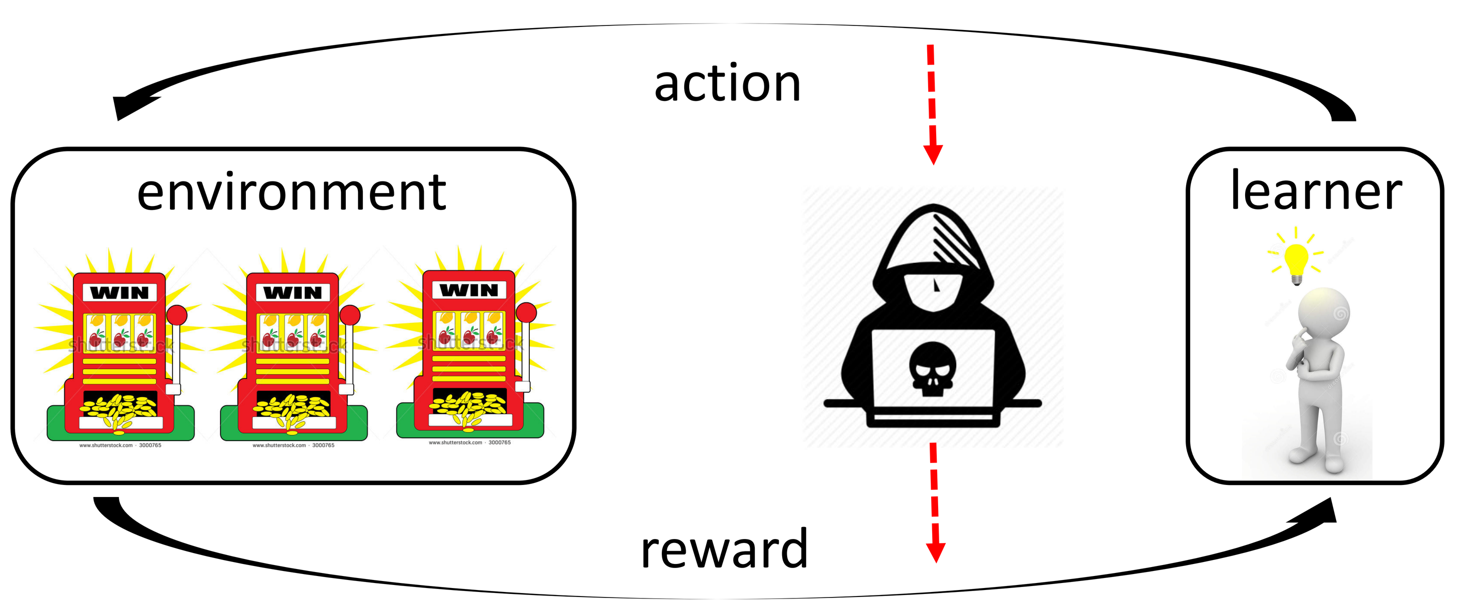
# Data Poisoning Attacks on Stochastic Bandits


Fang Liu and Ness Shroff

The Ohio State University


## Background




Adversarial learning has been well-studied in deep learning

How robust are bandit learning?

- They are vulnerable in some cases [1, 2]
- Behavior be hijacked by the attacker
- If under attack, hard to recognize

There is an urge to understand

- How does attack work?
- Is there any robust bandit algorithm?

## Data Poisoning Attacks

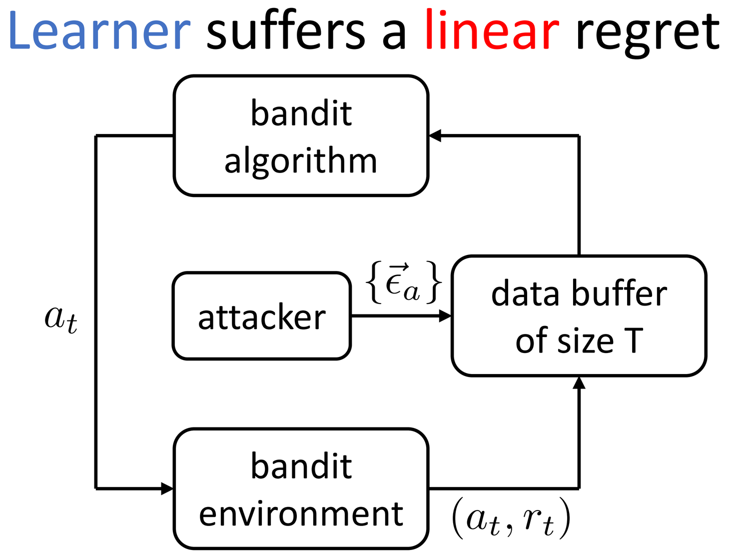
Data poisoning attacks on stochastic bandits:

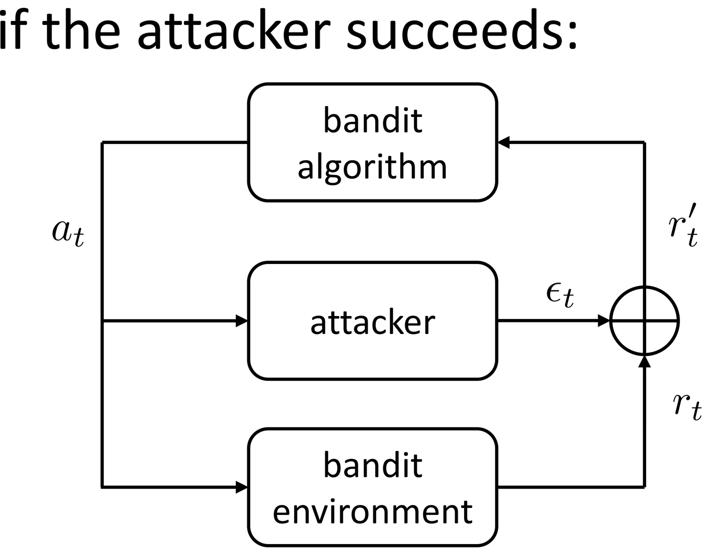
- At time *t*, the learner chooses an action $a_t$ from *K* actions
- The environment outputs an *i.i.d.* reward $r_t$ *drawn from* $a_t$
- The attacker observes $(a_t, r_t)$ and decides $\epsilon_t$
- The learner observes the poisoned feedback $r_t + \epsilon_t$

Attacker Performance measure:

- Target arm $a^*$, suboptimal (WLOG)
- Number of playing $a^*$, $N_{a^*}(T) = T - o(T)$ in expectation or with high probability $(1-\delta)$
- Attack cost is sublinear in *T*: $C(T) = \left(\sum_{t=1}^{T} |\epsilon_t|^p\right)^{1/p}$

Learner suffers a linear regret if the attacker succeeds:




## Offline Attacks

$\epsilon$-greedy algorithm:

$$\text{play action } a_t = \begin{cases} \text{draw uniformly over } \mathcal{A}, & \text{w.p. } \alpha_t \\ \arg\max_{a\in\mathcal{A}} \tilde{\mu}_a(t-1), & \text{otherwise} \end{cases}.$$

Post-attack empirical mean: $\tilde{\mu}_a(t)$

Quadratic program with linear constraints

$$P_1: \min_{\vec{\epsilon}_a : a\in\mathcal{A}} \sum_{a\in\mathcal{A}} ||\vec{\epsilon}_a||_2^2 \quad s.t. \quad \tilde{\mu}_{a^*}(T) \ge \tilde{\mu}_a(T) + \xi, \quad \forall a \ne a^*$$

UCB algorithm:

$$a_t = \arg\max_{a\in\mathcal{A}} u_a(t) := \tilde{\mu}_a(t-1) + 3\sigma\sqrt{\frac{\log t}{N_a(t-1)}}.$$

Quadratic program with linear constraints

$$P_2: \min_{\vec{\epsilon}_a : a\in\mathcal{A}} \sum_{a\in\mathcal{A}} ||\vec{\epsilon}_a||_2^2 \quad s.t. \quad u_{a^*}(T+1) \ge u_a(T+1) + \xi, \quad \forall a \ne a^*$$

Thomson Sampling:

$$a_t = \arg\max_{a\in\mathcal{A}} \theta_a(t) \sim \mathcal{N}(\tilde{\mu}_a(t-1)/\sigma^2, \sigma^2/N_a(t-1))$$

Quadratic program with convex constraints

$$P_3: \min_{\vec{\epsilon}_a : a\in\mathcal{A}} \sum_{a\in\mathcal{A}} ||\vec{\epsilon}_a||_2^2 \quad s.t. \quad \sum_{a\ne a^*} \Phi\left(\frac{\tilde{\mu}_a(T) - \tilde{\mu}_{a^*}(T)}{\sigma^3\sqrt{1/m_a + 1/m_{a^*}}}\right) \le \delta$$

$$\tilde{\mu}_a(T) - \tilde{\mu}_{a^*}(T) \le 0, \quad \forall a \ne a^*$$

## Online Attacks

Oracle attacks:

$$\epsilon_t = -\mathrm{I}\{a_t \ne a^*\}[\mu_{a_t} - \mu_{a^*} + \xi]^+$$

Not practical due to unknown expectations

(Prop. 1 in [1]) Assume that the bandit algorithm achieves an O(log *T*) regret. Then the oracle attack succeeds and the expected attack cost is $O(\sum_{i\ne a^*}[\mu_i - \mu_{a^*} + \xi]^+ \log T)$

Adaptive attack by constant estimation:

$$\epsilon_t = -\mathrm{I}\{a_t \ne a^*\}[\hat{\mu}_{a_t}(t) - \hat{\mu}_{a^*}(t) + \beta(N_{a_t}(t)) + \beta(N_{a^*}(t))]^+$$

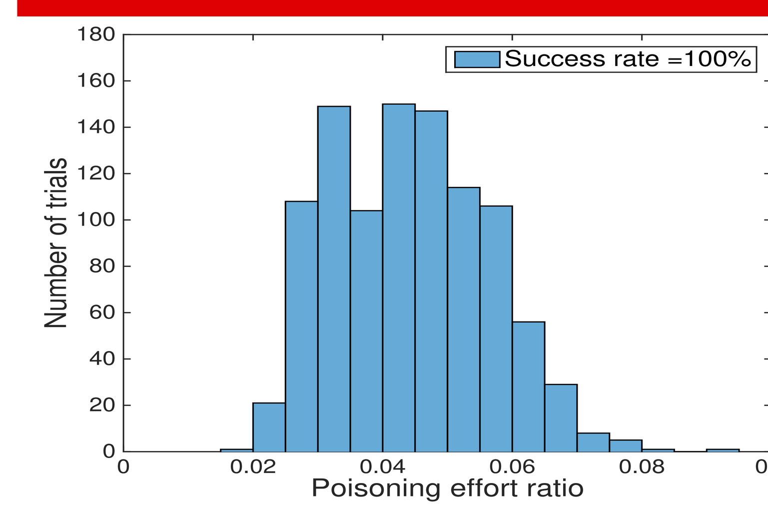
Where $\beta(n) = \sqrt{\frac{2\sigma^2}{n}\log\frac{\pi^2 K n^2}{3\delta}}$ is decreasing in n

Pre-attack empirical mean: $\hat{\mu}_a(t)$

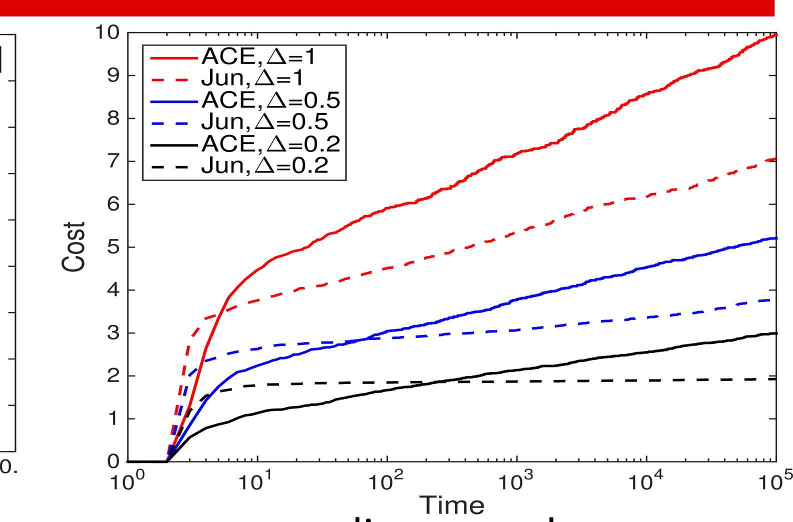
Theorem. Assume that the bandit algorithm achieves an O(log *T*) regret. Then the oracle attack succeeds and the expected attack cost is

$$\sum_{t=1}^{T} |\epsilon_t| \le O\left(\sum_{a\ne a^*}([\mu_a - \mu_{a^*}]^+ + 4\beta(1))\log T\right).$$

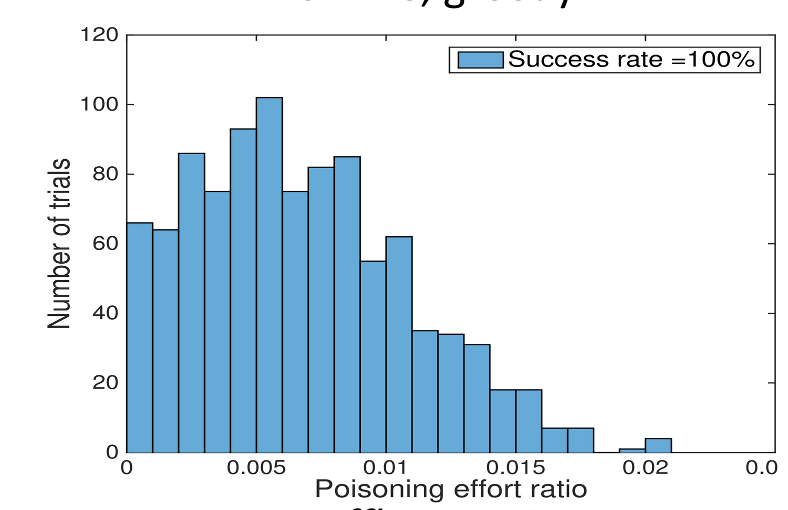
## Experiments



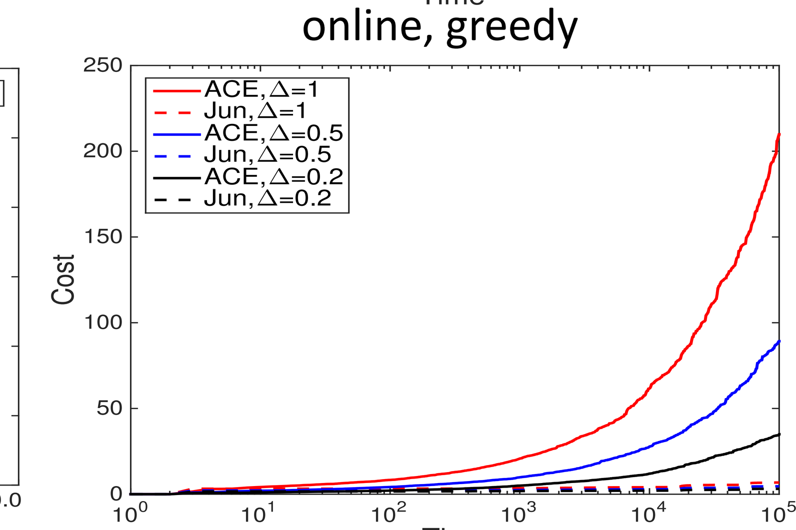

offline, greedy



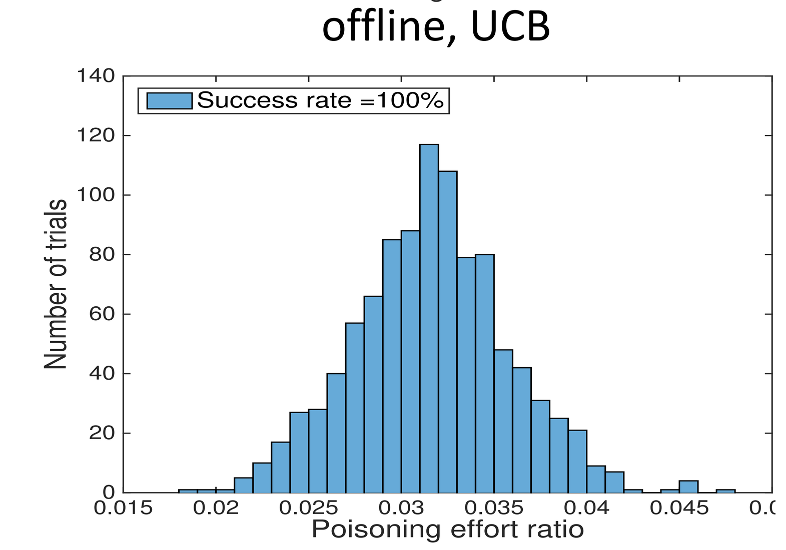

online, greedy



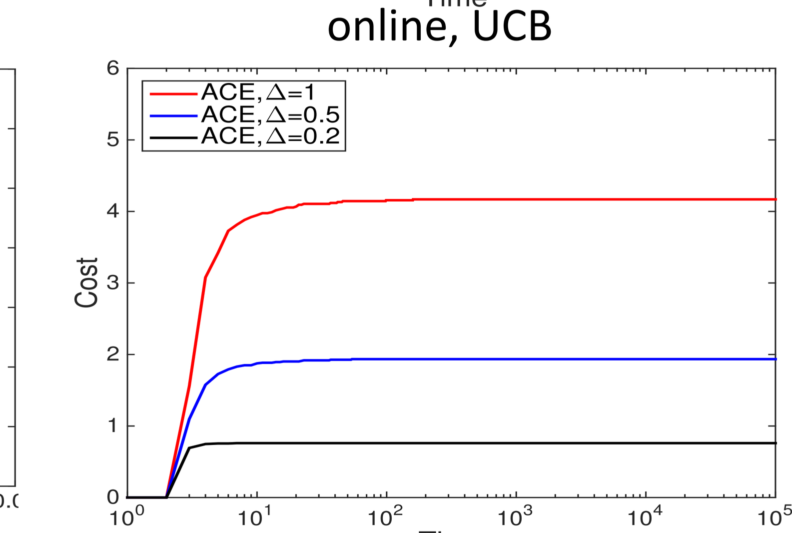

offline, UCB




online, UCB




offline, TS




online, TS

## Reference


[1] Jun, Kwang-Sung, et al. "Adversarial attacks on stochastic bandits." Advances in Neural Information Processing Systems. 2018.

[2] Ma, Yuzhe, et al. "Data poisoning attacks in contextual bandits." International Conference on Decision and Game Theory for Security. 2018.