Private Zeroth-Order Optimization with Public Data

Xuchen Gong and Tian Li (University of Chicago)

Challenge

- First-order differentially private (DP) algorithms has high computation and memory cost
- Zeroth-order DP methods are efficient to privatize as they leverage function evaluations to approximate gradients
- However, zeroth-order approaches suffer from **low utilities** and **limited application scenarios**

Our insight: **leverage public information** to improve private zeroth-order gradient approximation

Benefits of our proposed framework (PAZO):

- Improved convergence guarantee
- Stronger privacy/utility tradeoffs across **vision and language tasks** in both **pre-training and fine-tuning settings**
- 16× runtime speedup

Background: DPZero

Sample a batch of private data B
$g \leftarrow 0^d$
for q queries:
  Sample perturbation u uniformly from the sphere $\sqrt{d}\mathbb{S}^{d-1}$
  $g \leftarrow g + \left(\frac{1}{|B|}\sum_{\xi \in B} \mathrm{clip}_C\left(\frac{f(x+\lambda u;\xi)-f(x-\lambda u;\xi)}{2\lambda}\right) + z\right) u,\ z \sim \frac{1}{|B|}\mathcal{N}(0, qC^2\sigma^2)$
$x \leftarrow \eta g/q$

PAZO-M: Mixing Zeroth and First-Order Gradients

Sample a batch of public data and compute its gradient g_{pub}
$g \leftarrow 0^d$
for q queries:
  Sample perturbation u uniformly from the sphere $d^{\frac{1}{4}}\mathbb{S}^{d-1}$
  $g \leftarrow g + \left(\frac{1}{|B|}\sum_{\xi \in B} \mathrm{clip}_C\left(\frac{f(x+\lambda u;\xi)-f(x-\lambda u;\xi)}{2\lambda}\right) + z\right) u,\ z \sim \frac{1}{|B|}\mathcal{N}(0, qC^2\sigma^2)$
$x \leftarrow \eta(\alpha g_{\text{pub}} + (1-\alpha) g/q)$

PAZO-P: Sampling in Public Gradient Subspace

Sample k batches of public data and (ortho)normalize gradients $G \leftarrow [g_1, \ldots, g_k]$
$g \leftarrow 0^d$
for q queries:
  Sample perturbation u uniformly from the sphere $\sqrt{k}\mathbb{S}^{k-1}$
  $g \leftarrow g + \left(\frac{1}{|B|}\sum_{\xi \in B} \mathrm{clip}_C\left(\frac{f(x+\lambda Gu;\xi)-f(x-\lambda Gu;\xi)}{2\lambda}\right) + z\right) Gu,\ z \sim \frac{1}{|B|}\mathcal{N}(0, qC^2\sigma^2)$
$x \leftarrow \eta g/q$

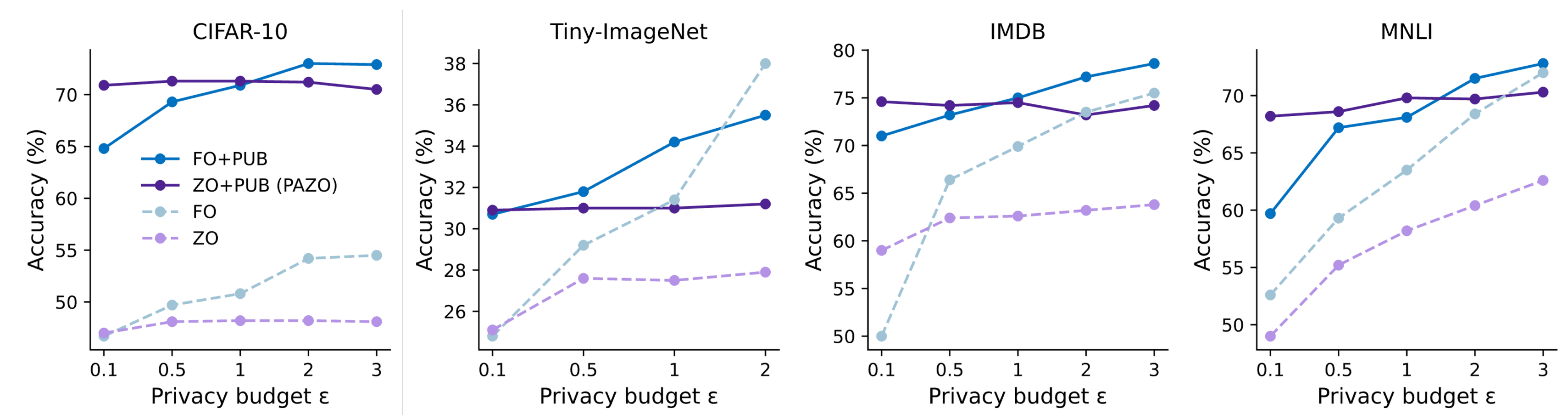
PAZO-S: Select the Best Public Gradient

Sample k batches of public data and compute gradients $\{g_1, \ldots, g_k\}$

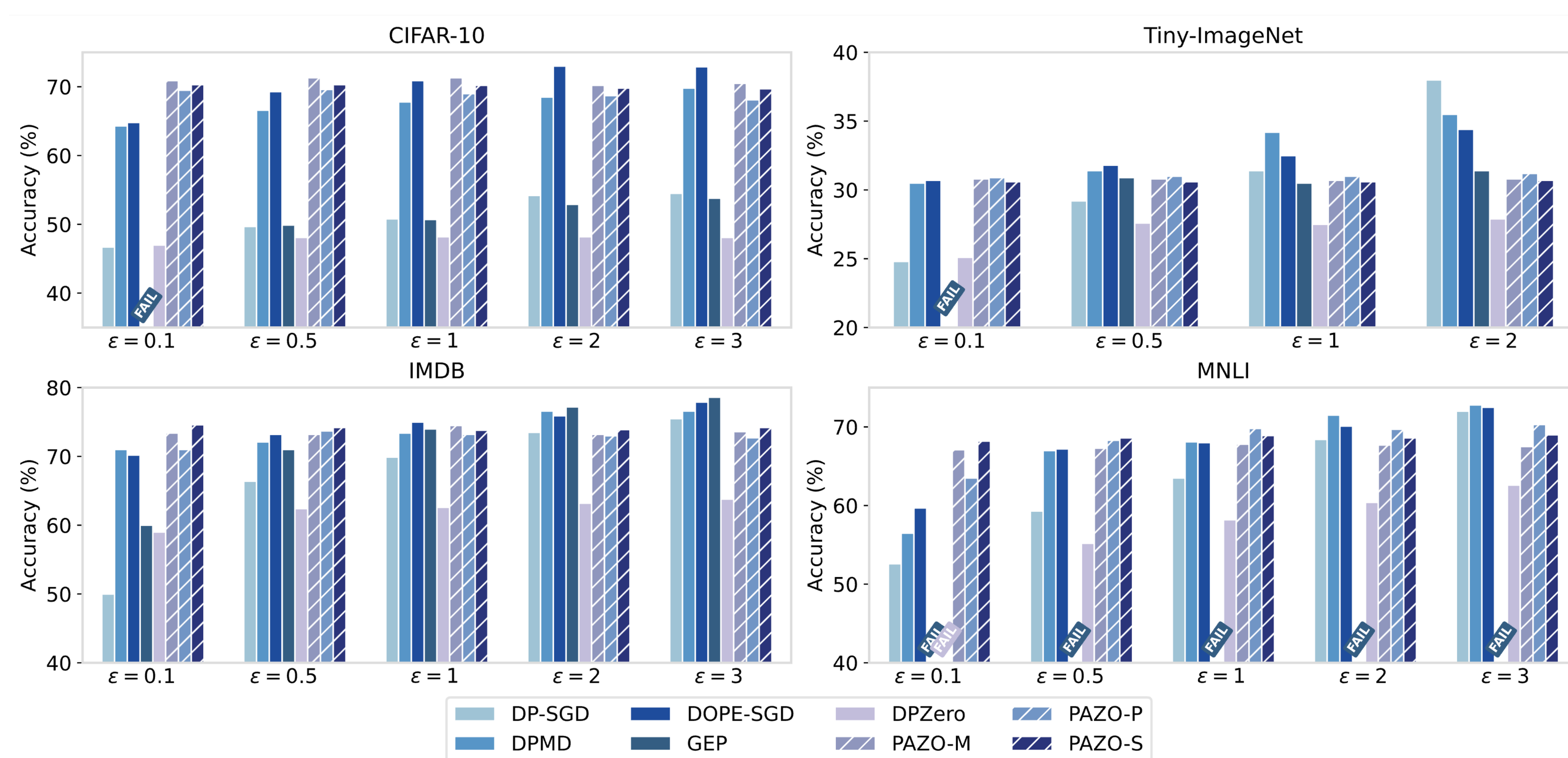
// find the best public descent direction
for $j = 1 \ldots k$:
  $f_j \leftarrow \frac{1}{|B|}\sum_{\xi \in B} \mathrm{clip}_C\left(f(x - \eta g_j; \xi)\right) + z,\ z \sim \frac{1}{|B|}\mathcal{N}(0, (k+1)C^2\sigma^2)$
$j^\star \leftarrow \arg\min_{j \in [k]} f_j$

// perturb the best candidate and compare
$g_{k+1} \leftarrow g_{j^\star} + z'$ where $z' \sim \mathcal{N}(0, \epsilon^2 I_d)$
$f_{k+1} \leftarrow \frac{1}{|B|}\sum_{\xi \in B} \mathrm{clip}_C\left(f(x - \eta g_{k+1}; \xi)\right) + z,\ z \sim \frac{1}{|B|}\mathcal{N}(0, (k+1)C^2\sigma^2)$
$j^\star \leftarrow \arg\min_{j \in [k+1]} f_j$
$x \leftarrow x - \eta g_{j^\star}$

Improved Privacy/Utility Tradeoffs

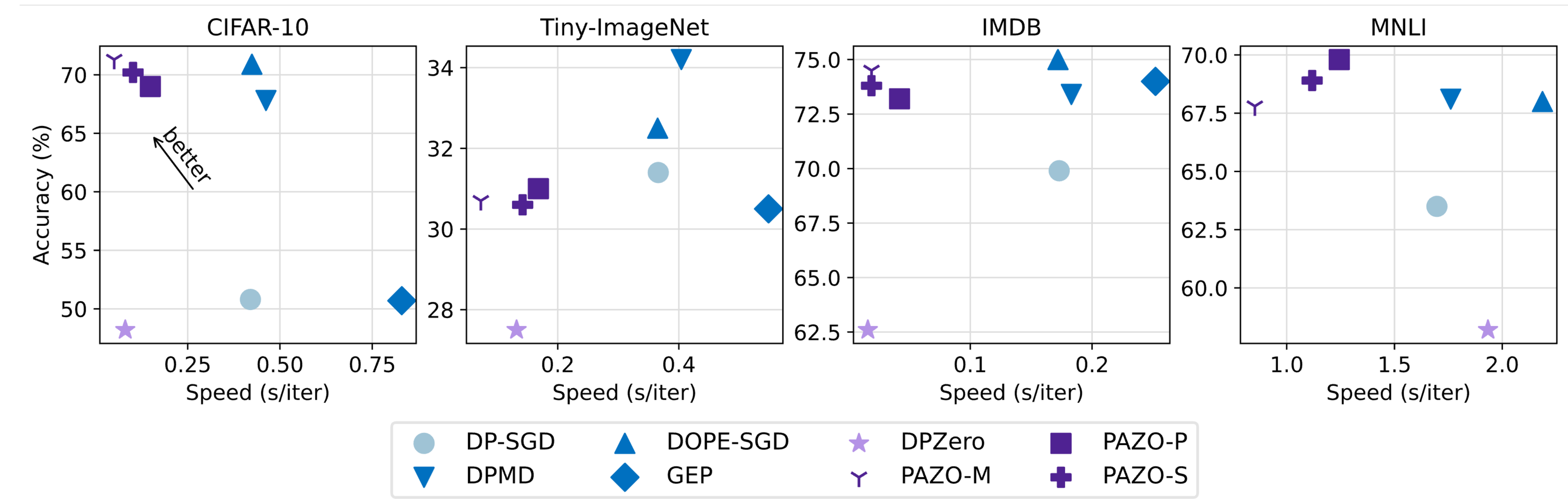


- **Without public data**, vanilla zeroth-order (ZO) underperforms first-order (FO)
- **With public data**, our method **(PAZO)** outperforms the best first-order with public data **(FO+PUB)**, especially in highly private regimes



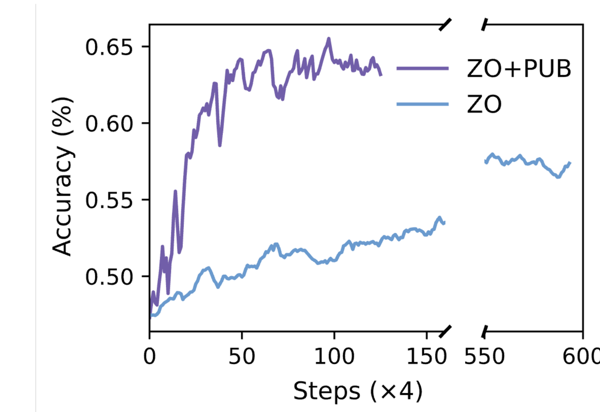
Detailed comparison between PAZO-* and all the baselines

Time Efficiency



PAZO is up to **16× faster in each training iteration** than FO and FO+PUB while staying performant

Slow convergence is a known disadvantage of zeroth-order methods, but PAZO **converges faster** than vanilla ZO (DPZero)



Convergence

[**γ-similar**] Public B' and private data B are γ-similar if $\|\nabla f(x;B) - \nabla f(x;B')\| \le \gamma, \forall x$

Our assumptions: L-smooth, M-lipschitz, γ-similar, and optionally $|f(x;B)| \le S, \forall x$

Method	wo. $\lvert f(x)\rvert \le S$	w. $\lvert f(x)\rvert \le S$
DP-SGD	$O(\sqrt{d})$	/
DPZero	/	$O(\sqrt{d}\log d)$
PAZO-M	$O(\frac{1-\alpha}{\alpha}\sqrt{d})$	$O(\frac{1-\alpha}{\alpha}d^{\frac{1}{4}})$
PAZO-P	$O(k)$	$O(\sqrt{k}\log k)$
PAZO-S	$O(c)$	

c is constant independent of k and d.

Takeaways:

- PAZO-M improves prior work by factor $d^{\frac{1}{4}}\log d$, and PAZO-{P,S} achieves d-independent rates
- Due to using biased public gradients, we additionally have an error $O(\gamma^2)$, which reduces as γ reduces

Future Work

- Sharpen the convergence bounds by considering other data similarity metrics
- Explore a broader set of (public, private) dataset pairs in practical DP applications
- Leverage insights from differential geometry

References

- [DPZero] Zhang, et al. "Dpzero: Private fine-tuning of language models without backpropagation." ICML 2024.
- [MeZO] Malladi, et al. "Fine-tuning language models with just forward passes." NeurIPS 2023.

Code: https://github.com/xuchengong/pazo