UNIVERSITY OF ALBERTA

Differentially Private Contextual Linear Bandits

Roshan Shariff and Or Sheffet
{rshariff,osheffet}@ualberta.ca
ALBERTA MACHINE INTELLIGENCE INSTITUTE
amii

STOCHASTIC BANDIT PROBLEMS

Sequential decision making with n rounds. At round t:

MULTI-ARMED BANDITS

- Choose *action* $A_t \in \{1, \ldots, k\}$
- Receive *reward* $Y_t \sim P_{A_t}$

CONTEXTUAL BANDITS

- Receive *context* $c_t \in \mathcal{C}$
- Choose *action* $A_t \in \mathcal{A}$
- Receive *reward* $Y_t \sim P_{c_t, A_t}$

LINEAR BANDITS

- Choose *action* $X_t \in \mathcal{D} \subset \mathbb{R}^d$
- Mean reward is $\langle \theta^*, X_t \rangle$ with unknown parameter $\theta^* \in \mathbb{R}^d$

CONTEXTUAL LINEAR BANDITS

- Known feature map $\varphi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$
- Mean reward is $\langle \theta^*, \varphi(c_t, A_t) \rangle$

LINEAR BANDITS WITH CHANGING DECISION SETS

$$\mathcal{D}_t \doteq \{\varphi(c_t, a) \mid a \in \mathcal{A}\}$$

- Choosing $X_t \in \mathcal{D}_t$ also chooses $A_t \in \mathcal{A}$
- $\mathcal{D}_t$ encodes everything about reward

REWARD VS. REGRET

Maximizing reward is equivalent to minimizing *regret*:

$$\hat{R}_n \doteq \sum_{t=1}^{n} \max_{x \in \mathcal{D}_t} \langle \theta^*, x - X_t \rangle$$

- Cost of learning: reward lost by having to learn unknown parameter θ^*
- Measures inherent difficulty of learning problem
- This is actually *pseudo-regret*: includes randomness in algorithm's actions but not unavoidable reward noise

MOTIVATION AND SUMMARY

Contextual bandits often use contexts and rewards that are **private information**.

For example, online shopping: **context** is user's past purchases; **actions** are recommendations; and **reward** is whether user accepted recommendation.

We present a contextual linear bandit algorithm that balances learning with privacy preservation.

DIFFERENTIAL PRIVACY

Outputs (actions) don't reveal too much about inputs (contexts, rewards)

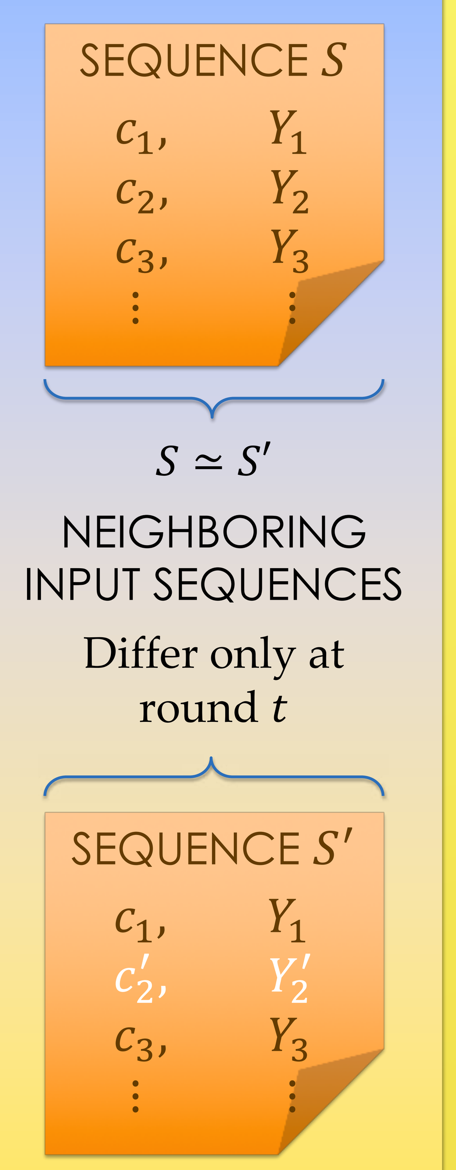
DEFINITION: (ε, δ)-Differential Privacy

Randomized algorithm $\mathcal{A}$ is (ε, δ)-DP for $\varepsilon \ge 0$ and $\delta \in [0,1]$ if for any subset of outputs $\mathcal{O}$,

$$\mathbb{P}(\mathcal{A}(S) \in \mathcal{O}) \le e^{\varepsilon} \mathbb{P}(\mathcal{A}(S') \in \mathcal{O}) + \delta$$

DEFINITION: (ε, δ)-Joint Differential Privacy

- Relaxation of (ε, δ)-DP for sequential tasks
- Context c_t revealed by action A_t, but not by *later* actions
- More suitable for contextual bandits (see lower bound below)



LOWER BOUNDS

DIFFERENTIAL PRIVACY REQUIRES IGNORING CONTEXT

Any (ε, δ)-DP contextual bandit algorithm must have linear regret

JOINT DIFFERENTIAL PRIVACY INCURS ADDITIONAL REGRET

Any ε-DP k-armed bandit algorithm must have $\Omega(k \log(n)/\varepsilon)$ regret

DIFFERENTIALLY PRIVATE LINEAR UCB

Modification of Linear Upper Confidence Bound (LinUCB) algorithm to maintain privacy

ELLIPSOIDAL CONFIDENCE SETS

Constructs Θ_t containing θ^* with high probability, based on:

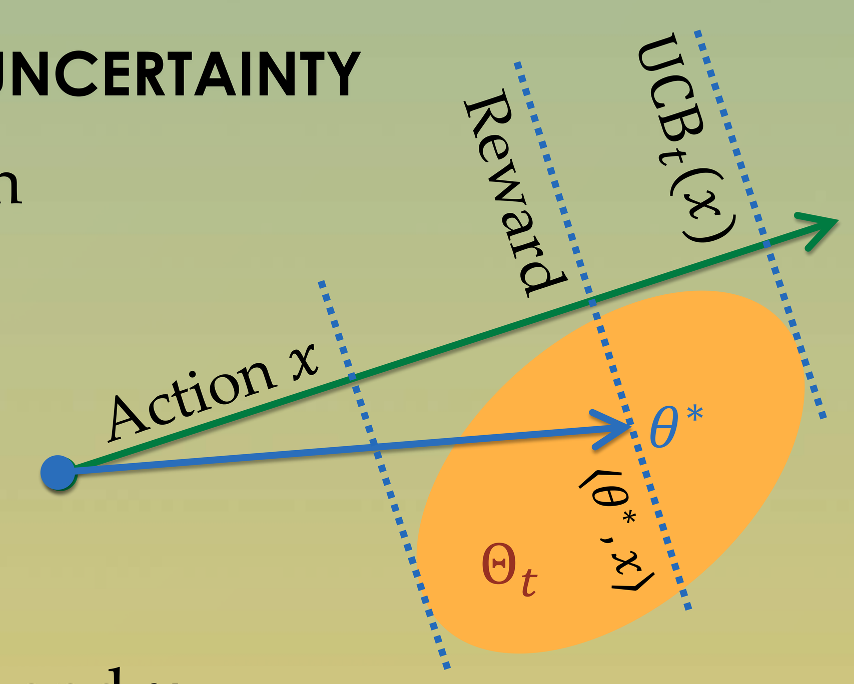
Gram matrix $V_t = \sum_{s<t} X_s X_s^{\mathsf{T}}$; vector $u_t = \sum_{s<t} X_s y_s$

OPTIMISM IN THE FACE OF UNCERTAINTY

Chooses "optimistic" action

$$X_t = \arg\max_{x \in \mathcal{D}_t} \mathrm{UCB}_t(x)$$

$$\mathrm{UCB}_t(x) \doteq \max_{\theta \in \Theta_t} \langle \theta, x \rangle$$



DIFFERENTIAL PRIVACY

Uses "noisy" versions of V_t and u_t

- Gaussian noise: variance $O(\log n \log^2(1/\delta)/\varepsilon^2)$
- Wishart noise: see details in paper

REGRET BOUNDS

- For both Wishart and Gaussian mechanisms, regret is
$$\mathbb{E}[\hat{R}_n] = \tilde{O}(\sqrt{n} \cdot d^{3/4}/\sqrt{\varepsilon})$$
- If suboptimal actions have a Δ reward gap, then
$$\mathbb{E}[\hat{R}_n] = O(\Delta^{-1} \operatorname{polylog}(n)\, d^2/\varepsilon)$$
- Both cases: multiplicative $\operatorname{polylog}(1/\delta)$ dependence
- See paper for details and high-probability bounds

EMPIRICAL RESULTS ON SYNTHETIC DATA