Toward Adversarial Learning as Control

Jerry Zhu

University of Wisconsin-Madison

The 2nd ARO/IARPA Workshop on Adversarial Machine Learning
May 9, 2018

Test time attacks

- Given classifier $f : \mathcal{X} \mapsto \mathcal{Y}$, $x \in \mathcal{X}$
- Attacker finds $x' \in \mathcal{X}$:

$$\begin{aligned} \min_{x'} \quad & \|x' - x\| \\ \text{s.t.} \quad & f(x') \neq f(x). \end{aligned}$$

"Large margin" defense against test time attacks

- Defender finds $f' \in \mathcal{F}$:

$$\begin{aligned} \min_{f'} \quad & \|f' - f\| \\ \text{s.t.} \quad & f'(x') = f(x), \forall \text{ training } x, \forall x' \in Ball(x, \epsilon). \end{aligned}$$

Heuristic implementation of large margin defense

Repeat:

- $(x, x') \leftarrow$ OracleAttacker(f)
- Add $(x', f(x))$ to (X, Y)
- $f \leftarrow A(X, Y)$

Training set poisoning attacks

- Given learner $A : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{F}$, data (X, Y), goal $\Phi : \mathcal{F} \mapsto bool$

Training set poisoning attacks

- Given learner $A : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{F}$, data (X, Y), goal $\Phi : \mathcal{F} \mapsto \mathit{bool}$
- Attacker finds poisoned data (X', Y')

$$\begin{aligned} \min_{(X',Y'),f} \quad & \|(X', Y') - (X, Y)\| \\ \text{s.t.} \quad & f = A(X', Y') \\ & \Phi(f) = \text{true}. \end{aligned}$$

defense = poisoning = machine teaching

[An Overview of Machine Teaching. ArXiv 1801.05927, 2018] defense = poisoning = machine teaching = control

[An Overview of Machine Teaching. ArXiv 1801.05927, 2018]

Attacking a sequential learner A =SGD

Learner A (plant):

- starts at $w_0 \in \mathbb{R}^d$
- $w_t \leftarrow w_{t-1} - \eta \nabla \ell(w_{t-1}, x_t, y_t)$

Attacking a sequential learner A =SGD

Learner A (plant):

- starts at $w_0 \in \mathbb{R}^d$
- $w_t \leftarrow w_{t-1} - \eta \nabla \ell(w_{t-1}, x_t, y_t)$

Attacker:

- designs $(x_1, y_1) \ldots (x_T, y_T)$ (control signal)
- wants to drive w_T to some w^*
- optionally minimizes T

Nonlinear discrete-time optimal control

...even for simple linear regression:

$$\ell(w, x, y) = \frac{1}{2}(x^\top w - y)^2$$

$$w_t \leftarrow w_{t-1} - \eta(x_t^\top w_{t-1} - y_t)x_t$$

Nonlinear discrete-time optimal control

...even for simple linear regression:

$$\ell(w, x, y) = \frac{1}{2}(x^\top w - y)^2$$

$$w_t \leftarrow w_{t-1} - \eta(x_t^\top w_{t-1} - y_t)x_t$$

Continuous version:

$$\dot{w}(t) = (y(t) - w(t)^\top x(t))x(t)$$

$$\|x(t)\| \leq 1, |y(t)| \leq 1, \forall t$$

Attack goal is to drive $w(t)$ from w_0 to w^* in minimum time.

Greedy heuristic

$$\begin{aligned} \min_{x_t, y_t, w_t} \quad & \|w_t - w^*\| \\ \text{s.t.} \quad & \|x_t\| \leq 1, |y_t| \leq 1 \\ & w_t = w_{t-1} - \eta(x_t^\top w_{t-1} - y_t)x_t \end{aligned}$$

... or further constrain x_t in the direction $w^* - w_{t-1}$

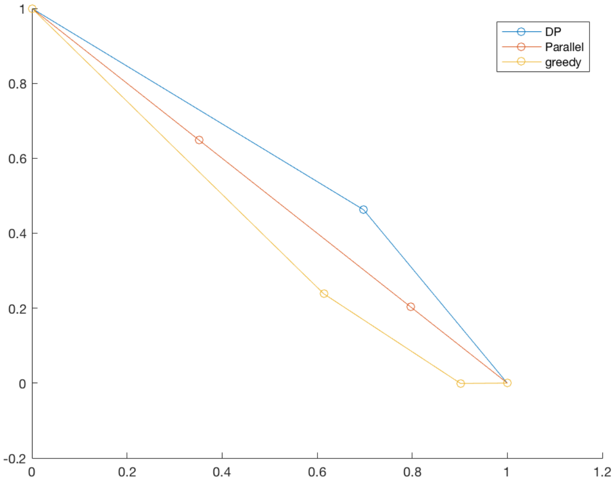
[Liu, Dai, Humayun, Tay, Yu, Smith, Rehg, Song. ICML'17]

Discrete-time optimal control

$$
\begin{aligned}
\min_{x_{1:T}, y_{1:T}, w_{1:T}} \quad & T \\
\text{s.t.} \quad & \|x_t\| \leq 1, |y_t| \leq 1, \quad t = 1 \ldots T \\
& w_t = w_{t-1} - \eta(x_t^\top w_{t-1} - y_t)x_t, \quad t = 1 \ldots T \\
& w_T = w^*.
\end{aligned}
$$

Controlling SGD squared loss

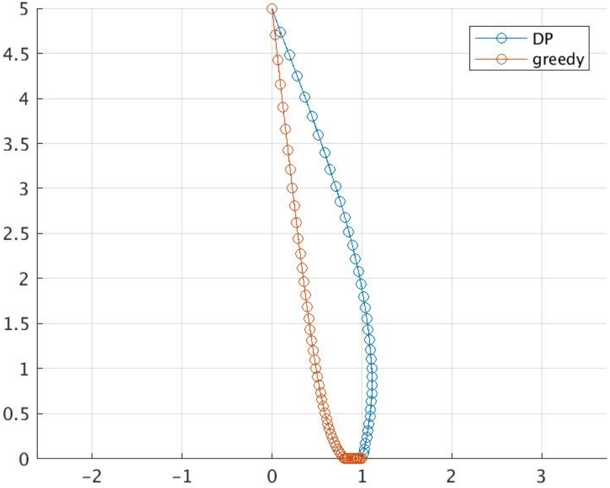
$T = 2$ (DTOC) vs. $T = 3$ (greedy)



$$w_0 = (0, 1, 0), w^* = (1, 0, 0), \|x\| \leq 1, |y| \leq 1, \eta = 0.55$$

Controlling SGD squared loss (2)

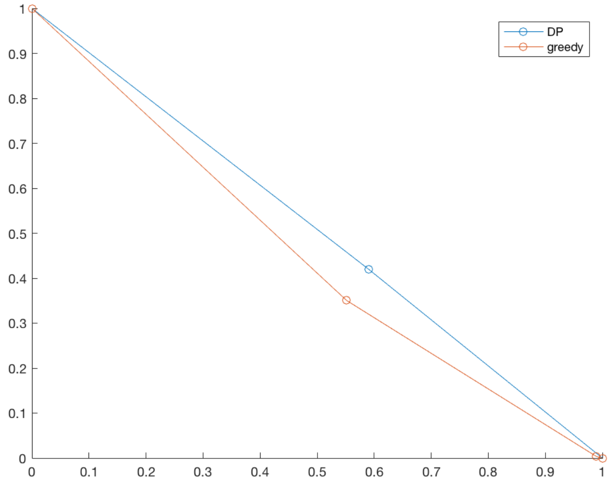
$T = 37$ (DTOC) vs. $T = 55$ (greedy)



$w_0 = (0, 5), w^* = (1, 0), \|x\| \leq 1, |y| \leq 1, \eta = 0.05$

Controlling SGD logistic loss

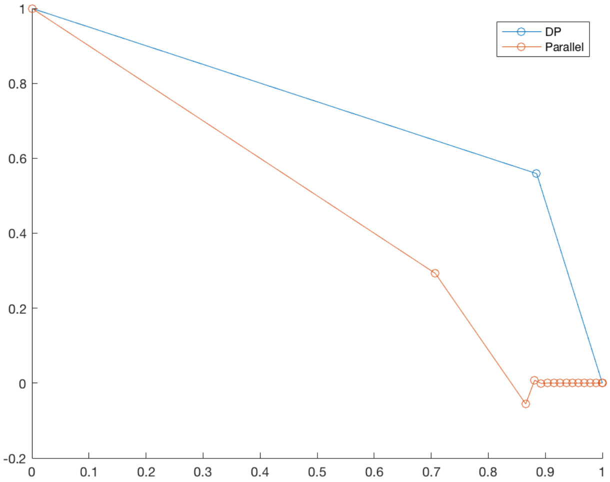
$T = 2$ (DTOC) vs. $T = 3$ (greedy)



$w_0 = (0, 1), w^* = (1, 0), \|x\| \le 1, |y| \le 1, \eta = 1.25$

Controlling SGD hinge loss

$T = 2$ (DTOC) vs. $T = 16$ (greedy)



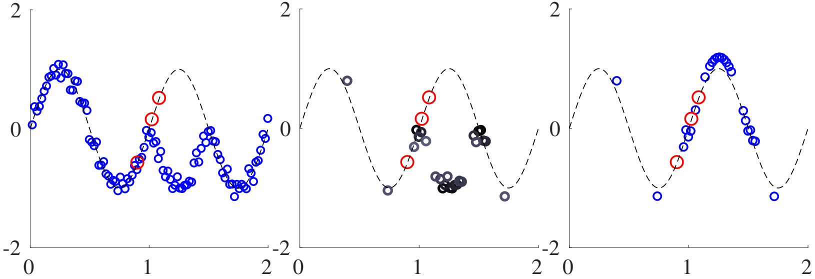
$$w_0 = (0, 1), w^* = (1, 0), \|x\| \leq 100, |y| \leq 1, \eta = 0.01$$

Detoxifying a poisoned training set

- Given poisoned (X', Y'), a small trusted $(\tilde{X}, \tilde{Y})$
- Estimate detox (X, Y):

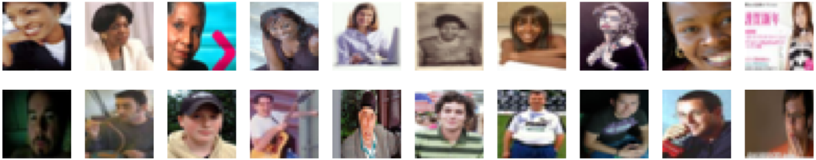
$$
\begin{aligned}
\min_{(X,Y),f} \quad & \|(X,Y) - (X',Y')\| \\
\text{s.t.} \quad & f = A(X,Y) \\
& f(\tilde{X}) = \tilde{Y} \\
& f(X) = Y.
\end{aligned}
$$

Detoxifying a poisoned training set



[Zhang, Zhu, Wright. AAAI 2018]

Training set camouflage: Attack on perceived intention



Alice $\longrightarrow$ Eve $\rightarrow$ Bob

Too obvious.

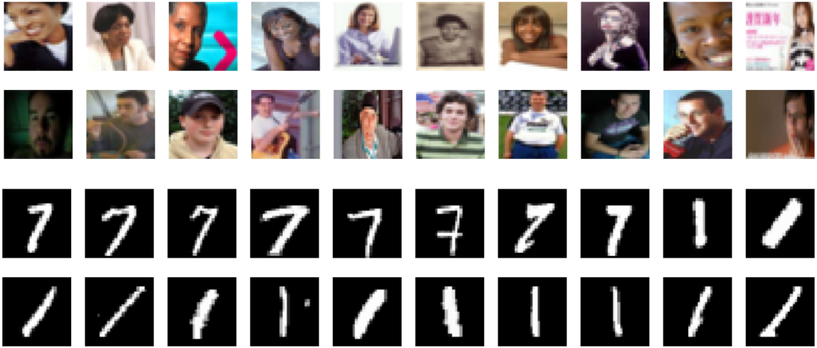
Training set camouflage: Attack on perceived intention

$f = A\Big($  $\Big)$

Alice $f \to$ Eve $\to$ Bob

Too suspicious.

Training set camouflage: Attack on perceived intention



Alice  $\longrightarrow$ Eve $\rightarrow$ Bob

- Less suspicious to Eve
- Bob learns $f' = A($$)$
- f' good at man vs. woman! $f' \approx f$.

Alice's camouflage problem

Given:

- sensitive data S (e.g. man vs. woman)
- public data P (e.g. the whole MNIST 1's and 7's)
- Eve's detection function Φ (e.g. two-sample test)
- Bob's learning algorithm A and loss ℓ

Alice's camouflage problem

Given:

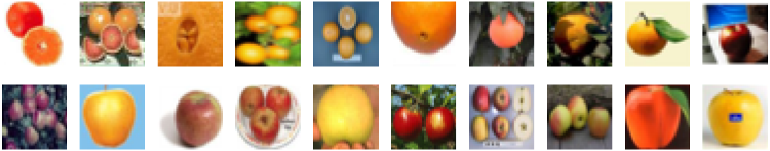
- sensitive data S (e.g. man vs. woman)
- public data P (e.g. the whole MNIST 1's and 7's)
- Eve's detection function Φ (e.g. two-sample test)
- Bob's learning algorithm A and loss ℓ

Find D:

$$\min_{D \subseteq P} \quad \sum_{(x,y) \in S} \ell(A(D), x, y)$$

$$\text{s.t.} \quad \Phi \text{ thinks } D, P \text{ from the same distribution.}$$

Camouflage examples



Camouflage examples

Sample of Sensitive Set		Sample of Camouflaged Training Set	
Class	**Article**	**Class**	**Article**
Christianity	. . .Christ that often causes critical of themselves . . .	Baseball	. . .The Angels won their Brewers today before 33,000+ . . .
	. . .I've heard it said of Christs life and ministry . . .		. . . interested in finding out to get two tickets . . .
Atheism	. . .This article attempts to introduction to atheism. . .	Hockey	. . . user and not necessarily the game summary for. . .
	. . .Science is wonderful to question scientific. . .		. . .Tuesday, and the isles/caps what does ESPN do. . .

Attack on stochastic multi-armed bandit

K-armed bandit

- ad placement, news recommendation, medical treatment ...
- suboptimal arm pulled $o(T)$ times

Attack goal:

- make the bandit algorithm almost always pull suboptimal arm (say arm K)

Shaping attack

```
Input: bandit algorithm A, target arm K
for t = 1, 2, ... do
    Bandit algorithm A chooses arm I_t to pull.
    World produces pre-attack reward r_t^0.
    Attacker decides the attacking cost α_t.
    Attacker gives r_t = r_t^0 − α_t to the bandit algorithm A.
end for
```

α_t chosen to make $\hat{\mu}_{I_t}$ look sufficiently small compared to $\hat{\mu}_K$.

Shaping attack

For ϵ-greedy algorithm:

- Target arm K is pulled at least

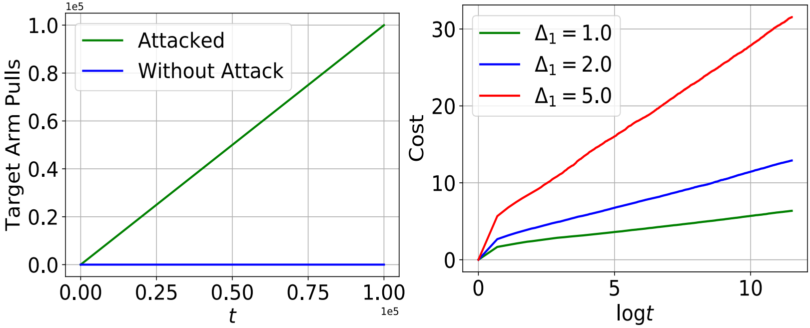
$$T - \left(\sum_{t=1}^{T} \epsilon_t\right) - \sqrt{3 \log\left(\frac{K}{\delta}\right)\left(\sum_{t=1}^{T} \epsilon_t\right)}$$

  times;
- Cumulative attack cost is

$$\sum_{t=1}^{T} \alpha_t = \widehat{O}\left(\left(\sum_{i=1}^{K} \Delta_i\right) \log T + \sigma \sqrt{\log T}\right).$$

Similar theorem for UCB1.

Shaping attack



Acknowledgments

Collaborators: Scott Alfeld, Ross Boczar, Yuxin Chen, Kwang-Sung Jun, Laurent Lessard, Lihong Li, Po-Ling Loh, Yuzhe Ma, Rob Nowak, Ayon Sen, Adish Singla, Ara Vartanian, Stephen Wright, Xiaomin Zhang, Xuezhou Zhang

Funding: NSF, AFOSR, University of Wisconsin