Private Reinforcement Learning with PAC and Regret Guarantees

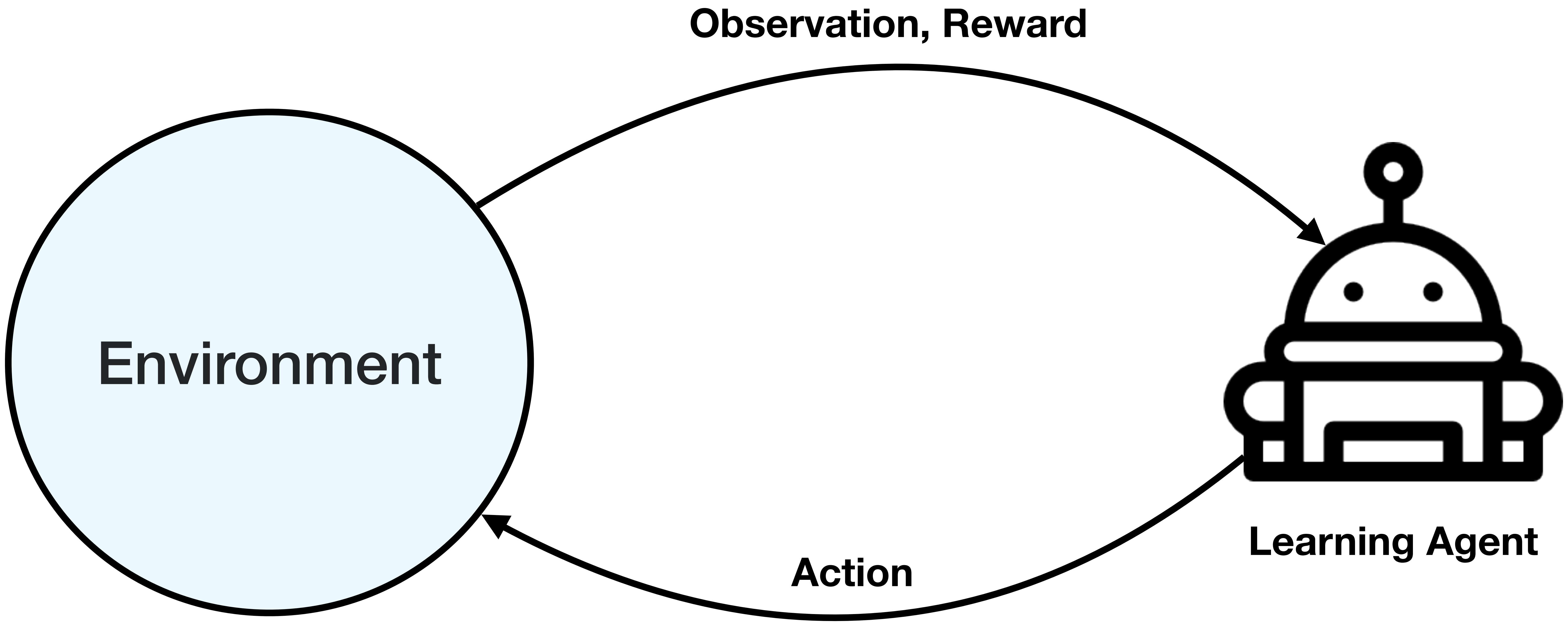
Giuseppe Vietri
University of Minnesota

Borja Balle
Deepmind

Akshay Krishnamurthy
Microsoft Research

Steven Wu
University of Minnesota

Reinforcement Learning

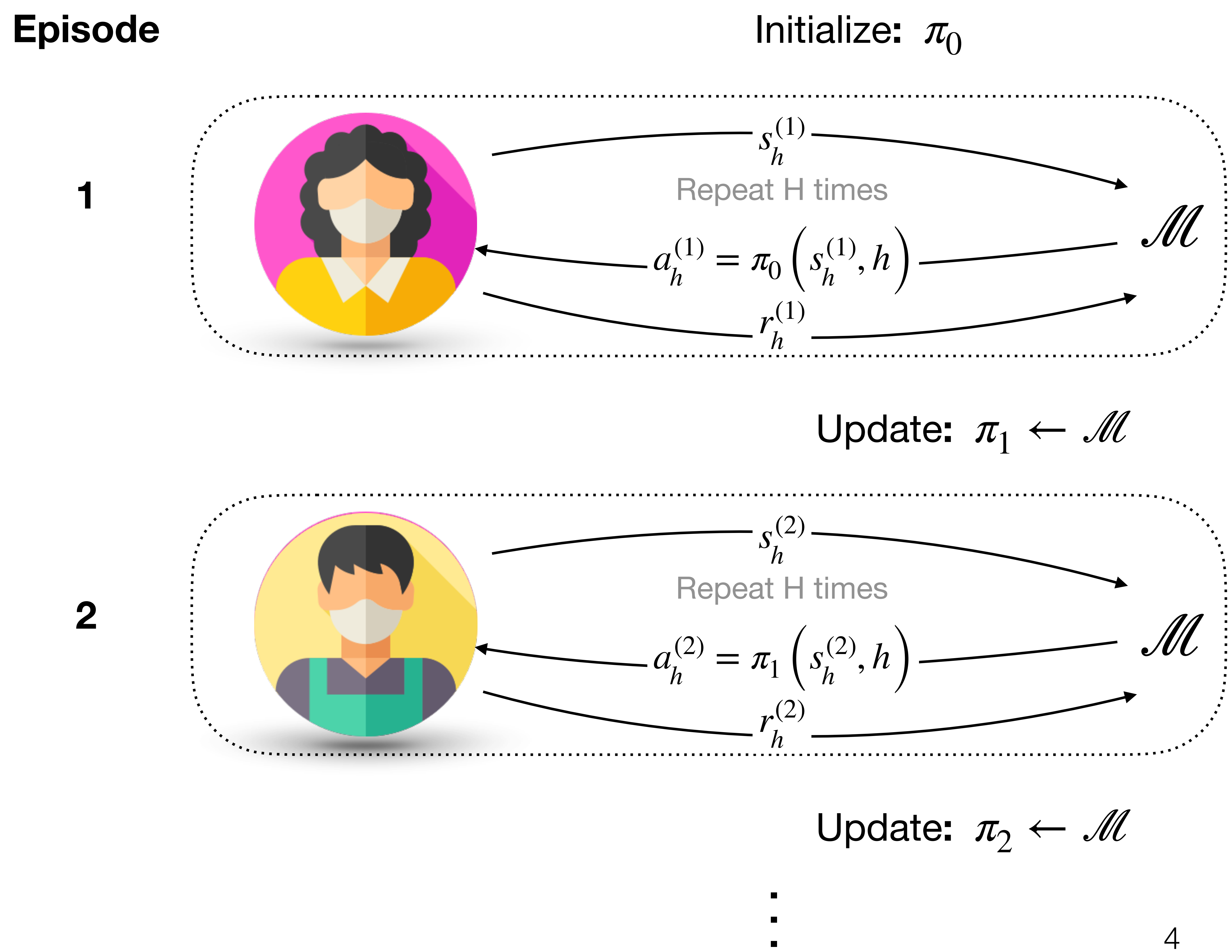


RL in healthcare

- Agent → Provider
- Environment → Patients
- Observations → Symptoms
- Actions → Treatments
- Reward →patient improves
- Privacy of patients?

Episodic RL Protocol

Input: Learning Agent $\mathscr{M}$, User sequence U= (, , …)



S = Number of States.

A = Number of Actions.

H = Time-steps per episode.

Main Results: A Privacy Formulation for RL

Let $\mathscr{M}$ be a RL algorithm.

t^{th} **users**

Input: $U = ($ , , …, , …, $)$

T is the total number of users/episodes.

Output:

$$\mathscr{M}(U) = \left(\begin{pmatrix} a_1^{(1)} \\ \vdots \\ a_H^{(1)} \end{pmatrix}, \begin{pmatrix} a_1^{(2)} \\ \vdots \\ a_H^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} a_1^{(t)} \\ \vdots \\ a_H^{(t)} \end{pmatrix}, \ldots, \begin{pmatrix} a_1^{(T)} \\ \vdots \\ a_H^{(T)} \end{pmatrix} \right)$$

$\mathscr{M}$ must satisfy ϵ-Joint Differential Privacy Under Continual Observation.

Main Results

Private

	PAC	Regret
Upper Bounds	$\widetilde{O}\left(\left(\frac{SAH^4}{\alpha^2}+\frac{S^2AH^4}{\epsilon\alpha}\right)\right)$	$\widetilde{O}\left(H^2\sqrt{SAT}+\frac{S^2AH^4}{\epsilon}\right)$
Lower Bound	$\widetilde{\Omega}\left(\left(\frac{SAH^2}{\alpha^2}+\frac{SAH}{\epsilon\alpha}\right)\right)$	$\widetilde{\Omega}\left(\sqrt{HSAT}+\frac{SAH}{\epsilon}\right)$

Prior work.
non-Private

	PAC	Regret
Upper Bounds	$\widetilde{O}\left(\frac{SAH^4}{\alpha^2}\right)$ [Dann et al. 2017]	$\widetilde{O}\left(\sqrt{SAHT}\right)$ [Azar et al. 2017]
Lower Bounds	$\widetilde{\Omega}\left(\frac{SAH^3}{\alpha^2}\right)$ [Dann et al. 2017]	$\widetilde{\Omega}\left(\sqrt{SAHT}\right)$ [Jaksch et al. 2010]

PUCB: A private RL Algorithm.

- **PUCB** is a private version of the **U**pper-**B**ound the **E**xpected next **V**alue algorithm (**UBEV**) [Dann et al., 2017].
- Compute an optimistic Q-value function using standard batch Q-learning updates with an optimism bonus.
- Keeps track of rewards and dynamics estimates.
- Follows a greedy policy: $\pi_t(s, h) = \arg\max_{a*} Q(s, a*, h)$

Q Learning

Without privacy:

For $t = 1\ldots, T$**:**

$Q \longleftarrow$ OptimismticPlanning $\left(\hat{n}, \hat{r}, \widehat{m}\right)$

$s_1^{(t)} \sim \langle$uniform distribution over states$\rangle$

For $h = 1\ldots, H$**:**

$a_h^{(t)} = \arg\max_{a*} Q(s_h^{(t)}, a*, h)$

$r_h^{(t)} \sim R(s_h^{(t)}, a_h^{(t)}, h), s_{h+1}^{(t)} \sim P(\cdot \mid s_h^{(t)}, a_h^{(t)}, h)$

Increment counters: $\hat{n}\left(s_h^{(t)}, a_h^{(t)}, h\right)$,

$\hat{r}\left(s_h^{(t)}, a_h^{(t)}, h\right)$, $\widehat{m}\left(s_h^{(t)}, a_h^{(t)}, h, s_{h+1}^{(t)}\right)$

With Privacy

For $t = 1\ldots, T$**:**

$Q \longleftarrow$ PrivateOptimismticPlanning $\left(\widetilde{n}, \widetilde{r}, \widetilde{m}\right)$

$s_1^{(t)} \sim \langle$uniform distribution over states$\rangle$

For $h = 1\ldots, H$**:**

$a_h^{(t)} = \arg\max_{a*} Q(s_h^{(t)}, a*, h)$

$r_h^{(t)} \sim R(s_h^{(t)}, a_h^{(t)}, h), s_{h+1}^{(t)} \sim P(\cdot \mid s_h^{(t)}, a_h^{(t)}, h)$

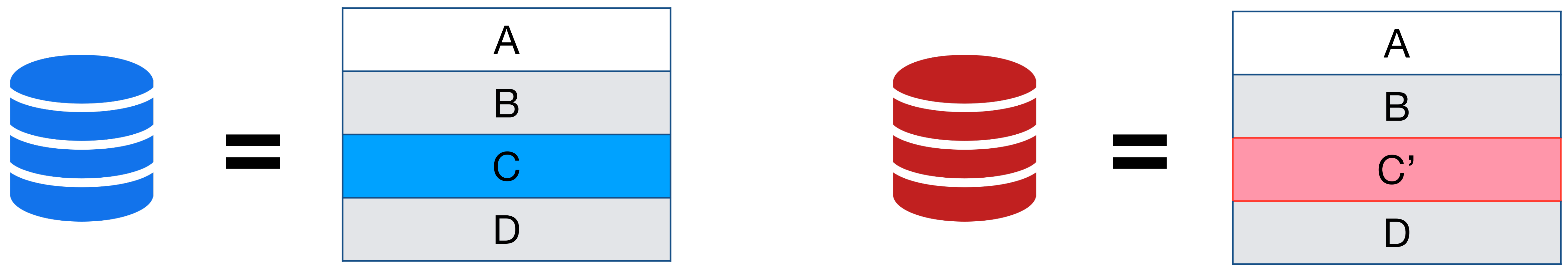
Increment private counters: $\widetilde{n}\left(s_h^{(t)}, a_h^{(t)}, h\right)$,

$\widetilde{r}\left(s_h^{(t)}, a_h^{(t)}, h\right)$, $\widetilde{m}\left(s_h^{(t)}, a_h^{(t)}, h, s_{h+1}^{(t)}\right)$

Standard Differential Privacy (DP)

[Dwork et al., 2006]

Two datasets



are **neighbors** if the are different on only one row.

Definition: Mechanism M satisfies ε-differential privacy if, for all neighboring datasets and for all $r \in \text{range}(M)$

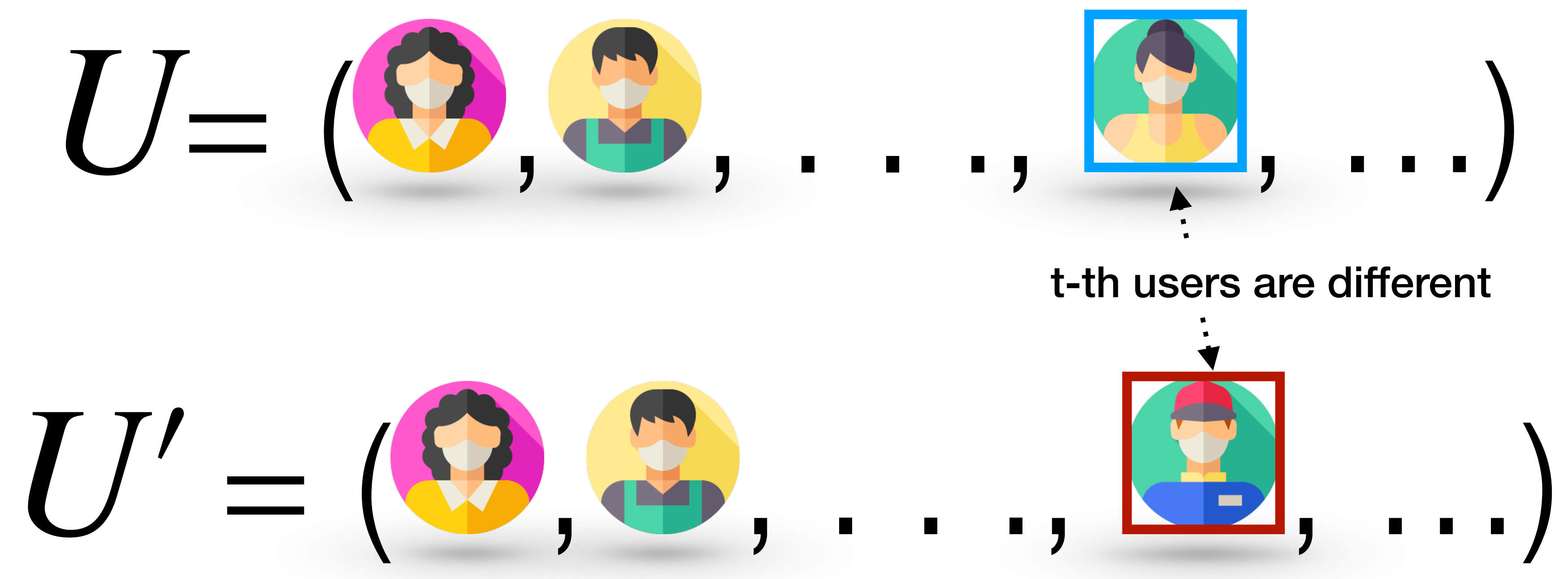
$\Pr[M($  $) = r] \le e^{\varepsilon} \Pr[M($  $) = r]$

Why Is DP not Applicable?

- If the algorithm must satisfy **D**ifferential **P**rivacy then for any two states s, s', it holds that

$$\text{Pr}\left[\pi_t\left(s, h\right) = a\right] \sim \text{Pr}\left[\pi_t\left(s', h\right) = a\right]$$

Joint Differential Privacy Definition



U and U' are *t-neighboring* user sequences

Definition: A mechanism $\mathscr{M}$ is **ε-jointly differentially private** if for all t, all t-neighboring user sequences U, U' and all future events $E \subseteq \mathscr{A}^{H\times[T-1]}$ we have

$$\Pr\left[\mathscr{M}_{-t}(U) \in E\right] \leq e^{\varepsilon}\Pr\left[\mathscr{M}_{-t}(U') \in E\right]$$

Event Counters

Total counters: $2SAH + S^2AH$

Use ***Binary mechanism*** from [Dwork et at., 2010] and [Chan et al., 2011].

If $\widetilde{n}(s, a, h) \longleftarrow$ Binary Mechanism With Privacy Parameter $\frac{\varepsilon}{H}$

The **composition** of all counters satisfies ε-DP [Hsu et al., 2014].

$$\left| \widetilde{n}(s, a, h) - \widehat{n}(s, a, h) \right| \leq \frac{H}{\varepsilon} \log(T)^{5/2} \log(2/\beta) := E_\varepsilon$$

Balancing Exploration/Exploitation with Optimism.

Confidence due to sampling error [Dann et al. 2017]

Confidence term due to privacy. [This work]

Without privacy: $\widehat{Q}^{+}(s,a,h) = \widehat{Q}(s,a,h) + \widehat{\phi}(s,a,h)$

With privacy: $\widetilde{Q}^{+}(s,a,h) = \widetilde{Q}(s,a,h) + \widetilde{\phi}(s,a,h) + \widetilde{\psi}(s,a,h)$

$$\widehat{\phi}(s,a,h) = (1+H)\sqrt{\frac{T/\beta}{\hat{n}(s,a,h)}} \qquad \widetilde{\psi}(s,a,h) = (1+SH)\left(\frac{3E_\varepsilon}{\widetilde{n}(s,a,h)} + \frac{2E_\varepsilon^2}{\widetilde{n}(s,a,h)^2}\right)$$

JDP Proof: Billboard Lemma

Theorem: Algorithm PUCB satisfies ε-joint differential privacy.

- We use the billboard lemma from [Hsu et al., 2016]
- The billboard lemma: An algorithm is JDP if the output sent to each user is a function of the user's private data and a common signal computed using standard differential privacy.
- For example:
  - The output for user t is: $a_h^{(t)} = \arg\max_{a^*} Q(s_h^{(t)}, a^*, h)$

$s_h^{(t)}$ is part of user t private data

The Q-function was computed using ε-differential privacy

PAC Upper Bound Proof: Optimism

$$\widehat{Q}^{+}(s,a,h) = \frac{\widehat{r}(s,a,h) + \sum_{s'} \widetilde{V}_{h+1}(s')\,\widehat{m}(s,a,h,s')}{\widehat{n}(s,a,h)} + \widehat{\phi}(s,a,h)$$

$$\widehat{Q}^{+}(s,a,h) \le \frac{\widetilde{r}(s,a,h) + E_\varepsilon + \sum_{s'} \widetilde{V}_{h+1}(s')(\widetilde{m}(s,a,h,s') + E_\varepsilon)}{\widetilde{n}(s,a,h) - E_\varepsilon} + \widehat{\phi}(s,a,h)$$

Case 1: If $\widetilde{n}(s,a,h) \ge 2E_\varepsilon$ the following holds: $\frac{1}{\widetilde{n}(s,a,h) - E_\varepsilon} \le \left(\frac{1}{\widetilde{n}(s,a,h)} + \frac{2E_\varepsilon}{\widetilde{n}(s,a,h)^2}\right)$

$$\widehat{Q}^{+}(s,a,h) \le \widetilde{Q}(s,a,h) + \left(\frac{1}{\widetilde{n}(s,a,h)} + \frac{\widetilde{2E_\varepsilon}}{\widetilde{n}(s,a,h)^2}\right)(1+SH)E_\varepsilon + \widetilde{\phi}(s,a,h) = \widetilde{Q}^{+}(s,a,h)$$

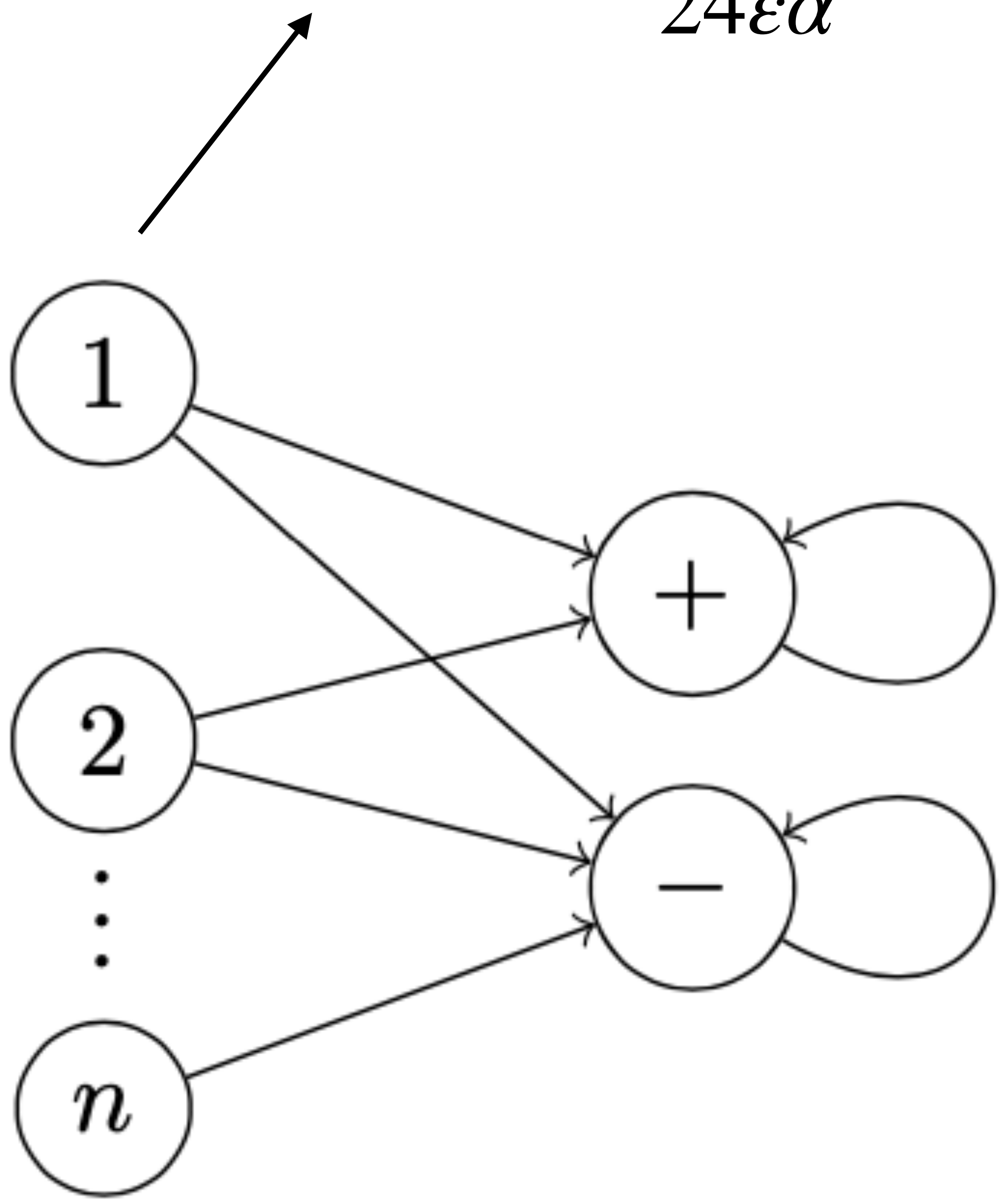
$\widetilde{\psi}(s,a,h)$

Case 2: If $\widetilde{n}(s,a,h) < 2E_\varepsilon$ then we make $\widetilde{Q}^{+}(s,a,h) = H$

PAC Lower Bound Proof

1. Lower bound for private-best-arm-identification problem.
    1. $\widetilde{\Omega}\left(\frac{A}{\varepsilon\alpha}\ln\frac{1}{4\beta}\right)$
2. We consider a simpler: *Public Initial State Setting.*
    1. Each initial states $\{1,\ldots,n\}$ is a private best-arm-identification MAB problem.
3. Therefore learner must make a total of at least $\frac{SAH}{24\varepsilon\alpha}\ln\frac{1}{4\beta}$ mistakes.
4. ε-JDP $\implies$ ε-JDP in the public initial state setting.

 # mistakes = $\frac{AH}{24\varepsilon\alpha}\ln\frac{1}{4\beta}$



Conclusion

- Introduced a meaningful formulation of privacy for RL.
- A private optimism based algorithm with PAC and regret Guarantees.
- First analysis of lower bounds for private RL.