CS 499/599: Machine Learning Security

02.14: Data Poisoning

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

sanghyun.hong@oregonstate.edu

SAIL

Secure AI Systems Lab

Notice

- Due dates
  - Checkpoint Presentation II (on the 16th)
    - *15-min* presentation + *3-5 min* Q&A
    - Presentation *MUST* cover:
      - 1 slide on your research topic
      - 1-2 slides on your motivation and goal(s)
      - 1-2 slides on your *ideas* (how do you plan to achieve your goals)
      - 1-2 slides on your *experimental design* (in detail)
      - 1-2 slides on your *hypotheses* and *preliminary results* [*very important*]
      - 1 slide on your *next steps* until the final presentation

- Sign-up (on Canvas)
  - Scribe Lecture Note [~5 more slots remain]
  - In-class Paper Presentation / Discussion [~4 more slots remain]

In-class Presentation (Quintin Pope) – Zoom-in: Introduction to Circuits

Topics for Today

- Motivation
  - Evade spam filter
  - DDoS detection
- Data Poisoning:
  - Attacks
    - Indiscriminate attacks on: SVMs and regression models
    - Targeted attacks on: DNNs (Poison Frogs and Meta-poison)
  - Defenses
    - Certified defenses
    - Differential privacy
  - Conclusion (and implications)

Steinhardt *et al.*, Certified Defenses for Data Poisoning Attacks
Ma *et al.*, Data Poisoning Attacks against Differentially-Private Learners: Attacks and Defenses

Traditionally, computer security seeks to ensure a system's integrity against attackers by creating clear boundaries between the system and the outside world (Bishop, 2002). In machine learning, however, the most critical ingredient of all–the training data–comes directly from the outside world.

– Steinhardt, Koh, and Liang, NeurIPS'17

Motivation

- Prior work
  - Many successful attacks, *e.g.*, [Biggio *et al.* 2012], on classification tasks
  - Defenses, *e.g.*, RONI, showed their effectiveness against those attacks

Wait, What's the Worst-case of Data Poisoning?

Threat Model

- Setup [binary classification task!]
  - **Data:** $x \in X$ (ex. R^d), $y \in Y = \{-1, +1\}$
  - **Clean train-set:** D_c of size n / **Test-set:** S
  - **Loss function:** $l(\theta; x, y) = \max(0, 1 - y\langle\theta, x\rangle)$
  - **Test-loss:** $\mathrm{L}(\theta) = \mathrm{E}_{(x,y)\sim S}[l(\theta; x, y)]$

- Attacker
  - **Goal:** Indiscriminate attack (increase the test-loss $\mathrm{L}(\theta)$)
  - **Capability:** D_p: inject ϵn poisons, where $\epsilon \in [0, 1]$, into D_c
  - **Knowledge:** D_c and the defense algorithm that will be used [white-box]

- Defender
  - **Goal:** Trains a model on $D_c \cup D_p$ and produce a model $\hat{\theta}$ that minimizes $\mathrm{L}(\hat{\theta})$

Threat Model: Defenses

- Setup [binary classification task!]
  - **Data:** $x \in X$ (ex. R^d), $y \in Y = \{-1, +1\}$
  - **Clean train-set:** D_c of size n / **Test-set:** S
  - **Loss function:** $l(\theta; x, y) = \max(0, 1 - y\langle\theta, x\rangle)$
  - **Test-loss:** $\mathrm{L}(\theta) = \mathrm{E}_{(x,y)\sim S}[l(\theta; x, y)]$

- Data sanitization defenses
  - **Goal:** Examine $D_c \cup D_p$ and remove poisons (*e.g.*, outliers)

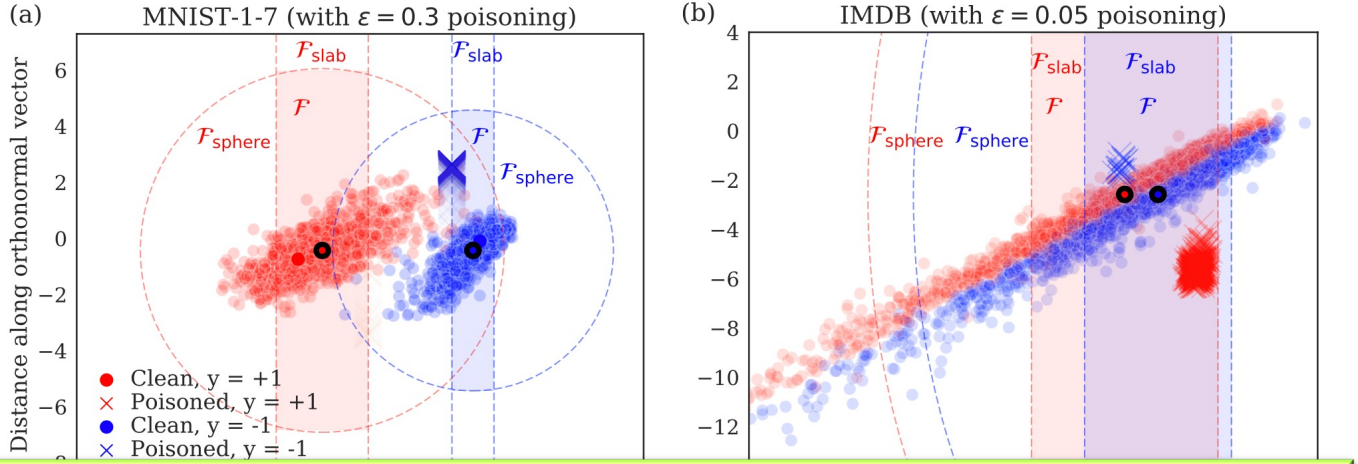
$$\hat{\theta} \overset{\text{def}}{=} \underset{\theta\in\Theta}{\operatorname{argmin}}\, L(\theta; (\mathcal{D}_{\mathrm{c}} \cup \mathcal{D}_{\mathrm{p}}) \cap \mathcal{F}), \quad \text{where } L(\theta; S) \overset{\text{def}}{=} \sum_{(x,y)\in S} \ell(\theta; x, y)$$

A filtered dataset

  - **Methods:**
    - *Fixed* (oracle) defense: when we know the true distribution of data (unrealistic)
    - *Data-dependent* defense: when we don't know the true distribution (real-world!)

Example Data Sanitization Defenses

- Data sanitization defenses
  - **Goal:** Examine $D_c \cup D_p$ and remove poisons (*e.g.*, outliers)
  - **Example defenses:**
    - *sphere* defense: removes points outside a spherical radius
    - *slab* defense: first project points onto the line btw. the centroids and then remove



Effectiveness of Defenses Largely Depends on the Data

The Worst-case Test Loss by Data Poisoning

- Upper-bound [refer to the paper for its derivation]

$$\max_{D_p} \mathrm{L}(\hat{\theta}) \;\leq\; \max_{D_p \subseteq F} \min_{\theta \in \Theta} \frac{1}{n} \mathrm{L}\left(\theta; D_c \cup D_p\right) \stackrel{\text{def}}{=} \mathbf{M}$$

  - **M:** the minimax loss
  - **It means:** the attack is bounded to a scenario where all poisons are alive under F!

The Worst-case Test Loss by Data Poisoning

- Upper-bound [refer to the paper for its derivation]

$$\max_{D_p} \mathrm{L}(\hat{\theta}) \;\leq\; \max_{D_p \subseteq F} \min_{\theta \in \Theta} \frac{1}{n} \mathrm{L}\left(\theta; D_c \cup D_p\right) \stackrel{\text{def}}{=} \mathbf{M}$$

  - **M:** the minimax loss
  - **It means:** the attack is bounded to a scenario where all poisons are alive under F!

- Two defense scenarios
  - ***Fixed* defense:** when we know the true distribution of data
  - ***Data-dependent* defense:** when we don't know the true distribution of data

The Worst-case Test Loss by Data Poisoning

- Upper-bound [refer to the paper for its derivation]

$$\max_{D_p} \mathrm{L}(\hat{\theta}) \;\leq\; \max_{D_p \subseteq F} \min_{\theta \in \Theta} \frac{1}{n} \mathrm{L}\left(\theta; D_c \cup D_p\right) \stackrel{\text{def}}{=} \mathbf{M}$$

  - **M:** the minimax loss
  - **It means:** the attack is bounded to a scenario where all poisons are alive under F!

- Two defense scenarios
  - ***Fixed* defense:** we can fix F regardless of poisoning samples
  - ***Data-dependent* defense:** when we don't know the true distribution of data

The Worst-case Test Loss by Data Poisoning

- Upper-bound [refer to the paper for its derivation]

$$\max_{D_p} \mathrm{L}(\hat{\theta}) \;\leq\; \max_{D_p \subseteq F} \min_{\theta \in \Theta} \frac{1}{n} \mathrm{L}\left(\theta; D_c \cup D_p\right) \stackrel{\text{def}}{=} \mathbf{M}$$

  - **M:** the minimax loss
  - **It means:** the attack is bounded to a scenario where all poisons are alive under F!

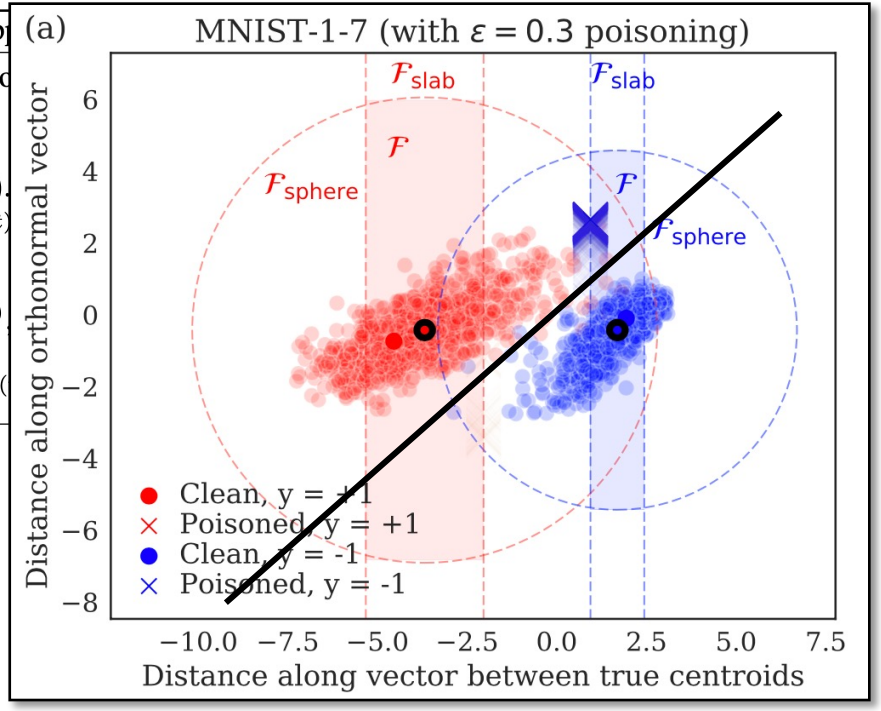
- Two defense scenarios
  - ***Fixed* defense:** we can fix F regardless of poisoning samples
  - ***Data-dependent* defense:** we cannot fix F (and hence can be influenced by the attacker)

Upper-bounds

- *Fixed* defense scenario
  - To simulate the worst-case, you craft poisons as follows and inject them

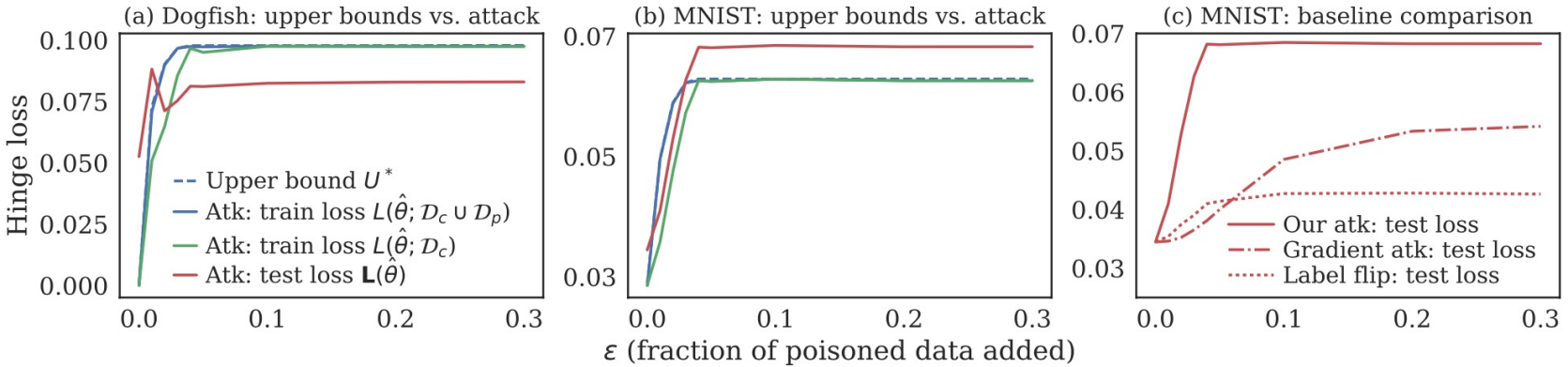
Algorithm 1 Online learning algorithm for generating an up

Input: clean data $\mathcal{D}_c$ of size n, feasible set $\mathcal{F}$, radius ρ, pc
Initialize $z^{(0)} \leftarrow 0$, $\lambda^{(0)} \leftarrow \frac{1}{\eta}$, $\theta^{(0)} \leftarrow 0$, $U^* \leftarrow \infty$.
for $t = 1, \ldots, \epsilon n$ **do**
  Compute $(x^{(t)}, y^{(t)}) = \text{argmax}_{(x,y)\in\mathcal{F}} \ell(\theta^{(t-1)}; x, y)$.
  $U^* \leftarrow \min\left(U^*, \frac{1}{n} L(\theta^{(t-1)}; \mathcal{D}_c) + \epsilon\ell(\theta^{(t-1)}; x^{(t)}, y^{(t)}\right.$
  $g^{(t)} \leftarrow \frac{1}{n}\nabla L(\theta^{(t-1)}; \mathcal{D}_c) + \epsilon\nabla\ell(\theta^{(t-1)}; x^{(t)}, y^{(t)})$.
  Update: $z^{(t)} \leftarrow z^{(t-1)} - g^{(t)}$, $\lambda^{(t)} \leftarrow \max(\lambda^{(t-1)}$
end for
Output: upper bound U^* and candidate attack $\mathcal{D}_p = \{(x^{($



Evaluations: Fixed Defense

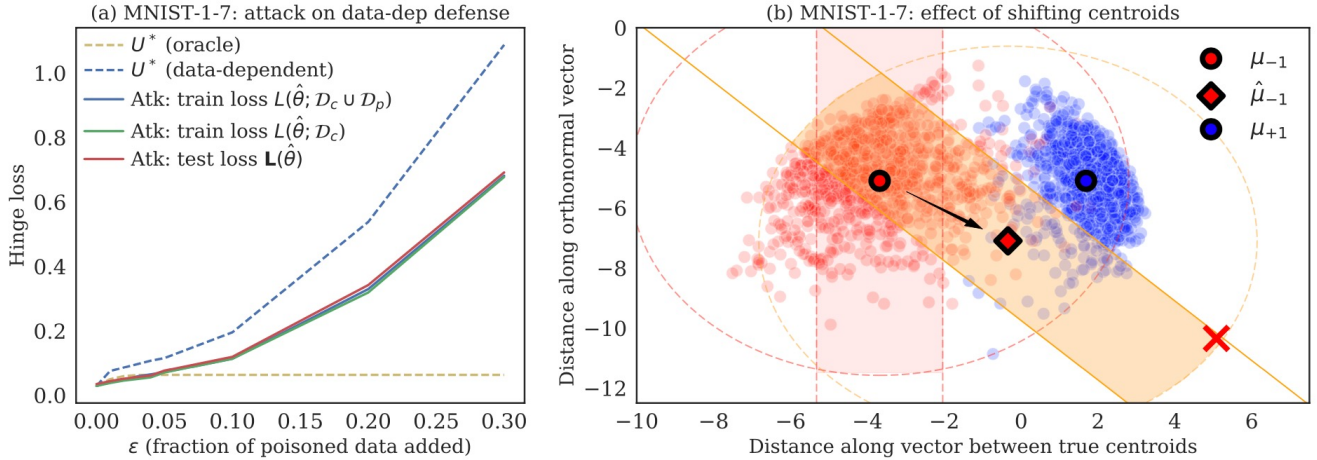
- On DogFish and MNIST-1/7



  - (a), (b), (c): oracle defenses are strong (the loss < 0.1…)
  - (a) and (b): the upper bound is *tight*
  - (c): the upper bound is tighter than what existing attacks can inflict

Evaluations: Data-Dependent Defense

- On MNIST-1/7 in 2-class SVMs



- (a): data-dependent defenses are much weaker (the bound increases exponentially…)
- (a): the upper-bound is still *tight*
- (b): in data-dependent defenses, the F is affected by the poisons

Topics for Today

- Motivation
  - Evade spam filter
  - DDoS detection
- Data Poisoning:
  - Attacks
    - Indiscriminate attacks on: SVMs and regression models
    - Targeted attacks on: DNNs (Poison Frogs and Meta-poison)
  - Defenses
    - Certified defenses
    - Differential privacy
  - Conclusion (and implications)

Motivation

- Steinhardt *et al.*
  - Fixed defenses are strong, but they are unrealistic
  - Data-depended defenses are largely affected by the poisons; thus, they are weak

How Can We Address Those Problems?

The Key Idea: Differential Privacy

- Differential Privacy
  - M $(D \times R^d \rightarrow \Theta)$ is (ϵ, δ)-differentially-private
    if $\forall D,\ \widetilde{D} \in D$ that **differ by one item**, and $\forall S \subset \Theta$,

$$\mathbf{P}\left(\mathcal{M}(D, b) \in \mathcal{S}\right) \leq e^{\epsilon}\mathbf{P}\left(\mathcal{M}(\tilde{D}, b) \in \mathcal{S}\right) + \delta$$

    where the probability is taken over $b \sim v$. When $\delta = 0$, M is ϵ-differentially-private.

The Key Idea: Differential Privacy

- Differential Privacy
  - M $(D \times R^d \rightarrow \Theta)$ is (ϵ, δ)-differentially-private
    if $\forall D,\ \widetilde{D} \in D$ that **differ by one item**, and $\forall S \subset \Theta$,

$$\mathbf{P}\left(\mathcal{M}(D, b) \in \mathcal{S}\right) \leq e^{\epsilon}\mathbf{P}\left(\mathcal{M}(\tilde{D}, b) \in \mathcal{S}\right) + \delta$$

    where the probability is taken over $b \sim v$. When $\delta = 0$, M is ϵ-differentially-private.

- Connection to Data Poisoning [**Conceptually!**]

Threat Model: Attacker

- Knowledge [white-box]
  - **Train-set:** D / **Poisoned train-set:** $\widetilde{D}$
  - **Differentially-private learner:** M
  - **Noise dist.:** v, but not the distribution b
- Capability
  - Modify k items in D
- Goals
  - Minimize the objective function $J(\widetilde{D})$ – attack cost!
  - **Objectives**
    - Parameter-targeting attack: make the model $\tilde{\theta}$ to *be close to* a target θ
    - Label-targeting attack: cause *small* prediction error on $\{z_i^*\}_{i\in[m]}$
    - Label-aversion attack: induce *large* prediction error on $\{z_i^*\}_{i\in[m]}$

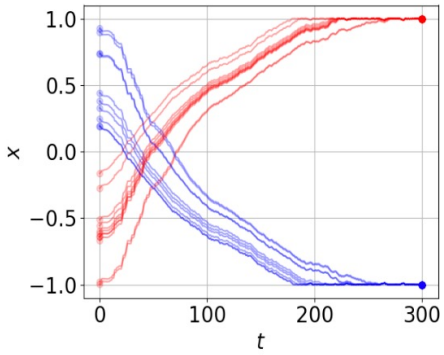
Impact of Differential Privacy

- Construct the lower-bound on $J(\widetilde{D})$
  - $J(\widetilde{D}) \geq e^{-k\epsilon} J(D)$
  - Data poisoning cannot make $J(\widetilde{D})$ infinitely small
- Lemma & Corollary
  - **Lemma 1:** If $k = 1$, it becomes $J(\widetilde{D}) \geq e^{-\epsilon} J(D)$
  - **Corollary 1:** To achieve $J(\widetilde{D}) \geq {}^{1}/_{\tau} J(D)$, $k \geq \lceil {}^{1}/_{\epsilon} \log \tau \rceil$ [Fun facts!]
  - …

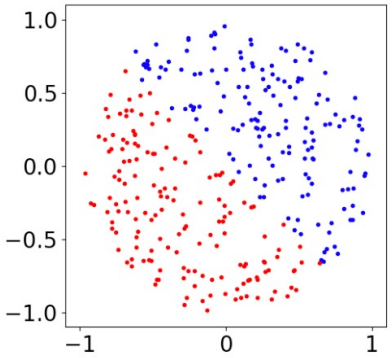
Evaluations

- Setup [binary classification tasks]
  - Dataset: Synthetic data | Real data (UCI ML Repo.)
  - Models: Logistic regression | Ridge-regression

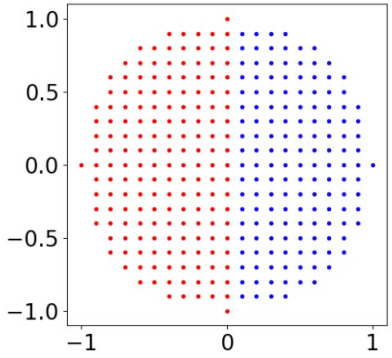
- Crafting poisons
  - Demonstrate on 2-D synthetic data



(a) poisoning trajectory



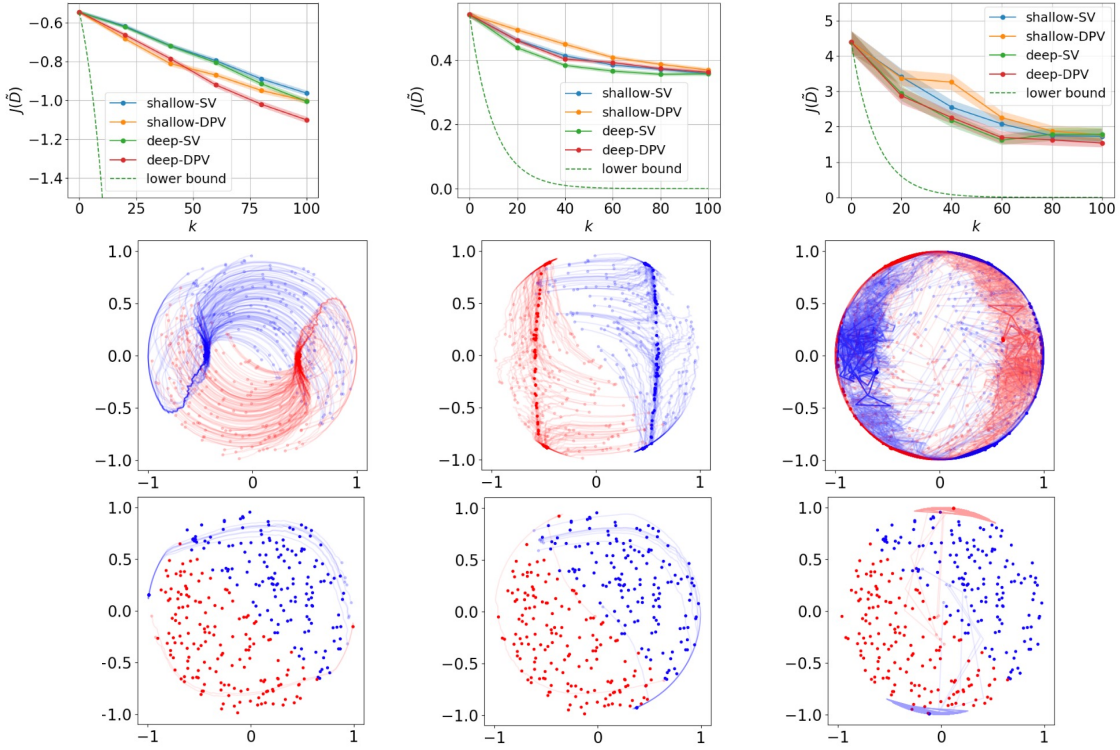
(b) training set



(c) evaluation set

Evaluations

- Results of the three attacks on 2-D artificial data

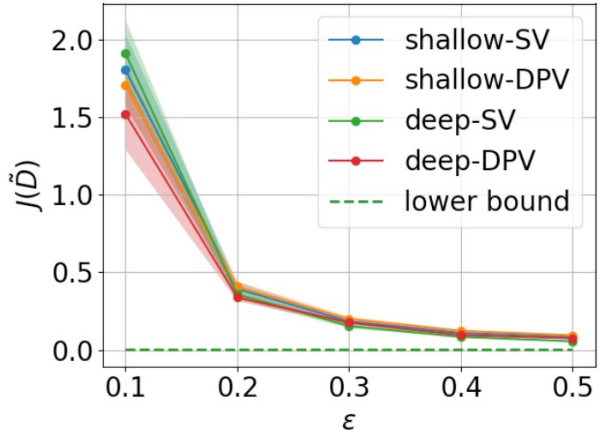


(d) label-aversion (e) label-targeting (f) parameter-targeting

Evaluations

- Results of the *label-targeting* attacks on real-world datasets
  - In DP, the attack costs significantly higher than the case w/o DP
  - ex. with 20 poisons, the cost w/o DP is almost zero whereas with DP, it's 0.4

- Interesting Observation!
  - Attacks are much easier with weak (small epsilon) privacy



Recap: Data Poisoning

- Motivation
  - Evade spam filter
  - DDoS detection
- Data Poisoning:
  - Attacks
    - Indiscriminate attacks on: SVMs and regression models
    - Targeted attacks on: DNNs (Poison Frogs and Meta-poison)
  - Defenses
    - Certified defenses
    - Differential privacy
  - Conclusion (and implications)

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