CS 499/599: Machine Learning Security

03.07: (Differential) Privacy

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

Notice

- Due dates (in Mar.)
  - 9th: Final project presentation
    - *11 min* presentation + *3-5 min* Q&A (strict)
    - Presentation *MUST* cover:
      - 1-2 slides on your research *motivation* and *goals*
      - 1-2 slides on your *ideas* (how do you plan to achieve your goals)
      - 1 slides on your *experimental design*
      - 2-3 slides on your *most interesting results*
      - 1 slides on your *conclusions* and *next steps*
  - 14th: Final exam (online, 24 hrs., unlimited trials)
  - 14th: Final project report (Template is on the website)
  - 16th: HW4 deadline (HW 1-3 late submissions are available; HW4 won't have late submissions)

- Sign-up (on Canvas)
  - Scribe lecture note [only 1 slots remain; today!]

In-class Presentation (Akshith Gunasekaran) – Red Teaming Language Models (LMs) with LMs

Topics for Today

- Privacy Attacks and Defenses
  - Non-ML: Data anonymization
  - Membership inference
    - Attacks: Yeom *et al.* and Shokri *et al.*
    - Defensive techniques
  - Model inversion
    - Attacks: Fredrikson *et al.* and Carlini *et al.*
    - Defensive techniques
  - Model extraction
    - Attacks: Tramer *et al*. and Jagielski *et al*.
    - Defensive techniques
  - Differential Privacy
    - DP-SGD
    - DP-SGD in Practice

What Can We Do To Reduce Privacy Risks of ML?

Abadi *et al.*, Deep Learning with Differential Privacy

Revisit'ed – Differential Privacy

- ϵ-Differential Privacy
  - A randomized algorithm $M: D \rightarrow R$ with domain D and a range R satisfies ϵ-differential privacy if for any two adjacent inputs $d, d' \in D$ and any subset of outputs $S \subset R$ it holds

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S]$$

- (ϵ, δ)-Differential Privacy

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta$$

  - δ: Represent some catastrophic failure cases [Link, Link]
  - $\delta < 1/|d|$, where $|d|$ is the number of samples in a database

Revisit'ed – Differential Privacy

- (ϵ, δ)-Differential Privacy **[Conceptually]**

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta$$

  - You have two databases d, d' differ by one item
  - You make the same query M to each and have results $M(d)$ and $M(d')$
  - You ensure the distinguishability between the two under a measure ϵ
    - ϵ is large: those two are distinguishable, less private
    - ϵ is small: the two outputs are similar, more private
  - You also ensure the catastrophic failure probability δ

Revisit'ed – Differential Privacy

- (ϵ, δ)-Differential Privacy

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta$$

- Mechanism for (ϵ, δ)-DP: Gaussian noise

$$\mathcal{M}(d) \triangleq f(d) + \mathcal{N}(0, S_f^2 \cdot \sigma^2)$$

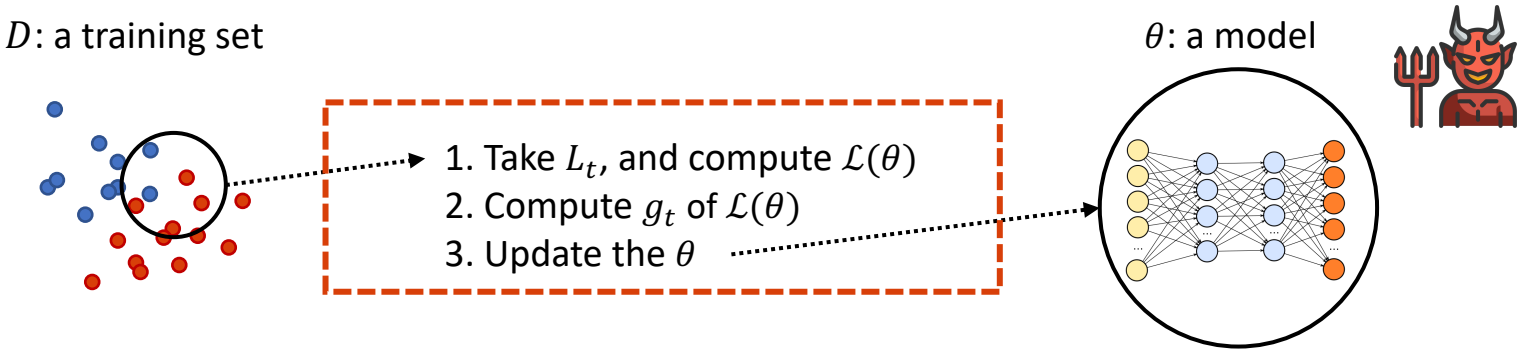
  - $M(d)$: (ϵ, δ)-DP query output on d
  - $f(d)$: non (ϵ, δ)-DP (original) query output on d
  - $N(0, S_f^2 \cdot \sigma^2)$: Gaussian normal distribution with mean 0 and the std. of $S_f^2 \cdot \sigma^2$

Post-hoc: Set the Goal ϵ and Calibrate the noise $S_f^2 \cdot \sigma^2$!

How Do We Use DP for ML?

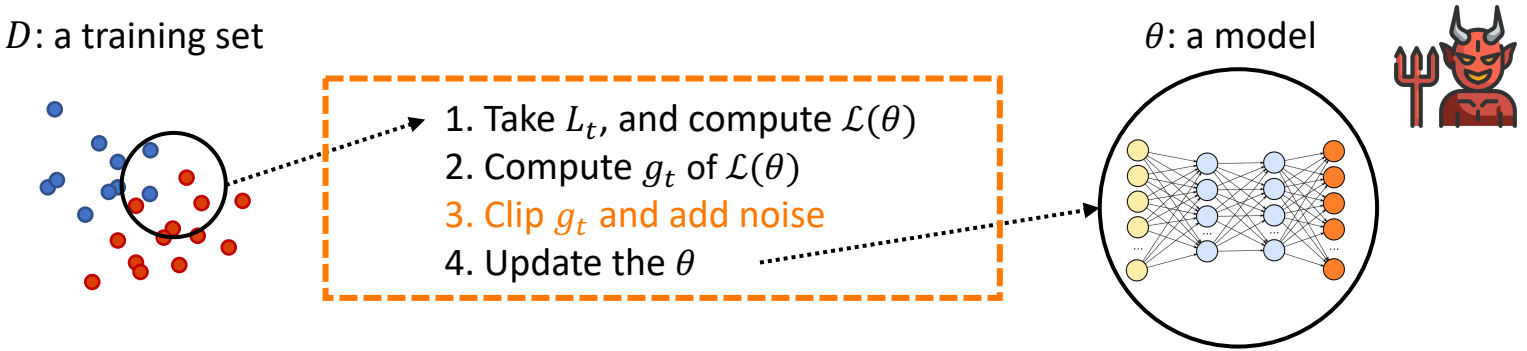
- Revisit'ed – Stochastic Gradient Descent (SGD)
  1. At each step t, it takes a mini-batch L_t
  2. Computes the loss $\mathcal{L}(\theta)$ over the samples in L_t, w.r.t. the label y
  3. Computes the gradients g_t of $\mathcal{L}(\theta)$
  4. Update the model parameters θ towards the direction of reducing the loss

This Process Should Be (ϵ, δ)-DP!



Make an SGD Step (ϵ, δ)-DP

- Stochastic Gradient Descent (SGD)
  1. At each step t, it takes a mini-batch L_t
  2. Computes the loss $\mathcal{L}(\theta)$ over the samples in L_t, w.r.t. the label y
  3. Computes the gradients g_t of $\mathcal{L}(\theta)$
  4. Clip (scale) the gradients to $1/C$, where $C > 1$
  5. Add Gaussian random noise $N(0, \sigma^2 C^2 \mathbf{I})$ to g_t
  6. Update the model parameters θ towards the direction of reducing the loss



Make the Whole SGD Process (ϵ, δ)-DP

- Stochastic Gradient Descent (SGD)
  - SGD iteratively computes the (ϵ, δ)-DP step T times
  - **Problem:** how do we compute the total privacy leakage ϵ_{tot} over T iterations?

- Privacy accounting with moment accountant
  - **Key intuition:** DP has the **composition** property
    - Suppose the two mechanism M_1 and M_2 satisfies $(\varepsilon_1, \delta_1)$- and $(\varepsilon_2, \delta_2)$-DP the composition of those mechanisms $M_3 = M_2(M_1)$ satisfies $(\varepsilon_1+\varepsilon_2, \delta_1+\delta_2)$-DP
    - If each step t satisfies (ε, δ)-DP, the total SGD process satisfies $(\varepsilon T, \delta T)$-DP
  - **Moment accountant:** tracking the total privacy leakage εT over T iterations

Putting All Together

- DP-Stochastic Gradient Descent (DP-SGD)

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do** // we train a model θ with the privacy budget ε_{budget}

Take a random sample L_t with sampling probability L/N // iterate over *T* mini-batches

Compute gradient // compute the gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$ // clip the magnitude of the gradients

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$ // add Gaussian random noise to the gradients

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

$\varepsilon, \delta \leftarrow$ compute the privacy cost (leakage) so far // compute the privacy cost (leakage) up to *t* iterations

If $\varepsilon > \varepsilon_{buget}$: then break; // if the cost is over the budget, then stop training

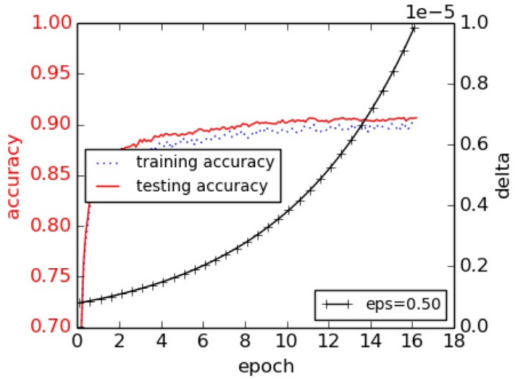
Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Evaluation

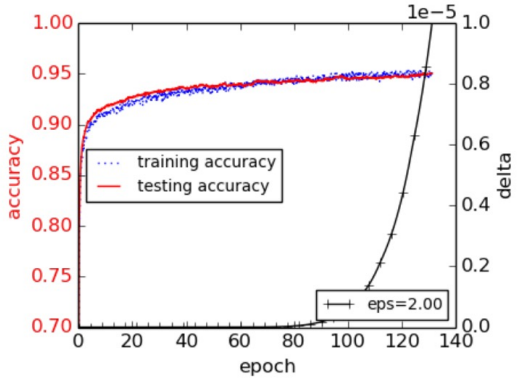
- Setup
  - Datasets: MNIST | CIFAR-10/100
  - Models:
    - MNIST: 2-layer feedforward NN on 60-dim. PCA projected inputs
    - CIFAR-10/100: A CNN with 2 conv. layers and 2 fully-connected layers
  - Metrics:
    - Classification accuracy
    - Privacy cost (ε_{budget})

Evaluation

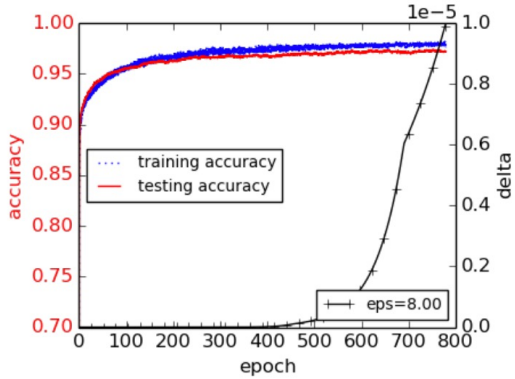
- Impact of Noise
  - Dataset, Models: MNIST, 2-layer feedforward NN
  - Setup: 60-dim PCA projected inputs | Clipping threshold ($\mathbf{C}$): 4 | Noise (σ): 8, 4, 2 (from the left)
  - **Summary:**
    - On MNIST, DP-SGD offers reasonable acc. under various privacy costs (**clean:** 98.3%)
    - The accuracy of private models decreases as we decrease the privacy cost



(1) Large noise



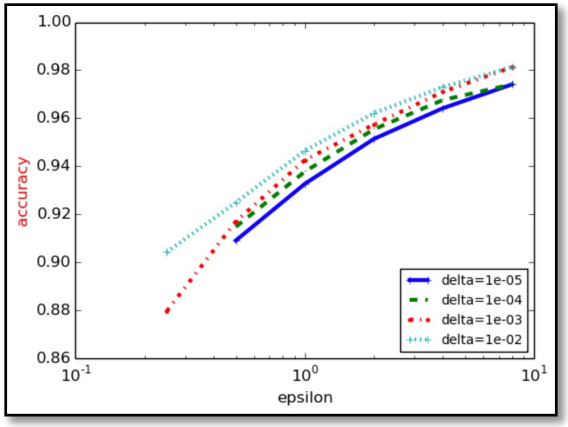
(2) Medium noise



(3) Small noise

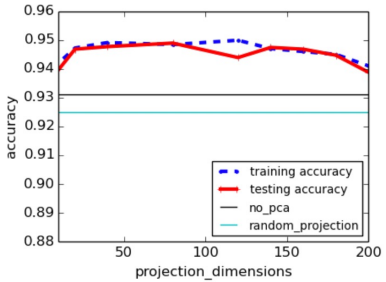
Evaluation

- Impact of Noise
  - Dataset, Models: MNIST, 2-layer feedforward NN
  - Setup: 60-dim PCA projected inputs | Clipping threshold ($\mathbf{C}$): 4 | Noise (σ): 8, 4, 2 (from the left)
  - **ummary:**
    - On MNIST, DP-SGD offers reasonable acc. under various privacy costs (**clean:** 98.3%)
    - The accuracy of private models decreases as we decrease the privacy cost

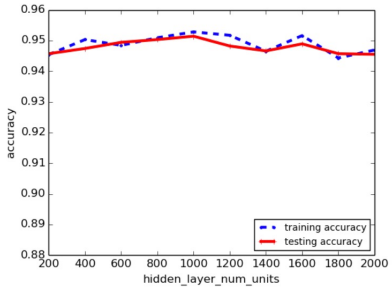


Evaluation

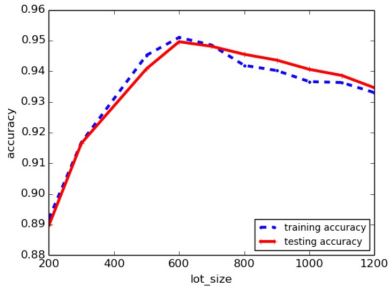
- Impact of Hyper-parameter Choices
  - Dataset, Models: MNIST, 2-layer feedforward NN
  - Setup: 60-dim PCA projected inputs



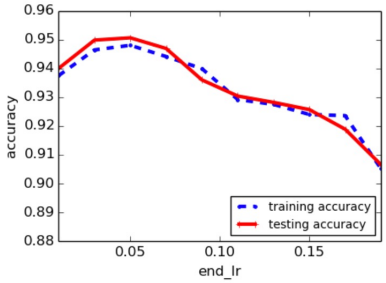
(1) variable projection dimensions



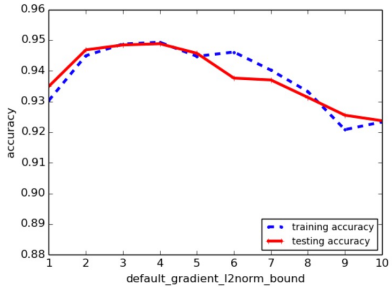
(2) variable hidden units



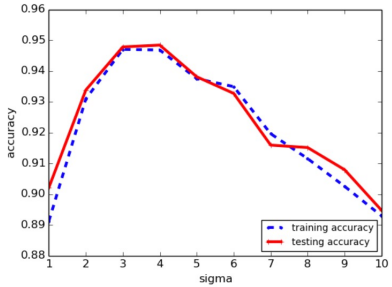
(3) variable lot size



(4) variable learning rate



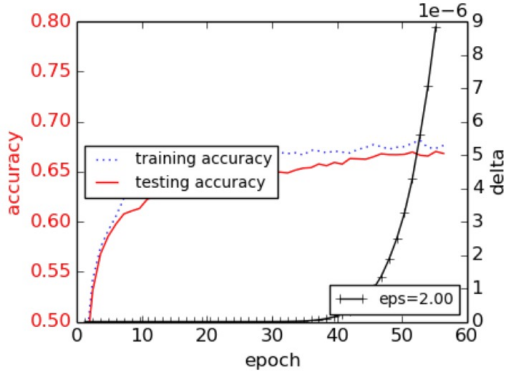
(5) variable gradient clipping norm



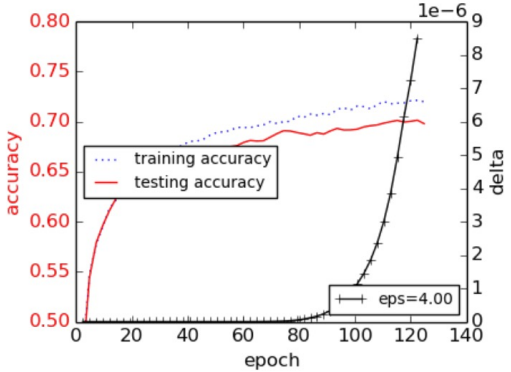
(6) variable noise level

Evaluation

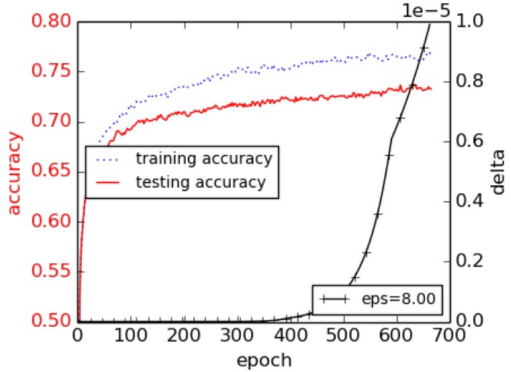
- Impact of Noise
  - Dataset, Models: CIFAR-10, CNN
  - Setup: Clipping threshold ($\mathbf{C}$): 3 | Noise (σ): 6
  - **Summary:**
    - On CIFAR-10, DP-SGD offers reasonable acc. under various privacy costs (**clean:** 80%)
    - The accuracy of private models decreases as we decrease the privacy cost



(1) $\varepsilon = 2$



(2) $\varepsilon = 4$



(3) $\varepsilon = 8$

Topics for Today

- Privacy Attacks and Defenses
  - Non-ML: Data anonymization
  - Membership inference
    - Attacks: Yeom *et al.* and Shokri *et al.*
    - Defensive techniques
  - Model inversion
    - Attacks: Fredrikson *et al.* and Carlini *et al.*
    - Defensive techniques
  - Model extraction
    - Attacks: Tramer *et al*. and Jagielski *et al*.
    - Defensive techniques
  - Differential Privacy
    - DP-SGD
    - DP-SGD in Practice

What Does It Mean by Epsilon = 2/4/6 in CIFAR-10?

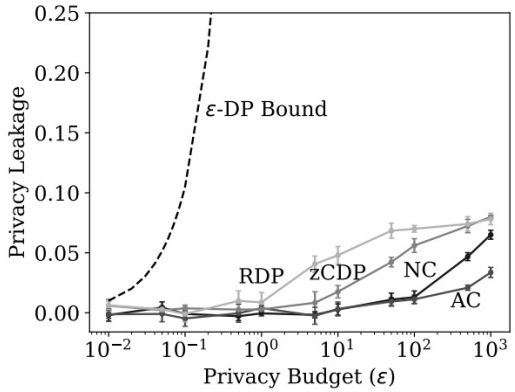
Jayaraman *et al.*, Evaluating Differentially Private Machine Learning in Practice

Empirical Evaluations of Privacy Risks in DP-Models

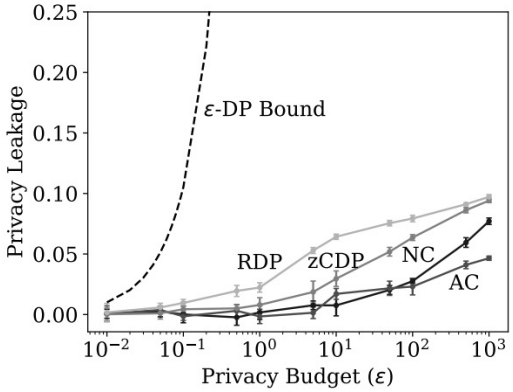
- Setup
  - **Datasets:** Purchase-100 | CIFAR-100 (on 50-dim PCA projected inputs)
  - **Models:** Logistic regressions | 2-layer feedforward NNs
  - **Privacy Attacks:**
    - Membership inference: Yeom *et al*. and Shokri *et al*.
  - **DP-SGD:**
    - Set the clipping norm ($\mathbf{C}$) to 1
    - Set the prob. of catastrophic failures ($\boldsymbol{\delta}$) to $10^{-5} < 1/|N|$ (N~60k in MNIST and 50k in CIFAR)
    - Set the batch size to 200
    - Set the learning rate to 0.01 for Adam optimizer
    - Vary ε from 0.01 to 1000
    - Compare (ϵ, δ)-DP with other DP-mechanisms: AC, CDP, zCDP, and RDP
    - Run 5-times and measure the (TPR – FPR) and accuracy loss on average

Evaluation on CIFAR-100, LRs

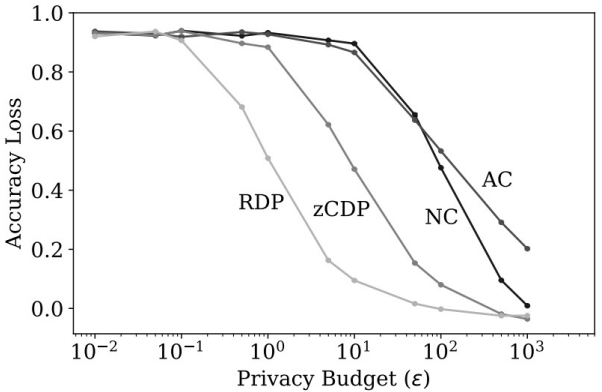
- Summary
  - Yeom *et al*. and Shokri *et al*. are weak privacy attacks
  - In other words, (ϵ, δ)-DP theoretically offers very strong privacy bounds
  - If a DP-mechanism offers stronger bound, the acc. of models decrease accordingly



(a) Shokri et al. membership inference



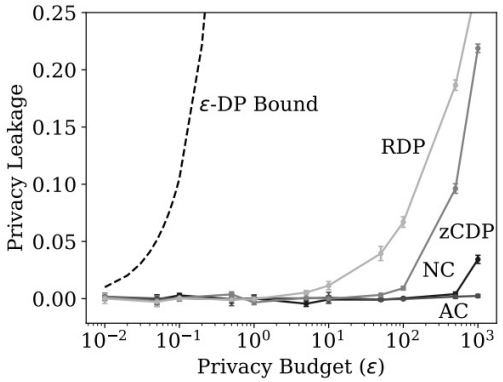
(b) Yeom et al. membership inference



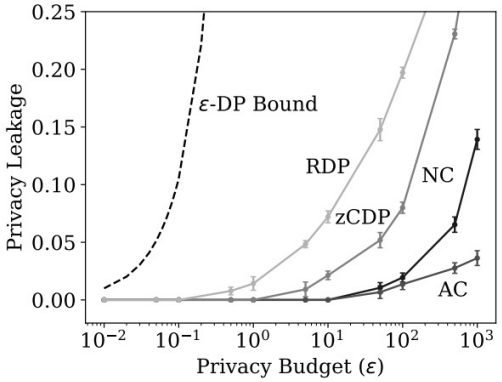
(b) Per-instance gradient clipping

Evaluation on CIFAR-100, NNs

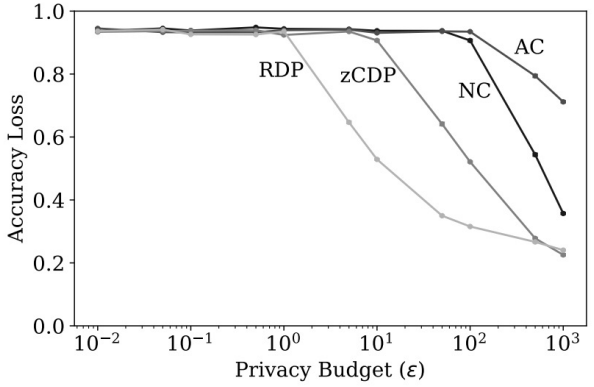
- Summary
  - Yeom *et al*. and Shokri *et al*. are weak privacy attacks
  - In other words, (ϵ, δ)-DP theoretically offers very strong privacy bounds
  - If a DP-mechanism offers stronger bound, the acc. of models decrease accordingly
  - Compared to LRs, NNs leak more in higher privacy budgets



(a) Shokri et al. membership inference



(b) Yeom et al. membership inference



(a) CIFAR-100

Evaluation on MI Predictions: LRs vs. NNs

- Summary
  - Yeom *et al*. and Shokri *et al*. are weak privacy attacks
  - In other words, (ϵ, δ)-DP theoretically offers very strong privacy bounds
  - If a DP-mechanism offers stronger bound, the acc. of models decrease accordingly
  - Compared to LRs, NNs leak more in higher privacy budgets
  - Predictions (TPRs and FPRs) are more consistent in LRs than NNs in CIFAR-100

Figure 3: Overlap of membership predictions across two runs of logistic regression with RDP at $\epsilon = 1000$ (CIFAR-100)

(a) Overlap of membership predictions across two runs

Recap: Privacy!

- Privacy Attacks and Defenses
  - Non-ML: Data anonymization
  - Membership inference
    - Attacks: Yeom *et al.* and Shokri *et al.*
    - Defensive techniques
  - Model inversion
    - Attacks: Fredrikson *et al.* and Carlini *et al.*
    - Defensive techniques
  - Model extraction
    - Attacks: Tramer *et al*. and Jagielski *et al*.
    - Defensive techniques
  - Differential Privacy
    - DP-SGD
    - DP-SGD in Practice

Thank You!

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL

Secure AI Systems Lab