Do Not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning

Da Yu*,[1], Huishuai Zhang*,[2], Wei Chen[2], Tie-Yan Liu[2]

* denotes equal contribution

[1]Sun Yat-sen University

[2]Microsoft Research

Corresponding to:

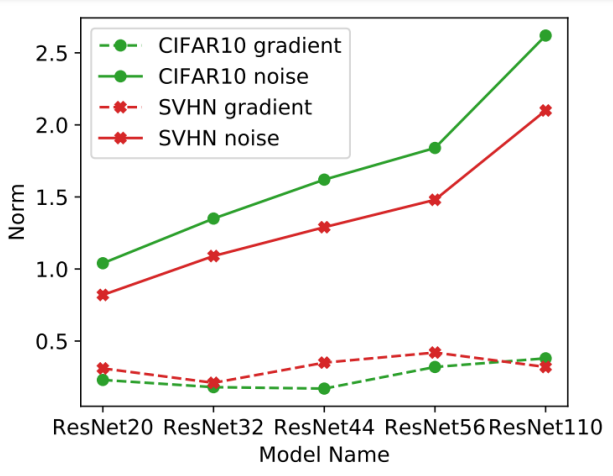
yuda3@mail2.sysu.edu.cn or

Huishuai.Zhang@microsoft.com

BACKGROUND

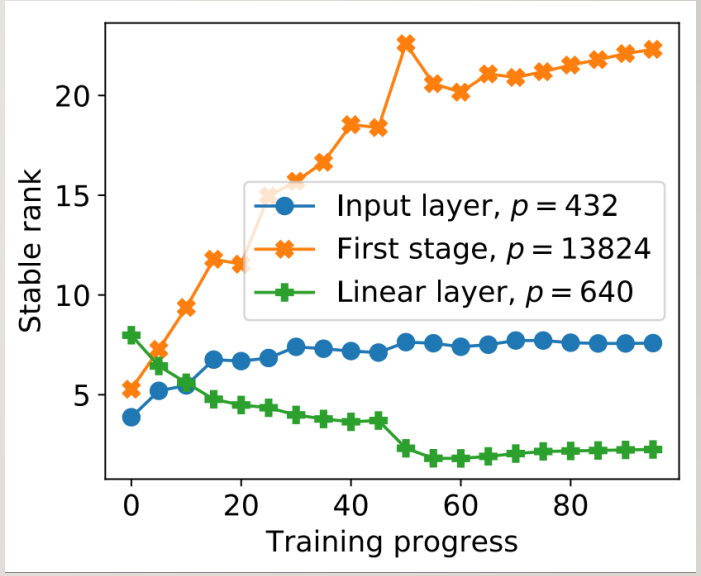
- Gradient perturbation:
  - $\tilde{g} = g + z,\ g \in \mathbb{R}^p$ and $z \sim N(0, \sigma^2 I_{p\times p})$.

- The curse of dimensionality:
  - The intensity of the added noise scales with p.



MOTIVATION

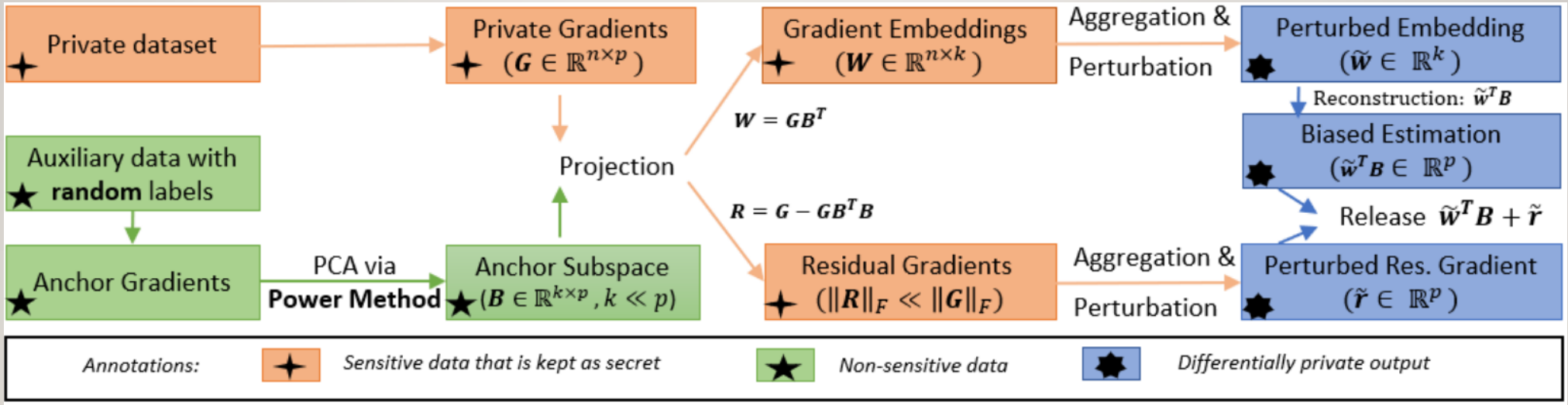
- Gradients have low stable rank!
  - Energy concentrates on low-dimensional subspace.



- To reduce the influence of noise:
  - Find a proper subspace for projection,
  - Deal with the projection error.

METHODOLOGY OVERVIEW

- Flow chart of Gradient Embedding Perturbation (GEP):



ANALYSIS OF GEP

- Compared with GP, the output of GEP
  - has smaller variance because we use low-dimensional noise for gradient embeddings,
  - is an unbiased estimator because of the residual gradients (if without clipping).

- For convex DP-ERM, we show

$$\mathbb{E}[L(\bar{\theta})] - L(\theta_*) \leq O\left(\frac{\sqrt{k\,log\left(\frac{1}{\delta}\right)}}{n\epsilon} + \frac{\bar{r}\sqrt{p\,log\left(\frac{1}{\delta}\right)}}{n\epsilon}\right)$$

, where k is the projection dimension and $\bar{r}$ denotes the norm of residual gradients.

EXPERIMENTS

GP: standard gradient perturbation.
PATE: private learning with ensemble of teachers.
B-GEP: Biased-GEP, GEP without residual gradients.

Table 1: Test accuracy (in %) with varying choices of privacy bound ϵ. The numbers under symbol Δ denote the improvement over GP baseline.

Dataset	Algorithm	$\epsilon = 2$	Δ	$\epsilon = 5$	Δ	$\epsilon = 8$	Δ
MNIST	GP	94.7	+0.0	96.8	+0.0	97.2	+0.0
	PATE	98.5	**+3.8**	98.5	**+1.7**	98.6	**+1.4**
	B-GEP	93.1	-1.6	94.5	-2.3	95.9	-1.3
	GEP	96.3	+1.6	97.9	**+1.1**	98.4	**+1.2**
SVHN	GP	87.1	+0.0	91.3	+0.0	91.6	+0.0
	PATE	80.7	-6.4	91.6	+0.3	91.6	+0.0
	B-GEP	88.5	+1.4	91.8	+0.5	92.3	+0.7
	GEP	92.3	**+5.2**	94.7	**+3.4**	95.1	**+3.5**
CIFAR-10	GP	43.6	+0.0	52.2	+0.0	56.4	+0.0
	PATE	34.2	-9.4	41.9	-10.3	43.6	-12.8
	B-GEP	50.3	+6.7	59.5	+7.3	63.0	+6.6
	GEP	59.7	**+16.1**	70.1	**+17.9**	74.9	**+18.5**