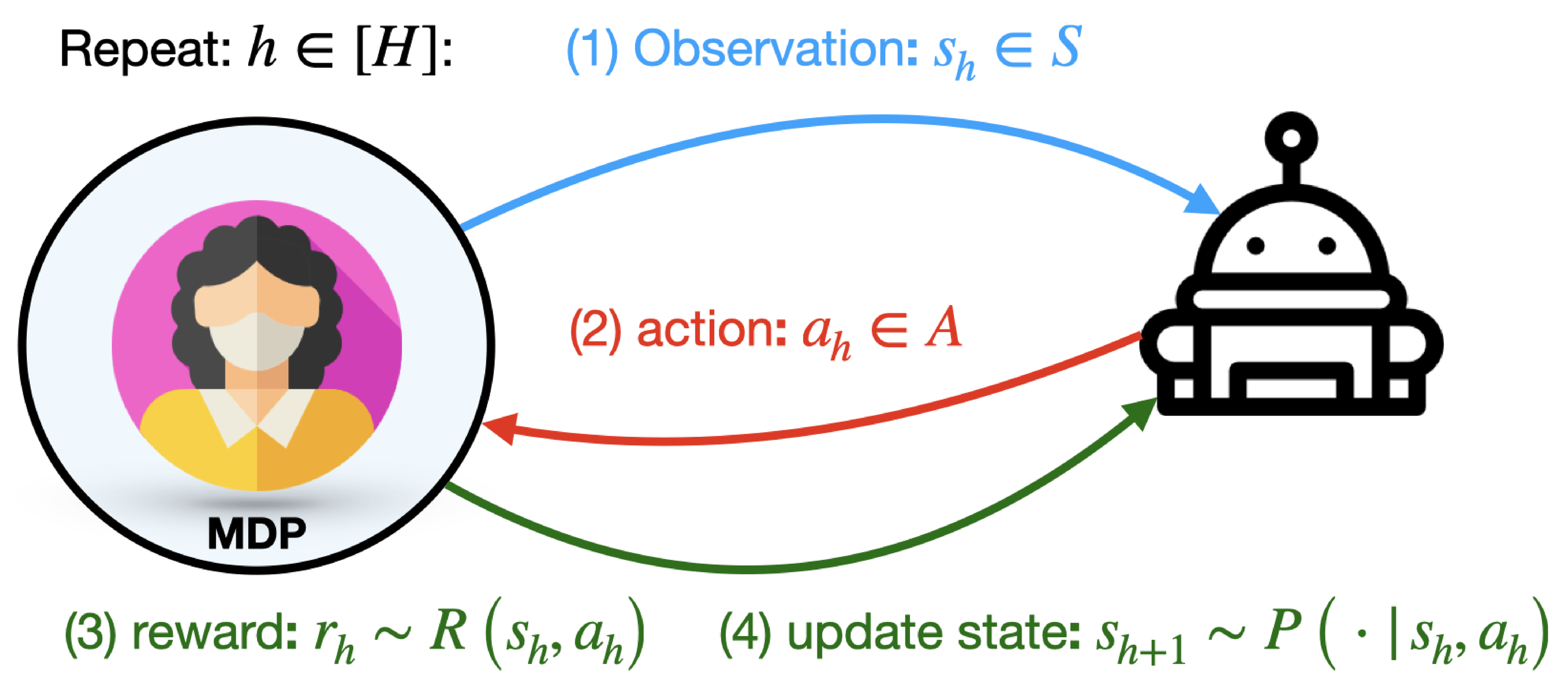
Private Reinforcement Learning with PAC and Regret Guarantees

Giuseppe Vietri[1], Borja Balle[2], Akshay Krishnamurthy[3], Zhiwei Steven Wu[4]
[1]University of Minnesota, [2]Deepmind, [3]Microsoft Research, [4]Carnegie Mellon University

Online Learning MDP

1. An agent interacts with a fix horizon MDP over a sequence of episodes:
2. $M = (S, A, R, P, p_0, H)$ // Environment
3. **Privacy in RL**: we assume that sequence of states and rewards is sensitive data belonging to users.



Contributions

1. Definition of Joint Differential Privacy (**JDP**) in Reinforcement Learning (**RL**).
2. Almost optimal ε-**JDP** **RL** algorithm with bounded *sample complexity* and bounded *regret*.
3. Lower bound for any **RL** algorithm that satisfies **JDP**.

Results

1. Let α be a target accuracy and ε the privacy parameter.
   - *How many episodes does it take to learn an α-optimal policy?*
   - A smaller ε means the algorithm is more private.
2. Upper bound sample complexity:

$$\widetilde{O}\left(\frac{SAH^4}{\alpha^2} + \frac{S^2AH^4}{\epsilon\alpha}\right) \tag{1}$$

3. Lower bound sample complexity:

$$\widetilde{\Omega}\left(\frac{SAH^2}{\alpha^2} + \frac{SAH}{\epsilon\alpha}\right) \tag{2}$$

Privacy Definition

- We represent a user as a tree of depth H encoding all possible state and reward paths.
- Let $\mathcal{M}$ be a **RL** algorithm:
  - The INPUT is a sequence of T users.
  - The OUTPUT is a sequence of T actions and a final policy π_T.

$$\mathcal{M}(\ldots) = \left(a^{(1)}, \ldots, a^{(T)}, \pi_T\right) \qquad a^{(t)} = \left(a_1^{(t)}, \ldots, a_h^{(t)}\right)$$

- U and U' are t-neighboring if they differ only on user at position t:

$$U = (\ldots, \ldots, \ldots),\ U' = (\ldots, \ldots, \ldots)$$

Definition (KPRU14)**.** A mechanism $\mathcal{M}$ is ε-**JDP** if for all t, all t-neighboring user sequences U, U' and all events $E \subseteq A^{H\times[T-1]} \times \Pi$ we have

$$\Pr\left[\mathcal{M}_{-t}(U) \in E\right] \le e^{\varepsilon} \Pr\left[\mathcal{M}_{-t}(U') \in E\right] \tag{3}$$

PUCB

- **PUCB** is a **JDP** version of UBEV [DLB 2017].
- Non-private event counters: $\widehat{n}_t(s,a,h), \widehat{m}_t(s,a,h,s'), \widehat{r}_t(s,a,h)$.
- Private event counters: $\widetilde{n}_t(s,a,h), \widetilde{m}_t(s,a,h,s'), \widetilde{r}_t(s,a,h)$.
- Use Binary mechanism (**BM**) from [Dwork et at., 2010] and [Chan et al., 2011]. For any $t \in [T]$:

$$|\widehat{n}_t(s,a,h) - \widetilde{n}_t(s,a,h)| \le \tfrac{H}{\varepsilon}\log(T)^{5/2}\log(SAH/\beta) := E_\varepsilon \tag{4}$$

- Balance exploration/exploitation with optimism:
  - Compute $\widetilde{Q}_t^+$ with **DP**:

$$\widetilde{Q}_t^+(s,a,h) = (\text{Reward}) + (\text{Future reward}) + (\text{Bonus}) \tag{5}$$

  - Greedy policy: $\pi_t(s,h) = \arg\max_{a^*} \widetilde{Q}_t^+(s,a^*,h)$.

Private Optimistic Q-function

$$\widetilde{Q}^+(s,a,h) = \frac{\underbrace{\widetilde{r}(s,a,h)}_{\text{Immediate reward}} + \underbrace{\sum_{s'} \widetilde{V}_{h+1}(s')\widetilde{m}(s,a,h,s')}_{\text{Expected future reward}}}{\widetilde{n}(s,a,h)} + \underbrace{\widetilde{\mathrm{conf}}(s,a,h)}_{\text{Bonus term}}$$

$$\widetilde{\mathrm{conf}}(s,a,h) = \widetilde{\phi}(s,a,h) + \widetilde{\psi}(s,a,h)$$

$$\widetilde{\phi}(s,a,h) = (1+H)\sqrt{\frac{T/\beta}{\widetilde{n}(s,a,h)}}$$ sampling error [Dann et al. 2017]

$$\widetilde{\psi}(s,a,h) = (1+SH)\left(\frac{3E_\varepsilon}{\widetilde{n}(s,a,h)} + \frac{2E_\varepsilon^2}{\widetilde{n}(s,a,h)^2}\right)$$ privacy error [This work].

PUCB Sample Complexity Analysis

- Construct $\widetilde{Q}_t^+$ with DP counters and show that:

$$\widetilde{Q}_t^+(s,a,h) \ge Q^*(s,a,h) \tag{6}$$

- Optimality gap decomposition:

$$V^* - V^{\pi_t} := \Delta_t \le \sum_{(s,a,h)\in(S,A,[H])} w_t(s,a,h)\widetilde{\mathrm{conf}}_t(s,a,h) \tag{7}$$

- Bounding number of episodes where $\Delta_t > \alpha$.

Lower Bound Analysis

- Consider hard-MDP construction.
- Lower bound for **DP** best-arm-identification:
- We show that finding the the α-optimal arm takes $\widetilde{\Omega}\left(\frac{A}{\varepsilon\alpha}\ln\frac{1}{4\beta}\right)$ tries.
- We consider a Public Initial State (**PIS**) Setting:
  - And do a reduction to **DP** best-arm-identification.
  - We show that the learner makes at least $\frac{SAH}{24\varepsilon\alpha}\ln\frac{1}{4\beta}$ mistakes in the PIS setting.
  - If algorithm satisfies ε-**JDP** $\implies$ satisfies ε-**JDP** in the PIS setting.

Acknowledgements

Giuseppe Vietri has been supported by the GAANN fellowship from the U.S. Department of Education. We want to thank Matthew Joseph, whose comments improved our definition of joint-differential-privacy.