# Active coverage for PAC Reinforcement Learning

Aymen Al Marjani[1,2], Andrea Tirinzoni[3], Emilie Kaufmann[2,3],

[1]UMPA, ENS de Lyon
[2]Inria Lille, Equipe SCOOL
[3]CNRS, CRIStAL [4]META AI, Paris

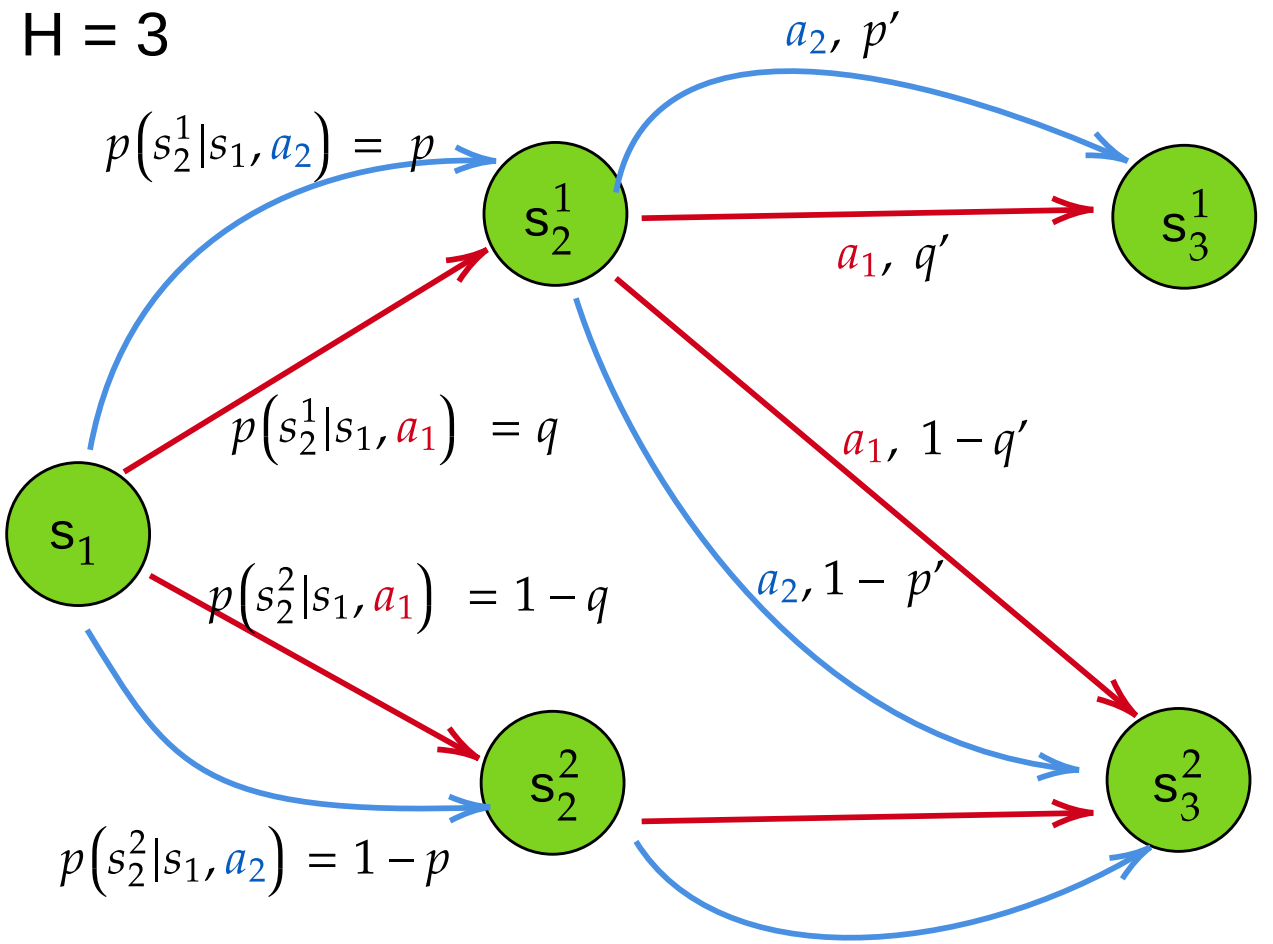
28 Avril 2023


# Outline

# Outline

# Finite-horizon Tabular MDPs

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



# Active coverage: learning problem

**The Active Coverage problem:** Design an algorithm $\mathbb{A}$ which takes as input a target function $c : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ and a failure probability $\delta$. $\mathbb{A}$ should interact with $\mathcal{M}$ and return a dataset $\mathcal{D}_t$ such that

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\left(\exists t \geq 1,\ \forall (h,s,a),\ n_h(s,a;\mathcal{D}_t) \geq c_h(s,a)\right) \geq 1-\delta.$$

# Active coverage: learning problem

**The Active Coverage problem:** Design an algorithm $\mathbb{A}$ which takes as input a target function $c : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ and a failure probability $\delta$. $\mathbb{A}$ should interact with $\mathcal{M}$ and return a dataset $\mathcal{D}_t$ such that

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\left(\exists t \geq 1,\ \forall (h,s,a),\ n_h(s,a;\mathcal{D}_t) \geq c_h(s,a)\right) \geq 1-\delta.$$

**Performance criteria:**

$$\tau := \inf\left\{t \geq 1,\ \forall (h,s,a),\ n_h(s,a;\mathcal{D}_t) \geq c_h(s,a)\right\}$$

$\Longrightarrow$ We want $\mathbb{A}$ to make $\tau$ as small as possible.

# Active coverage: Protocol of interaction

**Algorithm 1** Protocol of interation

**Input:** target function $c$, risk $\delta \in (0,1)$.
Initialize dataset $\mathcal{D}_0 \leftarrow \emptyset$
Set target set $\mathcal{X} = \{(s,a,h) \in [H] \times \mathcal{S} \times \mathcal{A} : c_h(s,a) > 0\}$
**for** $t = 1, 2, \ldots$ **do**
    $\pi^t \leftarrow$ CoverageAlgorithm()
    Play $\pi^t$ and observe trajectory $\mathcal{H}_t := \{(s_h^t, a_h^t, s_{h+1}^t)\}_{1 \le h \le H-1}$
    Update dataset $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \mathcal{H}_t$.
    **If** $\forall (h,s,a),\ \ n_h(s,a;\mathcal{D}_t) \ge c_h(s,a)$**:**
        Stop and return $\mathcal{D}_t$
**end for**

where the counts $n_h(s,a;\mathcal{D}_t) = \sum_{u=1}^{t} \mathbb{1}(s_h^t = s, a_h^t = a)$.

# Outline

# Applications

- Best-Policy Identification: Given some reward function $r : [H] \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, find an $\varepsilon$-optimal policy, i.e. $V_1^{\widehat{\pi}}(s_1) \geq V_1^{\pi^\star}(s_1) - \varepsilon$.
- Reward-free Exploration: Collect Dataset $\mathcal{D}$ such that you can find $\varepsilon$-optimal policy for any reward function $r$ using $\mathcal{D}$.
- Hybrid RL (Offline + Online): Complete some dataset $\mathcal{D}$ that was given to you.

# Outline

# Lower bound: Intuition

Imagine that the agent chooses to explore using the policy $\pi^{exp}$, for all episodes.

Let $p_h^{\pi^{exp}}(s,a) = \mathbb{P}(s_h = s, a_h = a | s_1, \pi^{exp})$ be the probability of visiting $(h,s,a)$ in one episode.

# Lower bound: Intuition

Imagine that the agent chooses to explore using the policy $\pi^{exp}$, for all episodes.

Let $p_h^{\pi^{exp}}(s,a) = \mathbb{P}(s_h = s, a_h = a | s_1, \pi^{exp})$ be the probability of visiting $(h,s,a)$ in one episode.

The expected number of episodes before the first visit is $1/p_h^{\pi^{exp}}(s,a)$.

# Lower bound: Intuition

Imagine that the agent chooses to explore using the policy $\pi^{exp}$, for all episodes.

Let $p_h^{\pi^{exp}}(s,a) = \mathbb{P}(s_h = s, a_h = a | s_1, \pi^{exp})$ be the probability of visiting $(h,s,a)$ in one episode.

The expected number of episodes before the first visit is $1/p_h^{\pi^{exp}}(s,a)$.

To satisfy $n_h^\tau(s,a) \geq c_h(s,a)$, the agent would need at least $\max_{h,s,a} \frac{c_h(s,a)}{p_h^{\pi^{exp}}(s,a)}$

# Lower bound: Intuition

Imagine that the agent chooses to explore using the policy $\pi^{exp}$, for all episodes.

Let $p_h^{\pi^{exp}}(s,a) = \mathbb{P}(s_h = s, a_h = a | s_1, \pi^{exp})$ be the probability of visiting $(h,s,a)$ in one episode.

The expected number of episodes before the first visit is $1/p_h^{\pi^{exp}}(s,a)$.

To satisfy $n_h^\tau(s,a) \geq c_h(s,a)$, the agent would need at least $\max_{h,s,a} \frac{c_h(s,a)}{p_h^{\pi^{exp}}(s,a)}$

Since the agent has the choice of which policy to play,

$$\mathbb{E}[\tau] \gtrapprox \inf_{\pi^{exp} \in \Pi^S} \max_{h,s,a} \frac{c_h(s,a)}{p_h^{\pi^{exp}}(s,a)}.$$

# Lower bound: Statement

## Theorem 1

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

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

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

# Outline

# Vanilla CovGame: Intuition

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

# Vanilla CovGame: Intuition

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

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

with $\mathcal{X} := \{(h,s,a) : c_h(s,a) > 0\}$ and $\Delta_{\mathcal{X}} =$ the simplex over $\mathcal{X}$.

# Vanilla CovGame: Intuition

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

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

$$= \text{A matrix game !}$$

with $\mathcal{X} := \{(h,s,a) : c_h(s,a) > 0\}$ and $\Delta_{\mathcal{X}} =$ the simplex over $\mathcal{X}$.

# Vanilla CovGame: Intuition

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

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

$$= \text{A matrix game !}$$

with $\mathcal{X} := \{(h,s,a) : c_h(s,a) > 0\}$ and $\Delta_{\mathcal{X}} =$ the simplex over $\mathcal{X}$.

$$\sum_{h,s,a} \frac{p_h^{\pi^{exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} = \mathbb{E}_{\mathcal{M},\pi^{exp}}\left[\sum_{h,s,a} \frac{\mathbb{1}(s_h = s, a_h = a)\lambda_h(s,a)}{c_h(s,a)}\right] = \text{a value functio}$$

$$\sum_{h,s,a} \frac{p_h^{\pi^{exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} = \lambda^\top(p^{\pi^{exp}}/c) = \text{linear loss of a forecaster}$$

# Vanilla CovGame: pseudo-code

**Algorithm 2** Vanilla CovGame

**Input:** target function $c$, Adversarial RL algo $\mathcal{A}^{\Pi}$, Online learner $\mathcal{A}^{\lambda}$, risk $\delta$.
Initialize dataset $\mathcal{D}_0 \leftarrow \emptyset$
Set target set $\mathcal{X} = \{(s, a, h) \in [H] \times \mathcal{S} \times \mathcal{A} : c_h(s, a) > 0\}$
Normalize targets $\widetilde{c}_h(s, a) = c_h(s, a)/c_{min} \quad (c_{min} := \min_{(h,s,a)\in\mathcal{X}} c_h(s, a))$
Initialize challenger weights $\lambda_h^1(s, a) \leftarrow \mathbb{1}((h, s, a) \in \mathcal{X})/|\mathcal{X}|$ for all $h, s, a$
**for** $t = 1, 2, \ldots$ **do**
    Define reward $R_h^t(s, a) = \mathbb{1}((h, s, a) \in \mathcal{X})\lambda_h^t(s, a)/\widetilde{c}_h(s, a)$ for all $h, s, a$
    Feed $\mathcal{A}^{\Pi}$ with $R^t$, confidence $\delta/2$ and get exploration policy $\pi^t$
    Play $\pi^t$ and observe trajectory $\mathcal{H}_t := \{(s_h^t, a_h^t, s_{h+1}^t)\}_{1\le h\le H-1}$
    Update dataset $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \mathcal{H}_t$.
    Feed $\mathcal{A}^{\lambda}$ with loss

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

    and get new weight vector $\lambda^{t+1}$
    **If** $\forall(h, s, a),\ n_h(s, a; \mathcal{D}_t) \ge c_h(s, a)$**:** Stop and return $\mathcal{D}_t$

# Vanilla CovGame: Sample Complexity

## Theorem 2

Let $\mathcal{R}^{\lambda}(T)$ be an anytime bound on the regret of the online learning algorithm $\mathcal{A}^{\lambda}$:

$$\forall T \in \mathbb{N}^*, \ \sum_{t=1}^{T} \ell^t(\lambda^t) - \min_{\lambda \in \Delta_{\mathcal{X}}} \sum_{t=1}^{T} \ell^t(\lambda) \leq \mathcal{R}^{\lambda}(T) \, . \tag{1}$$

Let $\mathcal{R}^{\Pi}(T, \delta)$ be a high-probability anytime bound on the adversarial regret of $\mathcal{A}^{\Pi}$:

$$\mathbb{P}\left( \forall T \in \mathbb{N}^*, \ \sum_{t=1}^{T} \sup_{\pi} V_1^{\pi}\left(s_1; R^t\right) - \sum_{t=1}^{T} V_1^{\pi^t}\left(s_1; R^t\right) \leq \mathcal{R}^{\Pi}(T, \delta) \right) \geq 1 - \delta \, . \tag{2}$$

Then w.p at least $1 - \delta$, for all $T \geq 1$,

$$\min_{(h,s,a) \in \mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^{\star}(c)} - \frac{1}{c_{min}} \left[ \mathcal{R}^{\lambda}(T) + \mathcal{R}^{\Pi}(T, \delta/2) + \sqrt{T \log\left(\frac{4T^2}{\delta}\right)} \right]$$

# Vanilla CovGame: Instanciation

Using $\mathcal{A}^{\Pi} = \mathrm{UCBVI}$ (for changing rewards) and $\mathcal{A}^{\lambda} = \mathrm{HEDGE}$, we have

$$\mathcal{R}^{\Pi}(T, \delta) \leq 32 \log(T+1) \sqrt{SAH^2 T(\log(2SAH/\delta) + S)}$$
$$\mathcal{R}^{\lambda}(T) \leq \sqrt{2T \log(SAH)} + \sqrt{\log(SAH)/8}$$

## Corallary 1

With probability at least $1-\delta$ CovGame instantiated with the algorithms above has sample complexity

$$\tau \leq 2\varphi^{\star}(c) + \widetilde{\mathcal{O}}\left( \left( \frac{\varphi^{\star}(c)}{c_{\min}} \right)^2 SH^2 A(\log(1/\delta) + S) \right)$$

# Vanilla CovGame: Instanciation

Using $\mathcal{A}^{\Pi} = \mathrm{UCBVI}$ (for changing rewards) and $\mathcal{A}^{\lambda} = \mathrm{HEDGE}$, we have

$$\mathcal{R}^{\Pi}(T, \delta) \leq 32 \log(T+1) \sqrt{SAH^2 T (\log(2SAH/\delta) + S)}$$
$$\mathcal{R}^{\lambda}(T) \leq \sqrt{2T \log(SAH)} + \sqrt{\log(SAH)/8}$$

## Corallary 1

With probability at least $1-\delta$ CovGame instantiated with the algorithms above has sample complexity

$$\tau \leq 2\varphi^{\star}(c) + \widetilde{\mathcal{O}}\left( \left( \frac{\varphi^{\star}(c)}{c_{\min}} \right)^2 SH^2 A (\log(1/\delta) + S) \right)$$

When $c_h(s,a) = N\mathbb{1}\left((h,s,a) \in \mathcal{X}\right)$ and $N \gg 1$, $\tau \approx 2\varphi^{\star}(c)$

# Vanilla CovGame: Proof sketch 1

**Step 1:** from the counts to the loss of the $\lambda$ player.

$$
\begin{aligned}
c_{\min} \min_{(h,s,a)\in\mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} &= \inf_{\lambda\in\Delta_{\mathcal{X}}} \lambda \cdot (n^T/\widetilde{c}) && \text{(definitions of } \widetilde{c} \text{ and } \Delta_{\mathcal{X}}) \\
&= \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{(h,s,a)\in\mathcal{X}} \lambda_h(s,a) \sum_{t=1}^{T} \frac{\mathbb{1}\left(s_h^t = s, a_h^t = a\right)}{\widetilde{c}_h(s,a)} \\
&= \inf_{\lambda\in\Delta_{\mathcal{X}}} \sum_{t=1}^{T} \ell^t(\lambda) && \text{(definition of } \ell_t(\lambda)) \\
&\geq \sum_{t=1}^{T} \ell^t(\lambda^t) - \mathcal{R}^{\lambda}(T), && \text{(regret bound of } \lambda \text{ player)}
\end{aligned}
$$

# Vanilla CovGame: Proof sketch 2

**Step 2:** from the loss of the $\lambda$ player to the optimal value function of $\pi$ player.

$$\sum_{t=1}^{T} \ell^t(\lambda^t) = \sum_{t=1}^{T} \sum_{h,s,a} \frac{\mathbb{1}((h,s,a) \in \mathcal{X}) \lambda_h^t(s,a)}{\widetilde{c}_h(s,a)} \left( \mathbb{1}\left(s_h^t = s, a_h^t = a\right) \pm p_h^{\pi^t}(s,a) \right)$$

(definition of $\ell_t(\lambda^t)$)

$$= \sum_{t=1}^{T} \sum_{h,s,a} p_h^{\pi^t}(s,a) R_h^t(s,a) + \sum_{t=1}^{T} \sum_{h,s,a} R_h^t(s,a) \left( \mathbb{1}\left(s_h^t = s, a_h^t = a\right) - p_h^{\pi^t}(s,a) \right)$$

$$= \sum_{t=1}^{T} V_1^{\pi^t}\left(s_1; R^t\right) + M_T$$

(definition of $V_1^{\pi}(s_1; R)$ + martingale)

$$\geq \sup_{\pi} \sum_{t=1}^{T} V_1^{\pi}\left(s_1; R^t\right) - \mathcal{R}^{\Pi}(T, \delta/2) - \sqrt{T \log\left(\frac{4T^2}{\delta}\right)},$$

(Regret of $\pi$ player + Azuma Hoeffding)

# Proof sketch 3

**Step 3:** from the optimal value function of $\pi$ player to the minimum flow.

$$\sup_{\pi} \sum_{t=1}^{T} V_1^{\pi}\left(s_1; R^t\right) = \sup_{\pi} \sum_{t=1}^{T} \sum_{h,s,a} p_h^{\pi}(s,a) \frac{\mathbb{1}((h,s,a) \in \mathcal{X}) \lambda_h^t(s,a)}{\widetilde{c}_h(s,a)}$$

$$= T \sup_{\pi} \sum_{h,s,a} \left( p_h^{\pi}(s,a) \frac{\mathbb{1}((h,s,a) \in \mathcal{X})}{\widetilde{c}_h(s,a)} \right) \left( \frac{\sum_{t=1}^{T} \lambda_h^t(s,a)}{T} \right)$$

$$\geq T \sup_{\pi} \min_{(h,s,a) \in \mathcal{X}} \frac{p_h^{\pi}(s,a)}{\widetilde{c}_h(s,a)} = c_{min} \frac{T}{\varphi^{\star}(c)}.$$

Wrapping up everything, we get

$$c_{\min} \min_{(h,s,a) \in \mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} \geq c_{min} \frac{T}{\varphi^{\star}(c)} - \mathcal{R}^{\lambda}(T) - \mathcal{R}^{\Pi}(T, \delta/2) - \sqrt{T \log\left(\frac{4T^2}{\delta}\right)}$$

$$\implies \min_{(h,s,a) \in \mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^{\star}(c)} - \frac{1}{c_{min}} \left[ \mathcal{R}^{\lambda}(T) + \mathcal{R}^{\Pi}(T, \delta/2) + \sqrt{T \log\left(\frac{4T^2}{\delta}\right)} \right]$$

# Outline

# Improved CovGame: Idea

Recall the bound for Vanilla CovGame:

$$\tau \leq 2\varphi^\star(c) + \widetilde{\mathcal{O}}\left(\left(\frac{\varphi^\star(c)}{c_{\min}}\right)^2 S^2 H^2 A \log(1/\delta)\right),$$

$\Longrightarrow$ the second-order term is no longer negligible if $\frac{c_{max}}{c_{min}} \gg 1$ !!

# Improved CovGame: Idea

Recall the bound for Vanilla CovGame:

$$\tau \leq 2\varphi^\star(c) + \widetilde{\mathcal{O}}\left(\left(\frac{\varphi^\star(c)}{c_{\min}}\right)^2 S^2 H^2 A \log(1/\delta)\right),$$

$\implies$ the second-order term is no longer negligible if $\frac{c_{max}}{c_{min}} \gg 1$ !!

**Idea:** Cluster triplets $(h, s, a)$ into sets

$\mathcal{Y}_j = \{(h, s, a) : c_h(s, a) \in [c_{min}2^j, c_{min}2^{j+1}]\}$ so that $\frac{\max_{(h,s,a)\in\mathcal{Y}_j} c_h(s,a)}{\min_{(h,s,a)\in\mathcal{Y}_j} c_h(s,a)} \leq 2$, then run Vanilla CovGame on each cluster separately.

# Improved CovGame: pseudo-code

**Algorithm 3** Improved CovGame

**Input:** Target function $c$, Adversarial RL algo $\mathcal{A}^{\Pi}$, Online learner $\mathcal{A}^{\lambda}$, risk $\delta$.
  Let $\mathcal{X}_0 := \mathcal{X}$ and $\mathcal{X}_k := \{(h,s,a) : c_h(s,a) > c_{\min} 2^k\}$ for all $k \in \mathbb{N}^*$
  Initialize counts $n_h^0(s,a) = 0$ for all $h, s, a$
  Reset $\mathcal{A}^{\lambda}$ on $\mathcal{P}(\mathcal{X})$, set $\lambda_h^1(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X})/|\mathcal{X}|$ for all $h, s, a$
  Initialize $k_1 \leftarrow 0$
  **for** $t = 1, 2, \ldots$ **do**
    Get $\pi^t$ from $\mathcal{A}^{\Pi}$ given reward function $\lambda^t$ and confidence $1 - \delta/2$
    Generate a trajectory $\{(s_h^t, a_h^t)\}_{h \in [H]}$ using policy $\pi^t$ and update counts $n^t$
    **if** $n_h^t(s,a) \geq c_h(s,a)$ for all $h, s, a$ **then** stop and return all sampled trajectories
    Update $k_{t+1} \leftarrow \max\{j \in \mathbb{N} : n_h^t(s,a) \geq c_h(s,a)\ \forall (h,s,a) \in \mathcal{X} \setminus \mathcal{X}_j\}$
    **if** $k_{t+1} \neq k_t$ **then**
      Reset $\mathcal{A}^{\lambda}$ on $\mathcal{P}(\mathcal{X}_{k_{t+1}})$, set $\lambda_h^{t+1}(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X}_{k_{t+1}})/|\mathcal{X}_{k_{t+1}}|$
    **else**
      Feed $\mathcal{A}^{\lambda}$ with loss $\ell^t(\lambda) = \sum_{(h,s,a) \in \mathcal{X}_{k_t}} \lambda_h(s,a) \mathbb{1}(s_h^t = s, a_h^t = a)$, get weight $\lambda^{t+1}$

# Improved CovGame: sample complexity

## Corollary 2 (Under the same conditions of Corollary 1)

With probability at least $1-\delta$, the stopping time of Improved CovGame with HEDGE and UCBVI can be bounded by

$$\tau \leq 8m\varphi^\star(c) + \widetilde{O}\big(m^2\varphi^\star(\mathbb{1}_{\mathcal{X}})^2 SAH^2(\log(1/\delta)+S)\big),$$

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

# Improved CovGame: sample complexity

## Corollary 2 (Under the same conditions of Corollary 1)

With probability at least $1-\delta$, the stopping time of Improved CovGame with HEDGE and UCBVI can be bounded by

$$\tau \leq 8m\varphi^\star(c) + \widetilde{O}\big(m^2\varphi^\star(\mathbb{1}_{\mathcal{X}})^2 SAH^2(\log(1/\delta)+S)\big),$$

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

$\Longrightarrow$ The scaling in $c_{\max}/c_{\min}$ is only logarithmic.

# Outline

# Previous approaches

- (Tarbouriech et al. 2021) GOSPRL
- (Wagenmaker et al. 2021) Learn2Explore

# Previous approaches

- (Tarbouriech et al. 2021) GOSPRL
- (Wagenmaker et al. 2021) Learn2Explore

Complexity:

$$\sum_{h,s,a} \frac{c_h(s,a)}{\sup_{\pi^{exp}} p_h^{\pi^{exp}}(s,a)} \geq \varphi^\star(c)$$

# Conclusion and perspectives

1. Active coverage = how collect $c_h(s,a)$ observations from each $(h,s,a)$ of an MDP.
2. The minimum number of episode one needs is $\varphi^\star(c) = \inf_{\pi^{exp} \in \Pi^S} \max_{(s,a,h)\in\mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi^{exp}}(s,a)}$, a quantity that depends on the MDP.
3. A simple game-based strategy can (nearly) match this lower bound.

# Thanks !