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 $x' \in \mathcal{X}$:

\[
\min_{x'} \|x' - x\| \\
\text{s.t.} \quad f(x') \neq f(x).
\]
“Large margin” defense against test time attacks

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

$$\min_{f'} \|f' - f\|$$

s.t. $f'(x') = f(x), \forall \text{ training } x, \forall x' \in Ball(x, \epsilon)$. 
Heuristic implementation of large margin defense

Repeat:
- \((x, x') \leftarrow \text{OracleAttacker}(f)\)
- 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
  $\Phi : \mathcal{F} \mapsto \text{bool}$
Training set poisoning attacks

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

\[
\min_{(X', Y') , f} \| (X', Y') - (X, Y) \| \\
\text{s.t.} \quad f = A(X', Y') \\
\Phi(f) = true.
\]
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 $A = \text{SGD}$

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 $A = \text{SGD}$

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

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

... 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

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

\[ T = 37 \text{ (DTOC) vs. } T = 55 \text{ (greedy)} \]

\[ w_0 = (0, 5), \ w^* = (1, 0), \ ||x|| \leq 1, \ |y| \leq 1, \ \eta = 0.05 \]
Controlling SGD logistic loss

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

$w_0 = (0, 1), w^* = (1, 0), \|x\| \leq 1, |y| \leq 1, \eta = 1.25$
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 \((X, Y)\):

\[
\min_{(X, Y), f} \| (X, Y) - (X', Y') \|
\text{s.t.} \quad f = A(X, Y) \\
\quad f(\tilde{X}) = \tilde{Y} \\
\quad f(X) = Y.
\]
Detoxifying a poisoned training set

[Zhang, Zhu, Wright. AAAI 2018]
Training set camouflage: Attack on perceived intention

Alice → Eve → Bob

Too obvious.
Training set camouflage: Attack on perceived intention

\[ f = A (\text{Alice} \rightarrow \text{Eve} \rightarrow \text{Bob}) \]

Alice \( f \rightarrow \) Eve \( \rightarrow \) Bob

Too suspicious.
Training set camouflage: Attack on perceived intention

Alice → Eve → Bob

- Less suspicious to Eve
- Bob learns $f' = A(\text{___________})$
- $f'$ 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 $\Phi$ (e.g. two-sample test)
- Bob’s learning algorithm $A$ and loss $\ell$
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 $\Phi$ (e.g. two-sample test)
- Bob’s learning algorithm $A$ and loss $\ell$

Find $D$:

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

s.t. $\Phi$ thinks $D, P$ from the same distribution.
Camouflage examples
### Sample of Sensitive Set

<table>
<thead>
<tr>
<th>Class</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christianity</td>
<td>...Christ that often causes critical of themselves ...</td>
</tr>
<tr>
<td></td>
<td>...I've heard it said of Christ's life and ministry ...</td>
</tr>
<tr>
<td>Atheism</td>
<td>...This article attempts to introduction to atheism ...</td>
</tr>
<tr>
<td></td>
<td>...Science is wonderful to question scientific ...</td>
</tr>
</tbody>
</table>

### Sample of Camouflaged Training Set

<table>
<thead>
<tr>
<th>Class</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseball</td>
<td>...The Angels won their Brewers today before 33,000+ ...</td>
</tr>
<tr>
<td></td>
<td>...interested in finding out to get two tickets ...</td>
</tr>
<tr>
<td>Hockey</td>
<td>...user and not necessarily the game summary for ...</td>
</tr>
<tr>
<td></td>
<td>...Tuesday, and the isles/caps what does ESPN do ...</td>
</tr>
</tbody>
</table>
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

1: **Input**: bandit algorithm $A$, target arm $K$
2: **for** $t = 1, 2, \ldots$ **do**
3: Bandit algorithm $A$ chooses arm $I_t$ to pull.
4: World produces pre-attack reward $r_t^0$.
5: Attacker decides the attacking cost $\alpha_t$.
6: Attacker gives $r_t = r_t^0 - \alpha_t$ to the bandit algorithm $A$.
7: **end for**

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

For $\epsilon$-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

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

Similar theorem for UCB1.
Shaping attack

- **Target Arm Pulls vs. Time (t)**
  - Green line: Attacked
  - Blue line: Without Attack

- **Cost vs. Logarithm of Time (logt)**
  - Green line: \( \Delta_1 = 1.0 \)
  - Blue line: \( \Delta_1 = 2.0 \)
  - Red line: \( \Delta_1 = 5.0 \)
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