Private zeroth-order optimization

Sewoong Oh (University of Washington)

Joint work with Liang Zhang (ETH), Kiran Koshy Thekumparampil (Amazon), and Niao He (ETH)

3 years ago…

- DP-SGD (Differentially Private Stochastic Gradient Descent) or ZO-SGD (Zeroth-order Stochastic Gradient Descent) methods were thought to be unfit for large scale optimization.

- Because, unlike (S)GD, DP-SGD and ZO-SGD suffer from dimension dependence for solving

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

Private first-order method: DP-SGD

- (ε,δ)-differential privacy achieved with a choice of noise $z_t \sim \mathcal{N}(0, (4\sqrt{2T\log(1.25/\delta)}/\varepsilon)^2 \mathbf{I}_{d\times d})$

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

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

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

“Differentially private empirical risk minimization revisited: Faster and more general.”, Wang et al. NeurIPS’17

Experiments seems to contradict theory

- DP-SGD does not suffer from high-dimensionality

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

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

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

“Differentially Private Fine-tuning of Language Models” Yu, et al. ICLR’22

DP-SGD does not suffer from high-dimensionality

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

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

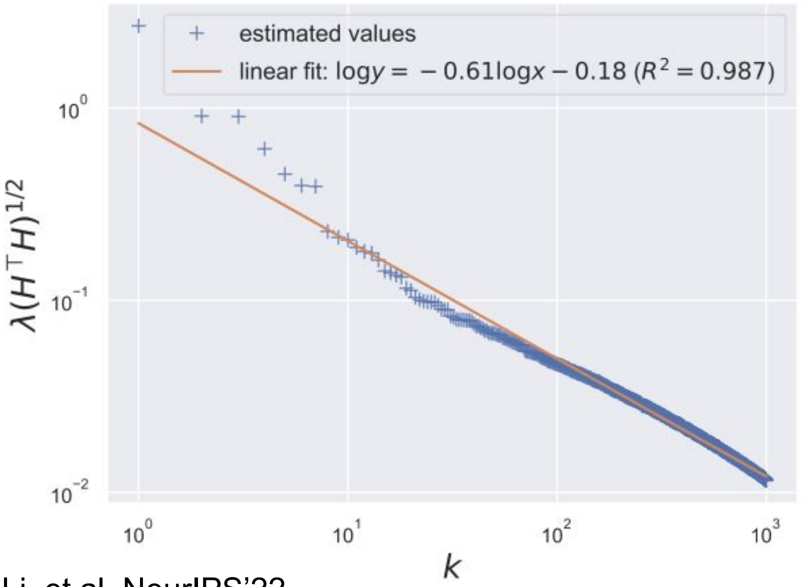
“Evading the curse of dimensionality in unconstrained private GLMs.”, Song et al., AISTATS’21

DP-SGD does not suffer from high-dimensionality

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

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

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



“When Does Differentially Private Learning Not Suffer in High Dimensions?” Li, et al. NeurIPS’22

Remaining bottlenecks in (private) fine-tuning of LLMs

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

Remaining bottlenecks in (private) fine-tuning of LLMs

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

- 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:

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

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

Zeroth-order optimization

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

Dimension independence rate with low effective rank

Zeroth-order optimization

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

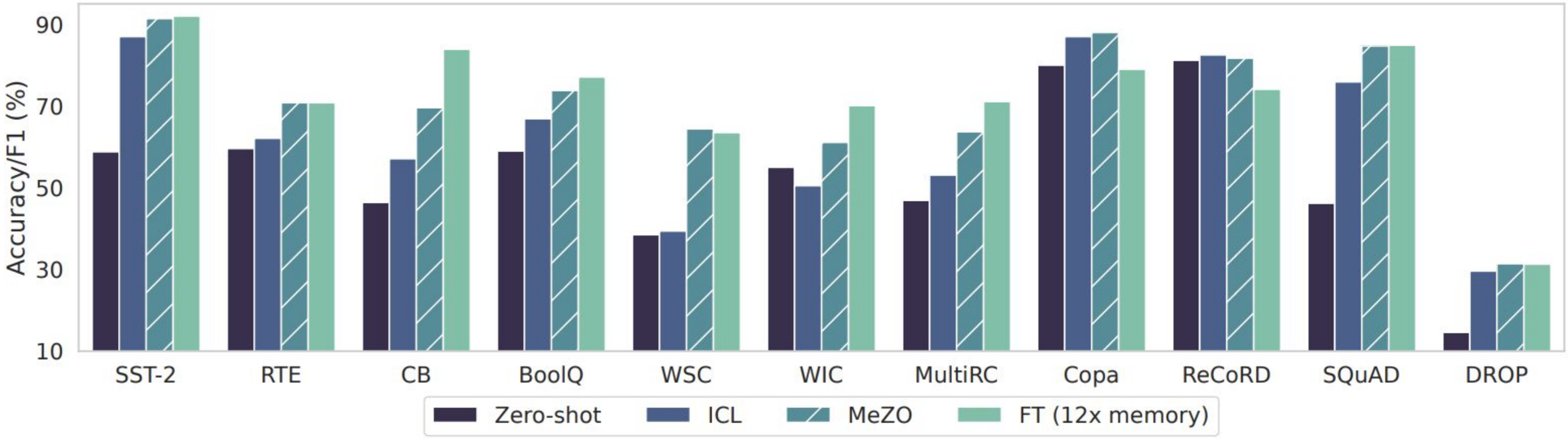
Assume

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

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

"Zeroth-order Optimization with Weak Dimension Dependency", Yue et al., COLT'23

Zeroth-order optimization: MeZO does not suffer from high-dimensionality



"Fine-Tuning Language Models with Just Forward Passes", Malladi et al., NeurIPS'23

Private Zeroth-order Optimization with dimension independent rates

First attempt: replace gradient with 0-th order approximation

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

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

First attempt: replace gradient with 0-th order approximation

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

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

- Zeroth-order update

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

First attempt: replace gradient with 0-th order approximation

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

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

- Zeroth-order update

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

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

Theorem

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

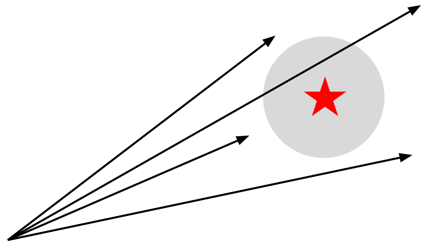
$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2\right] \lesssim \left(\left(F_S(x_0) - F_S^*\right)\ell + L^2\right)\frac{d\sqrt{r\log(1/\delta)}}{n\varepsilon},$$

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

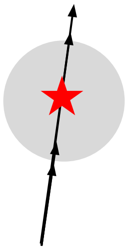
Improved private 0th-order method: DPZero

- 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

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



1st-order DPSGD



0th-order DPSGD

DPZero

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 $\sqrt{d}\mathbb{S}^{d-1}$

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

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

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 $\{0, 1, \cdots, T-1\}$.

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

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 $\boxed{\sigma = 4\sqrt{2T \log(e + (\varepsilon/\delta))}/\varepsilon,}$ 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: $\boxed{x_\tau}$ for τ sampled uniformly at random from $\{0, 1, \cdots, T-1\}$.

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

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

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

- DPZero achieves (ε,δ)-DP and

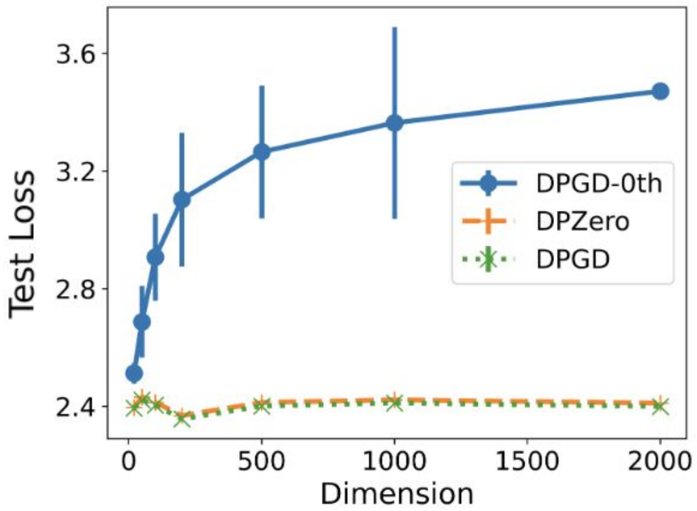
$$\mathbb{E}\left[\|\nabla F_S(x_\tau)\|^2\right] \lesssim \left(\left(F_S(x_0) - F_S^*\right)\ell + L^2\right)\frac{\sqrt{r\log(1/\delta)}}{n\varepsilon},$$

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

Empirical results in toy examples

- n = 10,000, (ε=2,δ=10^{-6})-DP, A = diag(1,½,⅓,...,1/d)

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x-x_i)^T A(x-x_i)}$$



Conclusion

- Zeroth-order optimization allows one to fine-tune larger language models
- **DPZero** is the first private zeroth-order optimization algorithm that achieves dimension-independence (under structured Hessian)
- "DPZero: Dimension-Independent and Differentially Private Zeroth-Order Optimization"
  Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, Niao He
  https://arxiv.org/abs/2310.09639

- Ongoing experiments on LLMs
- Future research directions
  - Stochastic mini-batch
  - Population guarantee
  - Convex, PL, nonsmooth
  - Potentially improved rates with tree-aggregation and variance reduction