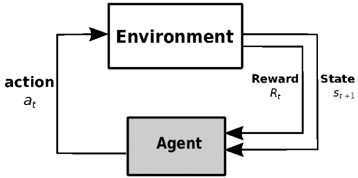
From regret to PAC RL

Emilie Kaufmann (CNRS, Univ. Lille, Inria Scool)

RL Theory Workshop, CWI, June 2025

Regret versus pure exploration

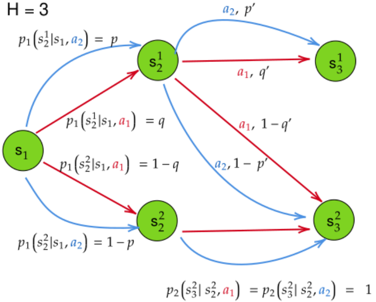
RL setup : an agent interacts with an environement (MDP)



Several Performance measures :

1. the agent should *adopt* a good behavior during learning
   ➜ maximize the total rewards (*regret minimization*)
2. the agent should *learn* a good behavior, regarless of rewards gathered
   ➜ Pure Exploration

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, s_1)$$



Value function

For a policy $\pi = \{\pi_h\}_{h\in[H]}$ for a reward function $r : [H] \times \mathcal{S} \times \mathcal{A} \to [0,1]$

$$V_h^\pi(s; r) = \mathbb{E}^\pi\left[\sum_{\ell=h}^{H} r_\ell(S_\ell, A_\ell) \middle| S_h = s\right] \qquad \begin{array}{rcl} A_\ell & \sim & \pi_\ell(S_\ell) \\ S_{\ell+1} & \sim & p_\ell(\cdot|S_\ell, A_\ell) \end{array}$$

Online episodic algorithm

In each episode $t = 1, 2, \dots,$ the agent

- selects an exploration policy π^t based on past data $\mathcal{D}_{t-1}$
- collects an episode under this policy

$$\mathcal{D}_t = \mathcal{D}_{t-1} \cup \left\{\left(s_1^t, a_1^t, s_2^t, a_2^t, \dots, s_H^t, a_H^t\right)\right\}$$

  where $s_1^t = s_1$, $a_h^t \sim \pi_h^t(s_h^t)$ and $s_{h+1}^t \sim p_h(\cdot | s_h^t, a_h^t)$
- can decide to stop exploration $\rightarrow$ adaptive stopping time τ
- if so, can output a prediction, e.g. a good policy $\widehat{\pi}$

Goal : make a Probaby Approximately Correct (PAC) prediction
Performance metric : Sample Complexity τ (number of episodes needed)

Best Policy Identification (BPI)

➜ Learn the optimal policy for a known reward function r

[Fiechter, 1994]

Algorithm :

- exploration policy π^t
- stopping rule τ
- $\widehat{\pi}$: guess for a good policy

(ε, δ)-PAC algorithm for Best Policy Identification

$$\mathbb{P}\left(V_1^\star(s_1; r) - V_1^{\widehat{\pi}}(s_1; r) \leq \varepsilon\right) \geq 1 - \delta$$

Reward Free Exploration (RFE)

➜ Learn the optimal policy for **any** reward function r given afterwards

[Jin et al., 2020]

Algorithm :

- exploration policy π^t
- stopping rule τ
- for any $r = (r_h(s,a)) \in [0,1]^{HSA}$, guess $\widehat{\pi}_r$ for a good policy

(ε, δ)-PAC algorithm for Reward-Free Exploration

$$\mathbb{P}\left(\text{for any } r \in \mathcal{B}, V_1^{\star}(s_1; r) - V_1^{\widehat{\pi}_r}(s_1; r) \leq \varepsilon\right) \geq 1 - \delta$$

Reward Free Exploration (RFE)

➜ Learn the optimal policy for **any** reward function r given afterwards

[Jin et al., 2020]

Algorithm :

- exploration policy π^t
- stopping rule τ
- for any $r \in \mathcal{B}$, guess $\widehat{\pi}_r$ for a good policy

(ε, δ)-PAC algorithm for Reward-Free Exploration

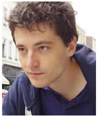
$$\mathbb{P}\left(\text{for any } r \in \mathcal{B}, V_1^\star(s_1; r) - V_1^{\widehat{\pi}_r}(s_1; r) \leq \varepsilon\right) \geq 1 - \delta$$

Outline

1. Minimax Sample Complexity : Optimism is Enough







Fast Active Learning for Pure Exploration in RL, ICML 2021
Adaptive Reward Free Exploration, ALT 2021

Optimistic RL algorithm

Bellman equation

$$\pi_h^\star = \text{greedy}(Q_h^\star) \quad \text{where} \quad Q_h^\star(s,a) = r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_b Q_{h+1}^\star(s',b)$$

Optimism : $\pi_h^{t+1} = \text{greedy}(\overline{Q}_h^t)$ where

$$\overline{Q}_h^t(s,a) = \max_{p \in \mathcal{M}_t} \left[r_h(s,a) + \sum_{s'} p_h(s'|s,a) \max_b \overline{Q}_{h+1}^t(s',b) \right]$$

where $\mathcal{M}_t$ is a set of plausible MDPs.

UCB-VI

UCB-VI style algorithm

$\pi_h^{t+1} = \text{greedy}\left(\overline{Q}_h^t\right)$ for the optimistic Q-function

$$\overline{Q}_h^t(s,a) = \left[r_h(s,a) + \mathrm{B}_h^t(s,a) + \sum_{s' \in \mathcal{S}} \hat{p}_h^t(s'|s,a) \max_b \overline{V}_{h+1}^t(s')\right] \wedge (H-h)$$

$$\overline{V}_h^t(s) = \max_b \overline{Q}_h^t(s,b).$$

Different exploration bonuses $\mathrm{B}_h^t(s,a)$ yield different guarantees

- Hoeffding bonuses $\mathrm{B}_h^t(s,a) \simeq \sqrt{\frac{\log(SAH/\delta)+S\log(n_h^t(s,a))}{n_h^t(s,a)}}$ ("*UCRL*")
- Bernstein bonuses (more complex) (*UCB-VI* [Azar et al., 2017])

$n_h^t(s,a)$: number of visits of (s,a) in step h in the first t episodes

Regret and PAC guarantees

The (pseudo)-regret of an episodic RL algorithm $\pi = (\pi^t)_{t\in\mathbb{N}}$ is

$$\mathcal{R}_T(\pi) = \sum_{t=1}^{T} \left[V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t) \right].$$

Regret of UCB-VI [Azar et al., 2017]

For appropriately chosen bonuses (depending on δ) UCB-VI satisfies

$$\mathbb{P}\left(\mathcal{R}_T(\pi) = \mathcal{O}\left(\sqrt{H^3SAT}\right)\right) \geq 1-\delta$$

which is minimax optimal in time-inhomogeneous MDPs.
[Domingues et al., 2021]

Regret and PAC guarantees

The (pseudo)-regret of an episodic RL algorithm $\pi = (\pi^t)_{t\in\mathbb{N}}$ is

$$\mathcal{R}_T(\pi) = \sum_{t=1}^{T} \left[V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t) \right].$$

Regret to PAC conversion [Jin et al., 2018]

Running UCB-VI for $T = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2\delta^2}\right)$ and outputting

$$\widehat{\pi} = \pi^N \quad \text{where} \quad N \sim \mathcal{U}(\{1,\dots,T\})$$

yields an (ε,δ)-PAC identification of the optimal policy.

Minimax lower bound : for any (ε,δ)-PAC BPI algorithm, there exists an MDP for which $\mathbb{E}[\tau] \geq c\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$ [Domingues et al., 2021]

Regret and PAC guarantees

The (pseudo)-regret of an episodic RL algorithm $\pi = (\pi^t)_{t\in\mathbb{N}}$ is

$$\mathcal{R}_T(\pi) = \sum_{t=1}^{T} \left[V_1^\star(s_1^t) - V_1^{\pi^t}(s_1^t) \right].$$

Regret to PAC conversion [Jin et al., 2018]

Running UCB-VI for $T = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2 \delta^2}\right)$ and outputting

$$\widehat{\pi} = \pi^N \quad \text{where} \quad N \sim \mathcal{U}(\{1,\dots,T\})$$

yields an (ε,δ)-PAC identification of the optimal policy.

Minimax lower bound : for any (ε,δ)-PAC BPI algorithm, there exists an MDP for which $\mathbb{E}[\tau] \geq c\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)$ [Domingues et al., 2021]

Minimax Optimal BPI Algorithm

Solution : UCB-VI using instead an adaptive stopping rule

BPI-UCRL [Kaufmann et al., 2021]

Using UCB-VI with Hoeffding bonuses together with

$$\tau = \inf\left\{t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \varepsilon\right\} \quad \hat{\pi} = \text{greedy}(\underline{Q}_1^\tau)$$

yields an (ε, δ)-PAC algorithm with $\mathbb{P}\left(\tau = \widetilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)\right) \geq 1 - \delta$.

➜ using Bernstein bonuses and a more sophisticated stopping rule yields a $\widetilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ sample complexity [Ménard et al., 2021]

Minimax Optimal BPI Algorithm

Solution : UCB-VI using instead an adaptive stopping rule

BPI-UCRL [Kaufmann et al., 2021]

Using UCB-VI with Hoeffding bonuses together with

$$\tau = \inf\left\{t \in \mathbb{N} : \overline{V}_1^t(s_1) - \underline{V}_1^t(s_1) \leq \varepsilon\right\} \quad \hat{\pi} = \text{greedy}(\underline{Q}_1^\tau)$$

yields an (ε, δ)-PAC algorithm with $\mathbb{P}\left(\tau = \widetilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)\right) \geq 1 - \delta$.

➜ using Bernstein bonuses and a more sophisticated stopping rule yields a $\widetilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ sample complexity [Ménard et al., 2021]

How about Reward Free Exploration ?

RF-UCRL [Kaufmann et al., 2021]

$$\pi_h^{t+1} = \text{greedy}\left(\overline{Q}_h^t\right) \text{ for}$$

$$\overline{Q}_h^t(s,a) = \left[\cancel{r_h^t(s,a)} + \mathrm{B}_h^t(s,a) + \sum_{s' \in \mathcal{S}} \hat{p}_h^t(s'|s,a) \max_b \overline{V}_{h+1}^t(s')\right] \wedge (H-h)$$

$$\overline{V}^t{}_h(s) = \max_b \overline{Q}_h^t(s,b).$$

Why does it work ? It greedily reduces the estimation error of the value of any policy for any reward function :

$$\forall \pi, \forall r, \forall h, s, a, t \quad |\hat{Q}_h^{t,\pi}(s,a;r) - Q_h^\pi(s,a;r)| \leq \overline{E}_h^t(s,a)$$

holds with high probability for some Hoeffding-type bonus B

How about Reward Free Exploration ?

RF-UCRL [Kaufmann et al., 2021]

$$\pi_h^{t+1} = \text{greedy}\left(\overline{E}_h^t\right) \text{ for}$$

$$\overline{E}_h^t(s,a) = \left[\cancel{r_h^t(s,a)} + \mathrm{B}_h^t(s,a) + \sum_{s' \in \mathcal{S}} \hat{p}_h^t(s'|s,a) \max_b \overline{V}_{h+1}^t(s')\right] \wedge (H-h)$$

$$\overline{V}^t{}_h(s) = \max_b \overline{E}_h^t(s,b).$$

Why does it work ? It greedily reduces the estimation error of the value of any policy for any reward function :

$$\forall \pi, \forall r, \forall h, s, a, t \quad |\hat{Q}_h^{t,\pi}(s,a;r) - Q_h^{\pi}(s,a;r)| \leq \overline{E}_h^t(s,a)$$

holds with high probability for some Hoeffding-type bonus B

How about Reward Free Exploration ?

Reward-Free UCRL

- **exploration policy** : π^{t+1} is the greedy policy wrt $\overline{E}^t(s,a)$:

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a\in\mathcal{A}} \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf\left\{t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

For a given reward function r

$$\begin{aligned} V_1^\star(s_1) - V_1^{\widehat{\pi}_r}(s_1) &= V_1^{\pi^\star}(s_1) - \widehat{V}_1^{\tau,\pi^\star}(s_1) + \underbrace{\widehat{V}_1^{\tau,\pi^\star}(s_1) - \widehat{V}_1^{\tau,\widehat{\pi}_r}(s_1)}_{\leq 0} + \widehat{V}_1^{\tau,\widehat{\pi}_r}(s_1) - V_1^{\widehat{\pi}_r}(s_1) \\ &\leq 2\max_a \overline{E}_1^\tau(s_1,a) \\ &\leq \varepsilon \end{aligned}$$

How about Reward Free Exploration ?

Reward-Free UCRL

- **exploration policy** : π^{t+1} is the greedy policy wrt $\overline{E}^t(s,a)$:

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a \in \mathcal{A}} \ \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf\left\{t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$
- **prediction** : $\forall r, \ \widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

For a given reward function r

$$\begin{aligned} V_1^\star(s_1) - V_1^{\widehat{\pi}_r}(s_1) &= V_1^{\pi^\star}(s_1) - \widehat{V}_1^{\tau,\pi^\star}(s_1) + \underbrace{\widehat{V}_1^{\tau,\pi^\star}(s_1) - \widehat{V}_1^{\tau,\widehat{\pi}_r}(s_1)}_{\leq 0} + \widehat{V}_1^{\tau,\widehat{\pi}_r}(s_1) - V_1^{\widehat{\pi}_r}(s_1) \\ &\leq 2\max_a \overline{E}_1^\tau(s_1, a) \\ &\leq \varepsilon \end{aligned}$$

How about Reward Free Exploration ?

Reward-Free UCRL

- **exploration policy** : π^{t+1} is the greedy policy wrt $\overline{E}^t(s,a)$:

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a\in\mathcal{A}} \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf\left\{t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2\right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

Theorem [Kaufmann et al. 2020]

RF-UCRL is (ε,δ)-PAC for Reward-Free Exploration and

$$\mathbb{P}\left(\tau^{\text{RF-UCRL}} = \tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\left[\log\left(\frac{1}{\delta}\right) + S\right]\right)\right) \geq 1-\delta.$$

How about Reward Free Exploration ?

Reward-Free UCRL

- **exploration policy** : π^{t+1} is the greedy policy wrt $\overline{E}^t(s,a)$:

$$\forall s \in \mathcal{S}, \forall h \in [h], \quad \pi_h^{t+1}(s) = \text{argmax}_{a \in \mathcal{A}} \overline{E}_h^t(s,a).$$

- **stopping rule** : $\tau = \inf \left\{ t \in \mathbb{N} : \overline{E}_1^t(s_1, \pi_1^{t+1}(s_1)) \leq \varepsilon/2 \right\}$
- **prediction** : $\forall r$, $\widehat{\pi}_r = \pi^\star(\hat{P}^\tau, r)$

➜ To get a near-optimal $\widetilde{O}\left(\frac{SAH^3}{\varepsilon^2}(\log(1/\delta) + S)\right)$ sample complexity the algorithm structure and bonus type has to be changed a bit [Ménard et al., 2021]

Summary

UCB-VI is (almost) enough to get minimax optimal sample complexity for both Best Policy Identification and Reward Free Exploration

➜ How about instance-dependent results ?

Ouline

2. Towards Instance Optimality

Active Coverage for PAC RL, COLT 2023
Near Instance-Optimal PAC RL for Deterministic MDPs, NeurIPS 2022
Optimistic PAC RL : the Instance-Dependent View, ALT 2022

Instance dependent results

Goal

Design (ε, δ)-PAC algorithms that adapt to the difficulty of each specific MDP $\mathcal{M}$ and get

$$\tau_\delta = \mathcal{O}\left(C_\varepsilon(\mathcal{M}) \log\left(1/\delta\right)\right)$$

where $C_\varepsilon(\mathcal{M})$ is some appropriate complexity term.

Reward Free Exploration : Given the worse-case nature of the problem, is it at all possible to get

$$\mathcal{C}_\varepsilon(\mathcal{M}) < \frac{SAH^3}{\varepsilon^2} \; ?$$

Instance dependent results

Goal

Design (ε, δ)-PAC algorithms that adapt to the difficulty of each specific MDP $\mathcal{M}$ and get

$$\tau_\delta = \mathcal{O}\left(C_\varepsilon(\mathcal{M}) \log\left(1/\delta\right)\right)$$

where $C_\varepsilon(\mathcal{M})$ is some appropriate complexity term.

Best Policy Identification :

- EPRL [Tirinzoni et al., 2022] for deterministic MDPs
- MOCA [Wagenmaker et al., 2022] (gap-visitation complexity)
- PEDEL [Wagenmaker and Jamieson, 2022]

Feature different complexity measures, and some mechanisms to visit certain triplets (h, s, a) proportionally to some instance-dependent quantity ("gap")

A Coverage Problem

Let $c : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ be a target function.

δ-correct c-coverage

An algorithm $(\pi^t)_{t\in\mathbb{N}}$ is a δ-correct c-coverage if it interacts with $\mathcal{M}$ and return a dataset $\mathcal{D}_t$ such that

$$\mathbb{P}\left(\exists t \geq 1,\ \forall (h,s,a),\ n_h^t(s,a) \geq c_h(s,a)\right) \geq 1 - \delta.$$

where $n_h^t(s,a)$ is the number of visits of (h,s,a) in $\mathcal{D}_t$

Sample complexity :

$$\tau = \inf\left\{t \in \mathbb{N} : \forall h,s,a, n_h^t(s,a) \geq c_h(s,a)\right\}$$

Lower bound

Theorem [Al Marjani et al., 2023]

For any target function c and $\delta \in [0, 1)$, the stopping time τ of any δ-correct c-coverage algorithm satisfies $\mathbb{E}[\tau] \geq (1 - \delta)\varphi^\star(c)$, where

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(s,a,h) \in \mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)},$$

with $\mathcal{X} := \{(h, s, a) : c_h(s, a) > 0\}$.

Intuition : $\frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$ is the expected number of episodes needed before getting $c_h(s, a)$ visits from (h, s, a) using exploration policy π_{exp}.

Insight on $\varphi^\star$

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

We prove the following bounds :

$$\max_h \sum_{s,a} c_h(s,a) \leq \varphi^\star(c) \leq \sum_h \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{s,a} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)} \leq \sum_{h,s,a} \frac{c_h(s,a)}{\max_\pi p_h^\pi(s,a)}$$

→ featured in the gap-visitation complexity in the sample complexity bound obtained for a BPI algorithm, MOCA [Wagenmaker et al., 2022]

Insight on $\varphi^\star$

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

We prove the following bounds :

$$\max_h \sum_{s,a} c_h(s,a) \leq \varphi^\star(c) \leq \sum_h \inf_{\pi_{\text{exp}}\in\Pi_S} \max_{s,a} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)} \leq \sum_{h,s,a} \frac{c_h(s,a)}{\max_\pi p_h^\pi(s,a)}$$

➜ featured in the gap-visitation complexity in the sample complexity bound obtained for a BPI algorithm, MOCA [Wagenmaker et al., 2022]

Towards a coverage algorithm

$$\varphi^\star(c) = \inf_{\pi_{\exp}\in\Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\exp}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\exp}\in\Pi_S} \min_{(s,a,h)\in\mathcal{X}} \frac{p_h^{\pi_{\exp}}(s,a)}{c_h(s,a)}$$

$$= \sup_{\pi_{\exp}\in\Pi_S} \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)}$$

$=$ value of a game !

where $\Delta_{\mathcal{X}}$ is the simplex over $\mathcal{X}$.

Towards a coverage algorithm

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\text{exp}} \in \Pi_S} \min_{(s,a,h)\in\mathcal{X}} \frac{p_h^{\pi_{\text{exp}}}(s,a)}{c_h(s,a)}$$

$$= \sup_{\pi_{\text{exp}} \in \Pi_S} \inf_{\lambda \in \Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\text{exp}}}(s,a)\lambda_h(s,a)}{c_h(s,a)}$$

$$= \text{value of a game!}$$

where $\Delta_{\mathcal{X}}$ is the simplex over $\mathcal{X}$.

Towards a coverage algorithm

$$\varphi^\star(c) = \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{(h,s,a)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$$

with $\mathcal{X} = \{(h,s,a) : c_h(s,a) > 0\}$

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\text{exp}} \in \Pi_S} \min_{(s,a,h)\in\mathcal{X}} \frac{p_h^{\pi_{\text{exp}}}(s,a)}{c_h(s,a)}$$

$$= \sup_{\pi_{\text{exp}} \in \Pi_S} \inf_{\lambda \in \Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\text{exp}}}(s,a)\lambda_h(s,a)}{c_h(s,a)}$$

$$= \text{value of a game!}$$

where $\Delta_{\mathcal{X}}$ is the simplex over $\mathcal{X}$.

Principle

$$\frac{1}{\varphi^\star(c)} = \sup_{\pi_{\exp}\in\Pi_S} \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)}$$

- $\sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} = V^{\pi_{\exp}}(s_1;\widetilde{r})$
  value function for the reward function $\widetilde{r}_h(s,a) = \frac{\lambda_h(s,a)}{c_h(s,a)}$
- $\sum_{h,s,a} \frac{p_h^{\pi_{\exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} = \lambda^\top\left(p^{\pi_{\exp}}/c\right)$
  linear loss function

⚠ unknown MDP : $V^{\pi_{\exp}}$ and $p^{\pi_{\exp}}$ cannot be computed

➜ use online learners !

[Degenne et al., 2019, Zahavy et al., 2021, Tiapkin et al., 2023]

CovGame

Algorithm 1 (Simplified) CovGame

Input : target function c, risk δ.
Adversarial RL algorithm $\mathcal{A}^{\Pi}$, Online learner $\mathcal{A}^{\lambda}$.
Initialize weights $\lambda_h^1(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X})/|\mathcal{X}|$ for all h, s, a
for $t = 1, 2, \dots$ **do**
    Define reward function $\tilde{r}_h^t(s,a) = \frac{\lambda_h^t(s,a)}{c_h(s,a)} \mathbb{1}((h,s,a) \in \mathcal{X})$
    Feed $\mathcal{A}^{\Pi}$ with $\tilde{r}^t$, confidence $\delta/2$ and get exploration policy π^t
    Play π^t and observe trajectory $\mathcal{H}_t := \{(s_h^t, a_h^t, s_{h+1}^t)\}_{1 \leq h \leq H-1}$
    Feed $\mathcal{A}^{\lambda}$ with linear loss ℓ^t and get new weight vector λ^{t+1}

$$\ell^t(\lambda) = \sum_{(h,s,a) \in \mathcal{X}} \lambda_h(s,a) \frac{\mathbb{1}(s_h^t = s, a_h^t = a)}{c_h(s,a)}$$

    If $\forall (h,s,a),\ n_h^t(s,a) \geq c_h(s,a)$: Stop and return $\mathcal{D}_t$

A key component : UCB-VI

Needed for the RL algorithm : If $\mathcal{A}^{\Pi}$ is run with confidence $1-\delta$ on a sequence of rewards $\{\lambda^t\}_{t\geq 1}$ with $\lambda^t \in \mathcal{P}(\mathcal{X})$, w.p. $1-\delta$, for all $T > 1$,

$$\sum_{t=1}^{T} V_1^{\star}(s_1;\lambda^t) - \sum_{t=1}^{T} V_1^{\pi_t}(s_1;\lambda^t) \leq \sqrt{\mathcal{R}_\delta(T)\sum_{t=1}^{T} V_1^{\pi^t}(s_1;\lambda^t) + \mathcal{R}_\delta(T)}$$

- first-order regret bounds
- ... with changing rewards

We prove that UCB-VI with Bernstein bonuses can be used with

$$\mathcal{R}_\delta(T) = cSAH^2\left(\log\left(\frac{2SAH}{\delta}\right) + S\right)\log^2(T)$$

(Full) CovGame

Extra trick to control the range of the rewards for the RL algorithm :

- Cluster triplets (h, s, a) by their order of magnitude

$$\mathcal{Y}_k = \{(h, s, a) : c_h(s, a) \in [c_{\min} 2^k, c_{\min} 2^{k+1}]\}$$

and restart the λ-learner when one of this set has been covered

Theorem [Al Marjani et al., 2023]

Let $m = \lceil \log_2(c_{\max}/c_{\min}) \rceil \vee 1$. With

- $\mathcal{A}^\lambda$: Weighted Majority Forecaster (WMF) with variance-dependent learning rate [Cesa-Bianchi et al., 2005]
- $\mathcal{A}^\pi$: UCB-VI

CovGame satisfies, with probability larger than $1 - \delta$,

$$\tau \leq 64 m \varphi^\star(c) + \widetilde{O}(m \varphi^\star(\mathbb{1}_{\mathcal{X}}) SAH^2(\log(1/\delta) + S)) .$$

UCB-VI for exploration

CovGame can be written

$$\pi^t(s) = \underset{a\in\mathcal{A}}{\mathrm{argmax}}\ \overline{Q}_h^t(s,a;\widetilde{r}_h^t)$$

$$\overline{Q}_h^t(s,a;r) = \left[r_h(s,a) + B_h^t(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a)\max_b \overline{Q}_{h+1}^t(s,b;r)\right] \wedge 1$$

for the a time-varying reward $\widetilde{r}^t \in \Delta_{\mathcal{X}}$

$$\widetilde{r}_h^t(s,a) \propto \exp\left(-\eta_t\left[n_h^t(s,a) - n_h^{m_t}(s,a)\right]\right)\mathbb{1}\left(c_h(s,a) > c_{\min}2^{k_t}\right)$$

Links with other exploration algorithms

- indicator-based rewards are more common in the literature, e.g. $\widetilde{r}_h^t(s,a) = \mathbb{1}\left(n_h^t(s,a) < c_h(s,a)\right)$ for GOSPRL [Tarbouriech et al., 2021a]
- other form of time-varying rewards proposed for entropy exploration [Tiapkin et al., 2023]

CoveGame for Reward Free Exploration

Proportional Coverage

Idea : visit each (h, s, a) in proportion to its maximum reachability :

$$\varphi^\star\left(\left[\max_\pi p_h^\pi(s,a)\right]_{h,s,a}\right)$$

Why Proportional Coverage ?

For RFE, we want a good estimate of the value functions of all policies, for all reward functions :

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{t,\pi}(s_1; r) - \widehat{V}_1^{t,\pi}(s_1; r) \right| \leq \frac{\varepsilon}{2}$$

Concentration inequality [Al Marjani et al., 2023]

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{\pi}(s_1; r) - \widehat{V}_1^{\pi,t}(s_1; r) \right| \leq \sqrt{\beta(t,\delta) \sum_{(h,s,a) \in \mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + \frac{\varepsilon}{4},$$

where $\mathcal{X}_\varepsilon \subseteq \left\{ (h,s,a) : \max_\pi p_h^\pi(s,a) \geq \frac{\varepsilon}{4SH^2} \right\}$

Why Proportional Coverage ?

For RFE, we want a good estimate of the value functions of all policies, for all reward functions :

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{t,\pi}(s_1; r) - \widehat{V}_1^{t,\pi}(s_1; r) \right| \leq \frac{\varepsilon}{2}$$

Concentration inequality [Al Marjani et al., 2023]

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{\pi}(s_1; r) - \widehat{V}_1^{\pi,t}(s_1; r) \right| \leq \sqrt{\beta(t,\delta) \sum_{(h,s,a) \in \mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + \frac{\varepsilon}{4},$$

where $\mathcal{X}_\varepsilon \subseteq \left\{ (h,s,a) : \max_\pi p_h^\pi(s,a) \geq \frac{\varepsilon}{4SH^2} \right\}$

Why Proportional Coverage?

For RFE, we want a good estimate of the value functions of all policies, for all reward functions :

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{t,\pi}(s_1; r) - \widehat{V}_1^{t,\pi}(s_1; r) \right| \leq \frac{\varepsilon}{2}$$

Concentration inequality [Al Marjani et al., 2023]

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{\pi}(s_1; r) - \widehat{V}_1^{\pi,t}(s_1; r) \right| \leq \sqrt{\beta(t,\delta) \sum_{(h,s,a) \in \mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{c \times p_h^\pi(s,a)}} + \frac{\varepsilon}{4},$$

where $\mathcal{X}_\varepsilon \subseteq \left\{ (h,s,a) : \max_\pi p_h^\pi(s,a) \geq \frac{\varepsilon}{4SH^2} \right\}$

Why Proportional Coverage?

For RFE, we want a good estimate of the value functions of all policies, for all reward functions :

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{t,\pi}(s_1; r) - \widehat{V}_1^{t,\pi}(s_1; r) \right| \leq \frac{\varepsilon}{2}$$

Concentration inequality [Al Marjani et al., 2023]

$$\forall \pi \in \Pi_D, \forall r \in \mathcal{B}, \quad \left| V_1^{\pi}(s_1; r) - \widehat{V}_1^{\pi,t}(s_1; r) \right| \leq \sqrt{\beta(t,\delta) \underbrace{\sum_{(h,s,a) \in \mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^{\not{2}}}{c \times \cancel{p_h^\pi(s,a)}}}_{=H/c}} + \frac{\varepsilon}{4},$$

where $\mathcal{X}_\varepsilon \subseteq \left\{ (h,s,a) : \max_\pi p_h^\pi(s,a) \geq \frac{\varepsilon}{4SH^2} \right\}$

Algorithm 2 Proportional Coverage Exploration

Input : Precision ε, Confidence δ.
For each (h, s), run EstimateReachability$((h, s))$ to get confidence intervals $\left[\underline{W}_h(s), \overline{W}_h(s)\right]$ on $\max_\pi p_h^\pi(s)$
Define $\widehat{\mathcal{X}} := \left\{(h, s, a) : \underline{W}_h(s) \geq \frac{\varepsilon}{32SH^2}\right\}$
for $k = 1, \dots$ **do**
    Compute targets $c_h^k(s, a) := 2^k \overline{W}_h(s) \mathbb{1}\left((h, s, a) \in \widehat{\mathcal{X}}\right)$ for all (h, s, a)
    Execute CovGame$\left(c^k,\ \delta/6(k+1)^2\right)$ to get dataset $\mathcal{D}_k$ of d_k episodes
    Update episode count $t_k \leftarrow t_{k-1} + d_k$ and statistics $n_h^k(s, a), \widehat{p}_h^k(.|s, a)$
    if $\sqrt{H\beta(t_k, \delta/3) 2^{4-k}} \leq \varepsilon$ **then** stop and return $\mathcal{D}_k$
end for

Sample complexity

Theorem [Al Marjani et al., 2023]

Proportional Coverage Exploration is (ε, δ)-PAC for reward free exploration. Moreover, with probability at least $1-\delta$ its sample complexity satisfies

$$\tau \leq \widetilde{\mathcal{O}}\Bigg(\left(H^3 \log(1/\delta) + SH^4\right) \underbrace{\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s) \mathbb{1}\left(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right)}_{\mathcal{C}(\mathcal{M},\varepsilon)} + \frac{S^3 A^2 H^5 (\log(1/\delta) + S)}{\varepsilon}\Bigg).$$

$$\mathcal{C}(\mathcal{M}, \varepsilon) \leq \frac{SAH}{\varepsilon^2}$$

Sample complexity

Theorem [Al Marjani et al., 2023]

Proportional Coverage Exploration is (ε, δ)-PAC for reward free exploration. Moreover, with probability at least $1-\delta$ its sample complexity satisfies

$$\tau \leq \widetilde{\mathcal{O}}\Bigg(\left(H^3 \log(1/\delta) + SH^4\right) \underbrace{\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2})}{\varepsilon^2}\right]_{h,s,a}\right)}_{\mathcal{C}(\mathcal{M},\varepsilon)}$$

$$+ \frac{S^3A^2H^5(\log(1/\delta) + S)}{\varepsilon}\Bigg).$$

$$\mathcal{C}(\mathcal{M}, \varepsilon) \leq \frac{SAH}{\varepsilon^2}$$

Beyond worse case

We always have

$$\tau \leq \widetilde{\mathcal{O}}\left(\frac{SAH^4 \log(1/\delta)}{\varepsilon^2} + \frac{S^2AH^5}{\varepsilon^2} + \frac{S^3A^2H^5(\log(1/\delta) + S)}{\varepsilon}\right)$$

- ➜ Only sub-optimal by H factors in the small (ε, δ) regime
- ➜ We exhibit a class of MDPs depending on $\alpha \in (0,1)$ such that $\mathcal{C}(\mathcal{M}, \varepsilon) \leq S^\alpha AH/\varepsilon^2$

$$\tau \leq \widetilde{\mathcal{O}}\left(\frac{S^\alpha AH^4 \log(1/\delta)}{\varepsilon^2} + \frac{S^{1+\alpha}AH^5}{\varepsilon^2} + \frac{S^3A^2H^5(\log(1/\delta) + S)}{\varepsilon}\right)$$

Summary

CovGame a.k.a. UCB-VI with (well designed) changing rewards

- provides a near-optimal solution to the coverage problem
- can be used to obtain a RFE algorithm "better than the worse case"
- can also be used as an ingredient for BPI algorithms

→ optimality ? computational efficiency ?

Ouline

3. Beyond Episodic MDPs

Finding good policies in average-reward MDPs without prior knowledge. NeurIPS 2024

Average Rewards MDPs

$$\text{gain :} \quad g^{\pi}(s) = \lim_{T\to\infty} \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} r_t \middle| s_1 = s\right]$$

Poisson equations :

$$g^{\star} + b^{\star}(s) = \max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) b^{\star}(s') \right\}$$

$$\pi^{\star}(s) = \arg\max_a \left\{ r(s,a) + \sum_{s'} p(s'|s,a) b^{\star}(s') \right\}$$

(in communicating MDPs, $g^{\star}(s) = g^{\star}$)

Best Policy Identification Algorithm

At step $t = 1, 2, \ldots,$ the agent

- selects an action a_t in its current state s_t based on past observations
- observes $s' \sim p(\cdot|s_t, a_t)$ and
  - generative model : selects s_{t+1}
  - online model : set $s_{t+1} = s'$
- can decide to stop exploration $\rightarrow$ adaptive stopping time τ
- if so, can output a guess for $\pi_\star$, $\widehat{\pi}$

(ε, δ)-PAC algorithm

$$\mathbb{P}\left(\tau < \infty, \exists s \in \mathcal{S} : g^{\widehat{\pi}}(s) < g^\star - \varepsilon\right) \leq \delta.$$

State-of-the-art

This problem has been mostly studies in the generative model setting.

Lower bound : [Wang et al., 2022] for any (ε, δ)-PAC algorithm, there exists an MDP $\mathcal{M}$ such that

$$\mathbb{E}_{\mathcal{M}}[\tau_\delta] = \Omega\left(\frac{SAH}{\varepsilon^2}\log(1/\delta)\right)$$

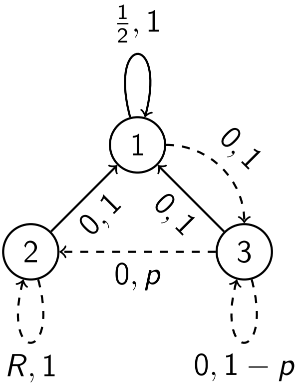
Upper bound : [Zurek and Chen, 2023] there exists an algorithm such that, for all MDPs,

$$\mathbb{E}_{\mathcal{M}}[\tau_\delta] = \widetilde{\mathcal{O}}\left(\frac{SAH}{\varepsilon^2}\log(1/\delta)\right)$$

... but it requires the knowledge of the optimal span bias, H

$$H = \max_s b^\star(s) - \min_s b^\star(s)$$

Estimating H is hard



- $R = 1/2 - \varepsilon \;\Rightarrow\; \pi^\star = \rightarrow \;\Rightarrow\; H = 1/2$
- $R = 1/2 + \varepsilon \;\Rightarrow\; \pi^\star = \dashrightarrow \;\Rightarrow\; H = (1/2 + \varepsilon)\frac{1+p}{p}$

Estimating H is hard

Theorem

For any $\delta < \frac{1}{2e^4}$, $T > 0$, Δ, there exists an MDP $\mathcal{M}$ with $H = 1/2$ such that any algorithm that computes a $\hat{H}$ satisfying $H \leq \hat{H} \leq H + \Delta$ with probability greater than $1 - \delta$ needs (in expectation) more than T samples in $\mathcal{M}$.

But estimating D is easy

Diameter (D) versus Optimal Bias Span (H) : $H \leq D$

$$\begin{aligned} D &= \max_{s \neq s'} \min_{\pi : \mathcal{S} \to \mathcal{A}} \mathbb{E}^{\pi}[\min\{t > 0, s_t = s'\} | s_0 = s] \\ H &= \max_s b^\star(s) - \min_s b^\star(s) \end{aligned}$$

A two-stage algorithm :

- Use an algorithm from [Tarbouriech et al., 2021b] that outputs $\widehat{D}$ such that $\mathbb{P}(D \leq \hat{D} \leq 4D) \geq 1 - \delta/2$ using $\widetilde{\mathcal{O}}(D^2 \log(1/\delta) + S)$ samples
- Use the algorithm of [Zurek and Chen, 2023] with $\widehat{D}$ as an upper bound on H, which uses $\widetilde{\mathcal{O}}\left(\frac{SA\widehat{D}}{\varepsilon^2} \log(1/\delta)\right)$ samples

Diameter Free Exploration

Entry : Accuracy $\varepsilon \in (0,1)$**, confidence level** $\delta \in (0,1)$

- $\widehat{D} = \mathrm{DiameterEstimation}(\delta/2)$
- $\hat{\pi} = \mathrm{BPI}(\widehat{D}, \varepsilon, \delta/2)$
- **Return** $\hat{\pi}$

Theorem

The algorithm above is (ε, δ)-PAC and

$$\mathbb{P}\left(\tau \leq \widetilde{\mathcal{O}}\left(\left[\frac{SAD}{\varepsilon^2} + D^2SA\right]\log(1/\delta) + D^2S^2A\right)\right) \geq 1 - \delta.$$

$\rightarrow$ optimal in the regime of small ε as the lower bound of [Wang et al., 2022] is for an instance with $H = D$!

Diameter Free Exploration

Entry : Accuracy $\varepsilon \in (0,1)$, confidence level $\delta \in (0,1)$

- $\widehat{D} = \mathrm{DiameterEstimation}(\delta/2)$
- $\hat{\pi} = \mathrm{BPI}(\widehat{D}, \varepsilon, \delta/2)$
- **Return** $\hat{\pi}$

Theorem

The algorithm above is (ε, δ)-PAC and

$$\mathbb{P}\left(\tau \leq \widetilde{\mathcal{O}}\left(\left[\frac{SAD}{\varepsilon^2} + D^2 SA\right]\log(1/\delta) + D^2 S^2 A\right)\right) \geq 1 - \delta.$$

$\rightarrow$ optimal in the regime of small ε as the lower bound of [Wang et al., 2022] is for an instance with $H = D$!

Without a generative model ?

Little is known in the online setting !

- we prove that H is definitely not the right complexity measure there
- using an online diameter estimation procedure, we propose an algorithm with a $\widetilde{O}_\delta\left(\frac{SAD^2}{\varepsilon^2} + S^2AD^3\right)$ sample complexity

➜ ... but more adaptive algorithms are needed

Conclusion

Episodic MDPs :

- Variants of UCB-VI (possibly with changing rewards) can solve different pure exploration tasks in a minimax sense
- The instance-dependent complexity of BPI remains hard to characterize
- ... and require complex algorithms

Average reward MDPs :

- Are (arguably) more meaningful in practice
- But there exists no minimax-optimal online algorithm yet
- ... and certainly no practical one, even with a generative model

Al Marjani, A., Tirinzoni, A., and Kaufmann, E. (2023).
Active coverage for PAC reinforcement learning.
In *Proceedings of the 36th Conference On Learning Theory (COLT)*.

Azar, M. G., Osband, I., and Munos, R. (2017).
Minimax regret bounds for reinforcement learning.
In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 263–272.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2005).
Improved second-order bounds for prediction with expert advice.
Machine Learning, 66 :321–352.

Degenne, R., Koolen, W. M., and Ménard, P. (2019).
Non-asymptotic pure exploration by solving games.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021).
Episodic reinforcement learning in finite mdps : Minimax lower bounds revisited.
In *Algorithmic Learning Theory (ALT)*.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018).
Is Q-learning provably efficient ?
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. (2021).
Adaptive reward-free exploration.
In *Algorithmic Learning Theory (ALT)*.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. (2021).
Fast active learning for pure exploration in reinforcement learning.
In *International Conference on Machine Learning (ICML)*.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. (2021a).
A provably efficient sample collection strategy for reinforcement learning.
Advances in Neural Information Processing Systems (NeurIPS).

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. (2021b).
Sample complexity bounds for stochastic shortest path with a generative model.
In *32nd International conference on Algorithmic learning theory*, volume 132. PMLR.

Tiapkin, D., Belomestny, D., Calandriello, D., Moulines, E., Munos, R., Naumov, A., Perrault, P., Tang, Y., Valko, M., and Ménard, P. (2023).
Fast rates for maximum entropy exploration.
In *International Conference on Machine Learning (ICML)*.

Tirinzoni, A., Marjani, A. A., and Kaufmann, E. (2022).
Near instance-optimal PAC reinforcement learning for deterministic mdps.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wagenmaker, A. and Jamieson, K. (2022).
Instance-dependent near-optimal policy identification in linear mdps via online experiment design.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wagenmaker, A. J., Simchowitz, M., and Jamieson, K. (2022).
Beyond no regret : Instance-dependent PAC reinforcement learning.
In *Conference On Learning Theory (COLT)*.

Wang, J., Wang, M., and Yang, L. F. (2022).
Near sample-optimal reduction-based policy learning for average reward MDP.
https://arxiv.org/abs/2212.00603.

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. (2021).
Reward is enough for convex mdps.
In *Neural Information Processing Systems (NeurIPS)*.

Zurek, M. and Chen, Y. (2023).
Span-based optimal sample complexity for average reward MDPs.
https://arxiv.org/abs/2311.13469.

Sample complexities bounds for BPI

For MOCA, PEDEL and PRINCIPLE we have

$$\tau = \widetilde{\mathcal{O}_{\varepsilon,\delta}}\left(\mathrm{Alg}(\mathcal{M},\varepsilon)\log\left(\frac{1}{\delta}\right)\right)$$

where

$$\begin{aligned}
\mathrm{MOCA}(\mathcal{M},\varepsilon) &= H^2\sum_{h=1}^{H}\min_{\rho\in\Omega}\max_{s,a}\frac{1}{\rho_h(s,a)}\min\left(\frac{1}{\widetilde{\Delta}_h(s,a)^2},\ \frac{W_h(s)^2}{\varepsilon^2}\right)\\
&\quad + \frac{H^4\left|(h,s,a):\ \widetilde{\Delta}_h(s,a)\leq 3\varepsilon/W_h(s)\right|}{\varepsilon^2}\\
\mathrm{PEDEL}(\mathcal{M},\varepsilon) &= H^4\sum_{h=1}^{H}\min_{\rho\in\Omega}\max_{\pi\in\Pi_D}\sum_{s,a}\frac{p_h^\pi(s,a)^2/\rho_h(s,a)}{\max(\varepsilon,\Delta(\pi),\Delta_{\min}(\Pi_D))^2}\\
\mathrm{PRINCIPLE}(\mathcal{M},\varepsilon) &= H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)
\end{aligned}$$