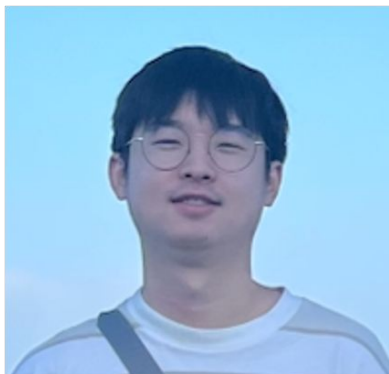


Private Fine-tuning of LLMs without Backpropagation

Sewoong Oh (University of Washington)

Joint work with

Liang Zhang (ETH), Bingcong Li (ETH), Kiran Koshy Thekumparampil (Amazon), and Niao He (ETH)



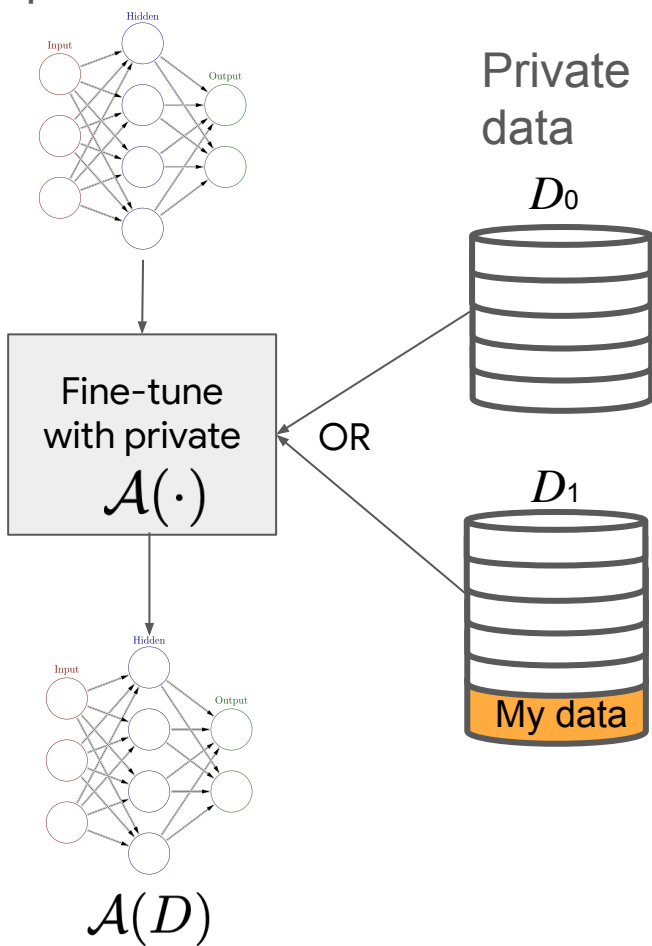
Data is central in machine learning research,
but high-quality data is often **private**

(ϵ, δ) -Differential Privacy definition

$$\underbrace{\mathbb{P}(\mathcal{A}(D_1) \in R)}_{\text{True Positive Rate}} \leq e^\epsilon \underbrace{\mathbb{P}(\mathcal{A}(D_0) \in R)}_{\text{False Positive Rate}} + \delta$$

- Equivalent to certain condition on the binary hypothesis testing on whether my data was in the dataset (D_1) or not (D_0)
- This gives plausible deniability

Pretrained on public data



Private optimization adapts to the intrinsic structure of **fine-tuning landscape** and scales to Billions-size LLMs

Clipping: “Exploring the Limits of Differentially Private Deep Learning with Group-wise Clipping” He, et al. ICLR’23

“Large Language Models Can Be Strong Differentially Private Learners” Li, et al. ICLR’22

Virtual batching: “Unlocking High-Accuracy Differentially Private Image Classification through Scale”, De, et al. 2022

A few years ago...

- Differentially Private Stochastic Gradient Descent (DP-SGD) was thought to be unfit for large scale optimization.
- Because, unlike SGD, DP-SGD suffer from dimension dependence for solving:

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

- SGD for Lipschitz and smooth non-convex f : $\|\nabla F_S(x)\|^2 \lesssim \frac{1}{T}$

Theoretically, Differentially Private SGD suffers in high dimensions

- DP-SGD:

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

(ϵ, δ) -differential privacy achieved with

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

Theoretically, Differentially Private SGD suffers in high dimensions

- DP-SGD:

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

(ϵ, δ) -differential privacy achieved with

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

- Under Lipschitz and smooth $f(\cdot)$, and $x \in \mathbb{R}^d$

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

Experiments seems to contradict theory

- DP-SGD does not suffer from high-dimensionality

	Model	BLEU (non-private)	BLEU (DP)	Drop due to privacy
345M	GPT-2-Medium	47.1	42.0	5.1
774M	GPT-2-Large	47.5	43.1	4.4
1.5B	GPT-2-XL	48.1	43.8	4.3

($\epsilon = 6.8, \delta = 1e-5$)

as long as we are **fine-tuning** a pretrained model.

- In practice, typical value of ϵ range from 1 to 10

Open question 1:

Why does DP-SGD with $\epsilon=10$ significantly reduce memorization?

Why does DP-SGD not suffer in high dimensions?

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r \|H\|_2$.

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

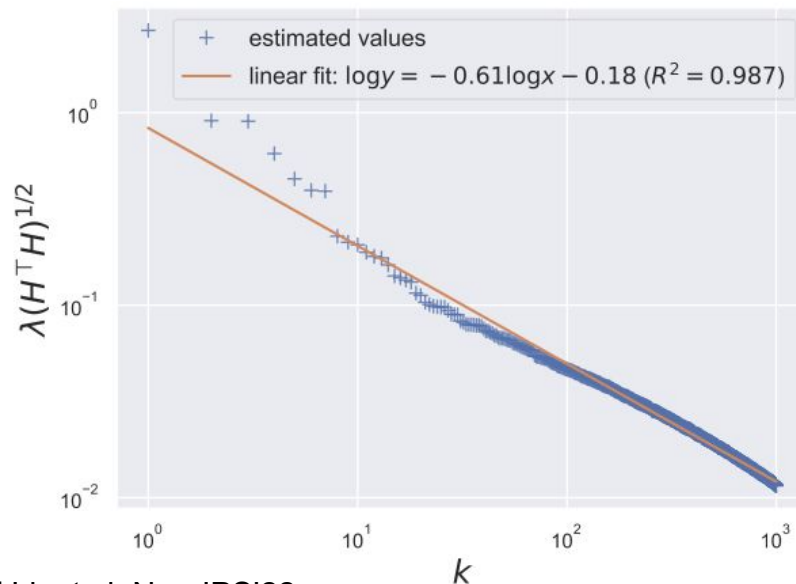
- The algorithm does not need to know the intrinsic dimension

Why does DP-SGD not suffer in high dimensions?

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r \|H\|_2$.

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

- Several variants of the above assumptions in the literature, such as singular value decay in the collection of the gradients



Zeroth-order optimization adapts to the intrinsic structure of **fine-tuning landscape** and scales to Billions-size LLMs

Bottleneck in (private) fine-tuning of LLMs

- As LLMs get larger, memory for **backpropagation** is becoming a bottleneck
- Can we finetune LLMs while running only **forward passes**?

Bottleneck in (private) fine-tuning of LLMs

- As LLMs get larger, memory for **backpropagation** is becoming a bottleneck
- Can we finetune LLMs while running only **forward passes**?
- Zeroth-order gradient estimate

$$\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t$$

- u_t Is drawn uniformly at random from $\sqrt{d}\mathbb{S}^{d-1}$
- Only requires forward passes
- Asymptotically unbiased:

$$\mathbb{E} \left[\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t \right] \xrightarrow{\lambda \rightarrow 0} \mathbb{E} [\langle \nabla f(x_t; \xi_i), u_t \rangle u_t] = d \nabla f(x_t, \xi_i)$$

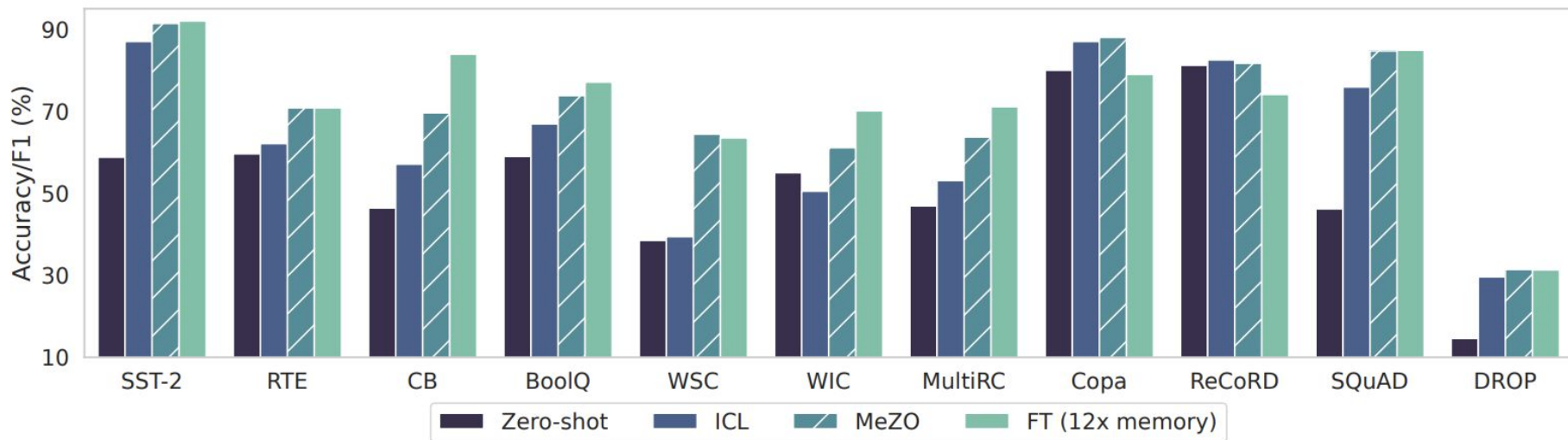
Theoretically,
ZO-SGD suffers in high-dimensions in the worst-case

- Gradient Descent: $\|\nabla F_S(x)\|^2 \lesssim \frac{1}{T}$
- ZO-SGD: $\|\nabla F_S(x)\|^2 \lesssim \frac{d}{T}$

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

Memory Efficient Zeroth-order Optimization: MeZO does not suffer from high-dimensionality

Number of parameters: $d=13\text{B}$



Dimension independence rate with low effective rank

- Gradient Descent: $\|\nabla F_S(x)\|^2 \lesssim \frac{1}{T}$
- ZO-SGD: $\|\nabla F_S(x)\|^2 \lesssim \frac{d}{T}$

Assume

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r\|H\|_2$.

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

How do we design a **private & zeroth-order optimization** that adapts to the intrinsic structure of **fine-tuning landscape** and scales to Billions-size LLMs?

First attempt: replace gradient with 0-th order approximation

- Zeroth-order gradient estimate

- Randomly draw direction u_t uniformly over the sphere $\sqrt{d}\mathbb{S}^{d-1}$

$$\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t \xrightarrow{\lambda \rightarrow 0} \langle \nabla f(x_t; \xi_i), u_t \rangle u_t$$

First attempt: replace gradient with 0-th order approximation

- Zeroth-order gradient estimate

- Randomly draw direction u_t uniformly over the sphere $\sqrt{d}\mathbb{S}^{d-1}$

$$\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t \xrightarrow{\lambda \rightarrow 0} \langle \nabla f(x_t; \xi_i), u_t \rangle u_t$$

- Substitute zeroth-order gradient estimate in DP-SGD

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

First attempt: replace gradient with 0-th order approximation

- Zeroth-order gradient estimate

- Randomly draw direction u_t uniformly over the sphere $\sqrt{d}\mathbb{S}^{d-1}$

$$\frac{f(x_t + \lambda u_t; \xi_i) - f(x_t - \lambda u_t; \xi_i)}{2\lambda} u_t \xrightarrow{\lambda \rightarrow 0} \langle \nabla f(x_t; \xi_i), u_t \rangle u_t$$

- Substitute zeroth-order gradient estimate in DP-SGD

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

- Clipping threshold $C = Ld$

- In practice, it is a hyperparameter to be tuned
- In theory, typical choice is to select worst-case “gradient” norm to avoid clipping bias

Degrades with dimension even under low effective rank

Assume

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r \|H\|_2$.

Theorem [Zhang, Thekumparampil, O., He 2023]

- First Attempt approach achieves (ε, δ) -DP and

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

with step-size $\alpha = \frac{1}{4\ell r}$, and $T = r \frac{n\varepsilon}{d\sqrt{r \log(1/\delta)}}$.

Degrades with dimension even under low effective rank

Assume

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r \|H\|_2$.

Theorem [Zhang, Thekumparampil, O., He 2023]

- First Attempt approach achieves (ε, δ) -DP and

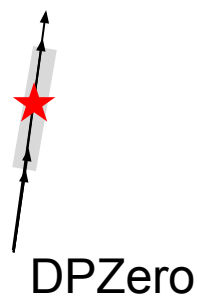
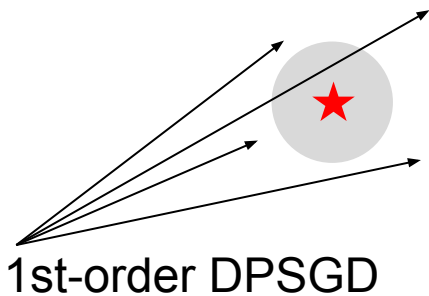
$$\mathbb{E}[\|\nabla F_S(x_\tau)\|^2] \lesssim \left((F_S(x_0) - F_S^*)\ell + L^2 \right) \frac{d\sqrt{r} \log(1/\delta)}{n\varepsilon},$$

with step-size $\alpha = \frac{1}{4\ell r}$, and $T = \frac{r}{\alpha} \frac{n\varepsilon}{\sqrt{r} \log(1/\delta)}$.

Improved private 0th-order method: DPZero

- The descent direction need not be private
 - u_t is drawn uniformly at random over the sphere $\sqrt{d}\mathbb{S}^{d-1}$, and does not touch the data

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



Improved private 0th-order method: DPZero

- Typical magnitude of the derivative is significantly smaller than the worst-case
 - u_t is drawn uniformly at random over the sphere $\sqrt{d}\mathbb{S}^{d-1}$

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

$\underbrace{\hspace{15em}}$
(approximate) directional derivative

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

DPZero

Algorithm 3 DPZERO

Input: Dataset $S = \{\xi_1, \dots, \xi_n\}$, 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** $t = 0, 1, \dots, T - 1$ **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

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

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

- With $C = \tilde{O}(L)$ and small enough $\lambda = O\left(\frac{L}{\ell d^{3/2}} \sqrt{\frac{r \log(1/\delta)}{n\varepsilon}}\right)$

DPZero

Algorithm 3 DPZERO

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

- 1: **for** $t = 0, 1, \dots, T - 1$ **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 $\sigma = 4\sqrt{2T \log(e + (\epsilon/\delta))}/\epsilon$, and

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

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

- With $C = \tilde{O}(L)$ and small enough $\lambda = O\left(\frac{L}{\ell d^{3/2}} \sqrt{\frac{r \log(1/\delta)}{n\epsilon}}\right)$

Nearly dimension independent guarantee

Assume

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r \|H\|_2$.

Theorem [Zhang, Thekumparampil, O., He 2023]

- DPZero achieves (ε, δ) -DP and

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

with step-size $\alpha = \frac{1}{4\ell r}$, and $T = r \frac{n\varepsilon}{\sqrt{r \log(1/\delta)}}$.

Nearly dimension independent guarantee

Assume

- Each $f(x; \xi_i)$ is L -Lipschitz and ℓ -smooth,
- (Effective rank r) $-H \preceq \nabla^2 F_S(x) \preceq H$, and $\text{Tr}(H) \leq r \|H\|_2$.

Theorem [Zhang, Thekumparampil, O., He 2023]

- DPZero achieves (ε, δ) -DP and

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

with step-size $\alpha = \frac{1}{4\ell r}$, and $T = \frac{n\varepsilon}{r \sqrt{r \log(1/\delta)}}$.

1st-order vs. 0-th order private SGD

DP-1st order

DPZero

- Error: $\mathbb{E}[\|\nabla F_S(x_\tau)\|^2]$

$$\frac{\sqrt{r}}{n\varepsilon}$$

$$\frac{\sqrt{r}}{n\varepsilon}$$

Accuracy on SST-2 with $\varepsilon=2$:

92.3%

91.8%

1st-order vs. 0-th order private SGD

DP-1st order

DPZero

- Error: $\mathbb{E}[\|\nabla F_S(x_\tau)\|^2]$

$$\frac{\sqrt{r}}{n\varepsilon}$$

$$\frac{\sqrt{r}}{n\varepsilon}$$

Accuracy on SST-2 with $\varepsilon=2$:

92.3%

91.8%

- Per iteration Memory
- Per iteration time

21,494 MB

2,668 MB

2.33 s

0.347 s





1st-order vs. 0-th order private SGD

DP-1st order

DPZero

• Error: $\mathbb{E}[\ \nabla F_S(x_\tau)\ ^2]$	$\frac{\sqrt{r}}{n\epsilon}$	$\frac{\sqrt{r}}{n\epsilon}$
Accuracy on SST-2 with $\epsilon=2$:	92.3%	91.8%
• Per iteration Memory	21,494 MB	2,668 MB
• Per iteration time	2.33 s	0.347 s
• # of iterations T :	$\frac{1}{\text{Error}}$	$\frac{r}{\text{Error}}$
	1,000	10,000

More gain under differential privacy

	Method	Acc.	Time per iteration		Memory per iteration	
Non-private gain	AdamW	94.4	0.425	 1.3x	16960	 6.4x
	MeZO	92.5	0.345		2668	
Private gain	DP-AdamW	92.3	2.33	 6.7x	21494	 8.1x
	DPZERO	91.8	0.347		2668	

(non-private) zeroth-order fine-tuning with **clipping**

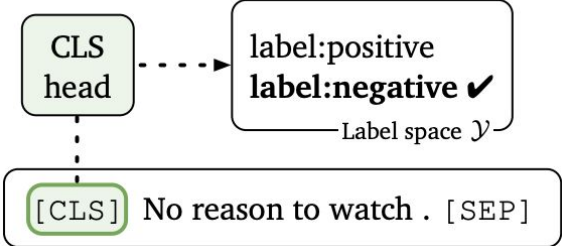
Stepsize \ Clip	1	10	50	100	200	500	No
10^{-5}	0.57	0.769	0.824	0.73	0.345	0.336	0.324
5×10^{-6}	0.54	0.688	0.825	0.826	0.782	0.332	0.338
10^{-6}	0.511	0.572	0.692	0.751	0.787	0.809	0.807

- Clipping average gradients sometimes helps,
e.g., “Eliminating sharp minima from SGD with truncated heavy-tailed noise”, Wang, O., Rhee ICLR’21
- But, here we are clipping each sample

Open question 2:

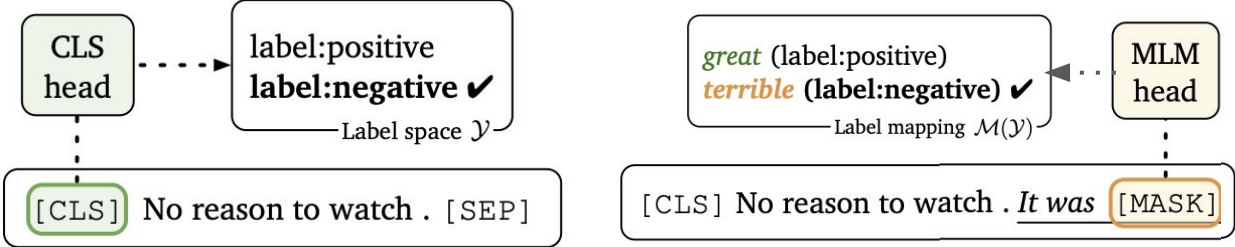
Why does sample-wise clipping help zeroth-order optimization?

How does prompt-based fine-tuning change the landscape?



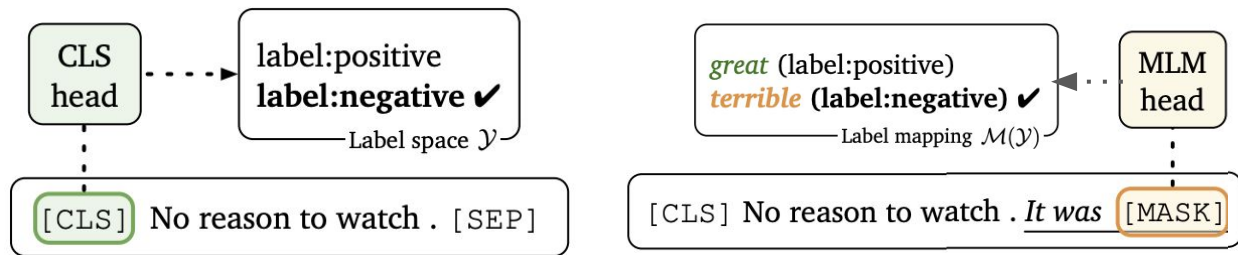
	Fine-tuning
1st-order method	81.4%

How does prompt-based fine-tuning change the landscape?



	Fine-tuning	Prompt-based FT
1st-order method	81.4%	92.7%

How does prompt-based fine-tuning change the landscape?



	Fine-tuning	Prompt-based FT
1st-order method	81.4%	92.7%
0th-order method	51.9%	89.6%

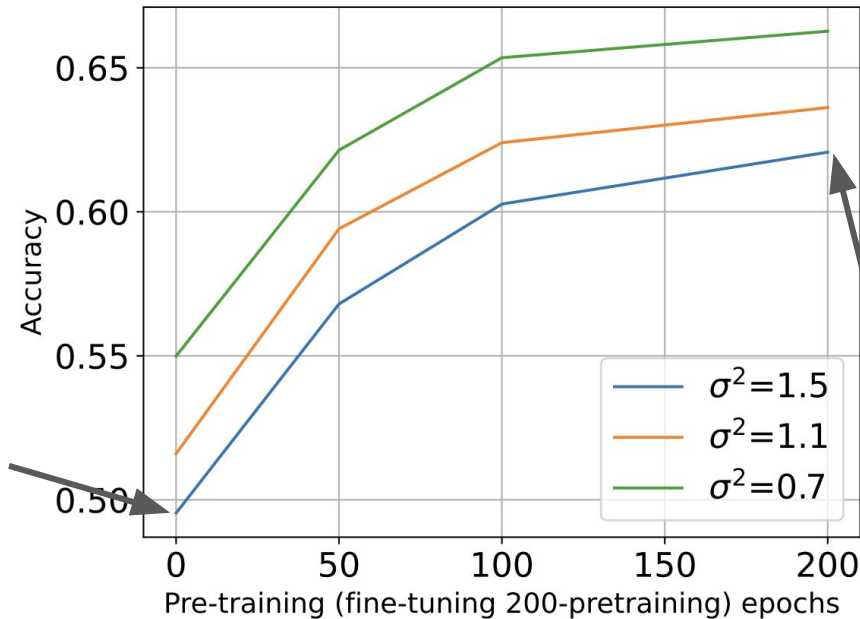
Open question 3:

Why does prompt-based fine-tuning help so much for 0-th order methods?

Why is pretraining necessary for DP-SGD?

Simple experiment on training from scratch on CIFAR-10 with DP-SGD,

- where we allowed 200 epochs of noiseless update



Noiseless updates
at the end

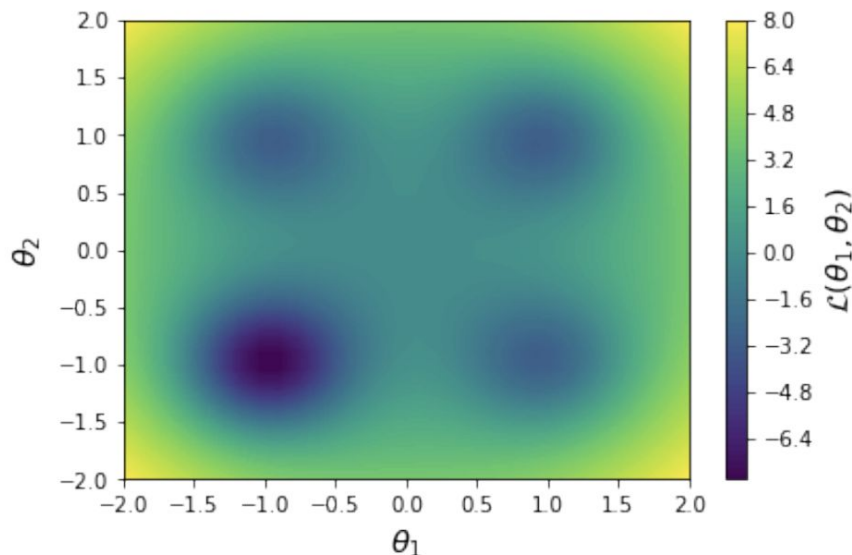
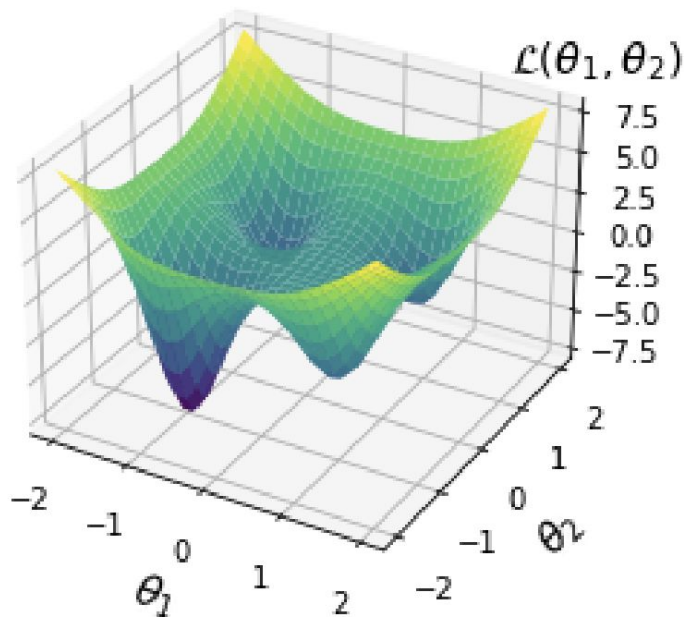
Noiseless updates
at the beginning

Open question 4:

Why does less noise in the beginning help DP-SGD more?

2-D Sketch of a conjectured landscape

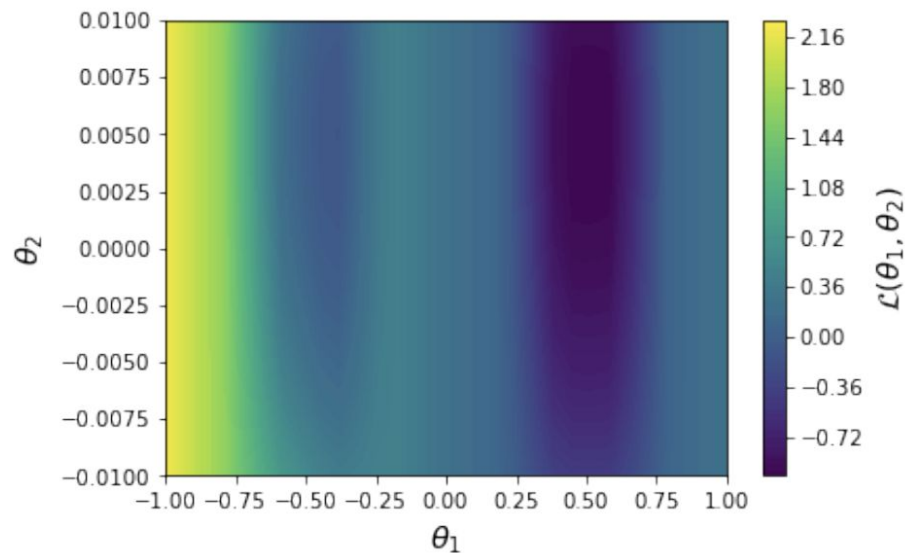
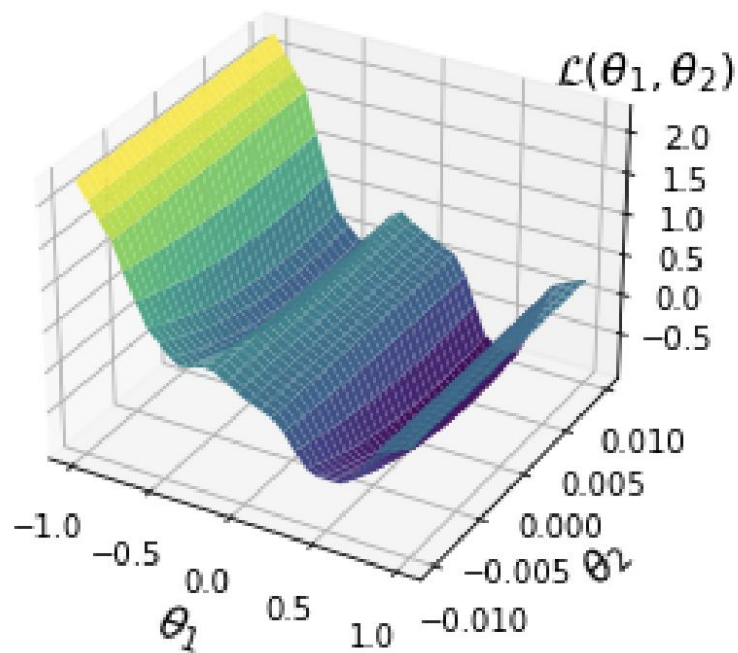
- At initialization, clean gradients are necessary to find the “good” basin



Constructing a lower bound that shows pretraining is necessary

- There exists a pair of datasets $(D_{\text{pub}}, D_{\text{pri}})$ of sizes $|D_{\text{pub}}| > |D_{\text{pri}}|$ such that
 - SGD on D_{pub} achieves Excess Risk = $\Omega(1)$
 - DP-SGD on D_{pri} achieves Excess Risk = $\Omega(1)$
 - Pretraining on D_{pub} with SGD followed by fine-tuning on D_{pri} with DP-SGD achieves Excess Risk = $O(1/|D_{\text{pri}}|)$

2-D projection of the construction



Conclusion

- The landscape of fine-tune LLMs is structured
- DP-SGD and ZO-SGD adapt to the structure and achieve dimension independent rates
- **DPZero** is the first private zeroth-order optimization algorithm that achieves dimension-independence
- There are many interesting and surprising observations that do not show up when training with typical SGD

- “Private Fine-Tuning of Language Models without Backpropagation”
Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, Niao He
<https://arxiv.org/abs/2310.09639>
- “Eliminating sharp minima from SGD with truncated heavy-tailed noise”
Xingyu Wang, Sewoong Oh, Chang-Han Rhee
ICLR 2021
- “Why Is Public Pretraining Necessary for Private Model Training?”
Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Thakurta, Lun Wang,
ICML 2023