Robust Learning for Data Poisoning Attacks

Yunjuan Wang, Poorya Mianjy, Raman Arora
Department of Computer Science, Johns Hopkins University

Background

- Machine Learning systems are fragile, susceptible to attacks.
- Types of attacks: inference-time attacks, data poisoning attacks.
- Data poisoning attacks: the adversary manipulates the training data.
- $\mathcal{S} = \{(\mathrm{x}_i, y_i)\}_{i=1}^n$ → $\tilde{\mathcal{S}} = \{(\tilde{\mathrm{x}}_i, \tilde{y}_i)\}_{i=1}^n$
  - Backdoor attack,
  - Clean label attack: $\tilde{\mathrm{x}}_i = \mathrm{x}_i + \delta_i, \tilde{y}_i = y_i$,
  - Label flip attack: $\tilde{\mathrm{x}}_i = \mathrm{x}_i, \tilde{y}_i = -y_i$ with probability β.

Notation: $B = \max \|\delta_i\|_2$ is per-sample perturbation; $S = \sum_{i=1}^n \|\delta_i\|_2$ is overall perturbation; β is probability of label flips; $\tilde{\mathcal{O}}$ hides poly-logarithmic dependence on n.

Convex Learning Problem (Warm-up)

- Input/Label spaces $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$; distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.
- *Goal:* solve the stochastic optimization problem

$$\min_{\mathrm{w} \in \mathsf{W}} F(\mathrm{w}) := \mathbb{E}_{(\mathrm{x},y)\sim\mathcal{D}}[\ell(yf(\mathrm{x};\mathrm{w}))],$$

where W is a convex set, ℓ is convex in w.

- Standard approach is to use SGD, where the learner takes w and gets access to a first order stochastic oracle for $\hat{\mathrm{g}}(\mathrm{w}) \in \partial F(\mathrm{w})$.

Observation: data poisoning attacks (δ_i) can be viewed as oracle poisoning attacks (ζ_i).

- $\delta_i = \tilde{\mathrm{x}}_i - \mathrm{x}_i$.
- $\zeta_i = \tilde{\mathrm{g}}(\mathrm{w}_i) - \hat{\mathrm{g}}(\mathrm{w}_i)$.

Theorem *(Robustness of SGD).* Excess risk bound for clean label attacks:

$$\mathbb{E}[F(\bar{\mathrm{w}})] - F(\mathrm{w}_*) \le O\left(\frac{1}{\sqrt{n}} + \frac{\sum_{i<n}\|\zeta_i\|}{n}\right)$$

where the expectation is w.r.t. the initialization and the training samples.

Remark: 1. $\sum_{i<n}\|\zeta_i\| = \mathcal{O}(\sqrt{n})$ gives no significant statistical overhead.

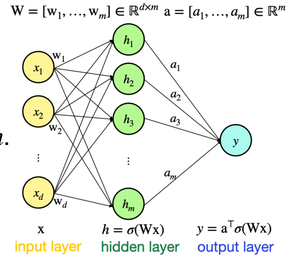
2. The above upper bound is tight in an information-theoretic sense (see paper for a lower bound).

Two-layer neural networks

- A two-layer ReLU net parameterized by (a, W),

$$f(\mathrm{x};\mathrm{a},\mathrm{W}) := \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \sigma(\mathrm{w}_s^\top \mathrm{x}),$$ ReLU: $\sigma(z)$, network width: m.

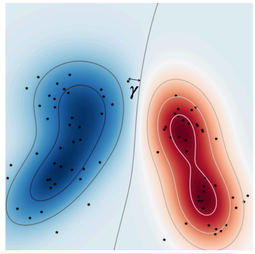
- Trained by online SGD using logistic loss.
- Goal: minimize $L(\mathrm{W}) := \mathbb{P}_{(\mathrm{x},y)\sim\mathcal{D}}(yf(\mathrm{x};\mathrm{a},\mathrm{W}) < 0)$.



Key Assumptions

There exists a *margin parameter* $\gamma > 0$, and a linear separator $\bar{v} : \mathbb{R}^d \to \mathbb{R}^d$, s.t.

- $\mathbb{E}_{\mathrm{z}}[\|\bar{v}(\mathrm{z})\|^2] < \infty$,
- $\|\bar{v}(\mathrm{z})\|_2 \le 1$ for all $\mathrm{z} \in \mathbb{R}^d$,
- $\mathbb{E}_{\mathrm{z}\sim\mathcal{N}(0,\mathrm{I}_d)}[y\langle\bar{v}(\mathrm{z}), \mathrm{x}\mathbb{I}[\mathrm{z}^\top\mathrm{x} \ge 0]\rangle] \ge \gamma$ for almost all $(\mathrm{x}, y) \sim \mathcal{D}$.



Main Results

Theorem: With probability at least $1-\delta$, we show the following for the iterates of SGD,

Regime A (clean label attack, *large* per-sample perturbation, *small* overall perturbation):

$$\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) \lesssim \frac{\ln^2(\sqrt{n}/4) + \ln(24n/\delta)}{\sqrt{n}\gamma^2}$$

provided that $B \le \tilde{\mathcal{O}}(\gamma/\sqrt{d})$, $\tilde{\mathcal{O}}(1/\gamma^8) \le m \le \tilde{\mathcal{O}}(n/\gamma^4 S^2)$.

Remark: Regime A requires $S \lesssim \gamma^2\sqrt{n}$ to allow a *non-empty width range*.

Regime B (clean label attack, *small* per-sample perturbation, *large* overall perturbation):

$$\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) \lesssim \frac{\ln^2(\sqrt{n}/4) + \ln(24n/\delta)}{n\gamma^2}$$

for $m \ge \tilde{\mathcal{O}}(1/\gamma^8)$, provided that $B \lesssim \min\{\frac{1}{\sqrt{md} + \sqrt{m\ln(m/\delta)}}, \frac{\gamma}{\gamma + \sqrt{d} + \sqrt{\ln(mn/\delta)}}\}$.

Remark: Regime B allows a larger overall perturbation budget of order $\mathcal{O}(n)$.

Regime C (label flip attack):

$$\frac{1}{n}\sum_{i<n} L(\mathrm{W}_i) \lesssim \frac{\ln^2(\sqrt{n}/4) + \ln(16n/\delta)}{\sqrt{n}\gamma^2}$$

for $m \ge \tilde{\mathcal{O}}(1/\gamma^8)$, provided that $\beta \lesssim \frac{\ln(n/\delta) + \ln^2(n)}{(\sqrt{\ln(n/\delta)} + \gamma^3\sqrt{m}(1+\gamma))\gamma^2\sqrt{n}}$.

Remark: 1. For regime A and C, the generalization bounds are of the same rate of $\mathcal{O}(1/\sqrt{n})$, the effective overall perturbation budget are almost of the same order $\tilde{\mathcal{O}}(\sqrt{n})$.

2. All three regimes require an upperbound and lowerbound on the network width.

Experiments

- Propose negated loss by flipping the sign on both the model prediction and the cross-entropy loss, i.e $\ell\_\_(z) := -\ell(-z)$, where z is the model prediction.
- Generate the poisoned data:

1) Use mini-batch SGD to learn the best model parameters w^* on the clean data.
2) Take a stochastic gradient ascent step on the negated loss to maximize the negated loss function $\ell\_\_(\mathrm{x};\mathrm{w}^*)$ with respect to x.
3) Project onto the constraints -- the $\ell_{2,1}$ (regime A) or $\ell_{2,\infty}$-norm ball (regime B).

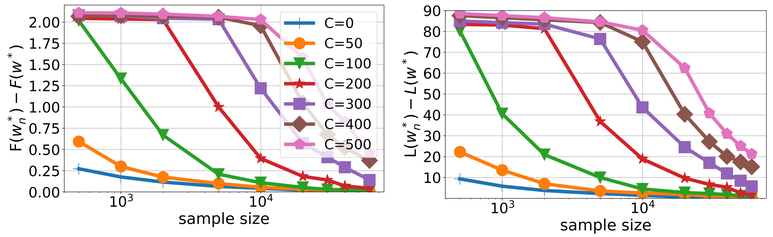


Figure: The excess loss $F(\mathrm{w}_n^*) - F(\mathrm{w}^*)$ (**left**); and the excess error $L(\mathrm{w}_n^*) - L(\mathrm{w}^*)$ (**right**), as a function of sample size n with different corruption parameter C under regime A. Specify overall budget $S = C\sqrt{n}$. Here, w_n^* denotes the optimal parameters on the given sample of size n.

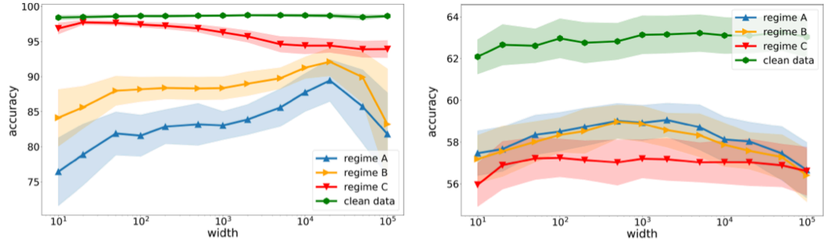


Figure: Clean test accuracy as a function of network width under clean data setting and poisoned data setting on MNIST (**left**) and CIFAR10 (**right**).

- The generalization accuracy decreases if the models are not wide enough or if they're too wide.
- The inverted U curve challenges the nascent view in the deep learning literature that larger models generalize better, at least under adversarial perturbation.

Main Takeaway: networks that are extremely over-parameterized are more susceptible to attacks.