Lecture 14

Modern Machine Learning

Neural Networks

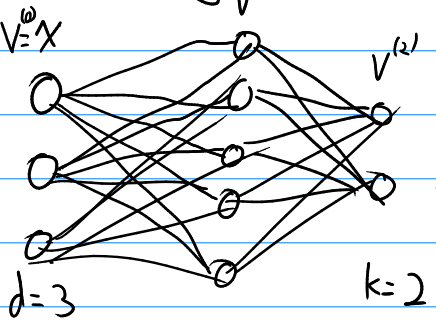
MLP

Dataset D, (x, y)

Loss fn l

$$\mathcal{L}(\theta, D) = \sum_{i \in [n]} l(\theta, x_i, y_i)$$

$x \in \mathbb{R}^d$, $y \in [k]$



$$v^{(i)} = f^{(i)}\left(W^{(i)} v^{(i-1)} + b^{(i)}\right)$$

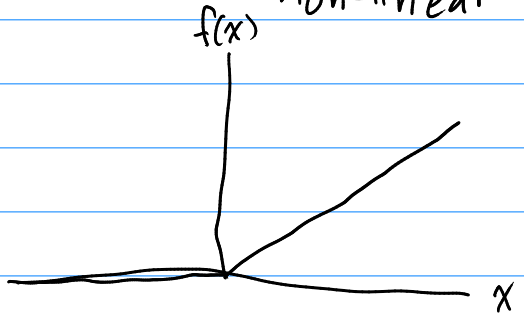
vector f applied entrywise

$v^{(i)} \in \mathbb{R}^{m_i}$

Params $\begin{cases} W^{(i)} \in \mathbb{R}^{m_i \times m_{i-1}} \\ b^{(i)} \in \mathbb{R}^{m_i} \end{cases}$

activation fn $f^{(i)} : \mathbb{R} \to \mathbb{R}$ non-linear

ReLU Rectified linear unit $f(x) = \max(0, x)$



Sigmoid : $f(x) = \frac{1}{1+e^{-x}}$

tanh : $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$dim(o) = k$

$$\hat{y} = \text{softmax}(o)$$

$$\hat{y}_i = \frac{\exp(o_i)}{\sum_{j \in [k]} \exp(o_j)}$$

Not convex

$$l(\theta, x, y) = -\sum_{j=1}^{k} y_j \log(\hat{y}_j)$$

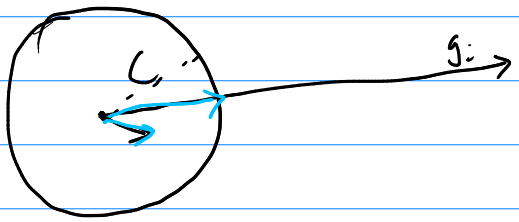
$\vec{y} = 1$ in coord of true label

0 elsewhere

Differentially Private Stochastic Gradient Descent Algorithm

BST '14
ACGMMT2 '16

1. Select (x_i, y_i) randomly ⟷ 1. "Lot" of points. size L (in exp.) each p.t. selected w.p. $\frac{L}{n}$
2. Compute $n\nabla l(\theta_t, x_i, y_i) + G_t$ ⟷ 2. Compute $g_i = \nabla l(\theta_t, x_i, y_i)$ $\forall i \in$ lot
3. Step in - direction ⟷ 3. Clip g_i to l_2 ball of rad. C, average them
4. Add noise
5. Step in - direction



Clipping

Sampling Schemes

Select 1 rnd pt. | Choose each p.t. w.p. $\frac{1}{n}$.

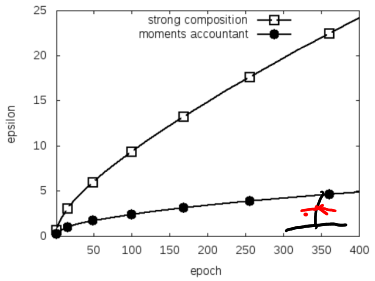
- Poisson Sampling
- Sampling w/o replacement
- Sampling w/ replacement
- Permute dataset, iterate over it, selecting B pts each time

Privacy Loss $q = \frac{L}{n}$

BST'14: Gauss + subsamp + adv. comp. (# of iters)

$(\varepsilon, \delta)\text{-DP} \Rightarrow \left(O\left(q\varepsilon\sqrt{T\log(1/\delta)}\right), T\delta\right)\text{-DP}$

$\left(O(q\varepsilon\sqrt{T}), \delta\right)\text{-DP}$



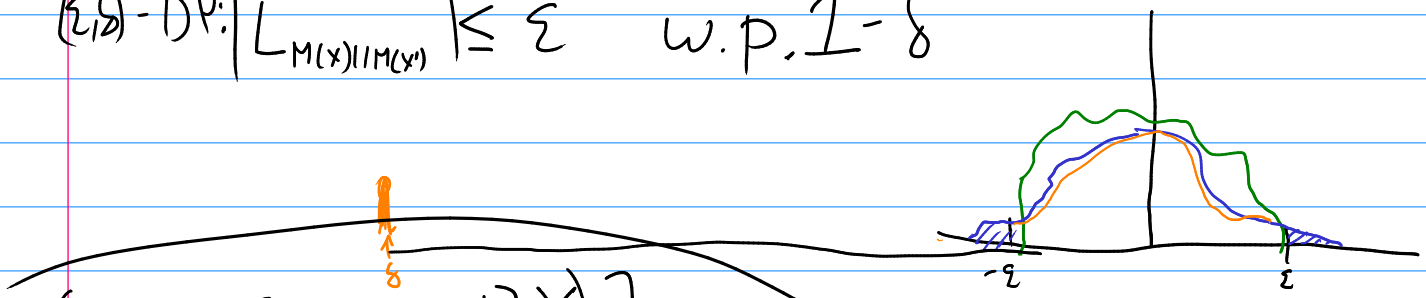
Jagielski, Ullman, Oprea

$$L_{Y\|Z} = \ln\left(\frac{\Pr[Y=t]}{\Pr[Z=t]}\right)$$

$t \sim Y$

ε-DP: $|L_{M(x)\|M(x')}| \le \varepsilon$ w.p. 1 $(\forall x, x' \text{ nbr})$

(ε,δ)-DP: $|L_{M(x)\|M(x')}| \le \varepsilon$ w.p. $1-\delta$



$$\left(\ln \mathop{E}_{t\sim M(x)}\left[\left(\frac{\Pr[M(x)=t]}{\Pr[M(x')=t]}\right)^{\lambda}\right] \le \gamma\right) \Rightarrow (\varepsilon,\delta)\text{-DP}$$

$$\delta = \Pr_{t\sim M(x)}\left[\ln\left(\frac{\Pr[M(x)=t]}{\Pr[M(x')=t]}\right) \ge \varepsilon\right]$$

$$= \Pr\left[\exp\left(\lambda \ln\left(\frac{\Pr[M(x)=t]}{\Pr[M(x')=t]}\right)\right) \ge \exp(\lambda\varepsilon)\right]$$

$$\le \frac{\exp(\gamma)}{\exp(\lambda\varepsilon)} = \exp(\gamma - \lambda\varepsilon)$$

RDP

Rényi-DP [Mironov '17]

Moments Accountant

1. $\forall \lambda$, compute γ for:
   Poisson subsample + Gaussian Mech.) Not data dependent
2. Apply comp. for T iterations
3. RDP $\rightarrow (\varepsilon, \delta)$-DP (use best λ)

0 1 3 5

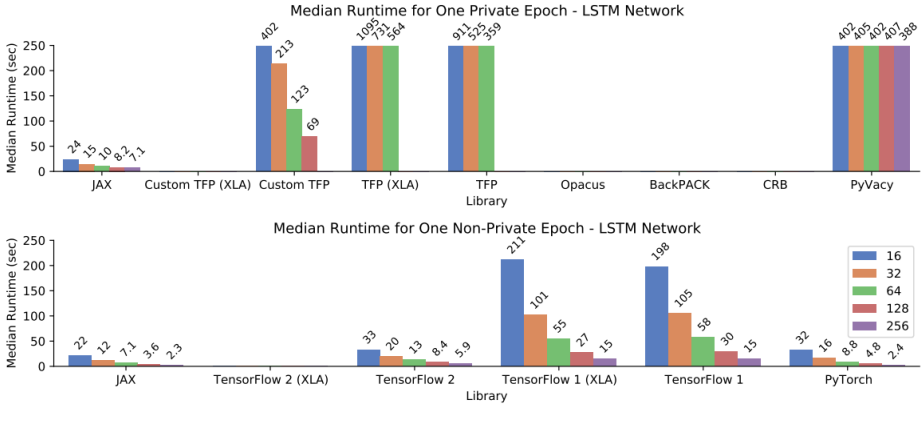
Accuracy?

MNIST 99.8% Non private
98.1% w/privacy. $(2.93, 10^{-5})$-DP

CIFAR-10 99.7% Non-privately
66.2% $(7.53, 10^{-5})$-DP

Speed of DPSGD

Clipping = Slow!

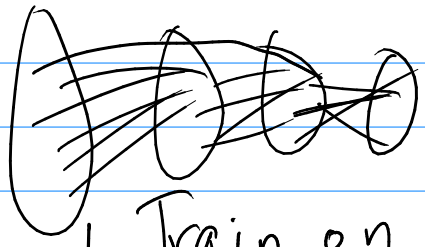


Changes in Architecture

- Activation fn's
  Non-private: ReLU
  Private: tanh!
- Size vs. Noise

Exploiting Public Data

- Hyperparameter tuning
  Liu Talwar '19
  Tune on public data

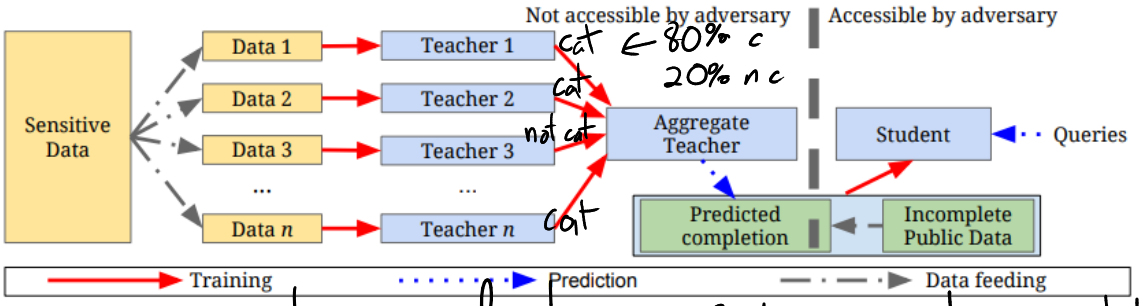


1. Train on a public dataset
2. Freeze most layers, privately train rest

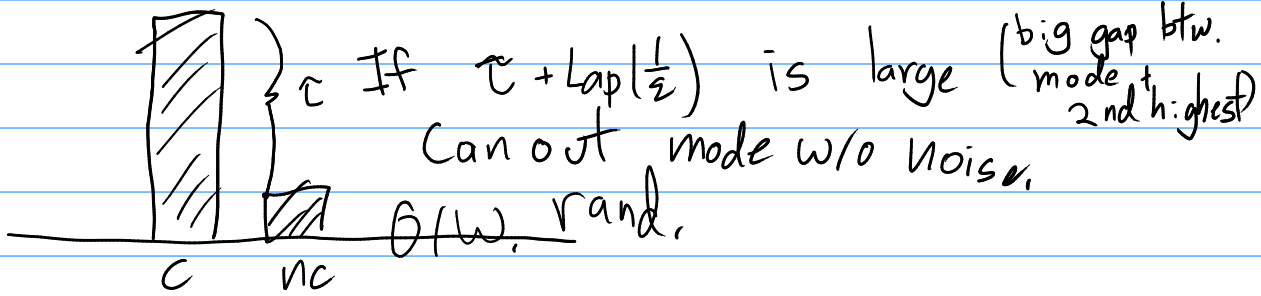
Private Aggregation of Teacher Ensembles

PATE — PAEGT '17 — Papernot et al '18
BTT '18

- Black Box - Non private learners at core
- Private labeled training data
  + Public unlabeled training data
- Stability - Sample and aggregate — Nissim Raskhodnikova Smith '07
  - "Sample" 1. Run np. algo on disjoint datasets
  - "Aggregate" 2. Combine results in a DP way.
    (np complexity) $\cdot \frac{1}{\varepsilon}$



1. One private prediction. (feature vector public)
Given x, output $\hat{y}$ privately — TS '13



$(x_i, \hat{y}_i)$

public — privacy-preserving

$(x_1, \hat{y}_1), (x_2, \hat{y}_2), \ldots$

synthetic private dataset

2. Many (T) classification tasks.

- Advanced comp → $\sqrt{T}$ cost
- Sparse vector → $\sqrt{c}$ cost

↑ #of > thresh.

which q's are unstable

3. A classifier??