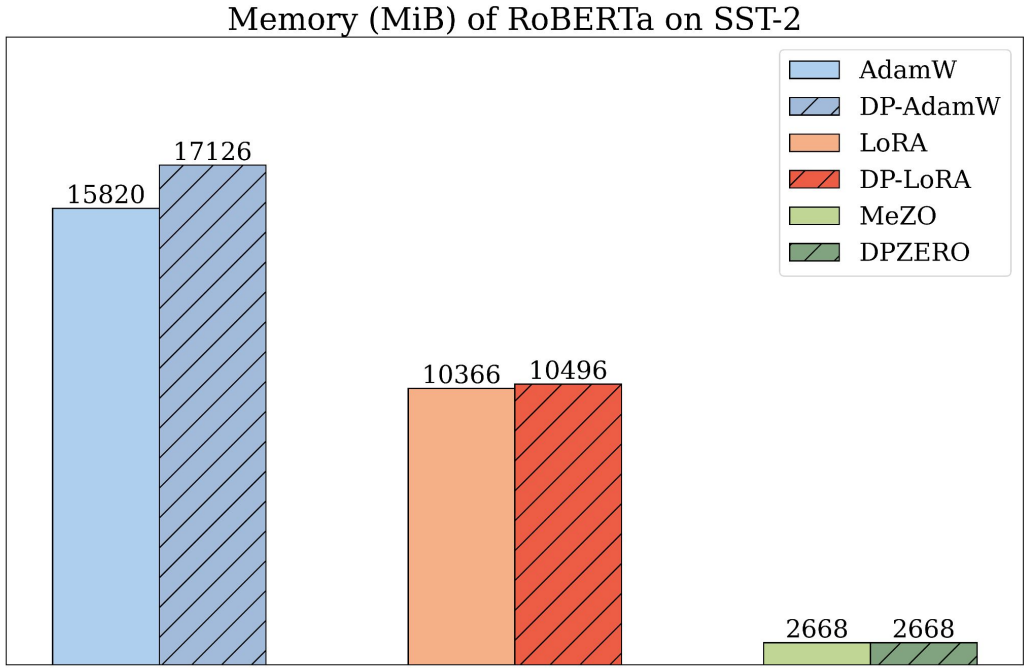


(c) $\mathrm{Tr}(A) = \mathcal{O}(\log d)$.

Private Fine-Tuning RoBERTa (355M)



Private Fine-Tuning RoBERTa (355M)



- **DPZero**: Nearly **no additional cost**
  - ✓ Efficient **per-sample loss clipping** (v.s. **per-sample gradient clipping** in DP-GD)

Private Fine-Tuning OPT