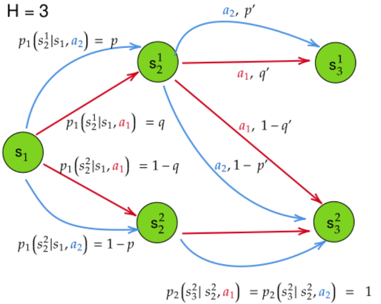
An instance-dependent view on PAC reinforcement Learning

Emilie Kaufmann (CNRS, Univ. Lille, Inria Scool)
based on joint works with
Aymen Al-Marjani (ENS Lyon) and Andrea Tirinzoni (Meta AI)

EWRL, Brussels, September 2023

Finite Horizon Tabular MDPs

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



Value function

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

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

Ouline

Ouline

1. Two PAC RL problems
2. Adaptive Coverage
3. CovGame
4. Reward Free Exploration Beyond the Worse Case
5. Best Policy Identification

Online episodic algorithm

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

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

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

  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

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

Worse case sample complexity : $\tau = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$, w.h.p.

[Dann et al., 2019, Ménard et al., 2021]

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

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

Worse-case sample complexity : $\tau = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right)$, w.h.p.

[Ménard et al., 2021] $\rightarrow$ Beyond worse case ?

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

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

Worse-case sample complexity : $\tau = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right)$, w.h.p.

[Ménard et al., 2021] $\rightarrow$ Beyond worse case ?

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

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

Worse-case sample complexity : $\tau = \mathcal{O}\left(\frac{SAH^3}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right)$, w.h.p.

[Ménard et al., 2021] $\rightarrow$ Beyond worse case ?

Ouline

Covering an MDP

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 :

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

Active coverage : Protocol of interaction

Algorithm 1 Protocol of interation

1: **Input** : target function c

2: Initialize dataset $\mathcal{D}_0 \leftarrow \emptyset$

3: Set target set $\mathcal{X} = \{(s, a, h) \in [H] \times \mathcal{S} \times \mathcal{A} : c_h(s, a) > 0\}$

4: **for** $t = 1, 2, \dots$ **do**

5: $\quad \pi^t \leftarrow \text{CoverageAlgorithm}()$

6: $\quad$ Play π^t and observe trajectory $\mathcal{H}_t := \{(s_h^t, a_h^t, s_{h+1}^t)\}_{1 \leq h \leq H-1}$

7: $\quad$ Update dataset $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \mathcal{H}_t$ and counts.

8: $\quad$ **If** $\forall (h, s, a) \in \mathcal{X}, \ n_h^t(s, a) \geq c_h(s, a)$ **:**

9: $\quad\quad$ Stop and return $\mathcal{D}_t$

10: **end for**

Lower bound : Intuition

Visitation probabilities

For a policy π, $p_h^\pi(s,a) := \mathbb{P}^\pi\left(S_h = s, A_h = a\right)$

- Imagine that the agent uses a fixed exploration policy $\pi^t = \pi_{\text{exp}}$.
- $p_h^{\pi_{\text{exp}}}(s,a)$ is the probability of visiting (h,s,a) in one episode.
- The expected number of episodes before its first visit is $1/p_h^{\pi_{\text{exp}}}(s,a)$.
- The expected number of episodes before getting $c_h(s,a)$ visits from all (h,s,a) is $\max_{h,s,a} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$.

Optimizing over π_{exp}, the agent may satisfy the sampling requirements with

$$\tau \simeq \inf_{\pi_{\text{exp}} \in \Pi_S} \max_{h,s,a} \frac{c_h(s,a)}{p_h^{\pi_{\text{exp}}}(s,a)}$$

Lower bound : Statement

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

proof : alternative LP formulation (stochastic minimum flow)

$$\varphi^\star(c) = \min_{\eta \in \mathbb{R}^{SAH}} \sum_{a\in\mathcal{A}} \eta_1(s_1,a),$$

$$\text{subject to} \quad \sum_{a\in\mathcal{A}} \eta_h(s,a) = \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} p_{h-1}(s|s',a')\eta_{h-1}(s',a') \quad \forall s\in\mathcal{S}, h>1,$$

$$\eta_h(s,a) \geq c_h(s,a) \quad \forall h\in[H], s\in\mathcal{S}, a\in\mathcal{A}, \quad \eta_1(s,a) = 0 \quad \forall s\in\mathcal{S}\setminus\{s_1\}, a\in\mathcal{A}.$$

Motivation for coverage

In MDPs with deterministic transitions and Gaussian rewards, we manage to prove that any (ε, δ)-PAC BPI algorithm satisfies

$$\forall h, s, a, \quad \mathbb{E}[n_h^\tau(s,a)] \geq \frac{C_0 \log(1/\delta)}{\overline{\Delta}_h^2(s,a) \vee \varepsilon^2}$$

for some appropriate (return) "gap" $\overline{\Delta}_h(s,a)$. [Tirinzoni et al., 2022]

➜ leads to the sample complexity lower bound

$$\mathbb{E}[\tau] \geq \varphi^\star(\underline{c})$$

➜ we match it by (adaptively) *covering* triplets with small (estimated) gaps

Motivation for coverage

In MDPs with deterministic transitions and Gaussian rewards, we manage to prove that any (ε, δ)-PAC BPI algorithm satisfies

$$\forall h, s, a, \quad \mathbb{E}[n_h^\tau(s,a)] \geq \frac{C_0 \log(1/\delta)}{\overline{\Delta}_h^2(s,a) \vee \varepsilon^2} := c_h(s,a)$$

for some appropriate (return) "gap" $\overline{\Delta}_h(s,a)$. [Tirinzoni et al., 2022]

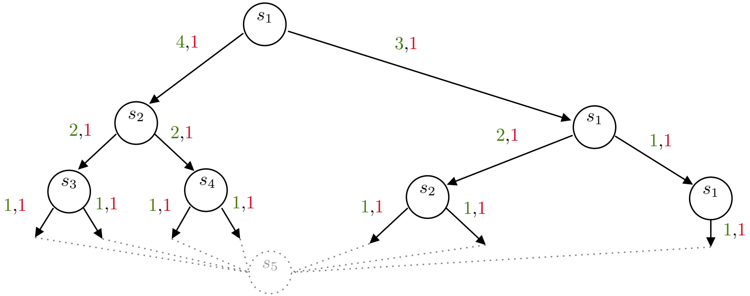
➜ leads to the sample complexity lower bound

$$\mathbb{E}[\tau] \geq \varphi^\star(\underline{c})$$

➜ we match it by (adaptively) *covering* triplets with small (estimated) gaps

Insight on $\varphi^\star$: Deterministic MDPs

Time inhomogeneous Deterministic MDPs $\equiv$ DAG with nodes $\mathcal{N}$ and arcs $\mathcal{E}$



Flow $\eta : \mathcal{E} \to [0,\infty) : \sum_{(s',a')\in\mathcal{I}_h(s)} \eta_{h-1}(s',a') = \sum_{a\in\mathcal{A}_h(s)} \eta_h(s,a) \quad \forall (s,h) \in \mathcal{N}$

Minimum flow for target function $c : \mathcal{E} \to [0,\infty)$

$$\varphi^\star(c) = \min_\eta \sum_{a\in\mathcal{A}_1(s_1)} \eta_1(s_1,a) \;\text{ s.t. }\; \eta_h(s,a) \geq c_h(s,a) \quad \forall (s,a,h) \in \mathcal{E}$$

Insight on $\varphi^\star$: General case

We prove the following bounds :

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

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

Insight on $\varphi^\star$: General case

We prove the following bounds :

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

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

Ouline

Principle

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

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 $\mathcal{X}$.

Principle

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

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

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

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

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

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

Principle

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

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

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

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

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

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

Principle

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

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

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

➜ use online learners !

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

Principle

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

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

Algorithm 2 (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 $\mathcal{H}_t := \{(s_h^t, a_h^t, s_{h+1}^t)\}_{1 \le h \le H-1}$
  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 $\forall (h,s,a),\ n_h^t(s,a) \ge c_h(s,a)$ **:** Stop and return $\mathcal{D}_t$

Why does it work ? From counts to values.

$$\min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} = \inf_{\lambda \in \Delta_{\mathcal{X}}} \sum_{(h,s,a)\in\mathcal{X}} \lambda_h(s,a) \frac{\sum_{t=1}^T \mathbb{1}(s_h^t = s, a_h^t = a)}{c_h(s;a)}$$

$$= \inf_{\lambda \in \Delta_{\mathcal{X}}} \sum_{t=1}^T \ell_t(\lambda)$$

$$\geq \sum_{t=1}^T \ell_t(\lambda^t) - \mathrm{Reg}(\mathcal{A}^\lambda, T) \qquad \text{//regret of the } \lambda \text{ learner}$$

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

$$\underset{w.h.p.}{\geq} \sum_{t=1}^T \sum_{(h,s,a)\in\mathcal{X}} p_h^{\pi_t}(s,a)\widetilde{r}_h^t(s,a) - \frac{1}{c_{\min}}\sqrt{T\log\left(\frac{4T^2}{\delta}\right)} - \mathrm{Reg}(\mathcal{A}^\lambda, T) \qquad \text{//Azuma}$$

$$= \sum_{t=1}^T V^{\pi_t}(s_1, \widetilde{r}_t) - \frac{1}{c_{\min}}\sqrt{T\log\left(\frac{4T^2}{\delta}\right)} - \mathrm{Reg}(\mathcal{A}^\lambda, T)$$

Why does it work ? From values to $\varphi^\star$.

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 $T > 1$,

$$\sum_{t=1}^{T} V_1^\star(s_1;\lambda^t) - \sum_{t=1}^{T} V_1^{\pi_t}(s_1;\lambda^t) \leq \mathrm{Reg}_\delta(\mathcal{A}^\pi, T)$$

$$\begin{aligned}
\sum_{t=1}^{T} V^{\pi_t}(s_1,\widetilde{r}_t) &\underset{w.h.p.}{\geq} \sum_{t=1}^{T} V^\star(s_1,\widetilde{r}_t) - \mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) \\
&= \sum_{t=1}^{T} \sup_{\pi\in\Pi_S} \sum_{h,s,a} p_h^\pi(s,a)\frac{\lambda_h^t(s,a)}{c_h(s,a)} - \mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) \\
&\geq T \sup_{\pi\in\Pi_S} \sum_{h,s,a} p_h^\pi(s,a)\frac{\frac{\sum_{t=1}^{T}\lambda_h^t(s,a)}{T}}{c_h(s,a)} - \mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) \\
&\geq T \inf_{\lambda\in\Delta_{\mathcal{X}}} \sup_{\pi\in\Pi_S} \sum_{h,s,a} p_h^\pi(s,a)\frac{\lambda_h(s,a)}{c_h(s,a)} - \mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T)
\end{aligned}$$

Putting things together

With probability larger than $1-\delta$, for all T,

$$\min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^\star(c)} - \underbrace{\left[\mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) - \mathrm{Reg}(\mathcal{A}^\lambda, T) - \frac{1}{c_{\min}}\sqrt{T\log\left(\frac{4T^2}{\delta}\right)}\right]}_{\text{for standard regret minimizers } \mathcal{O}\left(\sqrt{T}/c_{\min}\right)}$$

Hence

$$T = 2\varphi^\star(c) + \mathcal{O}\left(\left(\frac{\varphi^\star(c)}{c_{\min}}\right)^2\right) \quad \text{implies} \quad \min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq 1$$

Limitation : If $\frac{c_{\max}}{c_{\min}} \gg 1$, the second order term is not negligible...

$$c_{\min} = \min_{(h,s,a)\in\mathcal{X}} c_h(s,a) \quad \text{and} \quad c_{\max} = \max_{(h,s,a)\in\mathcal{X}} c_h(s,a)$$

Putting things together

With probability larger than $1-\delta$, for all T,

$$\min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^\star(c)} - \underbrace{\left[\mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) - \mathrm{Reg}(\mathcal{A}^\lambda, T) - \frac{1}{c_{\min}}\sqrt{T\log\left(\frac{4T^2}{\delta}\right)}\right]}_{\text{for standard regret minimizers } \mathcal{O}(\sqrt{T}/c_{\min})}$$

Hence

$$T = 2\varphi^\star(c) + \mathcal{O}\left(\left(\frac{\varphi^\star(c)}{c_{\min}}\right)^2\right) \quad \text{implies} \quad \min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq 1$$

Limitation : If $\frac{c_{\max}}{c_{\min}} \gg 1$, the second order term is not negligible...

$$c_{\min} = \min_{(h,s,a)\in\mathcal{X}} c_h(s,a) \quad \text{and} \quad c_{\max} = \max_{(h,s,a)\in\mathcal{X}} c_h(s,a)$$

Putting things together

With probability larger than $1-\delta$, for all T,

$$\min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^\star(c)} - \underbrace{\left[\mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) - \mathrm{Reg}(\mathcal{A}^\lambda, T) - \frac{1}{c_{\min}}\sqrt{T \log\left(\frac{4T^2}{\delta}\right)}\right]}_{\text{for standard regret minimizers } \mathcal{O}\left(\sqrt{T}/c_{\min}\right)}$$

Hence

$$T = 2\varphi^\star(c) + \mathcal{O}\left(\left(\frac{\varphi^\star(c)}{c_{\min}}\right)^2\right) \quad \text{implies} \quad \min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq 1$$

Limitation : If $\frac{c_{\max}}{c_{\min}} \gg 1$, the second order term is not negligible...

$$c_{\min} = \min_{(h,s,a)\in\mathcal{X}} c_h(s,a) \quad \text{and} \quad c_{\max} = \max_{(h,s,a)\in\mathcal{X}} c_h(s,a)$$

Putting things together

With probability larger than $1-\delta$, for all T,

$$\min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^\star(c)} - \underbrace{\left[\mathrm{Reg}_{\delta/2}(\mathcal{A}^\pi, T) - \mathrm{Reg}(\mathcal{A}^\lambda, T) - \frac{1}{c_{\min}}\sqrt{T\log\left(\frac{4T^2}{\delta}\right)}\right]}_{\text{for standard regret minimizers } \mathcal{O}(\sqrt{T}/c_{\min})}$$

Hence

$$T = 2\varphi^\star(c) + \mathcal{O}\left(\left(\frac{\varphi^\star(c)}{c_{\min}}\right)^2\right) \quad \text{implies} \quad \min_{h,s,a} \frac{n_h^T(s,a)}{c_h(s,a)} \geq 1$$

Limitation : If $\frac{c_{\max}}{c_{\min}} \gg 1$, the second order term is not negligible...

$$c_{\min} = \min_{(h,s,a)\in\mathcal{X}} c_h(s,a) \quad \text{and} \quad c_{\max} = \max_{(h,s,a)\in\mathcal{X}} c_h(s,a)$$

(Full) CovGame

Two modifications to obtain better guarantees :

- 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

- Rely on first-order regret bounds for $\mathcal{A}^\lambda$ and $\mathcal{A}^\pi$
  - $\mathcal{A}^\lambda$: Weighted Majority Forecaster (WMF) with variance-dependent learning rate
    [Cesa-Bianchi et al., 2005]
  - $\mathcal{A}^\pi$: variant of UCB-VI (new analysis)
    [Azar et al., 2017]

Near-Optimal Coverage

An optimistic algorithm

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

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

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

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

Theorem [Al Marjani et al., 2023]

Let $m = \lceil \log_2(c_{\max}/c_{\min})\rceil \vee 1$. With probability at least $1-\delta$, the stopping time of CovGame with WMF and UCBVI is bounded by

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

Near-Optimal Coverage

An optimistic algorithm

$$\pi^t(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \overline{Q}_h^t(s, a; \tilde{r}_h^t)$$

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

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

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

Links with existing exploration algorithms

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

Ouline

Proportional Coverage

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

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

Remark : link with the concentrability coefficient

$$\varphi^\star\left(\left[\max_\pi p_h^\pi(s,a)\right]_{h,s,a}\right) = \inf_{\rho\in\Omega} \underbrace{\max_{s,a,h} \frac{\max_\pi p_h^\pi(s,a)}{\rho_h(s,a)}}_{C_{\text{conc}}(\rho)}$$

a complexity measure in the offline RL literature
[Chen and Jiang, 2019, Xie et al., 2023]

Why Proportional Coverage ?

Sufficient condition for RFE [Jin et al., 2020] : have a good estimate of the value functions of all policies, for all reward functions

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

New 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 ?

Sufficient condition for RFE [Jin et al., 2020] : have a good estimate of the value functions of all policies, for all reward functions

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

New 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 ?

Sufficient condition for RFE [Jin et al., 2020] : have a good estimate of the value functions of all policies, for all reward functions

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

New 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 ?

Sufficient condition for RFE [Jin et al., 2020] : have a good estimate of the value functions of all policies, for all reward functions

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

New 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 3 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

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

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

Beyond worse case

- Minimax lower bound : $\Omega\left(\frac{SAH^3\log(1/\delta)}{\varepsilon^2} + \frac{S^2AH}{\varepsilon^2}\right)$
- As $\mathcal{C}(\mathcal{M},\varepsilon) \leq SAH/\varepsilon^2$, we always have

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

- For disguised contextual bandits ($p_h(s'|s,a) = p_h(s'|s)$), $\mathcal{C}(\mathcal{M},\varepsilon) = A/\varepsilon^2$

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

- A class of MDPs depending on $\alpha \in (0,1)$ such that $\mathcal{C}(\mathcal{M},\varepsilon) \leq S^\alpha AH/\varepsilon^2$

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

Ouline

Best Policy Identification

Idea : combine proportional coverage with eliminations

- Deterministic MDPs with random rewards

Theorem [Tirinzoni et al., 2022]

Elimination for PAC RL (EPRL) uses smart action eliminations and obtains a **near-optimal** sample complexity
(it matches the $\varphi^\star$ lower bound up to logarithmic terms and an H^2 factor)

Best Policy Identification

- General stochastic MDPs

MOCA [Wagenmaker et al., 2022]	PEDEL [Wagenmaker and Jamieson, 2022]	PRINCIPLE [Al Marjani et al., 2023]
proportional coverage	optimal design	proportional coverage
action elimination	policy elimination	**implicit** policy elimination
value gap based sample complexity	policy gap based sample complexity	policy gap based sample complexity
efficient	intractable	**efficient**

$$\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a) \quad \text{versus} \quad \Delta(\pi) = V_1^\star(s_1) - V_1^\pi(s_1)$$

Summary & Perspective

CovGame is a near-optimal algorithm algorithm for collecting a prescribed number of visits in an episodic MDP

Proportional Coverage Exploration is based on CovGame and achieves the first instance-dependent guarantees for Reward Free Exploration

The instance-dependent complexity of Best Policy Identification is still to be understood in stochastic MDPs

- different algorithms with different gap-dependent sample complexity
- ... that are essentially incomparable

➜ Lower bound ? A *computationally efficient* algorithm that attains it ?

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.

Chen, J. and Jiang, N. (2019).
Information-theoretic considerations in batch reinforcement learning.
In *International Conference on Machine Learning (ICML)*.

Dann, C., Li, L., Wei, W., and Brunskill, E. (2019).
Policy certificates : Towards accountable reinforcement learning.
In *International Conference on Machine Learning*.

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

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020).
Reward-free exploration for reinforcement learning.
In *International Conference on Machine Learning*, pages 4870–4879. PMLR.

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. (2021).
A provably efficient sample collection strategy for reinforcement learning.
Advances in Neural Information Processing Systems (NeurIPS).

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)*.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. (2023).
The role of coverage in online reinforcement learning.
In *ICLR*.

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

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

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