Privacy in Multi-armed Bandits

Fundamental Definitions & Lower Bounds on Performance

Debabrota Basu

Scool, Inria Lille - Nord Europe

What's Next?

1. Multi-armed Bandits: A Practitioner's View
2. Data Privacy: DP Framework
3. Private Bandits: Fundamental Definitions
4. Multi-armed Bandits: A Designer's View
5. Private Bandits: Regret Lower Bounds
6. Open Problems: Things to Work on

Sequential Decision Making

Medicine 1

$p_1^{\text{cured}} = 0.75$

Medicine 2

$p_2^{\text{cured}} = 0.95$

Medicine 3

$p_3^{\text{cured}} = 0.90$

...

Medicine K

$p_K^{\text{cured}} = 0.5$

Sequential Decision Making

under Incomplete Information: Multi-armed Bandits

Medicine 1

$p_1^{\text{cured}} = ?$

Medicine 2

$p_2^{\text{cured}} = ?$

Medicine 3

$p_3^{\text{cured}} = ?$

...

Medicine K

$p_K^{\text{cured}} = ?$

Sequential Decision Making

under Incomplete Information: Multi-armed Bandits




Medicine 1
$p_1^{\text{cured}} = ?$

Medicine 2
$p_2^{\text{cured}} = ?$

Medicine 3
$p_3^{\text{cured}} = ?$

$\dots$

Medicine K
$p_K^{\text{cured}} = ?$

For the t-th patient ($t \leq T$) in the study

1. the doctor π chooses a Medicine $A_t \in \{1, \dots, K\}$,
2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

Goal: Maximise the number of patients cured: $\sum_{t=1}^{T} R_t$.

A Fact Check

Multi-armed Bandits: A Practitioner's Perspective

- What is the Algorithm?
- What is the Input?
- What is the Output?
- What are the sources of Randomness?

A Fact Check

Multi-armed Bandits: A Practitioner's Perspective

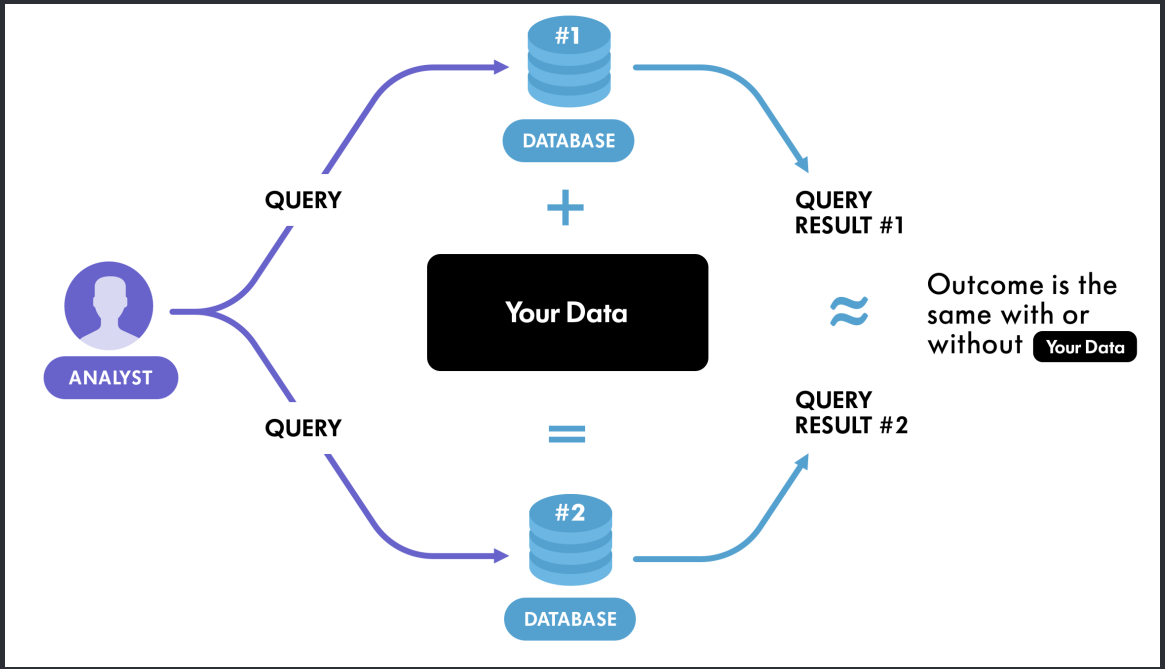
- What is the Algorithm?
  - The doctor or a digital assistant π
- What is the Input?
  - The sequence of observed responses from the patients $\{R_1, \ldots, R_T\}$
- What is the Output?
  - The sequence of chosen medicines by the algorithm $\{A_1, \ldots, A_T\}$
- What are the sources of Randomness?
  - The medical conditions of the patients and their reactions to the medicines, $\{\mathbb{P}_a\}_{a=1}^{K}$ (and a randomised algorithm π)

What's Next?

1. Multi-armed Bandits: A Practitioner's View
2. Data Privacy: DP Framework
3. Private Bandits: Fundamental Definitions
4. Multi-armed Bandits: A Designer's View
5. Private Bandits: Regret Lower Bounds
6. Open Problems: Things to Work on

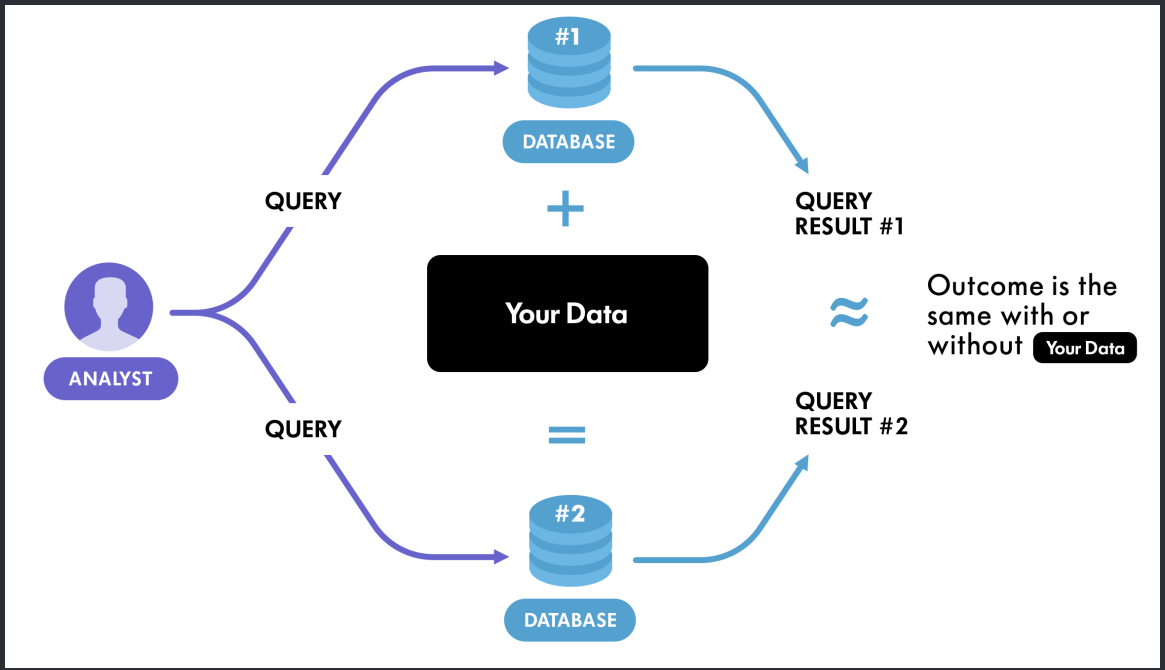
Data Privacy: ϵ-Differential Privacy [Dwork and Roth, 2014]

Information in input/database becomes private if it is indistinguishable from the output of a query/algorithm.



Data Privacy: ϵ-Differential Privacy [Dwork and Roth, 2014]

Information in input/database becomes private if it is indistinguishable from the output of a query/algorithm.



$$\frac{\mathbb{P}(\pi(\mathrm{DB} + \text{my data}) = \mathrm{Out})}{\mathbb{P}(\pi(\mathrm{DB}) = \mathrm{Out})} \leq e^{\epsilon} \longrightarrow \epsilon - \mathrm{DP}$$

Image Courtesy: www.winton.com

Differential Privacy (DP)

Ingredients:

- Input space: X (with symmetric neighbouring relation $\sim$)
- Output space: Y (with σ-algebra of measurable events)
- Privacy level: $\epsilon \geq 0$ (lower is better)

Differential Privacy (DP)

Ingredients:

- Input space: X (with symmetric neighbouring relation $\sim$)
- Output space: Y (with σ-algebra of measurable events)
- Privacy level: $\epsilon \geq 0$ (lower is better)

Formulation:
A randomised algorithm $\mathscr{A} : X \to Y$ is ϵ-differentially private if for all neighbouring inputs $x \sim x' \in X$ and for all subsets of outputs $O \subseteq Y$, we get

$$\mathbb{P}[\mathscr{A}(x) \in O] \leq e^{\epsilon} \mathbb{P}[\mathscr{A}(x') \in O].$$

- Neighbouring relation $\sim$ represents what is protected
- ϵ-DP is the worst-case guarantee

Differential Privacy (DP)

Ingredients:

- Input space: X (with symmetric neighbouring relation $\sim$)
- Output space: Y (with σ-algebra of measurable events)
- Privacy level: $\epsilon \geq 0$ (lower is better)

Formulation:
A randomised algorithm $\mathscr{A} : X \to Y$ is ϵ-differentially private if for all neighbouring inputs $x \sim x' \in X$ and for all subsets of outputs $O \subseteq Y$, we get

$$\mathbb{P}[\mathscr{A}(x) \in O] \leq e^{\epsilon} \mathbb{P}[\mathscr{A}(x') \in O].$$

- The slack on probability e^{ϵ} quantifies the amount of protection
- The randomness in the algorithm ensures the privacy

Why should we use Differential Privacy?

Fundamental Properties of DP

- Robustness to Post-processing: If $\mathscr{A}$ is (ϵ, δ)-DP, $f \circ \mathscr{A}$ is also (ϵ, δ)-DP for any $f : Y \to Z$.

Why should we use Differential Privacy?

Fundamental Properties of DP

- Robustness to Post-processing: If $\mathscr{A}$ is (ϵ, δ)-DP, $f \circ \mathscr{A}$ is also (ϵ, δ)-DP for any $f : Y \to Z$.

- Composition under Heterogeneity: If $\mathscr{A}_j$ are (ϵ_j, δ_j)-DP, aggregation of their outputs $(\mathscr{A}_1, \ldots, \mathscr{A}_K)$ is $(\sum_j \epsilon_j, \sum_j \delta_j)$-DP.

Why should we use Differential Privacy?

Fundamental Properties of DP

- Robustness to Post-processing: If $\mathscr{A}$ is (ϵ, δ)-DP, $f \circ \mathscr{A}$ is also (ϵ, δ)-DP for any $f : Y \to Z$.

- Composition under Heterogeneity: If $\mathscr{A}_j$ are (ϵ_j, δ_j)-DP, aggregation of their outputs $(\mathscr{A}_1, \ldots, \mathscr{A}_K)$ is $(\sum_j \epsilon_j, \sum_j \delta_j)$-DP.

- Group Privacy: If two inputs x and x' has t changes between them, a private algorithm $\mathscr{A}$ satisfies $(t\epsilon, te^{t\epsilon}\delta)$-DP for them.

Why should we use Differential Privacy?

Fundamental Properties of DP

- Robustness to Post-processing: If $\mathscr{A}$ is (ϵ, δ)-DP, $f \circ \mathscr{A}$ is also (ϵ, δ)-DP for any $f : Y \to Z$.

- Composition under Heterogeneity: If $\mathscr{A}_j$ are (ϵ_j, δ_j)-DP, aggregation of their outputs $(\mathscr{A}_1, \ldots, \mathscr{A}_K)$ is $(\sum_j \epsilon_j, \sum_j \delta_j)$-DP.

- Group Privacy: If two inputs x and x' has t changes between them, a private algorithm $\mathscr{A}$ satisfies $(t\epsilon, te^{t\epsilon}\delta)$-DP for them.

- Protection against Side-knowledge: If an attacker has prior knowledge $P_{prior}(x)$ and computes $P_{posterior}(x)$ after observing $\mathscr{A}(x)$ from an ϵ-DP algorithm, $P_{posterior}(x)$ still maintains the e^{ϵ} slack from $P_{prior}(x)$.

A Fact Check

Differential Privacy as a Data Privacy Framework

- What is privacy?
- What does DP definition encode?
- What are the benefits of using DP?

A Fact Check

Differential Privacy as a Data Privacy Framework

- What is privacy?
  - Indistinguishability from the mass in the eyes of a third-party.
- What does DP definition encode?
  - The idea of indistinguishability, the need of randomness for that, and the worst case loss of privacy for everyone involved.
- What are the benefits of using DP?
  - Flexible use of privatised data in future, linear mixture of multiple privacy levels and private mechanisms, and protection under prior information about the algorithm/individuals.

What's Next?

1. Multi-armed Bandits: A Practitioner's View
2. Data Privacy: DP Framework
3. Private Bandits: Fundamental Definitions
4. Multi-armed Bandits: A Designer's View
5. Private Bandits: Regret Lower Bounds
6. Open Problems: Things to Work on

Sequential Decision Making: Data Generation

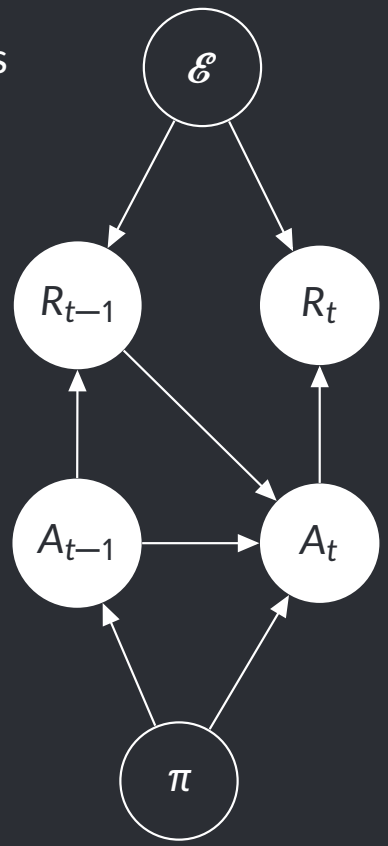
Data Privacy in Bandits

Response Distributions of Medicines on Patients $\mathscr{E} = \{\mathbb{P}(R|a)\}_{a=1}^{K}$

Observed Responses

Choice of Medicines

Doctor/ Algorithmic Assistant



The Bandit Game: For the t-th patient ($t \leq T$) in the study

1. the doctor π chooses a Medicine $A_t \in \{1, \ldots, K\}$,
2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

Sequential Decision Making: Data Generation

Data Privacy in Bandits

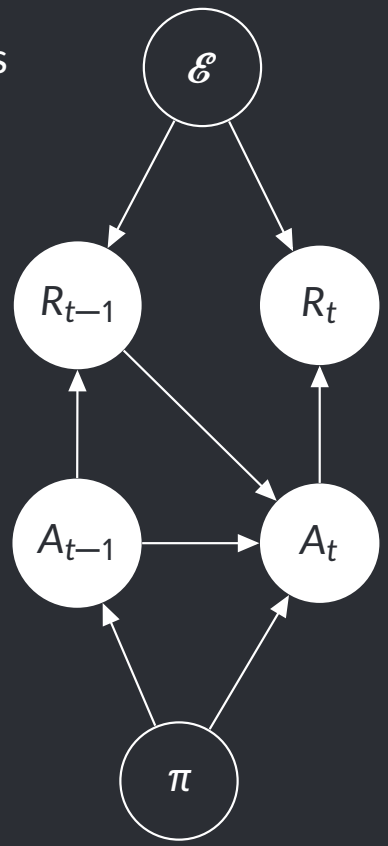
Response Distributions of Medicines on Patients

$\mathscr{E} = \{\mathbb{P}(R|a)\}_{a=1}^{K}$

Observed Responses

Choice of Medicines

Doctor/ Algorithmic Assistant



Input to π

Observed Responses: $R^T = \{R_1, \ldots, R_T\}$

Output of π

Decisions: $A^T = \{A_1, \ldots, A_T\}$

Data Privacy in Bandits

A patient t wants to keep her response R_t to a medicine A_t private.

Private Bandits: The History

Plethora of Claims, Plethora of Contradictions

1. DP on the sequence (Sequential DP)
   [Mishra and Thakurta, 2015, Tossou and Dimitrakakis, 2017]:

$$\mathbb{P}_\pi(A^T \mid r_1, \ldots, r_t, \ldots, r_T) \leq e^\epsilon \, \mathbb{P}_\pi(A^T \mid r_1, \ldots, r'_t, \ldots, r_T)$$

2. DP at every instance $t \leq T$ (Instantaneous DP)
   [Tossou and Dimitrakakis, 2016, Shariff and Sheffet, 2018]:

$$\mathbb{P}_\pi(a_{t+1} \mid r_1, \ldots, r_k, \ldots, r_t) \leq e^\epsilon \, \mathbb{P}_\pi(a_{t+1} \mid r_1, \ldots, r'_k, \ldots, r_t)$$

3. DP against external algorithm (Local DP) [Gajane et al., 2017]:

$$\mathbb{P}(input_t \mid r_t) \leq e^\epsilon \, \mathbb{P}(input_t \mid r'_t)$$

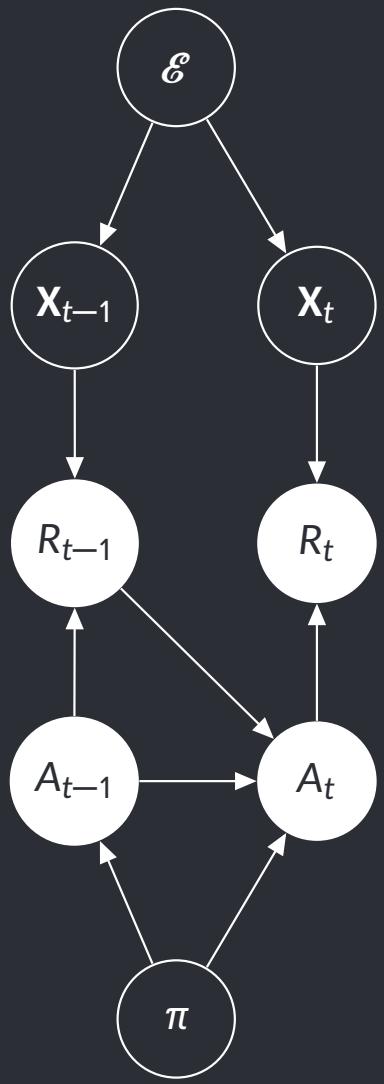
Privacy in Sequential Decision Making

Private Multi-armed Bandits: Differential Privacy

All Possible Responses

Observed Responses

Decision of Medicines



Generalising the Input

Make patient t's all possible responses

$$\mathbf{X}_t = [R_t^1, \ldots, R_t^A]$$

to all the A medicines private.

Generalised input: $\mathbf{X}^T = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$

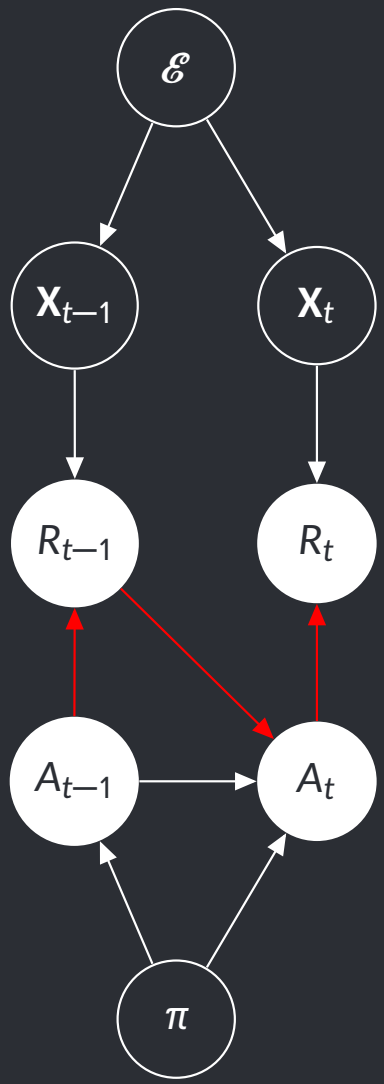
Privacy in Sequential Decision Making

Private Multi-armed Bandits: Global DP [Basu et al., 2020]

All Possible Responses

Observed Responses

Decision of Medicines



Input: $\mathbf{X}^T$

Output: A^T

Algorithm: π

My data: X_t

ϵ-Global DP

$$\frac{\mathbb{P}_\pi\left(\text{Set of Decisions} \,\middle|\, \text{Possible responses of T patients} + \text{my data}\right)}{\mathbb{P}_\pi\left(\text{Set of Decisions} \,\middle|\, \text{Possible responses of T patients}\right)} \leq e^\epsilon$$

Privacy in Sequential Decision Making

Private Multi-armed Bandits: Global DP [Basu et al., 2020]

ϵ-(global) DP for Bandits

A bandit algorithm π satisfies ϵ-DP if:

$$\mathbb{P}_\pi(a_1, \ldots, a_T \mid \mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T) \leq e^\epsilon \, \mathbb{P}_\pi(a_1, \ldots, a_T \mid \mathbf{x}_1, \ldots, \mathbf{x}'_t, \ldots, \mathbf{x}_T)$$

Privacy in Sequential Decision Making

Private Multi-armed Bandits: Global DP [Basu et al., 2020]

ϵ-(global) DP for Bandits

A bandit algorithm π satisfies ϵ-DP if:

$$\mathbb{P}_\pi(a_1, \ldots, a_T \mid \mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T) \leq e^\epsilon \, \mathbb{P}_\pi(a_1, \ldots, a_T \mid \mathbf{x}_1, \ldots, \mathbf{x}'_t, \ldots, \mathbf{x}_T)$$

The Unification of Existing Definitions:

- ϵ-(global) DP for bandits $\implies$ ϵ-Sequential DP
- ϵ-(global) DP for bandits $\implies$ 2ϵ-Instantaneous DP
- ϵ-Instantaneous DP $\implies$ $T\epsilon$-(global) DP for bandits
- ϵ-local DP $\implies$ ϵ-(global) DP for bandits

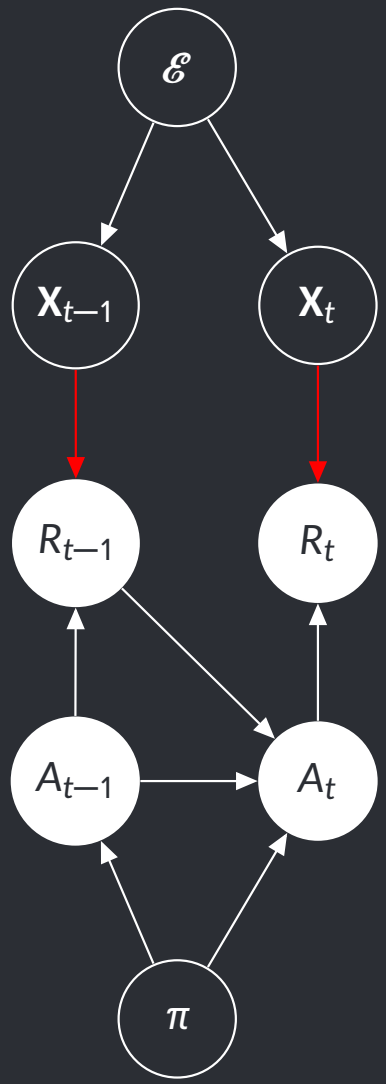
Privacy in Sequential Decision Making

Private Multi-armed Bandits: Local DP

All Possible Responses

Observed Responses

Decision of Medicines



ϵ-Local DP

$$\frac{\mathbb{P}\left(\begin{array}{l|l}\text{Observed} & \text{Possible responses} \\ \text{responses} & \text{of T patients} + \begin{array}{l}\text{my} \\ \text{data}\end{array}\end{array}\right)}{\mathbb{P}\left(\begin{array}{l|l}\text{Observed} & \text{Possible responses} \\ \text{responses} & \text{of T patients}\end{array}\right)} \leq e^{\epsilon}$$

Local DP $\Longrightarrow$ Global DP

while not constraining algorithm π.

(Post-processing Property of DP).

What did We Learn?

- What is the input for private bandit algorithm?
- What is the output for private bandit algorithm?
- What is the difference between local DP and other setups?
- What is the benefit of aiming for ϵ-global DP?

What did We Learn?

- What is the input for private bandit algorithm?
  - All possible generated responses of all the T patients against all the K decisions $\mathbf{X}^T = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$.
- What is the output for private bandit algorithm?
  - All the decisions for T patients $A^T = \{A_1, \ldots, A_T\}$.
- What is the difference between local DP and other setups?
  - In other DPs, the individual has to believe in the centralised algorithm. Local DP keeps the data private from individual level.
- What is the benefit of aiming for ϵ-global DP?
  - It provides a unified definition for privacy in bandits and satisfying this definition provides stronger guarantees than existing definitions.

What's Next?

1. Multi-armed Bandits: A Practitioner's View
2. Data Privacy: DP Framework
3. Private Bandits: Fundamental Definitions
4. Multi-armed Bandits: A Designer's View
5. Private Bandits: Regret Lower Bounds
6. Open Problems: Things to Work on

Sequential Decision Making

under Incomplete Information: Multi-armed Bandits




Distribution 1
$p_1^{\text{reward}} = ?$

Distribution 2
$p_2^{\text{reward}} = ?$

Distribution 3
$p_3^{\text{reward}} = ?$

$\dots$

Distribution K
$p_K^{\text{reward}} = ?$

In the t-th step ($t \in \{1, \dots, T\}$)

1. the algorithm π chooses a distribution $A_t \in \{1, \dots, K\}$,
2. Observes a reward $R_t \in \mathbb{R}$ such that $R_t \sim p_{A_t}^{\text{reward}}$.

Goal: Maximise the observed cumulative reward: $\sum_{t=1}^{T} R_t$.

Value of a Bandit Algorithm

Expected Cumulative Reward: A Theoretically Malleable Goal

- Maximise cumulative reward $\sum_{t=1}^{T} R_t$

Value of a Bandit Algorithm

Expected Cumulative Reward: A Theoretically Malleable Goal

- Maximise cumulative reward $\sum_{t=1}^{T} R_t \rightarrow$ a random variable

Value of a Bandit Algorithm

Expected Cumulative Reward: A Theoretically Malleable Goal

- Maximise cumulative reward $\sum_{t=1}^{T} R_t \rightarrow$ a random variable
- Maximise expected cumulative reward or value of π:

Value of a Bandit Algorithm

Expected Cumulative Reward: A Theoretically Malleable Goal

- Maximise cumulative reward $\sum_{t=1}^{T} R_t \rightarrow$ a random variable
- Maximise expected cumulative reward or value of π:

$$V_{\mathcal{E},\pi}(T) \triangleq \mathbb{E}_{\mathcal{E}}\left[\sum_{t=0}^{T} R_t \mid A_t \sim \pi\right]$$

$$= \sum_{a=1}^{K} \underbrace{\mathbb{E}_{\pi\mathcal{E}}\left[\sum_{t=1}^{T}\left(R_{A_t} \times \underbrace{\mathbb{1}(A_t = a)}_{\text{Arm } a \text{ is played}}\right)\right]}_{\text{Expected reward from arm } a \text{ by time } T} \quad \text{(the indicator allows the sum over a)}$$

$$= \sum_{a=1}^{K} \underbrace{\mathbb{E}_{\pi}\left[\sum_{t=1}^{T} \mathbb{1}(A_t = a)\right]}_{\text{Expected \#draws of } a \text{ by } T} \mathbb{E}_{\mathcal{E}}[R_a] \quad \triangleq \sum_{a=1}^{K} \mathbb{E}_{\pi}\left[N_T^a\right] \mu_a$$

Performance Metric Under Incomplete Information

Regret

Regret $\mathrm{Reg}_{\mathcal{E},\pi}(T)$

$\triangleq$ Value of Optimal Algorithm with Full Information

$-$ Value of Algorithm π with Incomplete Information

$$= T\mu^* - \sum_{a=1}^{K} \mathbb{E}_\pi \left[N_T^a \right] \mu_a$$

$$= \sum_{a=1}^{K} \mathbb{E}_\pi \left[N_T^a \right] (\mu^* - \mu_a) \qquad \left(\text{since, } T = \sum_{a=1}^{K} \mathbb{E}_\pi \left[N_T^a \right] \right)$$

$$= \sum_{a=1}^{K} \text{Expected number of time decision } a \text{ is taken}$$

$\times$ Expected suboptimality of arm a (Δ_a)

Two Faces of a Bandit: Exploration and Exploitation

Pure Exploration

Take each decision uniformly and accumulate empirical knowledge.

Pure Exploitation

Take the decision with maximum empirical reward as per present knowledge.

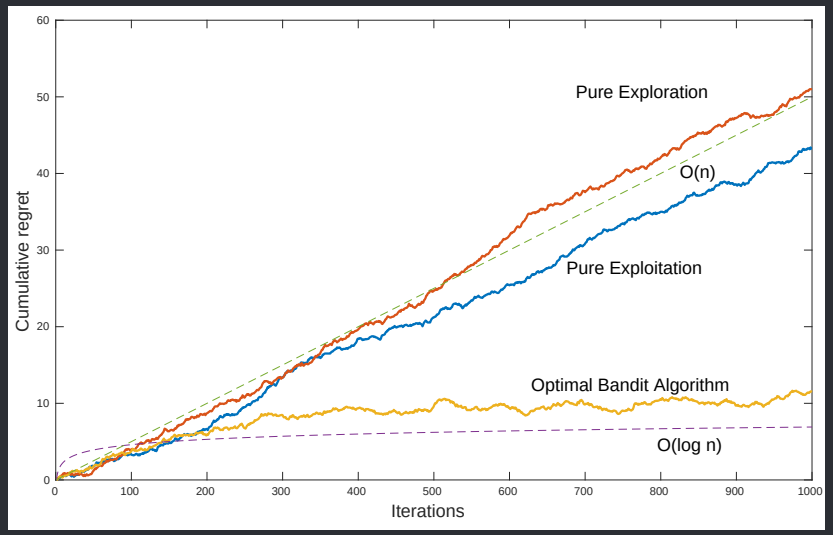
Two Faces of a Bandit: Exploration and Exploitation

Pure Exploration

Take each decision uniformly and accumulate empirical knowledge.

Pure Exploitation

Take the decision with maximum empirical reward as per present knowledge.



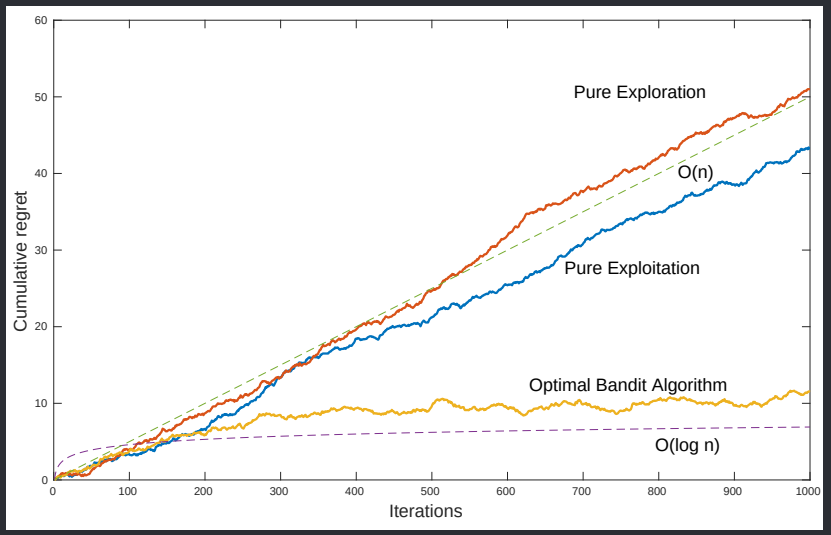
Two Faces of a Bandit: Exploration and Exploitation

Pure Exploration

Take each decision uniformly and accumulate empirical knowledge.

Pure Exploitation

Take the decision with maximum empirical reward as per present knowledge.



The Exploration–exploitation Trade-off

Exploration and exploitation should be adapted on-the-go to achieve the optimal regret.

Hardness of a Bandit Problem

Lower Bounds on Regret

Minimax Regret [Vogel, 1960]

$$\mathrm{Reg}^*_{\mathrm{Minimax}}(T) \triangleq \min_{\pi} \max_{\mathscr{E}} \mathrm{Reg}(\pi, \mathscr{E}, T)$$

- The best achievable regret in the worst-case scenario.
- The lower bound for non-private case is $\sqrt{(K-1)T}$

Hardness of a Bandit Problem

Lower Bounds on Regret

Bayesian Minimax Regret [Lattimore and Szepesvári, 2019]

In Bayesian setup, a prior distribution Q over environments $\mathscr{E}$ is assumed.

$$\mathrm{Reg}_{\mathrm{Bayes}}(\pi, T, Q) \triangleq \int_{\mathscr{E}^T} \mathrm{Reg}(\pi, \mathscr{E}, T) dQ(\mathscr{E}).$$

The *Bayesian minimax regret* is the worst possible regret over all priors Q:

$$\begin{aligned}\mathrm{Reg}^*_{\mathrm{Bayes}}(T) &\triangleq \min_{\pi} \max_{Q} \int_{\mathscr{E}^T} \mathrm{Reg}(\pi, \mathscr{E}, T) dQ(\mathscr{E}) \\ &= \min_{\pi} \max_{Q} \mathrm{Reg}_{\mathrm{Bayes}}(\pi, T, Q).\end{aligned}$$

- The best achievable regret for the worst-case prior.
- Lower bound for non-private case is $\sqrt{(K-1)T}$.

Hardness of a Bandit Problem

Lower Bounds on Regret

Problem-dependent Regret [Lai and Robbins, 1985]

$$\mathrm{Reg}^*_{\mathscr{E}}(T) \triangleq \min_{\pi} \mathrm{Reg}_{\mathscr{E}}(\pi, T)$$

- The best achievable regret for a specific environment $\mathscr{E}$.
- The lower bound for non-private case is

$$\sum_{a=1}^{K} \frac{\Delta_a}{D_{\mathrm{KL}}\left(f_a \| f^*\right)} \log T \triangleq c(\mathscr{E}) \log T.$$

What's Next?

1. Multi-armed Bandits: A Practitioner's View
2. Data Privacy: DP Framework
3. Private Bandits: Fundamental Definitions
4. Multi-armed Bandits: A Designer's View
5. Private Bandits: Regret Lower Bounds
6. Open Problems: Things to Work on

Preparing the Ingredients

The Probability Space of Observed Histories

- Random variable: Observed history $\mathscr{H}_T \triangleq \{(A_i, X_i)\}_{i=1}^{T}$
- Measurable space, σ-measure: $\left(([K] \times \mathbb{R})^T, \mathscr{B}([K] \times \mathbb{R})^T\right)$
- Probability measure: $\mathbb{P}_{\pi\mathscr{E}}^{T}$ induced by the algorithm π and environment $\mathscr{E}$

$$\mathbb{P}_{\pi\mathscr{E}}^{T} \triangleq \mathbb{P}_{\pi\mathscr{E}}(\mathscr{H}_T)$$

$$= \prod_{t=1}^{T} \underbrace{\pi(A_t|\mathscr{H}_{t-1})}_{\substack{\text{Chosen action depends only} \\ \text{on algorithm and history}}} \times \underbrace{f_{A_t}(X_t)}_{\substack{\text{Observed reward depends} \\ \text{only on the environment}}}$$

A Proof of Regret Lower Bounds

A Unified Framework

Step 1:

Choose two environments $\mathscr{E}_1$ and $\mathscr{E}_2$.

They are the same except that the arm 1 is optimal in $\mathscr{E}_1$ and arm i is optimal in $\mathscr{E}_2$.

Bad event for $\mathscr{E}_1$: $E \triangleq N_1(T) \leq T/2$

Bad event for $\mathscr{E}_2$: $E^C \triangleq N_1(T) > T/2$

A Proof of Regret Lower Bounds

A Unified Framework

Step 2:

Lower Bounding the Regrets of the Environments.

$$\begin{aligned}
\mathrm{Reg}(\pi, \mathscr{E}_1, T) &= \sum_{a=1}^{K} \mathbb{E}_\pi \left[N_T^a \right] (\mu^* - \mu_a) \\
&\geq \mathbb{P}^T_{\pi \mathscr{E}_1} \left(N_1(T) \leq T/2 \right) \frac{T}{2} (\mu_1 - \mu_i) \\
&= \mathbb{P}^T_{\pi \mathscr{E}_1} \left(E \right) \frac{T}{2} (\mu_1 - \mu_i) \\
\mathrm{Reg}(\pi, \mathscr{E}_2, T) &> \mathbb{P}^T_{\pi \mathscr{E}_2} \left(N_1(T) > T/2 \right) \frac{T}{2} (\mu_i' - \mu_1) \\
&= \mathbb{P}^T_{\pi \mathscr{E}_2} \left(E^C \right) \frac{T}{2} (\mu_i' - \mu_1)
\end{aligned}$$

A Proof of Regret Lower Bounds

A Unified Framework

Step 3:

From regret lower bounds to KL-divergence of observed histories.

$$
\begin{aligned}
&\mathrm{Reg}(\pi, \mathscr{E}_1, T) + \mathrm{Reg}(\pi, \mathscr{E}_2, T) \\
\geq &\frac{T}{2}\left(\mathbb{P}^T_{\pi\mathscr{E}_1}(E)(\mu_1 - \mu_i) + \mathbb{P}^T_{\pi\mathscr{E}_2}(E^C)(\mu_i' - \mu_1)\right) \\
\geq &\frac{T}{2}\left(\mathbb{P}^T_{\pi\mathscr{E}_1}(E) + \mathbb{P}^T_{\pi\mathscr{E}_2}(E^C)\right)\min\{(\mu_1 - \mu_i), (\mu_i' - \mu_1)\} \\
\geq &\frac{T}{4}\exp(- \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}^T_{\pi\mathscr{E}_1} \| \mathbb{P}^T_{\pi\mathscr{E}_2}\right)}_{\text{Dissimilarity of probability measures for two contrasting environments and a given algorithm}}) \underbrace{\min\{(\mu_1 - \mu_i), (\mu_i' - \mu_1)\}}_{\text{suboptimality of the environments}}
\end{aligned}
$$

Minimising regret is now equivalent to maximising $D_{\mathrm{KL}}\left(\mathbb{P}_{\pi\mathscr{E}_1} \| \mathbb{P}_{\pi\mathscr{E}_2}\right)$.

A Proof of Regret Lower Bounds

A Unified Framework

Step 4:

KL-divergence decomposition [Garivier et al., 2018] and upper bounding the divergence.

$$
\begin{aligned}
& D_{\mathrm{KL}}\left(\mathbb{P}^{T}_{\pi\mathscr{E}_1} \| \mathbb{P}^{T}_{\pi\mathscr{E}_2}\right) \\
&= \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi(A_t|\mathscr{H}_t, \mathscr{E}_1) \| \pi(A_t|\mathscr{H}_t, \mathscr{E}_2)\right) \\
&\qquad + \sum_{a=1}^{K} \mathbb{E}_{\mathscr{E}_1}\left[N_a(T)\right] D_{\mathrm{KL}}\left(f_a \in \mathscr{E}_1 \| f_a \in \mathscr{E}_2\right) \\
&= \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\pi(A_t|\mathscr{H}_t, \mathscr{E}_1) \| \pi(A_t|\mathscr{H}_t, \mathscr{E}_2)\right) + \mathbb{E}_{\mathscr{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right) \\
&\leq \mathrm{Upper\ Bound}_1 + \mathrm{Upper\ Bound}_2
\end{aligned}
$$

Upper Bounding the KL Divergence

Non-private

$$\text{Upper Bound}_1 = 0$$

$$\text{Upper Bound}_2 = \mathbb{E}_{\mathscr{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right)$$

Upper Bounding the KL Divergence

Local DP

$$\text{Upper Bound}_1 = 0$$

$$\begin{aligned}\text{Upper Bound}_2 &= 2\min\{4, e^{2\epsilon}\}(e^{\epsilon}-1)^2\mathbb{E}_{\mathscr{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right) \\ &= L^{-2}(\epsilon)\,\mathbb{E}_{\mathscr{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right)\end{aligned}$$

Upper Bounding the KL Divergence

Global DP

$$\text{Upper Bound}_1 = 2(\epsilon + L) = C$$

$$\begin{aligned}\text{Upper Bound}_2 &= \exp(2(\epsilon + L))\, \mathbb{E}_{\mathcal{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right) \\ &= e^C\, \mathbb{E}_{\mathcal{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right)\end{aligned}$$

- L is the Lipschitz constant of the log-density of the observed rewards

$$\ln \sup_{a, x_a, x_a'} \frac{\mathbb{P}_{\mathcal{E}}(x_a)}{\mathbb{P}_{\mathcal{E}}(x_a')} \leq L$$

This is a measure of smoothness on the probability of rewards.

Upper Bounding the KL Divergence

$$D_{\mathrm{KL}}\left(\mathbb{P}^{T}_{\pi\mathscr{E}_1} \| \mathbb{P}^{T}_{\pi\mathscr{E}_2}\right) \leq u_1 + u_2\, \mathbb{E}_{\mathscr{E}_1}\left[N_i(T)\right] D_{\mathrm{KL}}\left(f_i \| f_i'\right)$$

- For non-private bandit, $u_1 = 0$ and $u_2 = 1$
- For locally private bandit, $u_1 = 0$ and $u_2 = L^{-2}(\epsilon)$
- For globally private bandit, $u_1 = C = 2(\epsilon + L)$ and $u_2 = e^{C}$

Minimax Regret Bound I

Step 5:

Substitute environment parameters such that $\min\{(\mu_1 - \mu_i), (\mu_i' - \mu_1)\} = \Delta$ and $\mathbb{E}_{\mathscr{E}_1}[N_i(T)] \leq \frac{T}{K-1}$.

Thus, we get

$$
\begin{aligned}
&\max\{\mathrm{Reg}(\pi, \mathscr{E}_1, T), \mathrm{Reg}(\pi, \mathscr{E}_2, T)\} \\
&\geq \frac{1}{2}\left(\mathrm{Reg}(\pi, \mathscr{E}_1, T) + \mathrm{Reg}(\pi, \mathscr{E}_2, T)\right) \\
&\geq \frac{T\Delta}{4} \exp\left[u_1 + u_2 \frac{T}{K-1} D_{\mathrm{KL}}\left(f_K(0, l) \| f_K(2\Delta, l)\right)\right] \\
&\geq \frac{T\Delta}{4} \exp\left[u_1 + u_2 \frac{T}{K-1} \times 2\Delta^2\right].
\end{aligned}
$$

Minimax Regret Bound II

Step 6:

Boring algebra

$$\mathrm{Reg}^*_{\mathrm{Minimax}}(T) \geq \sqrt{G(\epsilon)(K-1)T} \quad \text{For global DP, } u_1 = C \text{ and } u_2 = e^C$$

$$\geq \sqrt{L^2(\epsilon)(K-1)T} \quad \text{For local DP, } u_1 = 0 \text{ and } u_2 = L^{-2}(\epsilon)$$

Here,

$$G(\epsilon) = \frac{\ln(\epsilon^2+1)}{e^{6\epsilon}\epsilon^{(1+\frac{2}{\epsilon})}} = O\left(\frac{1}{\epsilon}\right)$$

$$L^2(\epsilon) = \frac{1}{\min\{4, e^{2\epsilon}\}(e^\epsilon - 1)^2} = O\left(\frac{1}{\epsilon^2}\right)$$

Bayesian Minimax Regret

Theorem 1 in [Lattimore and Szepesvári, 2019]

For bounded rewards,

$$\mathrm{Reg}^*_{\mathrm{minimax}}(T) = \mathrm{Reg}^*_{\mathrm{Bayes}}(T).$$

Lower bounds are available for free here! :)

Problem-dependent Regret Bound

Step 5:

Substitute environment variables such that

$$D_{\mathrm{KL}}\left(f_i \| f_i'\right) \leq D_{\mathrm{KL}}\left(f_i \| f^*\right) + \delta.$$

For small δ, f_i' and f^* are similar and thus, hard to distinguish.

$$\begin{aligned}
&\mathrm{Reg}(\pi, \mathscr{E}_1, T) + \mathrm{Reg}(\pi, \mathscr{E}_2, T) \\
\geq &\frac{T}{4} \min\{(\mu_i - \mu^*), (\mu_i' - \mu^*)\} \\
&\exp\left[-u_1 - u_2 \mathbb{E}_{\pi \mathscr{E}_1}\left[N_i(T)\right]\left(D_{\mathrm{KL}}\left(f_i \| f^*\right) + \delta\right)\right].
\end{aligned}$$

Problem-dependent Regret Bound

Step 6:

Do some boring algebra, take limit $T \to \infty$, and assume that the regrets for both the environments are sublinear,

$$\liminf_{T\to\infty} \frac{\mathrm{Reg}(\pi, \mathscr{E}_1, T)}{\log T} = \liminf_{T\to\infty} \sum_{a\neq a^*} \frac{\mathbb{E}_{\pi\mathscr{E}_1}\left[N_a(T)\right](\mu_a - \mu^*)}{\log T} \quad \text{(Def. of regret)}$$

$$\geq \frac{1}{u_2} \sum_{a\neq a^*} \frac{(\mu_a - \mu^*)}{D_{\mathrm{KL}}\left(f_a \| f^*\right)} \quad \text{(upper bound on KL divergence)}$$

$$= \frac{1}{L^2(\epsilon)} \sum_{a\neq a^*} \frac{\Delta_a}{D_{\mathrm{KL}}\left(f_a \| f^*\right)} \quad \text{for local DP}$$

$$\geq \frac{1}{1+2\epsilon} \sum_{a\neq a^*} \frac{\Delta_a}{D_{\mathrm{KL}}\left(f_a \| f^*\right)} \quad \text{for global DP}$$

The Cost of Privacy

Regret Lower Bounds for Private Bandits [Basu et al., 2020]

Lower Bounds	Minimax Regret	Bayesian Minimax Regret	Problem-dependent Regret
Non-private	$\sqrt{(A-1)T}$	$\sqrt{(A-1)T}$	$c(\mathscr{E})\log T$
Global DP	$\sqrt{G(\epsilon)(A-1)T}$	$\sqrt{G(\epsilon)(A-1)T}$	$(1+\epsilon)^{-1}c(\mathscr{E})\log T$
Local DP	$L(\epsilon)\sqrt{(A-1)T}$	$L(\epsilon)\sqrt{(A-1)T}$	$L^2(\epsilon)c(\mathscr{E})\log T$

The Cost of Privacy

Regret Lower Bounds for Private Bandits [Basu et al., 2020]

Lower Bounds	Minimax Regret	Bayesian Minimax Regret	Problem-dependent Regret
Non-private	$\sqrt{(A-1)T}$	$\sqrt{(A-1)T}$	$c(\mathscr{E})\log T$
Global DP	$\sqrt{G(\epsilon)(A-1)T}$	$\sqrt{G(\epsilon)(A-1)T}$	$(1+\epsilon)^{-1}c(\mathscr{E})\log T$
Local DP	$L(\epsilon)\sqrt{(A-1)T}$	$L(\epsilon)\sqrt{(A-1)T}$	$L^2(\epsilon)c(\mathscr{E})\log T$

Lower bounds: Non-private (O(1)) < Global DP $(O(1/\epsilon))$ < Local DP $(O(1/\epsilon^2))$ ← Amount of Noise Injected

- As $\epsilon \to 0$, the lower bounds go to infinity but in practice regret in bandits is always $O(T)$.
- As $\epsilon \to \infty$, the lower bounds match with non-private lower bounds.

What's Next?

1. Multi-armed Bandits: A Practitioner's View
2. Data Privacy: DP Framework
3. Private Bandits: Fundamental Definitions
4. Multi-armed Bandits: A Designer's View
5. Private Bandits: Regret Lower Bounds
6. Open Problems: Things to Work on

Open Problems

(Dis)solving a Conjecture

Conjecture

The problem dependent lower bound for global DP will be

$$\left(c(\mathscr{E}) + \frac{1}{\epsilon}\right) \log(T).$$

- Our lower bound is different as $c(\mathscr{E}) + 1/\epsilon \geq \frac{c(\mathscr{E})}{1+\epsilon}$. We still don't know whether ours is achievable.

- The existing proof for contextual bandits by [Shariff and Sheffet, 2018] is not correct for all ϵ.

Open Problems

Designing Optimal Algorithms

- Designing optimal local DP algorithms, both UCB and Thompson sampling types, for bandits
  - Recent works in UCB type algorithms: [Ren et al., 2020, Zheng et al., 2020, Chen et al., 2020, Zhou and Tan, 2020]

- Designing optimal global DP algorithms, both UCB and Thompson sampling types, for bandits
  - Recent works in UCB type algorithms for linear bandits: [Sajed, 2019, Dubey and Pentland, 2020, Hannun et al., 2019, Malekzadeh et al., 2020]

- Designing optimal DP algorithms for general RL
  - Recent works with local DP: [Vietri et al., 2020]

Privacy in Multi-armed Bandits

Fundamental Definitions & Lower Bounds on Regret

Extended Paper: https://arxiv.org/abs/1905.12298

Co-creator: Christos Dimitrakakis

Chalmers University of Technology, Sweden & University of Oslo, Norway

References I

[Basu et al., 2020] Basu, D., Dimitrakakis, C., and Tossou, A. (2020).
Differential privacy for multi-armed bandits: What is it and what is its cost?
arXiv preprint arXiv:1905.12298.

[Chen et al., 2020] Chen, X., Zheng, K., Zhou, Z., Yang, Y., Chen, W., and Wang, L. (2020).
(locally) differentially private combinatorial semi-bandits.
arXiv preprint arXiv:2006.00706.

[Dubey and Pentland, 2020] Dubey, A. and Pentland, A. (2020).
Differentially-private federated linear bandits.
Advances in Neural Information Processing Systems, 33.

[Dwork and Roth, 2014] Dwork, C. and Roth, A. (2014).
The algorithmic foundations of differential privacy.
Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407.

[Gajane et al., 2017] Gajane, P., Urvoy, T., and Kaufmann, E. (2017).
Corrupt bandits for preserving local privacy.
arXiv preprint arXiv:1708.05033.

References II

[Garivier et al., 2018] Garivier, A., Ménard, P., and Stoltz, G. (2018).
Explore first, exploit next: The true shape of regret in bandit problems.
Mathematics of Operations Research.

[Hannun et al., 2019] Hannun, A., Knott, B., Sengupta, S., and van der Maaten, L. (2019).
Privacy-preserving contextual bandits.
arXiv preprint arXiv:1910.05299.

[Lai and Robbins, 1985] Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1):4–22.

[Lattimore and Szepesvári, 2019] Lattimore, T. and Szepesvári, C. (2019).
An information-theoretic approach to minimax regret in partial monitoring.
In *Conference on Learning Theory*, pages 2111–2139.

[Malekzadeh et al., 2020] Malekzadeh, M., Athanasakis, D., Haddadi, H., and Livshits, B. (2020).
Privacy-preserving bandits.
Proceedings of Machine Learning and Systems, 2:350–362.

References III

[Mishra and Thakurta, 2015] Mishra, N. and Thakurta, A. (2015).
(nearly) optimal differentially private stochastic multi-arm bandits.
In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 592–601. AUAI Press.

[Ren et al., 2020] Ren, W., Zhou, X., Liu, J., and Shroff, N. B. (2020).
Multi-armed bandits with local differential privacy.
arXiv preprint arXiv:2007.03121.

[Sajed, 2019] Sajed, T. (2019).
Optimal differentially private finite armed stochastic bandit.

[Shariff and Sheffet, 2018] Shariff, R. and Sheffet, O. (2018).
Differentially private contextual linear bandits.
In *Advances in Neural Information Processing Systems*, pages 4296–4306.

[Tossou and Dimitrakakis, 2016] Tossou, A. C. and Dimitrakakis, C. (2016).
Algorithms for differentially private multi-armed bandits.
In *Thirtieth AAAI Conference on Artificial Intelligence*.

References IV

[Tossou and Dimitrakakis, 2017] Tossou, A. C. Y. and Dimitrakakis, C. (2017).
Achieving privacy in the adversarial multi-armed bandit.
In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Vietri et al., 2020] Vietri, G., Balle, B., Krishnamurthy, A., and Wu, Z. S. (2020).
Private reinforcement learning with pac and regret guarantees.
arXiv preprint arXiv:2009.09052.

[Vogel, 1960] Vogel, W. (1960).
An asymptotic minimax theorem for the two armed bandit problem.
The Annals of Mathematical Statistics, 31(2):444–451.

[Zheng et al., 2020] Zheng, K., Cai, T., Huang, W., Li, Z., and Wang, L. (2020).
Locally differentially private (contextual) bandits learning.
arXiv preprint arXiv:2006.00701.

[Zhou and Tan, 2020] Zhou, X. and Tan, J. (2020).
Local differential privacy for bayesian optimization.
arXiv preprint arXiv:2010.06709.