Lecture 18:
Differentially Private Machine Learning

Foundations of Privacy

Carnegie Mellon University

Announcement

- HW 3 released. Due Nov. 14th

① Written Component (pdf)
② Programming component (ipynb)
} zip file for submission

Support Jeremy

https://gofund.me/d922f9f9

Search | How it works | Start a GoFundMe

gofundme

Support the Lacomis family after brain surgery

Afsoon Afzal is organizing this fundraiser on behalf of Jessica Lacomis.

Created 2 days ago | Medical, Illness & Healing

Model Training with DP

Given private data $x_1, \ldots, x_n$, solve

$$\min_{w\in\mathbb{R}^d} \underbrace{L(w)}_{\text{Empirical Risk.}} \equiv \frac{1}{n}\sum_{i=1}^{n} \ell(w; \overset{x_i}{\cancel{w_i}})$$

subject to differential privacy

DP-SGD (in Theory)

Differentially Private SGD [BST14, SCS13]

- At each iteration t,
  - Gradient estimate on a mini-batch B_t :

$$g_t = \left(\frac{1}{|B_t|} \sum_{i \in B_t} \nabla \ell(w_t; x_i) \right)$$

  - Noisy gradient update :

$$w_{t+1} = w_t - \eta \left(g_t + Z_t \right),$$
$$Z_t \sim \mathcal{N}(0, \sigma^2 I_d)$$

Privacy Proof
assumes ℓ is L-Lipschitz
for all x , for w

$$\|\nabla \ell(\overset{w_t; x_i}{\cancel{x_t; s_i}})\|_2 \leq L$$

Set σ to scale with L

DP-SGD (in Practice)

Differentially Private SGD [ACGMMTZ16]

- At each iteration t,
  - Average *clipped* gradient estimate:

$$g_t = \left(\frac{1}{|B_t|} \sum_{i \in B_t} \text{Clip}(\nabla \ell(w_t; x_i), G) \right)$$

  - Noisy gradient update :

$$w \quad \cancel{x}_{t+1} = \underset{w}{\cancel{x}_t} - \eta \left(g_t + Z_t\right),\ Z_t \sim \mathcal{N}(0, \sigma^2 I_d)$$

If $\|g\|_2 > G$,

what is $Clip(g, G)$?

$\left(g \cdot \frac{1}{\|g\|_2}\right) \cdot G$

Gradient Clipping:

$$\text{Clip}(\underset{\uparrow \atop \nabla \ell(w\,;\,x_i)}{g}, G) = g \min\left\{1, \frac{G}{\|g\|_2}\right\}$$

Set σ to scale with G

Privacy Guarantee for DP-SGD (with Clipping)

[BST14, ACGMMTZ16]

T: # iterations

Theorem: DP-SGD with gradient clipping of threshold G satisfies (ϵ, δ)-differential privacy, if the noise rate

$\sqrt{T}$ growth → advanced composition

$$\sigma \geq a \frac{G q \sqrt{T \ln(1/\delta)}}{\epsilon}$$

for some constant a and $\boxed{q = \frac{|B_t|}{n}}$.

→ sub-sampling rate.

↳ privacy amplification via sub-sampling.

How about convergence and optimality?

Gradient clipping can create bias

- *Xiangyi Chen, Z. S.W., Mingyi Hong*
  "Understanding Gradient Clipping in Private SGD: A Geometric Perspective"
  In NeurIPS 2020 (Spotlight)

Bad Example 1

$w, x_i \in \mathbb{R}$

Loss: $L(x) = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{2}(w - x_i)^2$

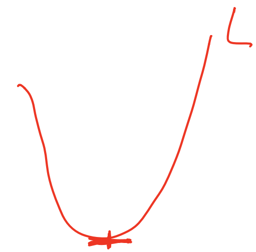
where $x_1 = x_2 = -3$ and $x_3 = 9$.

$\Rightarrow$ Optimum $w^\star = 1$

Clipped gradient at $w^\star$

$$\mathbb{E}[\text{Clip}(\nabla_x \ell(w^\star; x_i), 1)] = 1/3$$

$\Rightarrow$ push iterates away from opt



$w^* = 1$

$\nabla_w L(w^*; x) = 0$

Bad Example 2

$$\text{Loss: } L(w;x) = \frac{1}{2}\sum_{i=1}^{2}\frac{1}{2}(w - x_i)^2$$

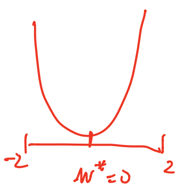
where $x_1 = 3,\ x_2 = -3$

$\Rightarrow$ Optimum $w^\star = 0$

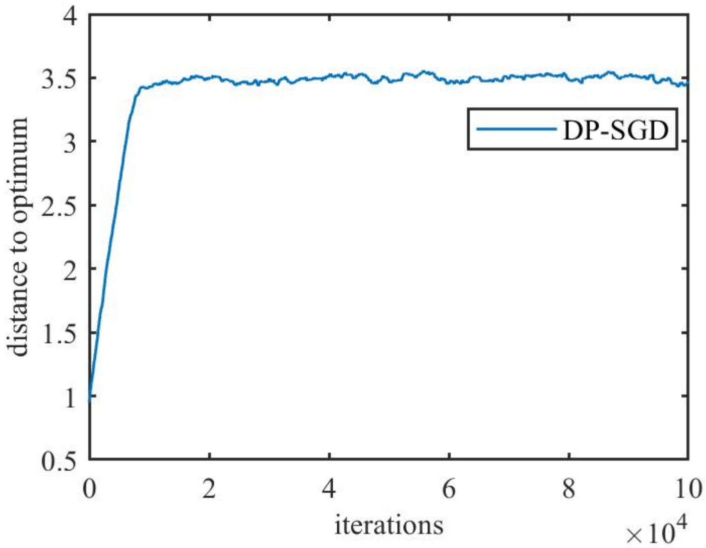
Clipped gradient for any $w \in [-2,2]$

$$\mathbb{E}[\text{Clip}(\nabla_x \ell(w; x_i), 1)] = 0$$

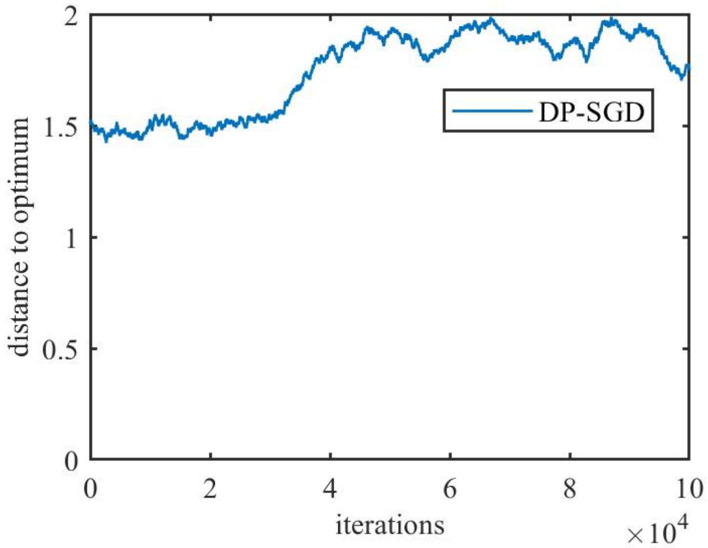
$\Rightarrow$ does not converge to opt



Adversarial Effects of Clipping



Example 1



Example 2

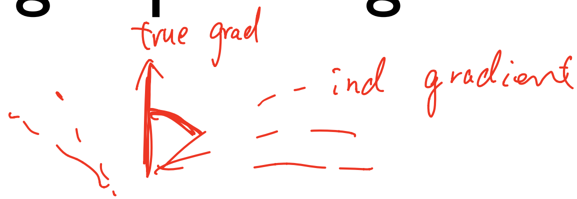
Do these occur in practical instances?

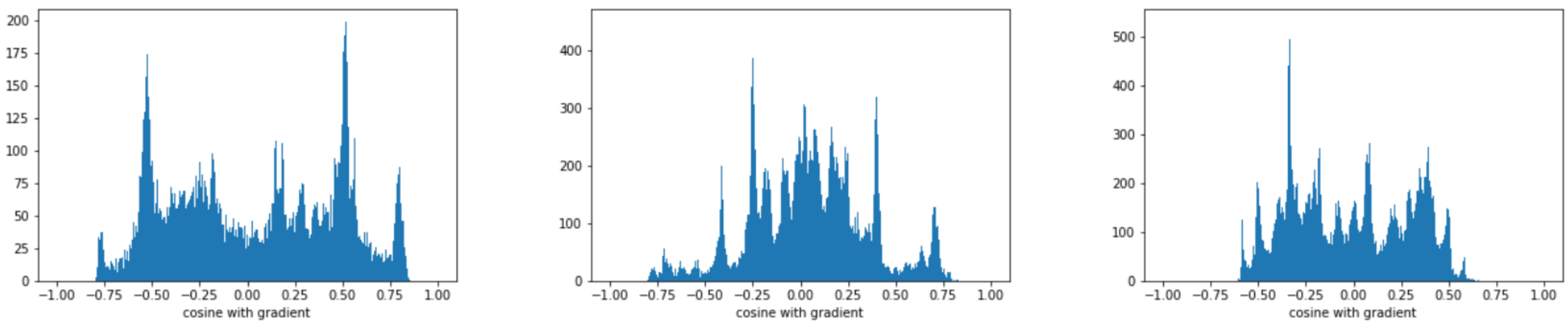
DP-SGD on MNIST

- DP-SGD with Clip norm $G = 1$
  60 epochs, $\epsilon \approx 3$, test accuracy $\approx 96.5\,\%$

- DP-SGD with Clip norm $G = 0.1$
  60 epochs, $\epsilon \approx 3$, test accuracy $\approx 92\,\%$

A glimpse of gradient distribution





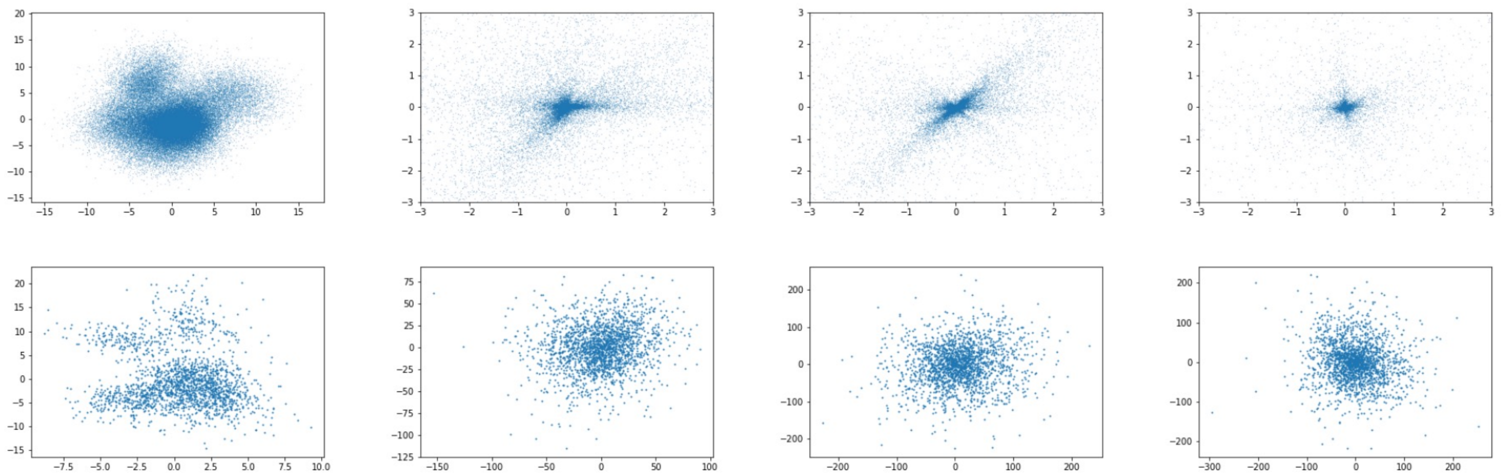
(a) Epoch 4 (b) Epoch 10 (c) Epoch 59

Histogram of cosine between
stochastic gradients and true gradient

Symmetric structures in gradients still lead to convergence under clipping.

Gradient Distribution of NN

Visualization with random projection

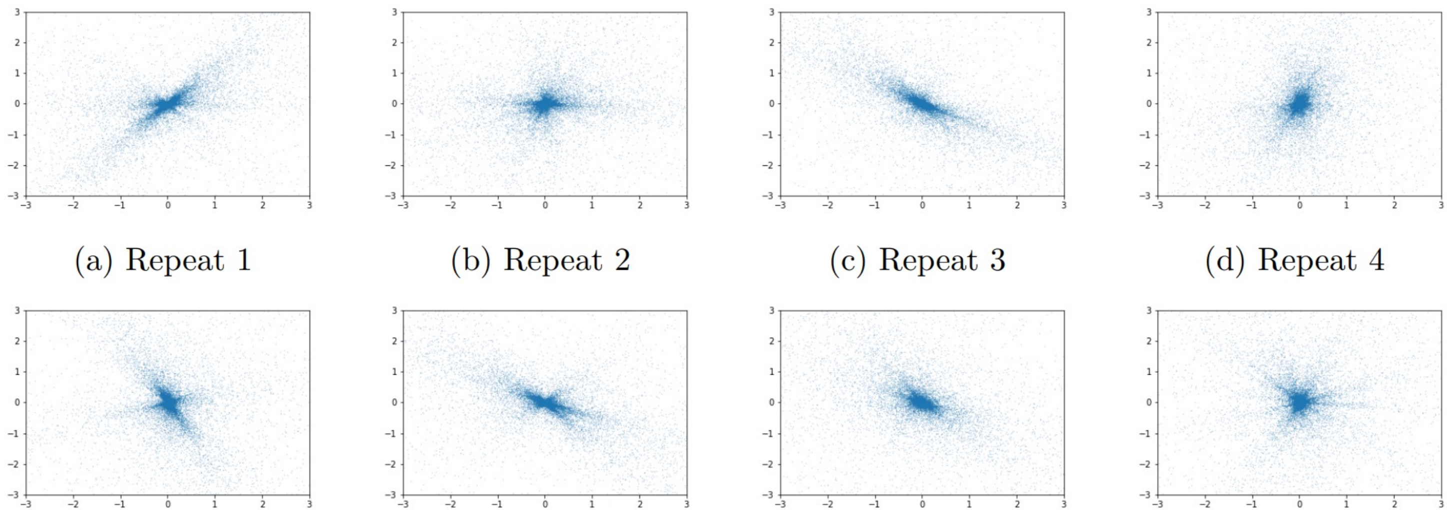


(a) Epoch 0 (b) Epoch 5 (c) Epoch 10 (d) Epoch 59

Figure 1: Gradient distributions on MNIST (top row) and CIFAR10 (bottom row) at the end of different epochs (indexed by columns). The gradients for epoch 0 are computed at initialization (before training).

Gradient Distribution of NN

Multiple random projections

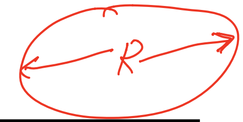


(a) Repeat 1 (b) Repeat 2 (c) Repeat 3 (d) Repeat 4

(e) Repeat 5 (f) Repeat 6 (g) Repeat 7 (h) Repeat 8

Figure 2: Gradient distributions on MNIST at the end of epoch 9 projected using different random matrices.

Convergence Guarantee for DP-SGD (in Theory)



Convex Set C

Consider DP-SGD with Projection

Theorem: Let $L: C \to \mathbb{R}$ be convex and L-Lipschitz.

Suppose $C \subseteq \mathbb{R}^d$ is a convex set with diameter R.

Let w^* be the minimizer of L in the set C.

- For regular SGD (w/ projection)

$$L(\hat{w}) - L(w^*) \leq \frac{RL}{\sqrt{T}}.$$

- For DP-SGD (w/ projection),

$$\mathbb{E}\left[L(\hat{w}) - L(w^*)\right] \leq O\left(\frac{RL\sqrt{d\ln(1/\delta)}}{n\varepsilon}\right)$$

Leveraging low-dimensional structure in gradients

- *Yingxue Zhou, Z. S.W., Arindam Banerjee*
  "Bypassing the Ambient Dimension: Private SGD with Gradient Subspace"
  In ICLR 2021

Dimensionality

$$O\left(\frac{C\sqrt{d\ln(1/\delta)}}{n\epsilon}\right) + \frac{1}{T}\sum_{t=1}^{T} W_{\nabla f(x_t),C}(\tilde{p}_t, p_t)$$

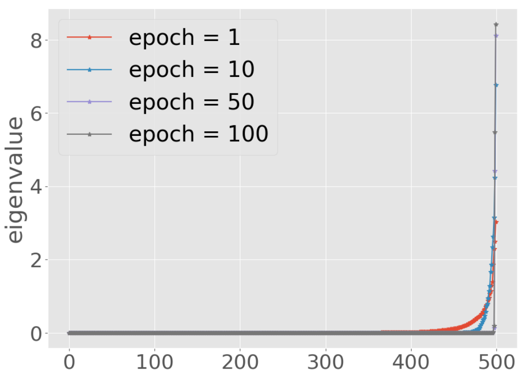
DP-SGD without clipping
Depends on ambient dimension d

Clipping bias

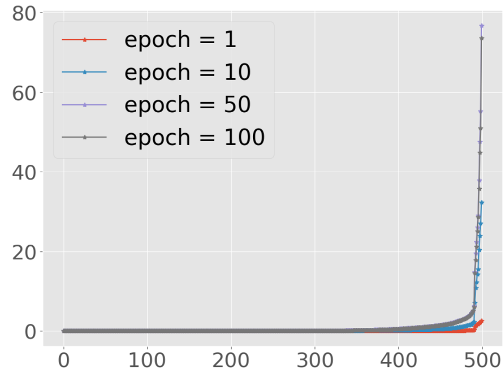
Spectrum of Gradient Second Moments

Eigenvalues of
$$M_t = \mathbb{E}[\nabla \ell(x_t, s_i) \nabla \ell(x_t, s_i)^\intercal]$$

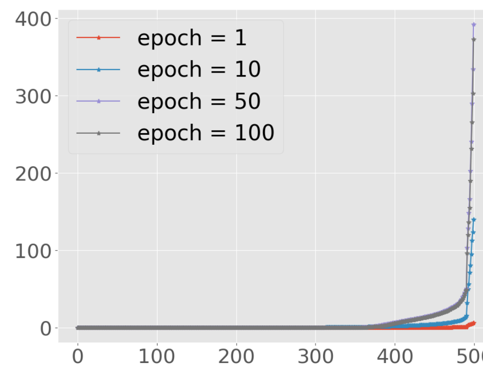
SGD



DP-SGD $\sigma = 1$



DP-SGD $\sigma = 2$



Order or eigenvalues from largest to smallest
Ambient dimension $d \approx$ 130,000

[ZWB21]

Projected DP-SGD (PDP-SGD)

Assume small amount of public data (no privacy concern)

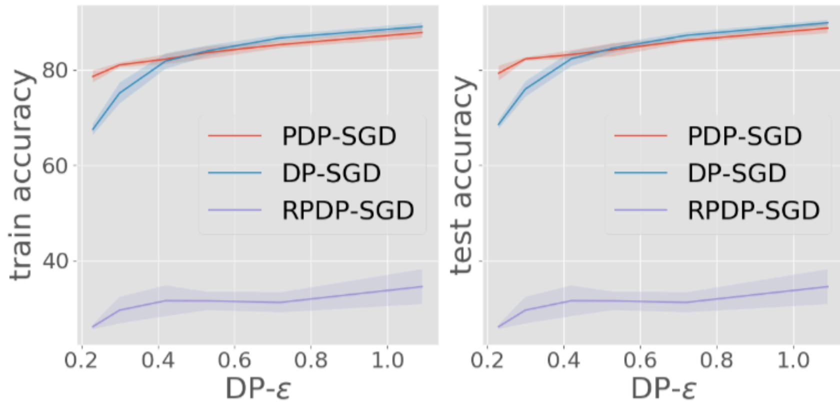
PDP-SGD [ZWB21]

- For $t = 1, \ldots, T$
  - Gradient estimate on a mini-batch B_t :
    $\tilde{g}_t \leftarrow$ noisy gradient estimate with Gaussian noise
  - Use public data to compute projection Π_k onto the top-k eigenspace of M_t
  - Update :
    $x_{t+1} = x_t - \eta \, \Pi_k \tilde{g}_t$

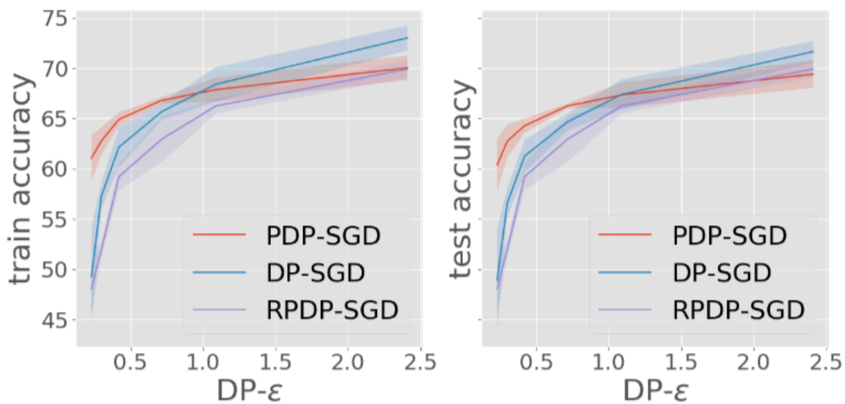
Balancing two sources of error

- Error due to projection
  $\|\Pi_k \nabla \ell(x; s_i) - \nabla \ell(x; s_i)\|$
- Gradient perturbation in the subspace $\approx \dfrac{\sqrt{k}}{n\epsilon}$
  (from $\sqrt{d}$ to $\sqrt{k}$)

PDP-SGD $\longrightarrow$ DP-SGD with gradient projection onto low-dim subspace.

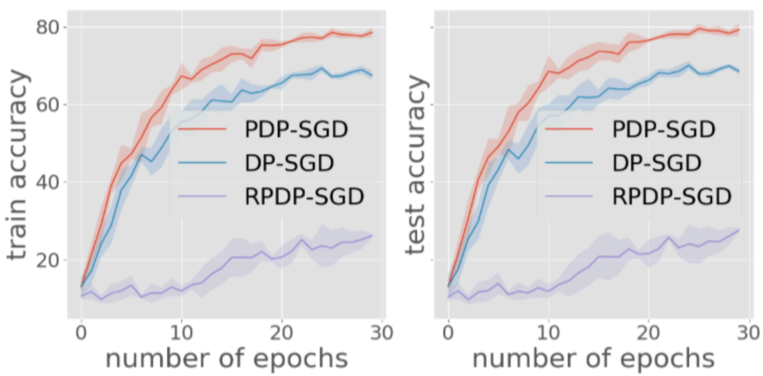


(a) MNIST

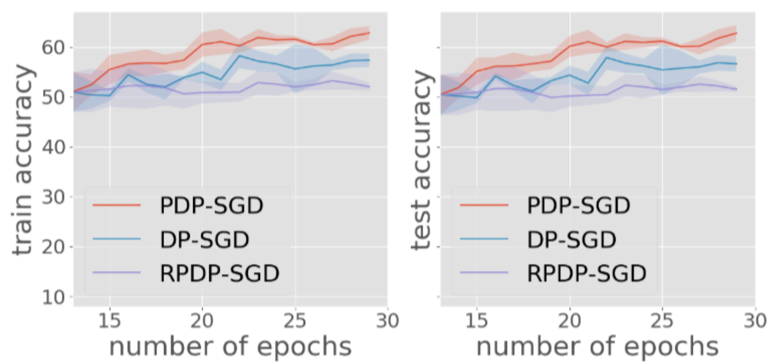


(b) Fashion MNIST

Training Dynamics



(a) MNIST, $\epsilon = 0.23$



(b) Fashion MNIST, $\epsilon = 0.30$

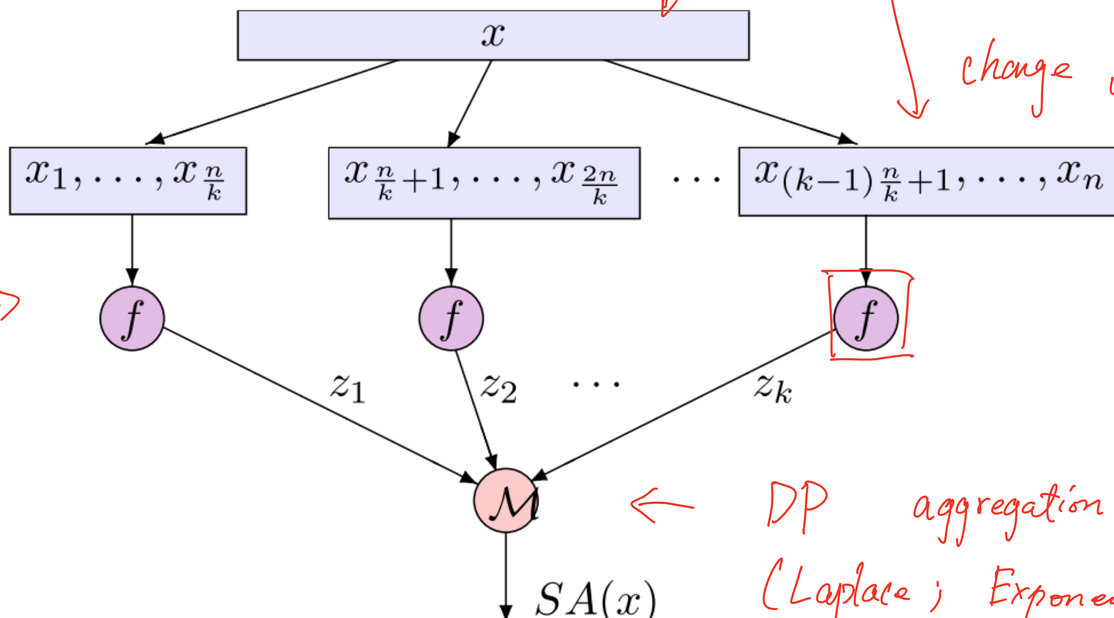
What if DPSGD is not applicable?

→ (generalized) linear { Regression
Classification (e.g. logistic, SVM)

→ Neural Networks.

Reduce the problem to Non-private ML.

Subsample and Aggregate



Change x_i

change one block.

k

Blocks

Run some Algo. (without DP)

DP aggregation. (Laplace ; Exponential)

Private Aggregation of Teacher Ensembles (PATE)