Active Coverage for PAC Reinforcement Learning

Aymen Al-Marjani, Andrea Tirinzoni, Emilie Kaufmann

Active Coverage for PAC Reinforcement Learning

Aymen Al-Marjani AYMEN.AL_MARJANI@ENS-LYON.FR
UMPA, ENS Lyon, Lyon, France

Andrea Tirinzoni TIRINZONI@META.COM
Meta AI, Paris, France

Emilie Kaufmann EMILIE.KAUFMANN@UNIV-LILLE.FR
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, Lille, France

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Collecting and leveraging data with good coverage properties plays a crucial role in different aspects of reinforcement learning (RL), including reward-free exploration and offline learning. However, the notion of "good coverage" really depends on the application at hand, as data suitable for one context may not be so for another. In this paper, we formalize the problem of *active coverage* in episodic Markov decision processes (MDPs), where the goal is to interact with the environment so as to fulfill given sampling requirements. This framework is sufficiently flexible to specify any desired coverage property, making it applicable to any problem that involves online exploration. Our main contribution is an *instance-dependent* lower bound on the sample complexity of active coverage and a simple game-theoretic algorithm, CovGame, that nearly matches it. We then show that CovGame can be used as a building block to solve different PAC RL tasks. In particular, we obtain a simple algorithm for PAC reward-free exploration with an instance-dependent sample complexity that, in certain MDPs which are "easy to explore", is lower than the minimax one. By further coupling this exploration algorithm with a new technique to do implicit eliminations in policy space, we obtain a computationally-efficient algorithm for best-policy identification whose instance-dependent sample complexity scales with gaps between policy values.

Keywords: Reinforcement learning, Coverage, Reward-free exploration, Best-policy identification

1. Introduction

The quality of the available data, whether it is actively gathered through *online* interactions with the environment or provided as a fixed *offline* dataset, plays a fundamental role in characterizing the performance of any reinforcement learning (RL, Sutton and Barto, 2018) agent. An important concept to quantify such quality is *coverage*, a property measuring the extent to which data spreads across the state-action space. The notion of coverage, through the so-called *concentrability coefficients*, is ubiquitous in the vast literature on offline RL (e.g., Munos, 2003; Munos and Szepesvári, 2008; Farahmand et al., 2009, 2010; Chen and Jiang, 2019; Xie and Jiang, 2020, 2021; Jin et al., 2021; Foster et al., 2022). Intuitively, the better data covers the state space, the better performance one can expect from an offline RL method. Recently, Xie et al. (2022) showed that a similar phenomenon also occurs in online RL: the sole existence of a good covering data distribution implies sample-efficient online RL with non-linear function approximation, even if such a distribution is unknown and inaccessible by the agent.

While these works treat coverage as a property of some *given* data or environment, a large body of literature focuses on *actively* collecting good covering data. This falls under the umbrella of

reward-free exploration (RFE, Jin et al., 2020), a setting where the agent interacts with an unknown environment without any reward feedback. The objective is typically to collect sufficient data to enable the computation of a near-optimal policy for any reward function provided at downstream, e.g., by planning on top of an estimated model of the environment or by running any off-the-shelf offline RL method. Many provably-efficient algorithms exist for this problem that mostly differ in their exploration strategy. Some try to gather a minimum number of samples from each reachable state (Jin et al., 2020; Zhang et al., 2021b), while others adaptively optimize a reward function proportional to their uncertainty over the environment (Kaufmann et al., 2021; Ménard et al., 2021) or more simply a zero reward (Chen et al., 2022). All these approaches provably guarantee that the collected data is sufficient to learn any reward function provided at test time. Another popular technique is to seek data distributions that maximize the entropy over the state-space (Hazan et al., 2019; Cheung, 2019; Zahavy et al., 2021; Mutti et al., 2022). Finally, there is a long recent line of empirical works focusing on RFE, where the problem is often called *unsupervised RL* (e.g., Laskin et al., 2021; Eysenbach et al., 2019; Burda et al., 2019; Yarats et al., 2021).

The RFE literature mostly focuses on collecting data with the *specific* properties needed for the task under consideration (e.g., achieving zero-shot RL at test time). Motivated by the crucial role of coverage in RL, in this paper we treat the problem at a higher level of generality. We formulate and study the problem of *active coverage* in episodic MDPs, where the goal is to interact online with the environment so as to collect data that satisfies some given coverage constraints. Following Tarbouriech et al. (2021) who considered a similar problem in reset-free MDPs, we formalize such constraints as a set of sampling requirements that the learner must fulfill during learning. This gives our framework a high flexibility, as one can require different notions of coverage simply by changing the sampling requirements. Moreover, the applications are numerous, as any active coverage algorithm yields an exploration strategy that can be readily plugged in to tackle different problems. In our specific case, we shall see how to apply it to design PAC algorithms for both RFE and best-policy identification (BPI, Fiechter, 1994; Dann and Brunskill, 2015; Dann et al., 2019; Wagenmaker et al., 2022; Wagenmaker and Jamieson, 2022; Tirinzoni et al., 2022, 2023).

Contributions First, we derive an *instance-dependent complexity measure* for the active coverage problem as a lower bound on the number of episodes that any algorithm must play in order to fulfill the sampling requirements on an MDP. We show interesting connections with existing coverage measures, especially the concentrability coefficients used in offline RL (e.g., Munos, 2003).

Then, we propose COVGAME, a novel approach for active coverage. COVGAME is based on a simple game-theoretic view of the problem, where an RL agent tries to optimize a sequence of rewards produced by an adversary that constantly challenges it to reach uncovered states. We show that the sample complexity of COVGAME scales with our complexity measure plus some lower order learning cost, hence making our approach near-optimal.

Finally, we show how active coverage can be readily applied to get PAC algorithms with *instance-dependent* sample complexity for both RFE and BPI. In particular, we show that an almost plug-and-play version of COVGAME solves RFE using a number of samples scaling with our *instance-dependent* coverage complexity, i.e., adapting to the complexity for navigating the underlying MDP. We show that this sample complexity can be smaller than the minimax one (Ménard et al., 2021; Zhang et al., 2021b), a perhaps surprising result given the worst-case nature of the problem (i.e., the agent aims at optimizing for *all* possible rewards). For BPI, we show how COVGAME can be sequentially applied to estimate the value function of all policies, while gradually focusing on poli-

cies with better performance. Notably, we obtain an instance-dependent sample complexity scaling with *policy gaps* (Tirinzoni et al., 2021; Dann et al., 2021) which is in line with the recent results of Wagenmaker and Jamieson (2022) and Tirinzoni et al. (2022) (the latter for the special case of deterministic MDPs). A key advantage is that our algorithm, as opposed to the one of Wagenmaker and Jamieson (2022), is computationally-efficient and does not need to enumerate all policies to perform explicit eliminations. This is obtained thanks to a novel scheme which instead sequentially constrains the set of state-action distributions corresponding to high-return and well-covered policies, a technique that we believe to be of broader interest. An important technical tool for both RFE and BPI is a novel concentration inequality for value functions (see Appendix D).

2. Active Coverage and its Complexity

We suppose that the learner interacts with an environment modeled as a tabular finite-horizon *Markov decision process* (MDP) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, s_1, H)$, where $\mathcal{S}$ is a finite set of S states, $\mathcal{A}$ is a finite set of A actions, $p_h : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$[1] denotes the transition function at stage $h \in [H]$, $s_1 \in \mathcal{S}$ is the initial state, and H is the horizon. The interaction with $\mathcal{M}$ proceeds through episodes of length H. In each episode, starting from the initial state $s_1 \in \mathcal{S}$, at each stage $h \in [H]$, the learner takes an action $a_h \in \mathcal{A}$ based on the current state $s_h \in \mathcal{S}$ and it observes a stochastic transition to a new state $s_{h+1} \sim p_h(s_h, a_h)$. We denote by $p_h(s'|s, a)$ the probability that the new state is s' when selecting action a in state s at step h of the episode.

The actions are chosen by a (possibly stochastic) policy $\pi = \{\pi_h\}_{h\in[H]}$, i.e., a sequence of mappings $\pi_h : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, where $\pi_h(a|s)$ denotes the probability that the learner takes action a in state s at stage h. With some abuse of notation, we shall use $\pi_h : \mathcal{S} \to \mathcal{A}$ to denote a deterministic policy, where $\pi_h(s)$ directly returns the action taken in state s at stage h. We denote by Π^{S} (resp. Π^{D}) the set of all stochastic (resp. deterministic policies).

Denoting by $\mathbb{P}^\pi$ (resp. $\mathbb{E}^\pi$) the probability (resp. expectation) operator induced by the execution of a policy $\pi \in \Pi^{\mathrm{S}}$ for an episode on $\mathcal{M}$, we define, for each (h, s, a), $p_h^\pi(s, a) := \mathbb{P}^\pi(s_h = s, a_h = a)$ and $p_h^\pi(s) := \mathbb{P}^\pi(s_h = s)$. We let $\Omega := \{p^\pi : \pi \in \Pi^{\mathrm{S}}\}$ denote the set of all valid state-action distributions. It is well known (e.g., Puterman, 1994) that any distribution $\rho \in \Omega$ satisfies $\rho_h \in \mathcal{P}(\mathcal{S} \times \mathcal{A})$ for all h and $\sum_a \rho_h(s, a) = \sum_{s',a'} \rho_{h-1}(s', a')p_{h-1}(s|s', a')$ for all s, a and $h > 1$. We make the following assumption to ensure that the whole state-space can be navigated.

Assumption 1 (Reachability) *Each state $s \in \mathcal{S}$ is reachable at any stage $h \in \{2, \dots, H\}$ by some policy, i.e., $\max_{\pi\in\Pi^{\mathrm{S}}} p_h^\pi(s) > 0$.*

Reachability conditions like Assumption 1 are standard in prior work. In non-episodic reset-free MDPs (e.g., Jaksch et al., 2010), the MDP is often required to be communicating to ensure learnability, i.e., any two states are reachable from each other by some policy. Assumption 1 is the analogous for episodic MDPs, where we only need reachability from the initial state. In episodic MDPs, reachability conditions have been used in different settings, including model-free learning (Modi et al., 2021) and reward-free exploration (Zanette et al., 2020).

Notation Throughout the paper, we shall use $\mathbb{1}_{\mathcal{X}}$ to denote an indicator function over some set $\mathcal{X}$, i.e., $\mathbb{1}_{\mathcal{X}}(h, s, a) := \mathbb{1}\{(h, s, a) \in \mathcal{X}\}$ for all h, s, a. We shall hide $\mathcal{X}$ whenever $\mathcal{X} = [H] \times \mathcal{S} \times \mathcal{A}$.

1. We use $\mathcal{P}(\mathcal{X})$ to denote the set of probability measures over a set $\mathcal{X}$.

2.1. Learning problem

The learner interacts with an MDP $\mathcal{M}$ with unknown transition probabilities in order to fulfill some given *sampling requirements*. In particular, it is given a *target function* $c : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, where $c_h(s,a)$ denotes the minimum number of samples that must be gathered from (s,a) at stage h. In each episode of interaction $t \in \mathbb{N}^*$, the learner plays a policy π^t and observes a corresponding trajectory $\{(s_h^t, a_h^t)\}_{h\in[H]}$. Let $n_h^t(s,a) := \sum_{j=1}^t \mathbb{1}(s_h^j = s, a_h^j = a)$ denote the number of times (s,a) has been visited at stage h up to episode t. The goal is to *minimize* the number of episodes required to collect at least $c_h(s,a)$ samples from each h,s,a with high probability.

Definition 1 (δ-correct c-coverage algorithm) *Fix $\delta \in (0,1)$ and a target function c. An algorithm is called δ-correct c-coverage if, with probability at least $1-\delta$, it stops after interacting with $\mathcal{M}$ for τ episodes and returns a dataset of transitions with visitation counts guaranteeing*

$$\forall (h,s,a),\ n_h^\tau(s,a) \geq c_h(s,a).$$

Examples While the definition of the active coverage problem gives complete freedom in choosing the target function c, for our applications we shall mostly be interested in two specific instances. In *uniform coverage*, we have $c_h(s,a) = N\mathbb{1}\left((h,s,a) \in \mathcal{X}\right)$ for some given set $\mathcal{X}$ and $N \in \mathbb{N}$. Intuitively, this requires collecting at least N samples from each state-action-stage triplet in $\mathcal{X}$, and the name suggests that the learner should explore $\mathcal{X}$ as uniformly as possible. Possible applications include estimating the transition model uniformly well across the state-action space (Tarbouriech et al., 2020) and discovering sparse rewards. In our applications to PAC RL, we will further explore the benefits of performing *proportional coverage*, which corresponds to setting $c_h(s,a) = N \max_\pi p_h^\pi(s,a)\mathbb{1}\left((h,s,a) \in \mathcal{X}\right)$ [2]. This requires collecting a number of samples from each $(h,s,a) \in \mathcal{X}$ that scales proportionally to its reachability.

2.2. The complexity of active coverage

Minimizing the sample complexity required to solve the active coverage problem requires the learner to properly plan how to distribute its exploration throughout the state-action space, hence accounting for the complex interplay between the MDP dynamics p and the target function c. The following theorem gives a precise characterization of the complexity of this problem.

Theorem 2 *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*

$$\varphi^\star(c) = \inf_{\rho\in\Omega} \max_{(s,a,h)\in\mathcal{X}} \frac{c_h(s,a)}{\rho_h(s,a)},$$

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

The quantity $\varphi^\star(c)$ of Theorem 2 provides an *instance-dependent* complexity measure for the active coverage problem. In particular, it depends on both the MDP $\mathcal{M}$ through the set of valid state-action distributions Ω and on the target function c. It can be interpreted as follows. Imagine that a learner repeatedly plays a policy which induces a state-action distribution $\rho \in \Omega$. Then, for any (h,s,a), the quantity $1/\rho_h(s,a)$ is roughly the expected number of episodes the learner takes to collect a single

2. To cope with unknown transitions, we will use an upper bound of $p_h^\pi(s,a)$ in the definition of proportional coverage.

sample from (h, s, a). This implies that $\max_{(s,a,h)\in\mathcal{X}} \frac{c_h(s,a)}{\rho_h(s,a)}$ is roughly the expected number of episodes needed to satisfy the sampling requirements across all (h, s, a) when playing distribution ω. Then, the complexity measure is intuitively the minimum of this quantity across all possible state-action distributions. In other words, any distribution $\rho^\star$ attaining the minimum in $\varphi^\star(c)$ denotes an *optimal* c-coverage distribution, i.e., generating data from $\rho^\star$ provably minimizes the time to satisfy all sampling requirements, in expectation.

We remark that the lower bound of Theorem 2 holds for any δ-correct algorithm, even for an oracle that knows the transition probabilities. In general, we do not believe it to be exactly matchable since (i) any algorithm must work with sample counts rather the expectations, (ii) the transition probabilities are unknown. However, $\varphi^\star(c)$ will appear as the leading order terms in our sample complexity, while these learning costs will be absorbed into lower order terms.

2.3. Links to existing measures of coverage

In Appendix B, we show that $\varphi^\star(c)$ can be reformulated as a *stochastic minimum flow*, a generalization of the minimum flow for directed acyclic graphs (DAGs), as used by Tirinzoni et al. (2022) in deterministic MDPs, to stochastic environments. In this reformulation, $\varphi^\star(c)$ is written as a linear program seeking the minimal allocation of visits to each (h, s, a) (i.e., a flow) that satisfies the sampling requirements while complying with the MDP dynamics.

In Appendix A, we prove that the complexity $\varphi^\star(c)$ satisfies the following inequalities

$$\underbrace{\max_h \sum_{s,a} c_h(s,a)}_{❶} \le \varphi^\star(c) \le \underbrace{\sum_h \inf_{\rho\in\Omega} \max_{s,a} \frac{c_h(s,a)}{\rho_h(s,a)}}_{❷} \le \underbrace{\sum_{h,s,a} \frac{c_h(s,a)}{\max_\pi p_h^\pi(s,a)}}_{❸}. \tag{1}$$

Interestingly, each of these terms relates to a complexity measure that appeared in previous works. Term ❶ is the complexity for covering a tree-based deterministic MDP (Tirinzoni et al., 2022), perhaps the easiest MDP topology to navigate. As $\varphi^\star(c)$ reduces to the complexity of Tirinzoni et al. (2022) in deterministic MDPs, we attain the equality $\varphi^\star(c) = ❶$ in this specific tree structure. For a specific choice of c, ❷ can be shown to be exactly the "gap visitation" complexity measure introduced by Wagenmaker et al. (2022) for BPI. As a component of their BPI algorithm MOCA, Wagenmaker et al. (2022) introduced Learn2Explore, a strategy that learns policies to reach all states in the MDP. While it may be possible to adapt Learn2Explore for our active coverage problem, one limitation is that it learns how to reach each layer independently, and this is reflected on the fact that ❷ is only a loose upper bound (up to a factor H larger) to the optimal complexity $\varphi^\star(c)$. Finally, ❸ can be related to the sample complexity for active coverage obtained by the GOSPRL algorithm of Tarbouriech et al. (2021)[3]. It can be interpreted as the complexity for learning how to reach each h, s, a independently, which makes it an even looser upper bound to $\varphi^\star(c)$.

Concentrability and coverability A definition of *concentrability coefficient* for data distribution ρ is $C_{\text{conc}}(\rho) := \max_{s,a,h} \frac{\max_\pi p_h^\pi(s,a)}{\rho_h(s,a)}$. This plays a fundamental role in characterizing the efficiency of offline RL methods (see, e.g., (Chen and Jiang, 2019; Xie et al., 2022) and references therein). It is easy to see that $\varphi^\star(c) = \inf_{\rho\in\Omega} C_{\text{conc}}(\rho)$ for the target function c of proportional

3. Since Tarbouriech et al. (2021) consider reset-free MDPs, their complexity actually scales as $\sum_{s,a} D_{s,a} c(s,a)$, where $D_{s,a}$ is the minimum expected time to reach s, a from any state. In episodic MDPs, the minimum expected number of episodes to reach some (h, s, a) is exactly $1/\max_\pi p_h^\pi(s,a)$, hence yielding ❸.

Algorithm 1 COVGAME

Input: Target function $c_h(s,a)$, RL algorithm $\mathcal{A}^{\Pi}$, online learning algorithm $\mathcal{A}^{\lambda}$, confidence parameter $\delta \in (0,1)$.
  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 $n^0_h(s,a) = 0$ for all h,s,a
  Reset $\mathcal{A}^{\lambda}$ on $\mathcal{P}(\mathcal{X})$, set $\lambda^1_h(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X})/|\mathcal{X}|$ for all h,s,a
  Initialize $k_1 \leftarrow 0$
  for $t = 1, 2, \dots$ **do**
    Get π^t from $\mathcal{A}^{\Pi}$ given reward function λ^t and confidence $1 - \delta/2$
    Generate a trajectory $\{(s^t_h, a^t_h)\}_{h\in[H]}$ using policy π^t and update counts n^t
    if $n^t_h(s,a) \geq c_h(s,a)$ for all h,s,a **then** stop and return all sampled trajectories
    Update $k_{t+1} \leftarrow \max\{j \in \mathbb{N} : n^t_h(s,a) \geq c_h(s,a)\ \forall (h,s,a) \in \mathcal{X} \setminus \mathcal{X}_j\}$
    if $k_{t+1} \neq k_t$ **then**
      Reset $\mathcal{A}^{\lambda}$ on $\mathcal{P}(\mathcal{X}_{k_{t+1}})$, set $\lambda^{t+1}_h(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X}_{k_{t+1}})/|\mathcal{X}_{k_{t+1}}|$ for all h,s,a
    else
      Feed $\mathcal{A}^{\lambda}$ with loss $\ell^t(\lambda) = \sum_{(h,s,a)\in\mathcal{X}_{k_t}} \lambda_h(s,a)\mathbb{1}(s^t_h = s, a^t_h = a)$, get weight λ^{t+1}

coverage. That is, our coverage complexity is equivalent to the minimum concentrability coefficient achievable by any distribution generated by some stochastic policy. Under a similar perspective, Xie et al. (2022) introduced the *coverability coefficient* $C_{\text{cov}} := \inf_{\rho_1,\dots,\rho_H \in \mathcal{P}(\mathcal{X}\times\mathcal{A})} \max_{s,a,h} \frac{\max_\pi p^\pi_h(s,a)}{\rho_h(s,a)}$ to characterize to what extent the best data distribution covers all policies. Noting that the infimum is taken across all probability distributions rather than valid state-action distributions, the optimal data distribution in C_{cov} may not be attained by the execution of any stochastic policy. This means that C_{cov} is not a valid complexity measure for active coverage in general, and it reduces exactly to ❶ for proportional coverage (see their Lemma 3), i.e., to a loose lower bound on $\varphi^\star(c)$.

3. Active Coverage by Solving Games

We propose COVGAME (Algorithm 1), which adopts a game-based perspective inspired by the bandit literature (Degenne et al., 2019). We first observe that the complexity $\varphi^\star(c)$ can be interpreted as a zero-sum game between a learner trying to produce the best sampling distribution $\rho \in \Omega$ and an adversary trying to challenge it with the tuple (h,s,a) whose sampling requirement is the hardest to meet under ρ. COVGAME does not directly solve the game in the definition of $\varphi^\star(c)$ but rather an equivalent formulation which simplifies learning. Thanks to the minmax theorem, we can write

$$\frac{1}{\varphi^\star(c)} = \sup_{\rho\in\Omega} \min_{(s,a,h)\in\mathcal{X}} \frac{\rho_h(s,a)}{c_h(s,a)} = \sup_{\rho\in\Omega} \inf_{\lambda\in\mathcal{P}(\mathcal{X})} \sum_{(h,s,a)\in\mathcal{X}} \lambda_h(s,a)\frac{\rho_h(s,a)}{c_h(s,a)}$$
$$= \inf_{\lambda\in\mathcal{P}(\mathcal{X})} \max_{\pi\in\Pi^{\text{D}}} \sum_{(h,s,a)\in\mathcal{X}} p^\pi_h(s,a)\frac{\lambda_h(s,a)}{c_h(s,a)},$$

where in the last equation we used that the inner maximization is a standard RL problem with reward function given by $\frac{\lambda_h(s,a)}{c_h(s,a)}\mathbb{1}\left((h,s,a) \in \mathcal{X}\right)$ and its optimum is known to be attained by a deterministic policy (e.g., Puterman, 1994).

COVGAME solves a variant of this minmax game that does not involve the target function c directly. The idea is to cluster the state-action pairs in $\mathcal{X}$ based on their sampling requirement. To this end, we define the sequence of sets $\{\mathcal{X}_k\}_{k\in\mathbb{N}}$ as $\mathcal{X}_0 := \mathcal{X}$ and $\mathcal{X}_k := \{(h,s,a) : c_h(s,a) > c^+_{\min} 2^k\}$ for all $k \in \mathbb{N}^*$, where $c^+_{\min} = \min_{(h,s,a)\in\mathcal{X}} c_h(s,a) \vee 1$. At each round $t \in \mathbb{N}^*$, COVGAME tries to solve the game $\inf_{\lambda\in\mathcal{P}(\mathcal{X}_{k_t})} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{h,s,a} p^\pi_h(s,a)\lambda_h(s,a)$, where k_t is the largest index such that all state-action pairs in $\mathcal{X} \setminus \mathcal{X}_{k_t} = \{(h,s,a) \in \mathcal{X} : c_h(s,a) \le c^+_{\min} 2^{k_t}\}$ have been already covered. Intuitively, COVGAME progressively focuses on covering state-action pairs with larger sampling requirement, while ignoring those that have already been covered. The main advantage over solving the initial formulation of $\varphi^\star(c)$ is two-fold. First, the learner is allowed to play only deterministic policies, each being the solution to an RL problem. Second, in the sequence of games that we consider, the objective function is independent of the scale of c, which avoids undesired dependencies (e.g., on the inverse of the minimum value of c) when the target function is unbalanced.

COVGAME approximately solves the sequence of games above by leveraging two online learning algorithms, $\mathcal{A}^\lambda$ and $\mathcal{A}^\Pi$. The one for the adversary ($\mathcal{A}^\lambda$) can be any method for online convex optimization on the simplex with linear losses. The one for the learner ($\mathcal{A}^\Pi$) can be any regret minimizer for RL that handles reward functions changing at each round (but observed at the beginning of the round). A simple approach like UCBVI (Azar et al., 2017) can be adapted to this purpose.

The final intuition behind COVGAME is quite simple: at each round t, the adversary produces a reward function λ^t supported over $\mathcal{X}_{k_t}$ (the current set to be covered) and the learner tries to find a good policy for maximizing it. This encourages the learner to visit uncovered state-action pairs, eventually meeting the sampling requirements.

In order to analyze the sample complexity of COVGAME, we make the following assumption on the adopted online learning algorithms, which will be satisfied by our specific instance.

Assumption 2 (First-order regret) *There exists a non-decreasing function $\mathcal{R}^\lambda(T)$ such that, if $\mathcal{A}^\lambda$ is instantiated on $\mathcal{P}(\mathcal{X}_k)$ for some k on a sequence of linear losses $\{\ell^t\}_{t\ge1}$ bounded in $[0,1]$,*

$$\forall T \in \mathbb{N}^*,\ \sum_{t=1}^T \ell^t(\lambda^t) - \min_{\lambda\in\Delta_{\mathcal{X}_k}} \sum_{t=1}^T \ell^t(\lambda) \le \sqrt{\mathcal{R}^\lambda(T) \sum_{t=1}^T \ell^t(\lambda^t)} + \mathcal{R}^\lambda(T). \tag{2}$$

There exists a non-decreasing function $\mathcal{R}^\Pi_\delta(T)$ such that, if $\mathcal{A}^\Pi$ is run with confidence $1-\delta$ on a sequence of rewards $\{\lambda^t\}_{t\ge1}$ with $\lambda^t \in \mathcal{P}(\mathcal{X})$ for all t, with probability $1-\delta$, for all $T \in \mathbb{N}^*$,

$$\sum_{t=1}^T V_1^\star\left(s_1;\lambda^t\right) - \sum_{t=1}^T V_1^{\pi_t}\left(s_1;\lambda^t\right) \le \sqrt{\mathcal{R}^\Pi_\delta(T) \sum_{t=1}^T V_1^{\pi^t}\left(s_1;\lambda^t\right)} + \mathcal{R}^\Pi_\delta(T), \tag{3}$$

where $V_1^\pi(s_1;\lambda) := \sum_{h,s,a} p^\pi_h(s,a)\lambda_h(s,a)$ and $V_1^\star(s_1;\lambda) := \max_\pi V_1^\pi\left(s_1;\lambda\right)$.

Theorem 3 (Sample complexity of COVGAME) *Under Assumption 1 and 2, with probability at least $1-\delta$, COVGAME satisfies $n^\tau_h(s,a) \ge c_h(s,a)$ for all h,s,a and its stopping time τ satisfies $\tau \le 64 m \varphi^\star(c) + T_1$, with $m := \lceil \log_2(c_{\max}/c^+_{\min})\rceil \vee 1$, $c_{\max} := \max_{h,s,a} c_h(s,a)$ and*

$$T_1 = \inf\left\{ T \in \mathbb{N}^* : \frac{T}{2} \ge m\varphi^\star(\mathbb{1}_{\mathcal{X}}) \left(3\mathcal{R}^\Pi_{\delta/2}(T) + 12\mathcal{R}^\lambda(T) + 24\log(4T/\delta)\right) + 1 \right\}.$$

While we require both learners to have first-order regret bounds (i.e., depending on the sum of observed losses), standard $\widetilde{O}(\sqrt{T})$ bounds can also be used at the cost of a larger second-order term T_1 in Theorem 3, from $T_1 = \widetilde{O}(\varphi^\star(\mathbb{1}_\mathcal{X}))$ as in our instantiation to $T_1 = \widetilde{O}(\varphi^\star(\mathbb{1}_\mathcal{X})^2)$. The key step in our proof is to show that first-order regret implies convergence to the value $\varphi^\star(c)$ of the game at a rate $\widetilde{O}(1/T)$ instead of the slower $\widetilde{O}(1/\sqrt{T})$ achieved with $\widetilde{O}(\sqrt{T})$ regret. As $\varphi^\star(\mathbb{1}_\mathcal{X})$ depends on the inverse visitation probabilities (see Theorem 2), this $\varphi^\star(\mathbb{1}_\mathcal{X})$ versus $\varphi^\star(\mathbb{1}_\mathcal{X})^2$ improvement will be crucial to avoid undesired scaling with these quantities in our applications to PAC RL.

3.1. Our instantiation

For $\mathcal{A}^\lambda$ we propose to use the weighted majority forecaster (WMF, Littlestone and Warmuth, 1994) with variance-dependent learning rate for which, for any sequence of losses bounded in $[0,1]$, we have by Theorem 5 of Cesa-Bianchi et al. (2005) that Assumption 2 is satisfied with

$$\mathcal{R}^\lambda(T) = 16\log(SAH). \tag{4}$$

For $\mathcal{A}^\Pi$ we propose to use a variant of UCBVI (Azar et al., 2017) that can cope with varying reward functions. The idea is that, since the reward function λ^t is revealed to $\mathcal{A}^\Pi$ at the beginning of round t, we can build an upper confidence bound $\overline{Q}_h^{t-1}(s,a;\lambda^t)$ to the optimal action-value function $Q_h^\star(s,a;\lambda^t)$ by estimating the transition probabilities with the data collected up to round $t-1$. Then, we play $\pi_h^t(s) = \arg\max_a \overline{Q}_h^{t-1}(s,a;\lambda^t)$, the greedy policy w.r.t. $\overline{Q}_h^{t-1}$. We build the UCBs by leveraging the same "monotonic value propagation" trick from Zhang et al. (2021c) and prove that Assumption 2 is satisfied with

$$\mathcal{R}_\delta^\Pi(T) = 65536SAH^2(\log(2SAH/\delta) + 6S)\log(T+1)^2. \tag{5}$$

See Appendix C for details. Notably, we manage to prove a similar first-order regret bound as the one derived by Jin et al. (2020) for EULER (Zanette and Brunskill, 2019b) with a remarkably simple analysis, without using any correction factor in the bonuses, and with improved dependences on H (from H^4 to H^2) and δ (from $\log(1/\delta)^3$ to $\log(1/\delta)$). As compared to the minimax regret rate (Azar et al., 2017), our resulting bound in (3) features a dependence on S instead of $\sqrt{S}$ in its leading-order term. This is the cost of handling changing rewards, which prevents us from building tight UCBs as commonly done for a fixed reward function. Instead, we build UCBs that hold for all rewards simultaneously using techniques from reward-free exploration (Ménard et al., 2021), a setting where an extra dependence on S is unavoidable in the worst case (Jin et al., 2020). Time-varying rewards, albeit under a weaker notion of regret, have also been studied in an adversarial setting in which the reward λ^t is not revealed prior to round t (Rosenberg and Mansour, 2019).

Corollary 4 (Sample complexity of CovGame with WMF and UCBVI) *With probability at least $1-\delta$, the stopping time of* CovGame *with* WMF *and* UCBVI *is bounded by*

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

where $m = \lceil \log_2(c_{\max}/c_{\min}^+)\rceil \vee 1$ and $\widetilde{O}$ hides poly-logarithmic factors in $S, A, H, \varphi^\star(\mathbb{1}_\mathcal{X}), \log(1/\delta)$.

The second term in the bound above can be interpreted as the cost incurred for learning the optimal coverage complexity $\varphi^\star(c)$ under *unknown* transition probabilities p. Still, this *learning*

cost depends at most logarithmically on the total sampling requirement $\|c\|_1 = \sum_{h,s,a} c_h(s,a)$. This implies that, for large $\|c\|_1$, this cost becomes negligible as compared to the first term and $\tau \leq \widetilde{O}(\varphi^\star(c))$, which matches the lower bound of Theorem 2 up to constant and logarithmic terms. We observe that if p is known, by replacing UCBVI with the computation of the optimal policy w.r.t. to λ^t, for which $\mathcal{R}^{\Pi}_{\delta/2}(T) = 0$, we get a smaller additive cost $\widetilde{O}(m\varphi^\star(\mathbb{1}_{\mathcal{X}})\log(SAH)\log(1/\delta))$ which is only due to the randomness in the collection of trajectories.

3.2. Comparison with prior work

While inspired by an original game perspective which is crucial in our analysis, the actual algorithmic approach of COVGAME has a similar flavor as existing algorithms for different exploration tasks: it runs a regret minimizer on different reward functions enforcing the visitation of uncovered states. Using WMF as the λ-learner, the reward function in round t is

$$\lambda_h^{t+1}(s,a) = \frac{\exp\left(-\xi_{t-i_t}\left(n_h^t(s,a) - n_h^{i_t}(s,a)\right)\right)\mathbb{1}\left((h,s,a)\in\mathcal{X}_{k_t}\right)}{\sum_{(h',s',a')\in\mathcal{X}_{k_t}}\exp\left(-\xi_{t-i_t}\left(n_{h'}^t(s',a') - n_{h'}^{i_t}(s',a')\right)\right)},$$

where i_t is the last restart of WMF that happened before t and ξ_t is the variance-dependent learning rate defined by Cesa-Bianchi et al. (2005). Our reward function is related to the number of prior visits and smoothly evolves over time, which is in contrasts with most prior approaches that rely on rewards of the form $r_h^{\mathcal{Y}}(s,a) = \mathbb{1}((h,s,a)\in\mathcal{Y})$ for some set $\mathcal{Y}$, For example, GOSPRL translated to our episodic setting would use $r_h^{t+1}(s,a) = \mathbb{1}\left(n_h^t(s,a) < c_h^t(s,a)\right)$. The Learn2Explore strategy (Wagenmaker et al., 2022) uses a subroutine to visit N times some of the state-action pairs in $\mathcal{Y}$: it runs EULER (Zanette and Brunskill, 2019a) on $r^{\mathcal{Y}}$ and restarts the algorithm with a reward function with reduced support whenever some new state-action pair has reached N visits. Several algorithms for RFE (Jin et al., 2020; Zhang et al., 2021a) also collect data using regret minimizers on top of indicator-based rewards. In Appendix B.3, we further discuss the connections between COVGAME and Frank-Wolfe approaches used in the convex RL literature.

4. Applications to PAC RL

A strategy for RFE should collect a dataset of trajectories from which it is possible to compute a near-optimal policy for any reward function. To be robust to any possible reward in the test phase, we intuitively need to gather sufficient samples everywhere in the MDP, which we propose to do explicitly by relying on COVGAME with proportional coverage (Section 4.1). By adding some ingredients to this exploration strategy, we further obtain a new algorithm for BPI (Section 4.2).

4.1. Proportional Coverage Exploration (PCE)

Algorithm 2 takes as input two parameters ε, δ and returns an estimate of the transition probabilities $\widehat{p}$ that, with probability $1-\delta$, yields an ε-optimal policy for any reward function bounded in $[0,1]$. The choice of proportional coverage is motivated by a novel *ellipsoid-shaped confidence region* for the value functions of all policies under any reward. Let $\widehat{p}^t$ denote the maximum likelihood estimator of p after observing t episodes. For any reward function r, let $V_1^\pi(s_1;r) := \sum_{h,s,a} p_h^\pi(s,a) r_h(s,a)$ be the expected return of π, and $\widehat{V}_1^{\pi,t}(s_1;r)$ be the same on the empirical MDP with transitions $\widehat{p}^t$.

Algorithm 2 PCE (Proportional Coverage Exploration)

1: **Input:** Precision ε, Confidence δ.
2: For each (h,s), run ESTIMATEREACHABILITY$((h,s); \frac{\varepsilon}{4SH^2}, \frac{\delta}{3SH})$ to get confidence intervals $\big[\underline{W}_h(s), \overline{W}_h(s)\big]$ on $\max_\pi p_h^\pi(s)$ (see Appendix G)
3: Define $\widehat{\mathcal{X}} := \{(h,s,a) : \underline{W}_h(s) \geq \frac{\varepsilon}{32SH^2}\}$
4: Define target function $c_h^0(s,a) = \mathbb{1}\big((h,s,a) \in \widehat{\mathcal{X}}\big)$ for all (h,s,a)
5: Execute COVGAME$\big(c^0,\ \delta/6\big)$ to get a dataset $\mathcal{D}_0$ of d_0 episodes // BURN-IN PHASE
6: Initialize episode count $t_0 \leftarrow d_0$ and statistics $n_h^0(s,a), \widehat{p}_h^0(.|s,a)$ using $\mathcal{D}_0$
7: **for** $k = 1, \dots$ **do**
8: // PROPORTIONAL COVERAGE
9: Compute targets $c_h^k(s,a) := 2^k \overline{W}_h(s) \mathbb{1}\big((h,s,a) \in \widehat{\mathcal{X}}\big)$ for all (h,s,a)
10: Execute COVGAME$\big(c^k,\ \delta/6(k+1)^2\big)$ to get dataset $\mathcal{D}_k$ and number of episodes d_k
11: Update episode count $t_k \leftarrow t_{k-1} + d_k$ and statistics $n_h^k(s,a), \widehat{p}_h^k(.|s,a)$ using $\mathcal{D}_k$
12: **if** $\sqrt{H\beta^{RF}(t_k, \delta/3)2^{4-k}} \leq \varepsilon$ **then** stop and return $\widehat{p}^k$
13: **end for**

Theorem 27 in Appendix D gives that, with probability $1-\delta$, jointly over all episodes t,

$$\forall r \in [0,1]^{SAH},\ \ \forall \pi \in \Pi^D,\ \ \big|V_1^\pi(s_1;r) - \widehat{V}_1^{\pi,t}(s_1;r)\big| \leq \sqrt{\beta^{\mathrm{RF}}(t,\delta) \sum_{(h,s,a)\in\mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + \frac{\varepsilon}{4}, \quad (6)$$

where $\beta^{\mathrm{RF}}(t,\delta) \propto H^2 \log(1/\delta) + SH^3 \log(A(1+t))$ and $\mathcal{X}_\varepsilon$ is a subset of triplets that are not too hard to reach: $\mathcal{X}_\varepsilon \subseteq \{(h,s,a) : \max_\pi p_h^\pi(s,a) \geq \frac{\varepsilon}{4SH^2}\}$. If we gather $c_h(s,a) = \mathcal{O}(H\beta^{\mathrm{RF}}(t,\delta) \sup_\pi p_h^\pi(s,a)/\varepsilon^2)$ visits from every $(h,s,a) \in \mathcal{X}_\varepsilon$, then the estimation error of $V_1^\pi(s_1;r)$ for any π and r is below $\varepsilon/2$, which is sufficient to solve RFE (Jin et al., 2020).

Yet as the visitation probabilities are unknown, neither $\mathcal{X}_\varepsilon$ nor $c_h(s,a)$ can actually be computed. To solve this issue, we rely on an initialization phase based on the ESTIMATEREACHABILITY subroutine (line 2 of Algorithm 2), described in Appendix G. This procedure, that is similar to the initialization phase in MOCA (Wagenmaker et al., 2022), outputs for each (h,s) an interval $[\underline{W}_h(s), \overline{W}_h(s)]$ to which $\max_\pi p_h^\pi(s)$ belongs with high probability using a low-order number of episodes of $\widetilde{O}(S^3AH^4/\varepsilon)$. The lower confidence bound is then used to build a set $\widehat{\mathcal{X}}$ that satisfies the requirements for $\mathcal{X}_\varepsilon$ and the upper bound is used to define the target function that is given as input to COVGAME in phase k of the algorithm: $c_h^k(s,a) := 2^k \overline{W}_h(s) \mathbb{1}\big((h,s,a) \in \widehat{\mathcal{X}}\big)$.

We remark that PCE is computationally-efficient as it inherits the complexity of COVGAME and ESTIMATEREACHABILITY, which both require to solve one dynamic program in every round to compute the optimistic policy used by UCBVI. We now present its theoretical properties.

Theorem 5 *Let $\widehat{p}$ be the estimate of the transition probabilities that* PCE *outputs. For any reward function r, let $\hat{\pi}_r$ be an optimal policy in the MDP $(\widehat{p}, r)$. Then,*

$$\mathbb{P}\left(\forall r \in [0,1]^{SAH}, |V_1^{\hat{\pi}_r}(s_1;r) - V_1^\star(s_1;r)| \leq \varepsilon\right) \geq 1-\delta.$$

Furthermore, with probability at least $1-\delta$, the total sample complexity of PCE *satisfies*

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

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in S, A, H, ε and $\log(1/\delta)$.

Perhaps the most interesting feature of this bound is that in the regime of small ε and small δ, the leading term is $H^3 \log(1/\delta)\varphi^\star\big([\sup_\pi p_h^\pi(s)\mathbb{1}(\sup_\pi p_h^\pi(s) \geq \varepsilon/(32SH^2))]_{h,s,a}\big)/\varepsilon^2$, which can be much smaller than the $(SAH^3/\varepsilon^2)\log(1/\delta)$ minimax rate (Ménard et al., 2021). First, using the inequality (1), this term is always smaller than $|\{(h,s) : \sup_\pi p_h^\pi(s) \geq \varepsilon/(32SH^2)\}|AH^3\log(1/\delta)$, which can be better than minimax in MDPs with many states that are hard to reach. Other examples of MDPs for which PCE is better than minimax in the small ε, δ regime are given in Appendix E.7. For any $\alpha \in [0,1)$, we notably propose a family of MDPs satisfying $\varphi^\star\big([\sup_\pi p_h^\pi(s)]_{h,s,a}\big) = \mathcal{O}(S^\alpha AH)$, leading to an asymptotic sample complexity of order $\big(S^\alpha AH^4/\varepsilon^2\big)\log(1/\delta)$. These examples suggest that, while RFE is by essence a worst-case problem, there is still hope to adapt to the "explorability" of the MDP. Beyond this asymptotic regime, a worst-case bound can be directly extracted from Theorem 5 for any ε, δ by using that the $\varphi^\star$ term is at most SAH/ε^2,

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

which is *minimax optimal* up to an H^2 factor and low-order terms scaling in $1/\varepsilon$.

Remark 6 (Reachability) *Thanks to its initialization phase, PCE can be used even when Assumption 1 is violated. All triplets that have zero probability to be reached are filtered out from the set $\widehat{\mathcal{X}}$, and* CovGame *always targets reachable states.*

4.2. PRINCIPLE: PRoportIoNal Coverage with Implicit PoLicy Elimination

Our second use-case of CovGame yields PRINCIPLE, an algorithm for BPI. Given an unknown reward distribution $\{\nu_h(s,a)\}_{h,s,a}$ with support in $[0,1]$ and mean $\{r_h(s,a)\}_{h,s,a}$, an (ε,δ)-PAC algorithm for BPI outputs a policy $\widehat{\pi}$ such that $\mathbb{P}\left(V_1^{\widehat{\pi}}(s_1;r) \geq V_1^\star(s_1;r) - \varepsilon\right) \geq 1-\delta$.

In the PCE algorithm, we sought to achieve good proportional coverage w.r.t. the set of all policies, i.e., by requiring that $n_h^k(s,a) \geq 2^k \sup_{\pi\in\Pi^D} p_h^\pi(s,a)$ for all h,s,a,k. This is due to the "worst-case" nature of RFE, where any policy can be potentially optimal for some reward function at test time. On the contrary, the mean-reward r is fixed in BPI, a property that we can leverage to perform more adaptive exploration. A natural idea, which led to tight theoretical guarantees in recent works (Tirinzoni et al., 2022; Wagenmaker and Jamieson, 2022), is to eliminate policies as soon as we are confident enough that they are sub-optimal, so that the algorithm can adapt its exploration to focus on policies of higher value. Unfortunately, while Tirinzoni et al. (2022) managed to achieve so in a computationally-efficient manner for deterministic MDPs, the approach of Wagenmaker and Jamieson (2022) needs to enumerate all policies to do the same in stochastic environments, hence yielding an exponential time-memory algorithm. Our method, PRINCIPLE, achieves the same while remaining computationally efficient. Due to space constraints, we report its full pseudo-code in Appendix F.2, while here we highlight its core technique.

Implicit policy elimination The key idea is to replace explicit policy eliminations by sequentially constraining the set of state-action distributions corresponding to high-reward policies. In particular, PRINCIPLE maintains, at each phase k, a high-probability lower bound $\underline{V}_1^k$ on the optimal expected

return $V_1^\star(s_1;r)$ computed as

$$\underline{V}_1^k := \sup_{\substack{\rho\in\Omega(\widehat{p}^k),\\ \max_{h,s,a}\rho_h(s,a)/n_h^k(s,a)\le 2^{-k}}} \sum_{h,s,a}\rho_h(s,a)\widehat{r}_h^k(s,a) - \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/3)},$$

where $\beta^{bpi}(t,\delta)\propto H^2\log(1/\delta)+SAH^3\log\log(t)$ and $\Omega(\widehat{p}^k)$ is the set of valid visitation probabilities in the empirical MDP with transition kernel $\widehat{p}^k$. As common, $\underline{V}_1^k$ is computed by subtracting a confidence interval to the maximum expected return estimated on the empirical MDP defined by $(\widehat{p}^k,\widehat{r}^k)$. A notable exception is that we focus only on state-action distributions that are *well-covered* by the current data. Then, PRINCIPLE defines a set of "active" state-action distributions as

$$\Omega^k := \left\{\rho\in\Omega(\widehat{p}^k) : \sum_{h,s,a}\rho_h(s,a)\widehat{r}_h^k(s,a)\ge \underline{V}_1^k,\ \max_{h,s,a}\rho_h(s,a)/n_h^k(s,a)\le 2^{-k}\right\}.$$

Intuitively, ρ is active at phase k if (1) it is a valid state-action distribution in the empirical MDP with transition probabilities $\widehat{p}^k$, (2) it induces an estimated expected return $\sum_{h,s,a}\rho_h(s,a)\widehat{r}_h^k(s,a)$ larger than $\underline{V}_1^k$, and (3) it is well-covered by the current data. Then, as compared to PCE, PRINCIPLE simply replaces the quantity $\sup_{\pi\in\Pi^D}p_h^\pi(s,a)$ in the target function used for COVGAME at phase k with $\sup_{\rho\in\Omega^{k-1}}\rho_h(s,a)$, i.e., it restricts the exploration to active state-action distributions. In our analysis, we show that, with high probability, state-action distributions corresponding to optimal policies are never eliminated from Ω^k and $\underline{V}_1^k$ gradually approaches $V_1^\star(s_1;r)$ from below. That is, Ω^k is dynamically pruned to contain only distributions corresponding to higher returns, hence achieving implicit eliminations of sub-optimal policies.

Computational complexity The computations of $\underline{V}_1^k$ and $\sup_{\rho\in\Omega^{k-1}}\rho_h(s,a)$ amount to solving standard constrained MDPs, which can be done by linear programming (e.g., Efroni et al., 2020). Moreover, PRINCIPLE does not store the set Ω^k but only its associated constraints, whose number is linear in SAH. This implies that PRINCIPLE requires polynomial (in SAH) time and memory.

Theoretical guarantees We prove that PRINCIPLE enjoys an instance-dependent complexity that scales with policy gaps and visitation probabilities.

Theorem 7 *PRINCIPLE is (ε,δ)-PAC for BPI and, with probability $1-\delta$, it has sample complexity*

$$\tau\le\widetilde{\mathcal{O}}\left((H^3\log(1/\delta)+SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)+\frac{\varphi^\star(\mathbb{1})}{\varepsilon}+\varphi^\star(\mathbb{1})\right]\right),$$

where $\Delta(\pi):=V_1^\star(s_1;r)-V_1^\pi(s_1;r)$ denotes the policy gap of π, $\mathbb{1}$ denotes a function equal to 1 for all h,s,a, and $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S,A,H,\varepsilon,\log(1/\delta)$ and $\varphi^\star(\mathbb{1})$.

Comparison with prior work Besides PRINCIPLE, there exist mostly two BPI algorithms with instance-dependent guarantees for MDPs with stochastic transitions: MOCA (Wagenmaker et al., 2022) and PEDEL (Wagenmaker and Jamieson, 2022). In the small (ε,δ) regime, the leading term in the sample complexity of these three algorithms is of the form $\mathrm{Alg}(\mathcal{M},\varepsilon)\log(1/\delta)$. We carefully compare these terms in Appendix F.3. Notably, while PRINCIPLE$(\mathcal{M},\varepsilon)$ and PEDEL$(\mathcal{M},\varepsilon)$ are both expressed with policy gaps, MOCA$(\mathcal{M},\varepsilon)$ depends on the value gaps $V_h^\star(s)-Q_h^\star(s,a)$. In

general, value gaps are known to be worse than policy gaps (Dann et al., 2021; Tirinzoni et al., 2021) and, while there is no clear ordering between PRINCIPLE and MOCA (just like PEDEL and MOCA, see Wagenmaker and Jamieson (2022)), we can exhibit instances in which the complexity of the former has a better scaling than that of the latter.

Lemma 8 *For any $\Delta \in (0, 1]$, there exists an MDP $\mathcal{M}$ where*

$$\textit{MOCA}(\mathcal{M}, \varepsilon) = \Omega\left(\frac{H^5 SA}{\varepsilon^2}\right) \textit{ while } \textit{PRINCIPLE}(\mathcal{M}, \varepsilon) = \mathcal{O}\left(\frac{H^4 SA}{\varepsilon\Delta} + \frac{H^4 \log(S)\log(A)}{\varepsilon^2}\right).$$

On the other hand, PEDEL directly minimizes the confidence interval (6) over all (active) policies, an objective that is always upper bounded by the complexity of proportional coverage:

$$\min_{\rho\in\Omega} \max_{\pi\in\Pi^D} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)} \leq \min_{\rho\in\Omega} \max_{h,s,a} \frac{\sup_\pi p_h^\pi(s,a)}{\rho_h(s,a)}.$$

We prove in Appendix F.3 that the complexity of PEDEL is indeed smaller (up to H factors) than that of PRINCIPLE. However, this objective may be intractable in general due to the maximization over all deterministic policies. On the other hand, proportional coverage is sufficient (though less statistically-efficient) to estimate the value of all policies and can be done in polynomial time. Besides optimistic algorithms whose sample complexity features policy gaps but with an extra suboptimal scaling in the *minimal* visitation probability (Tirinzoni et al., 2023), this makes PRINCIPLE the first computationally efficient BPI algorithm whose sample complexity scales with policy gaps.

5. Conclusion

We proposed COVGAME, a simple algorithm that adaptively collects episodes in an MDP to explicitly gather a required number of samples $c_h(s,a)$ from each triplet (h,s,a). We proved that its sample complexity scales with a new notion of optimal coverage $\varphi^\star(c)$, which is an instance-dependent lower bound on the sample complexity of *any* adaptive coverage algorithm. We then illustrated the use of COVGAME as a building block for PAC reinforcement learning algorithms. By relying on (an optimistic variant of) proportional coverage, we proposed an algorithm for reward-free exploration with an instance-dependent sample complexity bound. Further combining proportional coverage with an implicit policy elimination scheme, we obtained the first computationally efficient algorithm for best policy identification whose sample complexity scales with policy gaps. To assess the quality of these approaches, in future work we will investigate instance-dependent lower bounds on the sample complexity of PAC RL algorithms, that are currently missing in the literature.

Acknowledgments

Aymen Al-Marjani ackowledges the support of the Chaire SeqALO (ANR-20-CHIA-0020). Emilie Kaufmann acknoweldges the support of the French National Research Agency under the BOLD project (ANR-19-CE23-0026-04).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 24, 2011.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, 2017.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Nicolò Cesa-Bianchi, Y. Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2005.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. In *Advances in Neural Information Processing Systems*, 2022.

Wang Chi Cheung. Exploration-exploitation trade-off in reinforcement learning on online markov decision processes with global concave rewards. *arXiv preprint arXiv:1905.06466*, 2019.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5713–5723, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

Christoph Dann, Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14492–14501, 2019.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted q-iteration for planning in continuous-space markovian decision problems. In *2009 American Control Conference*, pages 725–730. IEEE, 2009.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems (NeurIPS)*, 23, 2010.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the Seventh Conference on Computational Learning Theory (COLT)*, 1994.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pages 3489–3489. PMLR, 2022.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *AAMAS*, 2022.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning (ICML)*, 2019.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory (ALT)*, 2021.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. Urlb: Unsupervised reinforcement learning benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2021.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*, 2021.

Rémi Munos. Error bounds for approximate policy iteration. In *International Conference on Machine Learning (ICML)*, 2003.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, pages 16223–16239. PMLR, 2022.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pages 8371–8380. PMLR, 2021.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA, 1st edition, 1994. ISBN 0471619779.

Clémence Réda, Andrea Tirinzoni, and Rémy Degenne. Dealing with misspecification in fixed-confidence linear top-m identification. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Aviv Rosenberg and Y. Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Jean Tarbouriech, Shubhanshu Shekhar, Matteo Pirotta, Mohammad Ghavamzadeh, and Alessandro Lazaric. Active model estimation in markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, pages 1019–1028. PMLR, 2020.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:7611–7624, 2021.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. Near instance-optimal pac reinforcement learning for deterministic mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. Optimistic PAC reinforcement learning: the instance-dependent view. In *Algorithmic Learning Theory (ALT)*, 2023.

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

Andrew Wagenmaker, Max Simchowitz, and Kevin G. Jamieson. Beyond no regret: Instance-dependent PAC reinforcement learning. In *Conference On Learning Theory (COLT)*, 2022.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pages 11920–11931. PMLR, 2021.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, 2019a.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 2019b.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:11756–11766, 2020.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning, (ICML)*, 2021a.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pages 12402–12412. PMLR, 2021b.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021c.

Contents

Appendix A. Optimal Coverage and Stochastic Minimum Flows

In this appendix, we present an equivalent linear programming formulation of the optimal coverage problem of Section 2.2 that we call *stochastic minimum flow*. It is a direct extension to stochastic MDPs of the minimum flows for directed acyclic graphs employed by Tirinzoni et al. (2022) in deterministic MDPs.

A.1. Stochastic minimum flows

We define a *flow* as a non-negative function $\eta : \mathcal{S} \times \mathcal{A} \times [H] \to [0, \infty)$ such that

$$\sum_{a \in \mathcal{A}} \eta_h(s,a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{h-1}(s|s',a') \eta_{h-1}(s',a') \quad \forall s \in \mathcal{S}, h > 1,$$
$$\eta_1(s,a) = 0 \quad \forall s \in \mathcal{S} \setminus \{s_1\}, a \in \mathcal{A}.$$

That is, a flow η is an allocation of visits to each state-action-stage triplet which satisfies the *navigation constraints* of the MDP. Note that the second constraint ensures that flow can only be created in the initial state s_1. The value of η is the total amount of flow leaving the initial state, i.e.,

$$\varphi(\eta) := \sum_{a \in \mathcal{A}} \eta_1(s_1, a).$$

Let $c : \mathcal{S} \times \mathcal{A} \times [H] \to [0, \infty)$ be a non-negative target function. We say that a flow η is *feasible* for c if

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

The *stochastic minimum flow* problem consists in finding a feasible flow of minimum value. It can be clearly solved as a linear program,

$$\begin{aligned}
&\underset{\eta \in \mathbb{R}^{SAH}}{\text{minimize}} \sum_{a \in \mathcal{A}} \eta_1(s_1, a), \\
&\text{subject to} \\
&\quad \sum_{a \in \mathcal{A}} \eta_h(s,a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{h-1}(s|s',a') \eta_{h-1}(s',a') \quad \forall s \in \mathcal{S}, h > 1, \\
&\quad \eta_1(s,a) = 0 \quad \forall s \in \mathcal{S} \setminus \{s_1\}, a \in \mathcal{A}, \\
&\quad \eta_h(s,a) \geq c_h(s,a) \quad \forall h \in [H], s \in \mathcal{S}, a \in \mathcal{A}.
\end{aligned} \tag{7}$$

We now prove that the optimal value of (7) is equal to $\varphi^\star(c)$, the optimal coverage complexity introduced in Section 2.2.

Lemma 9 *If there exists a feasible flow for the target function c, the optimal value of (7) is*

$$\varphi^\star(c) = \min_{\rho \in \Omega} \max_{h,s,a} \frac{c_h(s,a)}{\rho_h(s,a)}.$$

Proof Let us start from the linear programming formulation (7) and perform the change of variables $\rho_h(s,a) \leftarrow \frac{\eta_h(s,a)}{Z}$ and $Z \leftarrow \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \eta_h(s',a')$ for all h, s, a. Note that Z is the value of the original flow η (and thus it does not depend on the stage), while $\rho_h(s,a)$ is a probability distribution over the state-action space for each $h \in [H]$. We obtain the following optimization problem (no longer a linear program due to the presence of a bilinear constraint):

$$
\begin{aligned}
&\underset{Z \geq 0, \rho \in \mathbb{R}^{SAH}}{\text{minimize}} \; Z, \\
&\text{subject to} \\
&\quad \sum_{a \in \mathcal{A}} \rho_h(s,a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{h-1}(s|s',a') \rho_{h-1}(s',a') \quad \forall s \in \mathcal{S}, h > 1, \\
&\quad \rho_1(s,a) = 0 \quad \forall s \in \mathcal{S} \setminus \{s_1\}, a \in \mathcal{A}, \\
&\quad \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \rho_h(s,a) = 1 \quad \forall h \in [H], \\
&\quad \rho_h(s,a) \geq 0 \quad \forall h \in [H], s \in \mathcal{S}, a \in \mathcal{A}, \\
&\quad Z \geq \frac{c_h(s,a)}{\rho_h(s,a)} \quad \forall h \in [H], s \in \mathcal{S}, a \in \mathcal{A}.
\end{aligned}
$$

The optimal solution for Z is clearly $Z = \max_{h,s,a} \frac{c_h(s,a)}{\rho_h(s,a)}$, while the first four constraints define exactly the set of valid state-action distributions Ω. This proves the statement. ■

Lemma 10 *For any $\alpha, \beta \geq 0$ and target functions c_1, c_2, $\varphi^\star(\alpha c_1 + \beta c_2) \leq \alpha \varphi^\star(c_1) + \beta \varphi^\star(c_2)$.*

Proof Clearly, $\varphi^\star(\alpha c_1) = \alpha \varphi^\star(c_1)$ by definition for any $\alpha \geq 0, c_1$. From the LP formulation, we note that if $\eta_1^\star$ (resp. $\eta_2^\star$) is an optimal flow for c_1 (resp. c_2), then $\eta_1^\star + \eta_2^\star$ is a feasible flow for $c_1 + c_2$. This implies that $\varphi^\star(c_1 + c_2) \leq \varphi^\star(c_1) + \varphi^\star(c_2)$ for any c_1, c_2, which proves the statement. ■

A.2. Executing a minimum flow

Suppose we computed a solution $\eta_h^\star(s,a)$ to the stochastic minimum flow problem (7), or equivalently a solution $\rho_h^\star(s,a)$ to the coverage complexity $\varphi^\star(c)$. What policy should we execute in the MDP to realize the flow? The answer comes easily from standard MDP theory (Puterman, 1994): it is enough to execute a stochastic policy

$$
\pi_h(a|s) = \frac{\eta_h^\star(s,a)}{\sum_{b \in \mathcal{A}} \eta_h^\star(s,b)} = \frac{\rho_h^\star(s,a)}{\sum_{b \in \mathcal{A}} \rho_h^\star(s,b)} \; \forall h, s, a. \tag{8}
$$

It is then easy to prove that π realizes the the optimal distribution $\rho_h^\star(s,a)$.

Proposition 11 *Let π be the policy defined in (8), then, for each h, s, a,*

$$
p_h^\pi(s,a) = \frac{\eta_h^\star(s,a)}{\sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \eta_h^\star(s',a')} = \rho_h^\star(s,a).
$$

Proof This is a well-known result (e.g., Puterman, 1994). For completeness, let us prove it by induction. Note that $\rho^\star$ is the normalization of $\eta^\star$ by definition. Clearly, the statement holds at $h = 1$ since, for all actions $a \in \mathcal{A}$,

$$p_1^\pi(s_1, a) = \pi_1(a|s_1) = \frac{\rho_1^\star(s_1, a)}{\sum_b \rho_1^\star(s_1, b)} = \rho_1^\star(s_1, a),$$

and $p_1^\pi(s, a) = \rho_1^\star(s, a) = 0$ for all other states. Suppose the statement holds at $h - 1 \geq 1$. Then,

$$\begin{aligned} p_h^\pi(s, a) &= \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \underbrace{p_{h-1}^\pi(s', a')}_{=\rho_{h-1}^\star(s', a')} p_{h-1}(s|s', a') \pi_h(a|s) \\ &= \underbrace{\sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \rho_{h-1}^\star(s', a') p_{h-1}(s|s', a')}_{=\sum_{b \in \mathcal{A}} \rho_h^\star(s, b)} \frac{\rho_h^\star(s, a)}{\sum_{b \in \mathcal{A}} \rho_h^\star(s, b)} = \rho_h^\star(s, a). \end{aligned}$$

■

Note that the denominator in the expression of $p_h^\pi(s, a)$ is equal to $\varphi^\star(c)$ for any $h \in [H]$. Thus, we have $p_h^\pi(s, a) = \eta_h^\star(s, a)/\varphi^\star(c)$. If we execute π for $t = \lceil \varphi^\star(c) \rceil$ episodes, we have that

$$\mathbb{E}[n_h^t(s, a)] = \frac{\lceil \varphi^\star(c) \rceil}{\varphi^\star(c)} \eta_h^\star(s, a) \geq \eta_h^\star(s, a) \geq c_h(s, a) \; \forall h, s, a.$$

Hence, we realize the flow in expectation.

A.3. Bounding the minimum flow

We are interested in upper and lower bounding the value of the stochastic minimum flow $\varphi^\star(c)$ as a function of c. We start by deriving some simple (probably loose) bounds.

Lemma 12 *Suppose there exists a feasible flow for the target function c. Then,*

$$\max_{h \in [H]} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} c_h(s, a) \leq \varphi^\star(c) \leq \sum_{h \in [H]} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{c_h(s, a)}{\max_\pi p_h^\pi(s, a)}.$$

Proof The proof of the lower bound is trivial by noting that the value of any flow η can be written as $\varphi(\eta) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \eta_h(s, a)$ for all $h \in [H]$ and that any optimal flow satisfies $\eta_h^\star(s, a) \geq c_h(s, a)$ for all h, s, a. Let us prove the upper bound.

Let us define $w_h(s, a) := \frac{c_h(s,a)}{\max_{\pi \in \Pi} p_h^\pi(s,a)}$, with the convention that $w_h(s, a) = 0$ if $c_h(s, a) = 0$ regardless of the value of the denominator. Note that, if $\max_{\pi \in \Pi} p_h^\pi(s, a) = 0$, then (s, a, h) is unreachable and it must be that $c_h(s, a) = 0$ since we assumed the minimum flow problem to be feasible. For any reachable (s, a, h), let $\pi_{s,a,h} \in \arg\max_{\pi \in \Pi} p_h^\pi(s, a)$. For any unreachable (s, a, h), let $\pi_{s,a,h}$ be an arbitrary deterministic policy. Let us define the following mixed state-action distribution:

$$\forall h, s, a : \tilde{p}_h(s, a) := \sum_{l \in [H]} \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \frac{w_l(s', a')}{Z} p_h^{\pi_{s', a', l}}(s, a),$$

where $Z := \sum_{l\in[H]}\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} w_l(s',a')$. Since this is a convex combination of state-action distributions of deterministic policies (i.e., of $\{\pi_{s,a,h}\}_{s,a}$), $\tilde{p}\in\Omega$ (Puterman, 1994). Then,

$$\varphi^\star(c) = \min_{\rho\in\Omega}\max_{h,s,a}\frac{c_h(s,a)}{\rho_h(s,a)} \le \max_{h,s,a}\frac{c_h(s,a)}{\tilde{p}_h(s,a)} \le Z\max_{h,s,a}\frac{c_h(s,a)}{w_h(s,a)p_h^{\pi_{s,a,h}}(s,a)}$$
$$= \sum_{h\in[H]}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{c_h(s,a)}{\max_\pi p_h^\pi(s,a)}.$$

■

Lemma 13 *Suppose there exists a feasible flow for the lower bound function c. Then,*

$$\varphi^\star(c) \le \sum_{h\in[H]}\inf_{\pi\in\Pi^{\mathrm{S}}}\max_{s\in\mathcal{S}}\frac{1}{p_h^\pi(s)}\sum_{a\in\mathcal{A}}c_h(s,a).$$

Proof Fix any $h\in[H]$. Note that

$$\min_{\rho\in\Omega}\max_{s,a}\frac{c_h(s,a)}{\rho_h(s,a)} = \min_{\rho\in\Omega}\max_{s}\frac{1}{\rho_h(s)}\min_{\pi\in\mathcal{P}(A)}\frac{c_h(s,a)}{\pi(a)} = \min_{\rho\in\Omega}\max_{s}\frac{\sum_{a\in\mathcal{A}}c_h(s,a)}{\rho_h(s)}.$$

Now let ρ^h denote any solution to this optimization problem and define the mixed distribution $\tilde{\rho} := \sum_{l=1}^H \frac{Z_l}{Z}\rho^l$, where $Z_l := \min_{\rho\in\Omega}\max_s \frac{\sum_{a\in\mathcal{A}}c_l(s,a)}{\rho_l(s)}$ and $Z := \sum_{l=1}^H Z_l$. Then, $\tilde{\rho}\in\Omega$ and thus

$$\varphi^\star(c) \le \max_{h,s,a}\frac{c_h(s,a)}{\tilde{\rho}_h(s,a)} \le \max_h\frac{Z}{Z_h}\max_{s,a}\frac{c_h(s,a)}{\rho_h^h(s,a)} = \max_h\frac{Z}{Z_h}\min_{\rho\in\Omega}\max_{s,a}\frac{c_h(s,a)}{\rho^h(s,a)}$$
$$= \sum_{h\in[H]}\min_{\rho\in\Omega}\max_s\frac{\sum_{a\in\mathcal{A}}c_l(s,a)}{\rho_l(s)}.$$

■

A.4. Proof of Theorem 2

Define the coverage event $\mathcal{E}_{\mathrm{cov}} = \left(\forall(h,s,a)\in\mathcal{X},\ n_h^\tau(s,a)\ge c_h(s,a)\right)$. We have that for any δ-correct algorithm $\mathbb{P}(\mathcal{E}_{\mathrm{cov}})\ge 1-\delta$. Therefore, for any triplet $(h,s,a)\in\mathcal{X}$, we have that

$$\mathbb{E}[n_h^\tau(s,a)] \ge \mathbb{E}[n_h^\tau(s,a)\mathbb{1}(\mathcal{E}_{\mathrm{cov}})] \ge c_h(s,a)\mathbb{P}(\mathcal{E}_{\mathrm{cov}}) \ge (1-\delta)c_h(s,a). \tag{9}$$

Now consider the function $\eta_h(s,a) := \mathbb{E}[n_h^\tau(s,a)]$ for all h,s,a. We know that η satisfies the navigation constraints, hence it is a valid flow (see Appendix A). Moreover it satisfies the constraint (9). By definition of stochastic minimum flow, this means that

$$\mathbb{E}[\tau] = \sum_{a\in\mathcal{A}}\mathbb{E}[n_h^\tau(s_1,a)] = \varphi(\eta) \ge \varphi^\star\left([(1-\delta)c_h(s,a)]_{h,s,a}\right) = (1-\delta)\varphi^\star(c),$$

where in the last line we used that for any constant α, $\varphi^\star(\alpha c) = \alpha\varphi^\star(c)$.

■

Appendix B. CovGame

B.1. Proof of Theorem 3

Note that, at the beginning of any round $t \geq 1$, the learner $\mathcal{A}^\lambda$ works over the simplex $\mathcal{P}(\mathcal{X}_{k_t})$, hence $\lambda^t \in \mathcal{P}(\mathcal{X}_{k_t})$. Let m denote the number of times k_t changes value through the execution of the algorithm, that is $m = |\{t \leq \tau : k_t \neq k_{t+1}\}|$. Moreover, let $\tau_0 := 1$ and, for $i \in [m]$, let τ_i be the round at the beginning of which k_t has changed for the i-th time (i.e., $k_{\tau_i} \neq k_{\tau_i - 1}$). Note that, for any $i \geq 0$ and $t \in \{\tau_i, \dots, \tau_{i+1} - 1\}$, $k_t = k_{\tau_i}$. We start by bounding m.

Lemma 14 *It holds that* $m \leq \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$. *Moreover, for any* $i \in \{0, \dots, m-1\}$, *we have* $\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \leq c^+_{\min} 2^{k_{\tau_i}+2}$.

Proof By definition of the update rule, we have that $k_{t+1} \geq k_t$ for all $t \geq 1$. Now take any time t in which k_t has changed value m times. Since $k_1 \geq 0$, this means that $k_t \geq m$. By definition of k_t, we know that $n_h^{t-1}(s,a) \geq c_h(s,a)$ for all $(h,s,a) \in \mathcal{X} \setminus \mathcal{X}_j$ for some $j \geq m$. However, if $m \geq \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$, $\mathcal{X}_j = \emptyset$ and thus the algorithm must have stopped. This prove that $m \leq \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$.

To prove the second statement, we note that for any $i < m$, we have $k_{\tau_{i+1}-1} = k_{\tau_i}$ and $n_h^{\tau_{i+1}-2}(s,a) \geq c_h(s,a)$ for all $(h,s,a) \in \mathcal{X} \setminus \mathcal{X}_{k_{\tau_i}}$. Moreover, there must be some $(h,s,a) \in \mathcal{X} \setminus \mathcal{X}_{k_{\tau_i}+1}$ such that $n_h^{\tau_{i+1}-2}(s,a) < c_h(s,a)$. Indeed, if this was not the case, we would have an update of k at the end of round $\tau_{i+1} - 2$ instead of $\tau_{i+1} - 1$. Since all the triplets in $\mathcal{X}_{k_{\tau_i}}$ have been covered, the uncovered triplet must be in $\mathcal{X}_{k_{\tau_i}} \cap \mathcal{X} \setminus \mathcal{X}_{k_{\tau_i}+1} = \mathcal{X}_{k_{\tau_i}} \setminus \mathcal{X}_{k_{\tau_i}+1}$. By definition, all $(h,s,a) \in \mathcal{X}_{k_{\tau_i}} \setminus \mathcal{X}_{k_{\tau_i}+1}$ satisfy $c_h(s,a) \leq c^+_{\min} 2^{k_{\tau_i}+1}$. Hence,

$$\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \leq \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-2}(s,a) + 1 < c^+_{\min} 2^{k_{\tau_i}+1} + 1 \leq c^+_{\min} 2^{k_{\tau_i}+2}$$

where we use that $c^+_{\min} \geq 1$. ∎

Lemma 15 *Under Assumption 1 and 2, with probability at least* $1-\delta$, *for any* $i \in \{0, \dots, m-1\}$,

$$\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \geq \frac{1}{8}\sum_{j=0}^{i} \frac{\tau_{j+1} - \tau_j}{\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_j}}})} - \frac{3}{8}\mathcal{R}^\Pi_\delta(\tau_{i+1}) - \frac{3}{2}\sum_{j=0}^{i} \mathcal{R}^\lambda(\tau_{j+1} - \tau_j) - 3\log(4\tau_{i+1}/\delta).$$

$$\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \geq \frac{1}{8}\frac{\tau_{i+1} - \tau_i}{\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}})} - \frac{3}{8}\mathcal{R}^\Pi_\delta(\tau_{i+1}) - \frac{3}{2}\mathcal{R}^\lambda(\tau_{i+1}) - 3\log(4\tau_{i+1}/\delta).$$

Proof Take any $i \in \{0, \dots, m-1\}$. Note that

$$\begin{aligned}
\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) &= \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} \sum_{t=1}^{\tau_{i+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) && \text{(definition of counts)} \\
&= \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} \sum_{j=0}^{i} \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) && \text{(definition of } \{\tau_j\}_{j\geq 0}) \\
&\geq \sum_{j=0}^{i} \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_j}}} \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) && (\mathcal{X}_{k_{\tau_i}} \subseteq \mathcal{X}_{k_{\tau_j}} \text{ for all } j \leq i) \\
&= \sum_{j=0}^{i} \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_j}})} \sum_{(h,s,a)\in\mathcal{X}_{k_{\tau_j}}} \lambda_h(s,a) \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) \\
&= \sum_{j=0}^{i} \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_j}})} \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda). && \text{(definition of } \ell_t(\lambda))
\end{aligned}$$

For each j, by the regret bound of the λ player (Assumption 2),

$$\begin{aligned}
\min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_j}})} \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda) &\geq \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda^t) - \sqrt{\mathcal{R}^\lambda(\tau_{j+1}-\tau_j) \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda^t)} - \mathcal{R}^\lambda(\tau_{j+1}-\tau_j) \\
&\geq \frac{1}{2} \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda^t) - \frac{3}{2}\mathcal{R}^\lambda(\tau_{j+1}-\tau_j),
\end{aligned}$$

where in the last step we used the AM-GM inequality $\sqrt{xy} \leq \frac{x+y}{2}$ for $x, y \geq 0$. Summing over j,

$$\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \geq \frac{1}{2} \sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t) - \frac{3}{2} \sum_{j=0}^{i} \mathcal{R}^\lambda(\tau_{j+1}-\tau_j). \tag{10}$$

Let us now bound $\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t)$. Note that $\ell^t(\lambda^t) = \sum_{h,s,a} \lambda_h^t(s,a) \mathbb{1}\left(s_h^t = s, a_h^t = a\right)$ for all for all $t \in \{\tau_j, \dots, \tau_{j+1}-1\}$ since λ^t is equal to zero outside $\mathcal{X}_{k_{\tau_j}}$. Then,

$$\begin{aligned}
\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t) &= \sum_{t=1}^{\tau_{i+1}-1} \sum_{h,s,a} \lambda_h^t(s,a) \Big(\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \pm p_h^{\pi^t}(s,a) \Big) \\
&= \sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1; \lambda^t\right) + \underbrace{\sum_{t=1}^{\tau_{i+1}-1} \sum_{h,s,a} \lambda_h^t(s,a) \Big(\mathbb{1}\left(s_h^t = s, a_h^t = a\right) - p_h^{\pi^t}(s,a) \Big)}_{:= M_{\tau_{i+1}-1}}.
\end{aligned}$$

Since both λ^t and π^t are $\mathcal{F}_{t-1}$-measurable, $M_{\tau_{i+1}-1}$ is a martingale with differences bounded by 1 in absolute value. Therefore, by Freedman's inequality (e.g., Lemma 26 of Papini et al. (2021)),

with probability at least $1-\delta/2$,

$$\forall T \geq 1, \quad |M_T| \leq \sqrt{\sum_{t=1}^{T} V_t \times 4\log(4T/\delta)} + 4\log(4T/\delta)$$
$$\leq \sqrt{\sum_{t=1}^{T} V_1^{\pi_t}(s_1;\lambda^t) \times 4\log(4T/\delta)} + 4\log(4T/\delta),$$

where we defined $V_t := \mathrm{Var}[\sum_{h,s,a} \lambda_h^t(s,a)\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \mid \mathcal{F}_{t-1}]$ and used the simple bound $V_t \leq \mathbb{E}[\sum_{h,s,a} \lambda_h^t(s,a)\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \mid \mathcal{F}_{t-1}] = V_1^{\pi_t}\left(s_1;\lambda^t\right)$, which holds since $\sum_{h,s,a} \lambda_h^t(s,a)\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \leq 1$ almost surely by definition of λ^t. Plugging this into the initial decomposition of $\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t)$ and using the AM-GM inequality $\sqrt{xy} \leq \frac{x+y}{2}$ for $x, y \geq 0$,

$$\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t) \geq \sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1;\lambda^t\right) - \sqrt{\sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}(s_1;\lambda^t) \times 4\log(4\tau_{i+1}/\delta)} - 4\log(4\tau_{i+1}/\delta)$$
$$\geq \frac{1}{2}\sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1;\lambda^t\right) - 6\log(4\tau_{i+1}/\delta).$$

We finally bound $\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right)$ for any T. For all $T \geq 1$, with probability at least $1-\delta/2$ from Assumption 2,

$$\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right) \geq \sum_{t=1}^{T} V_1^{\star}\left(s_1;\lambda^t\right) - \sqrt{\mathcal{R}_\delta^{\Pi}(T)\sum_{t=1}^{T} V_1^{\star}(s_1;\lambda^t)} - \mathcal{R}_\delta^{\Pi}(T).$$

Applying once again the AM-GM inequality yields

$$\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right) \geq \frac{1}{2}\sum_{t=1}^{T} V_1^{\star}\left(s_1;\lambda^t\right) - \frac{3}{2}\mathcal{R}_\delta^{\Pi}(T)$$
$$= \frac{1}{2}\sum_{t=1}^{T} \sup_{\rho\in\Omega} \sum_{h,s,a} \rho_h(s,a)\lambda_h^t(s,a) - \frac{3}{2}\mathcal{R}_\delta^{\Pi}(T).$$

Now note that, since λ^t is supported on $\mathcal{X}_{k_{\tau_j}}$ for any $t \in \{\tau_j, \dots, \tau_{j+1}-1\}$,

$$\sum_{t=1}^{\tau_{i+1}-1} \sup_{\rho\in\Omega} \sum_{h,s,a} \rho_h(s,a)\lambda_h^t(s,a) = \sum_{j=0}^{i}\sum_{t=\tau_j}^{\tau_{j+1}-1} \sup_{\rho\in\Omega} \sum_{h,s,a} \rho_h(s,a)\lambda_h^t(s,a)$$
$$\geq \sum_{j=0}^{i}\sum_{t=\tau_j}^{\tau_{j+1}-1} \sup_{\rho\in\Omega} \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_j}}} \rho_h(s,a) = \sum_{j=0}^{i} \frac{\tau_{j+1}-\tau_j}{\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_j}}})}.$$

Plugging everything together proves the first statement. The second result can be proved analogously by simply using $\sum_{j=0}^{i} \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_j}})} \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda) \geq \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_i}})} \sum_{t=\tau_i}^{\tau_{i+1}-1} \ell^t(\lambda)$ in the

first series of inequalities and continuing with the same steps. This yields a single dependence on $\mathcal{R}^\lambda(\tau_{i+1} - \tau_i)$, which can be upper bounded by the stated $\mathcal{R}^\lambda(\tau_{i+1})$ by monotonicity of $\mathcal{R}^\lambda$. ■

We are now ready to prove Theorem 3

Proof [Proof of Theorem 3] Let m be the number of times k_t has changed throughout the execution of the algorithm. Note that, in the round τ in which the algorithm stops the last change must occur, thus $\tau_m = \tau + 1$, and $k_{\tau+1}$ is set to any value such that $\mathcal{X}_{k_{\tau+1}} = \emptyset$. Then,

$$\tau = \tau_m - 1 = \sum_{i=0}^{m-1} (\tau_{i+1} - \tau_i)\,.$$

By combining Lemma 14 with Lemma 15 and rearranging, with probability at least $1 - \delta$, for any $i \in \{0, \dots, m-1\}$,

$$\begin{aligned}\tau_{i+1} - \tau_i &\le 8\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) c_{\min}^+ 2^{k_{\tau_i}+2} + 8\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) \left(\frac{3}{8}\mathcal{R}_\delta^\Pi(\tau_{i+1}) + \frac{3}{2}\mathcal{R}^\lambda(\tau_{i+1}) + 3\log(4\tau_{i+1}/\delta)\right) \\ &\le 8\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) c_{\min}^+ 2^{k_{\tau_i}+2} + \varphi^\star(\mathbb{1}_{\mathcal{X}}) \left(3\mathcal{R}_\delta^\Pi(\tau_m) + 12\mathcal{R}^\lambda(\tau_m) + 24\log(4\tau_m/\delta)\right),\end{aligned}$$

where the second inequality is due to $\mathcal{X}_k \subseteq \mathcal{X}$ for all $k \in \mathbb{N}$ and $\tau_{i+1} \le \tau_m$ for $i \le m-1$. Then,

$$\tau_m \le 8 \sum_{i=0}^{m-1} c_{\min}^+ \varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) 2^{k_{\tau_i}+2} + m\varphi^\star(\mathbb{1}_{\mathcal{X}}) \left(3\mathcal{R}_\delta^\Pi(\tau_m) + 12\mathcal{R}^\lambda(\tau_m) + 24\log(4\tau_m/\delta)\right) + 1.$$

The first term can be bounded by

$$\begin{aligned}8 \sum_{i=0}^{m-1} c_{\min}^+ \varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) 2^{k_{\tau_i}+2} &= 8 \sum_{i=0}^{m-1} c_{\min}^+ 2^{k_{\tau_i}+2} \min_{\rho \in \Omega} \max_{s,a,h} \frac{\mathbb{1}((h,s,a) \in \mathcal{X}_{k_{\tau_i}})}{\rho_h(s,a)} \\ &\le 32 \sum_{i=0}^{m-1} c_{\min}^+ 2^{k_{\tau_i}} \min_{\rho \in \Omega} \max_{s,a,h} \frac{\mathbb{1}(c_{\min}^+ 2^{k_{\tau_i}} < c_h(s,a))}{\rho_h(s,a)} \\ &\le 32 \sum_{i=0}^{m-1} \min_{\rho \in \Omega} \max_{s,a,h} \frac{c_h(s,a)}{\rho_h(s,a)} = 32m\varphi^\star(c).\end{aligned}$$

Plugging this into the bound on τ_m, we obtain the inequality,

$$\tau_m \le 32m\varphi^\star(c) + m\varphi^\star(\mathbb{1}_{\mathcal{X}}) \left(3\mathcal{R}_\delta^\Pi(\tau_m) + 12\mathcal{R}^\lambda(\tau_m) + 24\log(4\tau_m/\delta)\right) + 1.$$

Thus, for $\tau_m \ge T_1$, we get that the sample complexity is bounded by $\tau \le 64m\varphi^\star(c)$. Thus, we conclude that $\tau \le \tau_m \le \max\{T_1, 64m\varphi^\star(c)\} \le 64m\varphi^\star(c) + T_1$. The proof is concluded by using Lemma 14 to bound m. ■

B.2. Proof of Corollary 4

We need to bound T_1 from Theorem 3 when using WMF and UCBVI. By definition of T_1 in Theorem 3,

$$\frac{T_1 - 1}{2} \leq m\varphi^\star(\mathbb{1}_\mathcal{X}) \left(3\mathcal{R}_\delta^\Pi(T_1) + 12\mathcal{R}^\lambda(T_1) + 24\log(4T_1/\delta)\right) + 1.$$

Recall that, by (4) and (5),

$$\mathcal{R}^\lambda(T) = 16\log(SAH) + 1 \quad \mathcal{R}_\delta^\Pi(T) = 65536SAH^2(\log(2SAH/\delta) + 6S)\log(T+1)^2.$$

For $T \geq 3$ and assuming $SAH \geq 2$ (otherwise the result is trivial), it is easy to see that $\mathcal{R}^\lambda(T) \leq \mathcal{R}_\delta^\Pi(T)$ and $24\log(4T/\delta) \leq \mathcal{R}_\delta^\Pi(T)$. Theorefore, for some numerical constant c_1,

$$T_1 \leq c_1 m\varphi^\star(\mathbb{1}_\mathcal{X})SAH^2(\log(2SAH/\delta) + 6S)\log(T_1+1)^2.$$

Solving the inequality in T_1 yields the stated bound.

■

B.3. Links with concave-utility reinforcement learning

The (inverse) complexity term $\varphi^\star(c)$ that we seek to approximate with COVGAME can be expressed as the maximization of a concave function of the visitation probabilities:

$$\frac{1}{\varphi^\star(c)} = \max_{\rho\in\Omega} f_c(\rho) \quad \text{where} \quad f_c(\rho) = \min_{(h,s,a)\in\mathcal{X}} \frac{\rho_h(s,a)}{c_h(s,a)}.$$

Computing the maximizer without the knowledge of the MDP falls in the framework of concave utility reinforcement learning (or convex reinforcement learning when we instead minimize a convex function (Zahavy et al., 2021)) which has attracted a lot of interest recently (Hazan et al., 2019; Zhang et al., 2020; Geist et al., 2022). Several authors proposed the use of a Frank-Wolfe approach, when the function f to maximize is smooth (which is not the case for f_c). Indeed, it was observed that in the Frank-Wolfe update the computation of

$$\arg\max_{\rho\in\Omega} \rho^\top \nabla f(\rho) = \arg\max_{\rho\in\Omega} \sum_{h,s,a} \rho_h(s,a)(\nabla f(\rho))_{h,s,a}$$

can be interpreted as solving the MDP when the reward function is $r_h(s,a) = (\nabla f(\rho))_{h,s,a}$. Different authors proposed to combine Frank-Wolfe with regret minimizers to cope for the unknown MDP (Cheung, 2019; Zahavy et al., 2021). For example Wagenmaker and Jamieson (2022) propose a generic algorithm for smooth experimental design in linear MDPs (which generalizes $\max_{\rho\in\Omega} f(\rho)$ to optimizing over possible covariance matrices) which runs a regret minimizer for a long time on a reward function given by the gradient of the objective. To tackle non-smooth objective, they further propose to use a log-sum-exp smoothening trick.

Interestingly, each phase of COVGAME may be interpreted as doing a Frank-Wolfe update on a *sequence* of smoothening of an objective of the form $g(\rho) = \min_{(h,s,a)\in\mathcal{X}_k} \rho_h(s,a)$, where the regret minimizer is further never restarted. Indeed, introducing

$$g_\eta(\rho) = \frac{1}{\eta}\log\left(\sum_{(h,s,a)\in\mathcal{X}} e^{\eta\rho_h(s,a)}\right),$$

we have

$$(\nabla g_\eta(\rho))_{h,s,a} = \frac{e^{\eta\rho_h(s,a)}}{\sum_{(h',s',a')\in\mathcal{X}} e^{\eta\rho_{h'}(s',a')}}$$

and the reward $\lambda_h^t(s,a)$ used by CovGame when $\mathcal{X}_k$ is the set to be covered and the last restart occured at time t_k can be written

$$\lambda_h^t(s,a) = \nabla g_{\eta_{t-t_k}}\left((n_h^t(s,a) - n_h^{t_k}(s,a))_{h,s,a}\right) = \nabla g_{\widetilde{\eta}_t}\left(\left(\frac{(n_h^t(s,a) - n_h^{t_k}(s,a)}{t-t_k}\right)_{h,s,a}\right)$$

where $\widetilde{\eta}_t = \xi_{t-t_k}$ is the (time-varying) smoothening parameter, with ξ_t the variance-dependent learning rate defined by Cesa-Bianchi et al. (2005).

Appendix C. UCBVI with Changing Rewards

In this appendix, we study the following regret minimization setting with changing rewards. At the beginning of each episode $t \geq 1$, the learner receives a known reward function $r_h^t(s,a)$. The learner does not know the transition probabilities p and its goal is to minimize the regret

$$\sum_{t=1}^{T}\left(V_1^\star(s_1;r^t) - V_1^{\pi^t}(s_1;r^t)\right),$$

where $V_1^\pi(s_1;r) := \sum_{h,s,a} p_h^\pi(s,a) r_h(s,a)$ and $V_1^\star(s_1;r) := \max_\pi V_1^\pi(s_1;r)$. We make the following assumption on the sequence of rewards.

Assumption 3 *For all $t \geq 1$, $r_h^t(s,a) \in [0,1]$ for all h,s,a, and $\sum_{h,s,a} r_h^t(s,a) \leq 1$.*

Note that this implies that $\sum_{h=1}^H r_h(s_h,a_h) \in [0,1]$ for any trajectory $\{(s_h,a_h)\}_{h\in[H]}$ almost surely.

C.1. Algorithm

We study a variant of the UCBVI algorithm (Azar et al., 2017) adapted to this setting. For any $h < H$, we define recursively upper confidence bounds over optimal value functions for any reward r as

$$\overline{Q}_h^t(s,a;r) = \left(r_h(s,a) + \widehat{P}_{h,s,a}^t \overline{V}_{h+1}^t(r) + B_h^t(s,a;r)\right) \wedge 1,$$

where $\overline{Q}_H^t(s,a;r) = r_H(s,a)$, $\overline{V}_{h+1}^t(s;r) := \max_a \overline{Q}_h^t(s,a;r)$, and

$$B_h^t(s,a;r) := \max\left\{\sqrt{\frac{8\mathbb{V}(\widehat{P}_{h,s,a}^t, \overline{V}_{h+1}^t(r))\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}, \frac{8\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}\right\}.$$

Note that this bonus is infinite for $n_h^t(s,a) = 0$. As we will only evaluate these quantities in the rewards observed at the corresponding round, we shall abbreviate $\overline{Q}_h^t(s,a) := \overline{Q}_h^t(s,a;r^{t+1})$, $\overline{V}_h^t(s) := \overline{V}_h^t(s;r^{t+1})$, and $B_h^t(s,a) := B_h^t(s,a;r^{t+1})$ for all $t \in \mathbb{N}$. UCBVI plays at each episode

$$\pi_h^t(s) \in \arg\max_a \overline{Q}_h^{t-1}(s,a),$$

which is thus greedy w.r.t. the optimistic value function for reward r^t.

C.2. Analysis

The analysis follows the one of EULER (Zanette and Brunskill, 2019b) and uses several technical results from Ménard et al. (2021) and Zhang et al. (2021c). Let us define the event

$$E := \left\{ \forall t \in \mathbb{N}, h, s, a : \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \leq \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\},$$

where $\beta(n,\delta) := \log(2SAH/\delta) + S\log(8e(n+1))$. Moreover, let

$$G := \left\{ \forall t \in \mathbb{N}, h, s, a : n_h^t(s,a) \geq \frac{1}{2}\overline{n}_h^t(s,a) - \beta^{\mathrm{cnt}}(\delta) \right\},$$

where $\beta^{\mathrm{cnt}}(\delta) := \log(2SAH/\delta)$ and $\overline{n}_h^t(s,a) = \sum_{j=1}^t p_h^{\pi^j}(s,a)$.

Lemma 16 (Bernstein-like bound) *Under event E, for all h, s, a, t and value function V s.t. $V_h(s) \in [0,1]$ for all h, s,*

$$\begin{aligned}|(P_{h,s,a} - \widehat{P}_{h,s,a}^t)V_{h+1}| &\leq \sqrt{\frac{2\mathbb{V}(\widehat{P}_{h,s,a}^t, V_{h+1})\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2\beta(n_h^t(s,a),\delta)}{3n_h^t(s,a)} \\ &\leq \max\left\{ \sqrt{\frac{8\mathbb{V}(\widehat{P}_{h,s,a}^t, V_{h+1})\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}, \frac{4\beta(n_h^t(s,a),\delta)}{3n_h^t(s,a)} \right\}.\end{aligned}$$

Proof This is immediate by combining the definition of E with Lemma 10 of Ménard et al. (2021) and $x + y \leq 2\max\{x, y\}$. ■

Lemma 17 (Optimism) *Under event E, $\overline{Q}_h^t(s,a;r) \geq Q_h^\star(s,a;r)$ for all t, h, s, a and any reward r satisfying Assumption 3.*

Proof By definition, $\overline{Q}_H^t(s,a;r) = Q_H^\star(s,a;r) = r_H(s,a)$. Thus, the statement holds at stage H. Now suppose it holds at stage $h+1$ for $h \in [H-1]$. This implies that $\overline{V}_{h+1}^t(s;r) \geq V_{h+1}^\star(s;r)$ for all s. Then,

$$\begin{aligned}&r_h(s,a) + \widehat{P}_{h,s,a}^t \overline{V}_{h+1}^t(r) + B_h^t(s,a;r) \\ &= r_h(s,a) + \widehat{P}_{h,s,a}^t \overline{V}_{h+1}^t(r) + \max\left\{ \sqrt{\frac{8\mathbb{V}(\widehat{P}_{h,s,a}^t, \overline{V}_{h+1}^t(r))\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}, \frac{8\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\} \\ &\geq r_h(s,a) + \widehat{P}_{h,s,a}^t V_{h+1}^\star(s;r) + \max\left\{ \sqrt{\frac{8\mathbb{V}(\widehat{P}_{h,s,a}^t, V_{h+1}^\star(r))\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}, \frac{8\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\} \\ &\geq r_h(s,a) + P_{h,s,a} V_{h+1}^\star(s;r) = Q_h^\star(s,a;r),\end{aligned}$$

where the first inequality uses the inductive hypothesis together with the monotonicity property in Lemma 14 of Zhang et al. (2021c), while the second inequality uses Lemma 16. The fact that $Q_h^\star(s,a;r) \in [0,1]$ for any h, s, a and r satisfying Assumption 3 concludes the proof. ■

Lemma 18 (Variance concentration) *Under event E, for any t, h, s, a, any reward r satisfying Assumption 3, and any value function V s.t. $V_h(s) \in [0,1]$ for all h, s,*

$$\mathbb{V}(\widehat{P}^t_{h,s,a}, \overline{V}^t_{h+1}(r)) \leq 4\mathbb{V}(P_{h,s,a}, V_{h+1}) + 4P_{h,s,a}|\overline{V}^t_{h+1}(r) - V_{h+1}| + 4\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}.$$

Proof By combining Lemma 11 and 12 of Ménard et al. (2021) together with the definition of E,

$$\begin{aligned}\mathbb{V}(\widehat{P}^t_{h,s,a}, \overline{V}^t_{h+1}(r)) &\leq 2\mathbb{V}(P_{h,s,a}, \overline{V}^t_{h+1}(r)) + 4\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)} \\ &\leq 4\mathbb{V}(P_{h,s,a}, V_{h+1}) + 4P_{h,s,a}|\overline{V}^t_{h+1}(r) - V_{h+1}| + 4\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)}.\end{aligned}$$

■

Theorem 19 *Under the assumptions above, with probability $1-\delta$, for any $T \in \mathbb{N}$, the regret of UCBVI for changing rewards is bounded by*

$$\sum_{t=1}^{T}\left(V_1^\star(s_1;r^t) - V_1^{\pi^t}(s_1;r^t)\right) \leq 5140SAH^2L_{T,\delta} + 256\sqrt{SAHL_{T,\delta}\sum_{t=1}^{T}V_1^{\pi^t}(s_1;r^t)},$$

whee $L_{T,\delta} := (\log(2SAH/\delta) + 6S)\log(T+1)^2$.

Proof Note that $\mathbb{P}(E,G) \geq 1-\delta$ by Lemma 3 of Ménard et al. (2021) and a union bound. We shall thus carry out the proof conditioned on E and G holding. Fix any $T \in \mathbb{N}$. We start from the same regret decomposition as in the proof of Theorem 2 of Zanette and Brunskill (2019b). First, by Lemma 17,

$$\sum_{t=1}^{T}\left(V_1^\star(s_1;r^t) - V_1^{\pi^t}(s_1;r^t)\right) \leq \sum_{t=1}^{T}\left(\overline{V}_1^{t-1}(s_1) - V_1^{\pi^t}(s_1;r^t)\right). \tag{11}$$

For any t, h, s, a,

$$\begin{aligned}\overline{Q}_h^{t-1}(s,a) - Q_h^{\pi^t}(s,a;r^t) &\leq \widehat{P}^t_{h,s,a}\overline{V}^{t-1}_{h+1} + B_h^{t-1}(s,a) \wedge 1 - P_{h,s,a}V^{\pi^t}_{h+1}(r^t) \\ &\leq P_{h,s,a}\overline{V}^{t-1}_{h+1} + |(\widehat{P}^t_{h,s,a} - P_{h,s,a})\overline{V}^{t-1}_{h+1}| + B_h^{t-1}(s,a) \wedge 1 - P_{h,s,a}V^{\pi^t}_{h+1}(r^t) \\ &\leq P_{h,s,a}\overline{V}^{t-1}_{h+1} + 2B_h^{t-1}(s,a) \wedge 1 - P_{h,s,a}V^{\pi^t}_{h+1}(r^t),\end{aligned}$$

where the last step uses Lemma 16 and the fact that values are all in $[0,1]$. Theorefore,

$$\begin{aligned}\overline{V}_h^{t-1}(s) - V_h^{\pi^t}(s;r^t) &= \overline{Q}_h^{t-1}(s,\pi^t_h(s)) - Q_h^{\pi^t}(s,\pi^t_h(s);r^t) \\ &\leq P_{h,s,\pi^t_h(s)}\left(\overline{V}^{t-1}_{h+1} - V^{\pi^t}_{h+1}(r^t)\right) + 2B_h^{t-1}(s,\pi^t_h(s)) \wedge 1.\end{aligned} \tag{12}$$

Enrolling this reasoning, we thus obtain

$$\begin{aligned}\overline{V}_1^{t-1}(s_1) - V_1^{\pi^t}(s_1; r^t) &\leq 2\sum_{h,s,a} p_h^{\pi^t}(s,a)\left(B_h^{t-1}(s,a) \wedge 1\right) \\ &= 2\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\left(B_h^{t-1}(s,a) \wedge 1\right) + 2\sum_{(h,s,a)\notin\mathcal{Z}_t} p_h^{\pi^t}(s,a)\left(B_h^{t-1}(s,a) \wedge 1\right),\end{aligned} \tag{13}$$

where $\mathcal{Z}_t := \{(h,s,a) : \overline{n}_h^{t-1}(s,a) \geq 4\beta^{\mathrm{cnt}}(\delta)\}$. Recall that $\overline{n}_h^{t-1}(s,a) := \sum_{i=1}^{t-1} p_h^{\pi^i}(s,a)$. Let $W_T := 2\sum_{t=1}^{T}\sum_{h,s,a} p_h^{\pi^t}(s,a)\left(B_h^{t-1}(s,a) \wedge 1\right)$. Summing the previous inequality over all time steps and using Lemma 20,

$$W_T \leq 2\underbrace{\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\left(B_h^{t-1}(s,a) \wedge 1\right)}_{①} + 10SAH\beta^{\mathrm{cnt}}(\delta). \tag{14}$$

Now recall that $B_h^{t-1}(s,a)$ depends on $\mathbb{V}(\widehat{P}_{h,s,a}^{t-1}, \overline{V}_{h+1}^{t-1})$. By Lemma 18, for any t, s, a, h,

$$\mathbb{V}(\widehat{P}_{h,s,a}^{t-1}, \overline{V}_{h+1}^{t-1}) \leq 4\mathbb{V}(P_{h,s,a}, V_{h+1}^{\pi^t}(r^t)) + 4P_{h,s,a}|\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}(r^t)| + 4\frac{\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a)}.$$

Plugging this into the definition of $B_h^{t-1}(s,a)$ and using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$,

$$\begin{aligned}B_h^{t-1}(s,a) \leq \sqrt{\frac{32\mathbb{V}(P_{h,s,a}, V_{h+1}^{\pi^t}(r^t))\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a)}} + \sqrt{\frac{32P_{h,s,a}|\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}(r^t)|\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a)}} \\ + \frac{16\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a)}.\end{aligned}$$

Back into ① and using that $\beta(x,\delta) \geq 1$ for all $x \geq 0$, we get

$$\begin{aligned}① \leq &\underbrace{\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\sqrt{\frac{32\mathbb{V}(P_{h,s,a}, V_{h+1}^{\pi^t}(r^t))\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a) \vee 1}}}_{②} \\ &+ \underbrace{\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\sqrt{\frac{32P_{h,s,a}|\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}(r^t)|\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a) \vee 1}}}_{③} \\ &+ 16\underbrace{\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\frac{\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a) \vee 1}}_{④}\end{aligned}$$

We bound these terms separately. By Lemma 21 and monotonicity of $\beta(\cdot,\delta)$,

$$④ \leq 16SAH\log(T+1)\beta(T,\delta).$$

By Cauchy-Schwartz inequality and the bound on ④,

$$\begin{aligned}
② &\le \sqrt{32\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\mathbb{V}(P_{h,s,a},V_{h+1}^{\pi^t}(r^t))\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\frac{\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a)\vee 1}} \\
&= \sqrt{32\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\mathbb{V}(P_{h,s,a},V_{h+1}^{\pi^t}(r^t))\times ④} \\
&\le 32\sqrt{SAH\log(T+1)\beta(T,\delta)\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\mathbb{V}(P_{h,s,a},V_{h+1}^{\pi^t}(r^t))} \\
&\le 64\sqrt{SAH\log(T+1)\beta(T,\delta)\sum_{t=1}^{T}V_1^{\pi^t}(s_1;r^t)},
\end{aligned}$$

where the last inequality uses Lemma 22. It only remains to bound ③. By Cauchy-Schwartz inequality and the bound on ④,

$$\begin{aligned}
③ &\le \sqrt{32\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)P_{h,s,a}|\overline{V}_{h+1}^{t-1}-V_{h+1}^{\pi^t}(r^t)|\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\frac{\beta(n_h^{t-1}(s,a),\delta)}{n_h^{t-1}(s,a)\vee 1}} \\
&= \sqrt{32\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)P_{h,s,a}|\overline{V}_{h+1}^{t-1}-V_{h+1}^{\pi^t}(r^t)|\times ④} \\
&\le 32\sqrt{SAH\log(T+1)\beta(T,\delta)\sum_{t=1}^{T}\sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)P_{h,s,a}|\overline{V}_{h+1}^{t-1}-V_{h+1}^{\pi^t}(r^t)|} \\
&\le 32\sqrt{SAH^2\log(T+1)\beta(T,\delta)\underbrace{2\sum_{t=1}^{T}\sum_{h,s,a} p_h^{\pi^t}(s,a)\left(B_h^{t-1}(s,a)\wedge 1\right)}_{=W_T}},
\end{aligned}$$

where the last inequality uses Lemma 23 together with $|\overline{V}_{h+1}^{t-1}(s)-V_{h+1}^{\pi^t}(s;r^t)| = \overline{V}_{h+1}^{t-1}(s)-V_{h+1}^{\pi^t}(s;r^t)$ for any s (due to optimism). Plugging the bounds on ②,③,④ into ① in (14),

$$\begin{aligned}
W_T \le 64\sqrt{SAH^2\log(T+1)\beta(T,\delta)W_T} + 128\sqrt{SAH\log(T+1)\beta(T,\delta)\sum_{t=1}^{T}V_1^{\pi^t}(s_1;r^t)} \\
+ 512SAH\log(T+1)\beta(T,\delta) + 10SAH\beta^{\mathrm{cnt}}(\delta).
\end{aligned}$$

The sum of the last two terms can be bounded by $522SAH\log(T+1)\beta(T,\delta)$ since $\beta^{\mathrm{cnt}}(\delta)\le\beta(T,\delta)$. Solving the quadratic inequality in $\sqrt{W_T}$, we get

$$\begin{aligned}
W_T \le 4096SAH^2\log(T+1)\beta(T,\delta) + 256\sqrt{SAH\log(T+1)\beta(T,\delta)\sum_{t=1}^{T}V_1^{\pi^t}(s_1;r^t)} \\
+ 1044SAH\log(T+1)\beta(T,\delta).
\end{aligned}$$

Finally, note that W_T bounds the regret by (11) and (13). The proof is concluded by using that $\beta(T,\delta) \leq (\log(2SAH/\delta) + 6S)\log(T+1)$ to simplify the expression. ■

Lemma 20 *For any $T \geq 1$, $\sum_{t=1}^{T} \sum_{(h,s,a)\notin\mathcal{Z}_t} p_h^{\pi^t}(s,a) \leq 5SAH\beta^{\mathrm{cnt}}(\delta)$.*

Proof By definition of $\mathcal{Z}_t$ and since $p_h^{\pi^t}(s,a) \leq 1$,

$$\sum_{t=1}^{T} \sum_{(h,s,a)\notin\mathcal{Z}_t} p_h^{\pi^t}(s,a) = \sum_{h,s,a}\sum_{t=1}^{T} p_h^{\pi^t}(s,a)\mathbb{1}\left(\overline{n}_h^{t-1}(s,a) < 4\beta^{\mathrm{cnt}}(\delta)\right) \leq SAH(4\beta^{\mathrm{cnt}}(\delta)+1).$$

The result is proved by noting that $1 \leq \beta^{\mathrm{cnt}}(\delta)$. ■

Lemma 21 *Under event G, for any $T \geq 1$,*

$$\sum_{t=1}^{T} \sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\frac{1}{n_h^{t-1}(s,a)\vee 1} \leq 16SAH\log(T+1).$$

Proof By definition of G and Z_t, if $(h,s,a) \in \mathcal{Z}_t$ then $n_h^{t-1}(s,a) \geq \overline{n}_h^{t-1}(s,a)/4$. Then,

$$\begin{aligned}\sum_{t=1}^{T} \sum_{(h,s,a)\in\mathcal{Z}_t} p_h^{\pi^t}(s,a)\frac{1}{n_h^{t-1}(s,a)\vee 1} &\leq 4\sum_{t=1}^{T}\sum_{h,s,a} p_h^{\pi^t}(s,a)\frac{1}{\overline{n}_h^{t-1}(s,a)\vee 1} \\ &= 4\sum_{h,s,a}\sum_{t=1}^{T}\frac{\overline{n}_h^{t}(s,a)-\overline{n}_h^{t-1}(s,a)}{\overline{n}_h^{t-1}(s,a)\vee 1} \leq 16SAH\log(T+1),\end{aligned}$$

where the last inequality uses Lemma 9 of Ménard et al. (2021). ■

Lemma 22 *For any $T \geq 1$,*

$$\sum_{t=1}^{T}\sum_{h,s,a} p_h^{\pi^t}(s,a)\mathbb{V}(P_{h,s,a}, V_{h+1}^{\pi^t}(r^t)) \leq 4\sum_{t=1}^{T} V_1^{\pi^t}(s_1;r^t).$$

Proof Starting from the well-known variance decomposition lemma (see, e.g., Lemma 7 of Ménard et al. (2021)) and following with the same bounds as in the proof of Lemma 3.4 of Jin et al. (2020),

$$\begin{aligned}\sum_{h,s,a} p_h^{\pi^t}(s,a)\mathbb{V}(P_{h,s,a}, V_{h+1}^{\pi^t}(r^t)) &= \mathbb{E}^{\pi^t}\left[\left(\sum_{h=1}^{H} r_h^t(s_h,a_h) - V_1^{\pi^t}(s_1;r^t)\right)^2\right] \\ &\leq 2\mathbb{E}^{\pi^t}\left[\left(\sum_{h=1}^{H} r_h^t(s_h,a_h)\right)^2\right] + 2V_1^{\pi^t}(s_1;r^t)^2 \\ &\leq 2\mathbb{E}^{\pi^t}\left[\sum_{h=1}^{H} r_h^t(s_h,a_h)\right] + 2V_1^{\pi^t}(s_1;r^t) \\ &= 4V_1^{\pi^t}(s_1;r^t)\end{aligned}$$

where the first inequality uses $(x+y)^2 \leq 2x^2+2y^2$ and the second one uses Assumption 3. ■

Lemma 23 *Under event E, for any t,*

$$\sum_{h,s,a} p_h^{\pi^t}(s,a) P_{h,s,a}(\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}(r^t)) \leq 2H \sum_{h,s,a} p_h^{\pi^t}(s,a) \left(B_h^{t-1}(s,a) \wedge 1\right).$$

Proof Since $\sum_{s,a} p_h^{\pi^t}(s,a) p_h(s'|s,a) = p_{h+1}^{\pi^t}(s')$ for any s',

$$\begin{aligned}
\sum_{h,s,a} p_h^{\pi^t}(s,a) P_{h,s,a}(\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}(r^t)) &= \sum_{h=2}^{H} \sum_{s} p_h^{\pi^t}(s)(\overline{V}_h^{t-1}(s) - V_h^{\pi^t}(s;r^t)) \\
&\leq \sum_{h=2}^{H} \sum_{s} p_h^{\pi^t}(s) P_{h,s,\pi_h^t(s)} \left(\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}(r^t)\right) + \sum_{h=2}^{H} \sum_{s} p_h^{\pi^t}(s) 2B_h^{t-1}(s,\pi_h^t(s)) \wedge 1 \\
&\leq H \sum_{h,s} p_h^{\pi^t}(s) 2B_h^{t-1}(s,\pi_h^t(s)) \wedge 1,
\end{aligned}$$

where the first inequality uses the decomposition in (12) while the second one applies this reasoning recursively. ■

Appendix D. Concentration of Value Functions

In this appendix, we derive the concentration bounds on value functions needed for our PAC RL algorithms. We shall assume that rewards lie in $[0,1]$ almost surely.

D.1. General results

Lemma 24 *[Concentration of $\widehat{p}^T V$] Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$, $Z := |\mathcal{Z}|$, and $\{V_h : \mathcal{S} \to [0,H]\}_{h \in [H+1]}$ be a collection of bounded functions. With probability at least $1-\delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$,*

$$\sum_{(h,s,a) \in \mathcal{Z}} n_h^t(s,a) \left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}\right|^2 \leq 4H^2 \log(1/\delta) + 2ZH^2 \log(1+t).$$

Proof We start by building a suitable stochastic process to apply Theorem 1 of Abbasi-Yadkori et al. (2011). Let $\mathcal{F}_{t,h}$ denote the filtration up to stage h of round t. For any $h \in [H], t \geq 1$, the random variable $\eta_h^t := V_{h+1}(s_{h+1}^t) - p_h(s_h^t, a_h^t)^T V_{h+1}$ is zero-mean and H^2-subgaussian conditionally on $\mathcal{F}_{t,h}$ due to the boundedness of the functions $\{V_h\}_{h \in [H]}$. Let X_h^t be a Z-dimensional vector containing a value 1 at position (h, s_h^t, a_h^t) if $(h, s_h^t, a_h^t) \in \mathcal{Z}$, and zero at all other positions. Note that X_h^t is $\mathcal{F}_{t,h}$-measurable, while η_h^t is $\mathcal{F}_{t,h+1}$-measurable. Let $Y_t := \sum_{j=1}^{t} \sum_{h=1}^{H} X_h^t \eta_h^t$. For all $(h,s,a) \in \mathcal{Z}$, we have

$$\begin{aligned}
[Y_t]_{h,s,a} &= \sum_{j=1}^{t} \mathbb{1}\left(s_h^j = s, a_h^j = a\right) \left(V_{h+1}(s_{h+1}^j) - p_h(s_h^j, a_h^j)^T V_{h+1}\right) \\
&= n_h^t(s,a)(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}.
\end{aligned}$$

Let $D_t := \sum_{j=1}^{t} \sum_{h=1}^{H} X_h^t (X_h^t)^T = \mathrm{diag}([n_h^t(s,a)]_{(h,s,a)\in\mathcal{Z}})$. Theorem 1 of Abbasi-Yadkori et al. (2011) combined with Equation 20.9 from Lattimore and Szepesvari (2019) yield that

$$\mathbb{P}\left(\forall t \geq 1,\ \left\|Y^t\right\|^2_{(I+D_t)^{-1}} \leq 2H^2 \log(1/\delta) + ZH^2 \log(1+t/Z)\right) \geq 1-\delta.$$

Since $n_h^t(s,a) \geq 1$ for any $t \geq t_0$ and $(h,s,a) \in \mathcal{Z}$, following Corollary 3 in Réda et al. (2021),

$$D_t = \mathrm{diag}\big([n_h^t(s,a)]_{(h,s,a)\in\mathcal{Z}}\big) \succeq (I+D_t)/2,$$

which implies $\left\|Y^t\right\|^2_{D_t^{-1}} \leq 2\left\|Y^t\right\|^2_{(I+D_t)^{-1}}$ for any $t \geq t_0$. Plugging this into the probability above and using that $\left\|Y^t\right\|^2_{D_t^{-1}}$ is exactly the left-hand side of the statement concludes the proof. ■

Lemma 25 *[Concentration of $\widehat{p}^T V$ for all V] Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$, $Z := |\mathcal{Z}|$, and $\mathcal{V} := \{V : \mathcal{S} \to [0,H]\}$ be the set of all bounded functions mapping $\mathcal{S}$ into $[0,H]$. With probability at least $1-\delta$, for any functions $\{V_h \in \mathcal{V}\}_{h=2}^{H+1}$ and $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$,*

$$\sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}\right|^2 \leq 4H^2\log(1/\delta) + 12(SH+Z)H^2\log(1+t).$$

Proof Let $Y_t(V_2,\dots,V_{H+1}) := \sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}\right|^2$ denote the quantity to be bounded for fixed functions $V_h \in \mathcal{V}$ for all $2 \leq h \leq H+1$. Let $\{\xi_t\}_{t\geq 1}$ be a sequence of positive values to be specified later. For all t, let $\Xi_t := \{\xi_t, 2\xi_t, \dots \lfloor H/\xi_t \rfloor \xi_t\}$. Note that $|\Xi_t| = \lfloor H/\xi_t \rfloor$ and, for all $x \in [0,H]$, there exists $y \in \Xi_t$ s.t. $|x-y| \leq \xi_t$. For all t, we build a discrete cover $\overline{\mathcal{V}}_t$ of $\mathcal{V}$ as $\overline{\mathcal{V}}_t := \{V : \mathcal{S} \to [0,H] \mid \forall s : V(s) \in \Xi_t\}$. For any t, $\{V_h \in \mathcal{V}\}_{h=2}^{H+1}$, and $\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}$, using $x^2 - y^2 = (x+y)(x-y)$ and abbreviating $p_h(s,a)$ and $\widehat{p}_h^t(s,a)$ respectively as $p_{h,s,a}$ and $\widehat{p}_{h,s,a}^t$,

$$\begin{aligned}
&\left|Y_t(V_2,\dots,V_{H+1}) - Y_t(\overline{V}_2,\dots,\overline{V}_{H+1})\right| \\
&\quad= \Big|\sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)(\widehat{p}_{h,s,a}^t - p_{h,s,a})^T(V_{h+1}+\overline{V}_{h+1})(\widehat{p}_{h,s,a}^t - p_{h,s,a})^T(V_{h+1}-\overline{V}_{h+1})\Big| \\
&\quad\leq 2H \sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)\left|(\widehat{p}_{h,s,a}^t - p_{h,s,a})^T(V_{h+1}-\overline{V}_{h+1})\right| \\
&\quad\leq 4Ht\|V_{h+1}-\overline{V}_{h+1}\|_\infty.
\end{aligned}$$

Therefore,

$$\min_{\{\overline{V}_h\in\overline{\mathcal{V}}_t\}_{h=2}^{H+1}} \left|Y_t(V_2,\dots,V_{H+1}) - Y_t(\overline{V}_2,\dots,\overline{V}_{H+1})\right| \leq 4H\xi_t t. \tag{15}$$

Now let $\alpha_t := 4H^2 \log(1/\delta_t) + 2ZH^2 \log(1+t) + 4H\xi_t t$ for a sequence $\{\delta_t\}_t$ of values in $(0,1)$ to be defined. We have

$$\begin{aligned}
&\mathbb{P}\left(\exists t \geq t_0, \{V_h \in \mathcal{V}\}_{h=2}^{H+1} : Y_t(V_2, \dots, V_{H+1}) \geq \alpha_t\right) \\
&\quad \leq \mathbb{P}\left(\exists t \geq t_0, \{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1} : Y_t(\overline{V}_2, \dots, \overline{V}_{H+1}) \geq \alpha_t - 4H\xi_t t\right) \\
&\quad \leq \sum_{t=t_0}^{\infty} \sum_{\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}} \mathbb{P}\left(Y_t(\overline{V}_2, \dots, \overline{V}_{H+1}) \geq 4H^2 \log(1/\delta_t) + 2ZH^2 \log(1+t)\right) \\
&\quad \leq \sum_{t=t_0}^{\infty} \sum_{\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}} \delta_t = \sum_{t=t_0}^{\infty} \delta_t \lfloor H/\xi_t \rfloor^{SH},
\end{aligned}$$

where the first inequality uses (15), the second one uses a union bound and the definition of α_t, the third one uses Lemma 24, and the equality uses the sizes of the two sets in the sums. Setting $\xi_t = H/t$ and $\delta_t = \frac{\delta}{2t^{SH+2}}$,

$$\sum_{t=t_0}^{\infty} \delta_t \lfloor H/\xi_t \rfloor^{SH} \leq \frac{\delta}{2} \sum_{t=t_0}^{\infty} \frac{1}{t^2} \leq \delta.$$

Finally, with these choices we have

$$\begin{aligned}
\alpha_t &= 4H^2 \log(1/\delta) + 4H^2 \log(2) + 4H^2 \log(t^{SH+2}) + 2ZH^2 \log(1+t) + 4H^2 \\
&\leq 4H^2 \log(1/\delta) + 4H^2 \log(2) + 12SH^3 \log(t) + 2ZH^2 \log(1+t) + 4H^2 \\
&\leq 4H^2 \log(1/\delta) + 12SH^3 \log(t) + 12ZH^2 \log(1+t).
\end{aligned}$$

This implies the statement. ■

Lemma 26 *[Concentration of $\widehat{r}$] Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$ and $Z := |\mathcal{Z}|$. With probability at least $1-\delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$,*

$$\sum_{(h,s,a) \in \mathcal{Z}} n_h^t(s,a) \left(\widehat{r}_h^t(s,a) - r_h(s,a)\right)^2 \leq 4 \log(1/\delta) + 2Z \log(1+t).$$

Proof Following the proof of Lemma 24, we build a suitable stochastic process to apply Theorem 1 of Abbasi-Yadkori et al. (2011). We define $\mathcal{F}_{t,h}, X_h^t, Y_t, D_t$ exactly as in the proof of Lemma 24, while we redefine $\eta_h^t := r_h^t - r_h(s_h^t, a_h^t)$, with r_h^t the random reward sample observed at stage h of episode t. Since rewards lie in $[0,1]$ almost surely, η_h^t is zero-mean and 1-subgaussian conditionally on $\mathcal{F}_{t,h}$. Moreover, it is easy to see that, for all $(h,s,a) \in \mathcal{Z}$,

$$[Y_t]_{h,s,a} = n_h^t(s,a)(\widehat{r}_h^t(s,a) - r_h(s,a)).$$

Theorem 1 of Abbasi-Yadkori et al. (2011) combined with Equation 20.9 from Lattimore and Szepesvari (2019) yield that

$$\mathbb{P}\left(\forall t \geq 1,\ \left\|Y^t\right\|^2_{(I+D_t)^{-1}} \leq 2\log(1/\delta) + Z\log(1+t/Z)\right) \geq 1-\delta.$$

We can then conclude exactly as in Lemma 24 by showing that $\|Y^t\|^2_{D_t^{-1}} \leq 2\|Y^t\|^2_{(I+D_t)^{-1}}$ for any $t \geq t_0$, which implies the statement. ■

D.2. Concentration results for RFE

For reward-free exploration, it is sufficient to concentrate the values of all *deterministic* policies. Our concentration result stated below features the threshold function

$$\beta^{RF}(t,\delta) := 4H^2 \log(1/\delta) + 24SH^3 \log(A(1+t)).$$

Theorem 27 *Let $\mathcal{Z} \subseteq [H]\times\mathcal{S}\times\mathcal{A}$ and $Z := |\mathcal{Z}|$. Suppose that, for some $\varepsilon_0 > 0$, $\max_\pi p_h^\pi(s,a) \leq \varepsilon_0$ for all $(h,s,a) \notin \mathcal{Z}$. With probability at least $1-\delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$, $\pi \in \Pi^{\mathrm{D}}$, and reward function $r \in [0,1]^{SAH}$,*

$$\Big|\sum_{h,s,a}\big(\widehat{p}_h^{\pi,t}(s,a) - p_h^\pi(s,a)\big)r_h(s,a)\Big| \leq \sqrt{\beta^{RF}(t,\delta)\sum_{(h,s,a)\in\mathcal{Z}}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + (SH - Z_\pi)H\varepsilon_0,$$

where $Z_\pi := |\mathcal{Z} \cap \{(h,s,\pi_h(s)) : h \in [H], s \in \mathcal{S}\}|$.

Proof Fix any reward r and deterministic policy π. Let V_h^π and $\widehat{V}_h^{\pi,t}$ denote the value functions of π under (p,r) and $(\widehat{p}^t,r)$, respectively. By Lemma 29 and the assumption on the set $\mathcal{Z}$,

$$\begin{aligned}\Big|\sum_{h,s,a}\big(\widehat{p}_h^{\pi,t}(s,a) - p_h^\pi(s,a)\big)r_h(s,a)\Big| &\leq \sum_{h,s,a} p_h^\pi(s,a)\big|(\widehat{p}_h^t(s,a) - p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\big| \\ &\leq \sum_{(h,s,a)\in\mathcal{Z}} p_h^\pi(s,a)\big|(\widehat{p}_h^t(s,a) - p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\big| + (SH - Z_\pi)H\varepsilon_0.\end{aligned}$$

By applying Lemma 25 on the set $\mathcal{Z}_\pi = \mathcal{Z} \cap \{(h,s,\pi_h(s)) : h \in [H], s \in \mathcal{S}\}$, whose cardinality is at most SH, and union bounding over all A^{SH} deterministic policies, with probability at least $1-\delta$, the following holds for all $t \geq t_0$, $\pi \in \Pi^{\mathrm{D}}$, and value functions bounded in $[0,H]$:

$$\sum_{(h,s,\pi_h(s))\in\mathcal{Z}} n_h^t(s,\pi_h(s))\big|(\widehat{p}_h^t(s,\pi_h(s)) - p_h(s,\pi_h(s)))^T V_{h+1}\big|^2 \leq \beta^{RF}(t,\delta).$$

Thus, by Lemma 30,

$$\begin{aligned}\sum_{(h,s,a)\in\mathcal{Z}} p_h^\pi(s,a)\big|(\widehat{p}_h^t(s,a) - p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\big| &= \sum_{(s,\pi_h(s),h)\in\mathcal{Z}} p_h^\pi(s)\big|(\widehat{p}_h^t(s,\pi_h(s)) - p_h(s,\pi_h(s)))^T\widehat{V}_{h+1}^{\pi,t}\big| \\ &\leq \sup_{\substack{u\in\mathbb{R}^{SH},\\ \sum_{(s,\pi_h(s),h)\in\mathcal{Z}} n_h^t(s,\pi_h(s))u_{s,h}^2 \leq \beta^{RF}(t,\delta)}} \sum_{(s,\pi_h(s),h)\in\mathcal{Z}} p_h^\pi(s)u_{s,h} \\ &= \sqrt{\beta^{RF}(t,\delta)\sum_{(h,s,a)\in\mathcal{Z}}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}}.\end{aligned}$$

■

D.3. Concentration results for BPI

For BPI, we need concentration bounds on $\left|\widehat{V}_1^{\pi,t} - V_1^{\pi}\right|$ that hold uniformly across all time steps and *stochastic* policies. Here $\widehat{V}_1^{\pi,t} := \sum_{h,s,a} \widehat{p}_h^{\pi,t}(s,a)\widehat{r}_h^t(s,a)$, where $\widehat{r}_h^t(s,a)$ is the MLE of $r_h(s,a)$ and $\widehat{p}_h^{\pi,t}(s,a)$ is an estimator of $p_h^{\pi}(s,a)$ computed from the MLEs $\{\widehat{p}_h(s'|s,a)\}_{h,s,a,s'}$ of the transition probabilities. To this end, we shall define the thresholds

$$\begin{aligned}
\beta^r(t,\delta) &:= 4\log(2/\delta) + 2SAH\log(1+t),\\
\beta^p(t,\delta) &:= 4H^2\log(2/\delta) + 24SAH^3\log(1+t),\\
\beta^{bpi}(t,\delta) &:= 16H^2\log(2/\delta) + 96SAH^3\log(1+t).
\end{aligned}$$

Compared to $\beta^{RF}(t,\delta)$, we note that $\beta^{bpi}(t,\delta)$ features larger multiplicative constants but also a dependency in A instead of $\log(A)$ in its second term which comes from the need to concentrate the values of all stochastic policies.

Theorem 28 *With probability at least $1-\delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall(h,s,a)\}$ and $\pi \in \Pi^{\mathrm{S}}$, the following holds:*

$$\left|\widehat{V}_1^{\pi,t} - V_1^{\pi}\right| \leq \sqrt{\beta^{bpi}(t,\delta)\min\Big(\sum_{h,s,a}\frac{p_h^{\pi}(s,a)^2}{n_h^t(s,a)}, \sum_{h,s,a}\frac{\widehat{p}_h^{\pi,t}(s,a)^2}{n_h^t(s,a)}\Big)}.$$

Moreover, for any $\widetilde{r} \in [0,1]^{SAH}$,

$$\Big|\sum_{h,s,a}\big(\widehat{p}_h^{\pi,t}(s,a) - p_h^{\pi}(s,a)\big)\widetilde{r}_h(s,a)\Big| \leq \sqrt{\beta^p(t,\delta)\sum_{h,s,a}\frac{p_h^{\pi}(s,a)^2}{n_h^t(s,a)}}\,.$$

Proof Fix any stochastic policy π. By Lemma 29,

$$\left|\widehat{V}_1^{\pi,t} - V_1^{\pi}\right| \leq \sum_{h,s,a} p_h^{\pi}(s,a)\left|\widehat{r}_h^t(s,a) - r_h(s,a)\right| + \sum_{h,s,a} p_h^{\pi}(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\right|.$$

By applying Lemma 26 and Lemma 25 for the set $\mathcal{Z} = \{(h,s,a) : h \in [H], s \in \mathcal{S}, a \in \mathcal{A}\}$, which is of cardinality SAH, with probability at least $1-\delta$, the following hold for all $t \geq t_0$ and for all value functions $(V_h)_{h\in[H]}$ supported in $[0,H]$:

$$\begin{aligned}
\sum_{h,s,a} n_h^t(s,a)\left|\widehat{r}_h^t(s,a) - r_h(s,a)\right|^2 &\leq \beta^r(t,\delta),\\
\sum_{h,s,a} n_h^t(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}^{\pi,t}\right|^2 &\leq \beta^p(t,\delta).
\end{aligned} \tag{16}$$

Thus, by Lemma 30, optimizing over the deviations as in the proof of Lemma 27,

$$\left|\widehat{V}_1^{\pi,t} - V_1^{\pi}\right| \leq \sqrt{\beta^r(t,\delta)\sum_{h,s,a}\frac{p_h^{\pi}(s,a)^2}{n_h^t(s,a)}} + \sqrt{\beta^p(t,\delta)\sum_{h,s,a}\frac{p_h^{\pi}(s,a)^2}{n_h^t(s,a)}}.$$

Using that $\beta^r(t,\delta) \leq \beta^p(t,\delta)$ and noting that $\beta^{bpi}(t,\delta) = 4\beta^p(t,\delta)$ proves the first statement with the first term in the minimum only. To prove it with the second term as well, it is enough to use Lemma 29 with the roles of the two value functions swapped and repeat the same steps as above.

To prove the second statement, we proceed as in the proof of Theorem 27 and write

$$\begin{aligned}\Big|\sum_{h,s,a}\big(\widehat{p}_h^{\pi,t}(s,a)-p_h^\pi(s,a)\big)\widetilde{r}_h(s,a)\Big| &\leq \sum_{h,s,a} p_h^\pi(s,a)\big|(\widehat{p}_h^t(s,a)-p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\big| \\ &\leq \sup_{\substack{u\in\mathbb{R}^{SH},\\ \sum_{h,s,a} n_h^t(s,a)u_{h,s,a}^2\leq\beta^p(t,\delta)}} \sum_{h,s,a} p_h^\pi(s,a)u_{h,s,a} \\ &= \sqrt{\beta^p(t,\delta)\sum_{h,s,a}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}},\end{aligned}$$

where we used Lemma 30 and together with inequality (16). ■

D.4. Auxiliary results

Lemma 29 (Lemma E.15 of Dann et al. (2017)) *Consider two MDPs with transitions $p, \widehat{p}$ and rewards $r, \widehat{r}$, respectively. Let $V_h^\pi, \widehat{V}_h^\pi$ denote the value function of a (possibly stochastic) policy π in these two MDPs. Then, for any s, h,*

$$V_h^\pi(s) - \widehat{V}_h^\pi(s) = \widehat{\mathbb{E}}^\pi\left[\sum_{\ell=h}^H \Big(r_\ell(s_\ell,a_\ell) - \widehat{r}_\ell(s_\ell,a_\ell) + \big(p_\ell(s_\ell,a_\ell) - \widehat{p}_\ell(s_\ell,a_\ell)\big)^T V_{\ell+1}^\pi\Big)\middle| s_h = s\right].$$

Lemma 30 *Let $n \in \mathbb{N}$, $p, b \in \mathbb{R}^n$ with b having strictly positive entries, and $c \in \mathbb{R}_{\geq 0}$. Then,*

$$\sup_{\substack{x\in\mathbb{R}^n:\\ \sum_{i=1}^n b_i x_i^2 \leq c}} \sum_{i=1}^n p_i x_i = \sqrt{c\sum_{i=1}^n \frac{p_i^2}{b_i}}.$$

Proof Let v be the value of the optimization program. Then we know that

$$-v = \inf_{\substack{x\in\mathbb{R}^n:\\ \sum_{i=1}^n b_i x_i^2 \leq c}} -\sum_{i=1}^n p_i x_i. \tag{17}$$

The Lagrangian of the quadratic program above writes as

$$\mathcal{L}(x,\lambda) = -\sum_{i=1}^n p_i x_i + \lambda\Big(\sum_{i=1}^n b_i x_i^2 - c\Big),$$

where $\lambda \geq 0$. The KKT conditions then yield that the optimal solution satisfies that

$$\forall i \in [|1,n|], \quad x_i = -\frac{p_i}{2\lambda b_i}$$
$$\sum_{i=1}^n b_i x_i^2 = c$$

Solving this system yields that the optimal Lagrange multiplier $\lambda = \sqrt{\frac{c}{\sum_{i=1}^n \frac{p_i^2}{b_i}}}$ which implies that the value of (17) is $-\sqrt{c\sum_{i=1}^n \frac{p_i^2}{b_i}}$. ■

Appendix E. Analysis of PCE

To simplify the presentation of the algorithm and the analysis, we index the counts as well as the empirical estimates of transitions and rewards by their phase number. Hence, for each triplet (h,s,a), $n_h^k(s,a)$ and $\widehat{p}_h^k(.|s,a)$ will refer to the number of visits and the empirical transition kernel respectively after t_k episodes, i.e. at the end of the k-th phase. Finally, for a dataset of episodes $\mathcal{D}$, $n_h(s,a;\mathcal{D})$ denotes the number of visits of (h,s,a) in the episodes stored in $\mathcal{D}$.

E.1. Good event

We introduce the following events

$$\mathcal{E}_{vis} := \Bigg(\text{The set built using } \textsc{EstimateReachability}\left((h,s);\frac{\varepsilon}{4SH^2},\frac{\delta}{3SH}\right) \text{ for all } (h,s)$$
$$\text{satisfies } \left\{(h,s):\sup_\pi p_h^\pi(s)\geq\frac{\varepsilon}{4SH^2}\right\}\subseteq\widehat{\mathcal{X}}\subseteq\left\{(h,s):\sup_\pi p_h^\pi(s)\geq\frac{\varepsilon}{32SH^2}\right\}$$
$$\text{and } \forall(h,s)\in\widehat{\mathcal{X}},\ \sup_\pi p_h^\pi(s)\leq\overline{W}_h(s)\leq 36\sup_\pi p_h^\pi(s)\Bigg),$$
$$\mathcal{E}_p^{RF} := \Bigg(\forall k\in\mathbb{N}^\star,\forall\pi\in\Pi^D,\forall r\in[0,1]^{SAH},$$
$$\left|\sum_{s,a,h}\left(\widehat{p}_h^{\pi,k}(s,a)-p_h^\pi(s,a)\right)r_h(s,a)\right|\leq\sqrt{\beta^{RF}(t_k,\delta/3)\sum_{(s,a,h)\in\widehat{\mathcal{X}}}\frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}}+\frac{\varepsilon}{4}\Bigg),$$
$$\mathcal{E}_{cov} := \Bigg(\forall k\in\mathbb{N},\ \text{CovGame run with inputs } (c^k,\delta/6(k+1)^2) \text{ terminates after at most}$$
$$64m_k\varphi^\star(c^k)+\widetilde{\mathcal{O}}\big(m_k\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6(k+1)^2/\delta)+S)\big) \text{ episodes and returns a dataset } \mathcal{D}_k$$
$$\text{such that for all } (h,s,a)\in\widehat{\mathcal{X}}, n_h(s,a;\mathcal{D}_k)\geq c_h^k(s,a)\Bigg),$$

where $m_k=\log_2\left(\frac{\max_{s,a,h}c_h^k(s,a)}{\min_{s,a,h}c_h^k(s,a)\vee 1}\right)\vee 1$ and β^{RF} is defined in appendix D.2. Then our good event is defined as the intersection

$$\mathcal{E}_{good}^{RF} := \mathcal{E}_{vis}\cap\mathcal{E}_p^{RF}\cap\mathcal{E}_{cov}.$$

Lemma 31 *We have that* $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_{good}^{RF})\geq 1-\delta$.

Proof Let $\overline{\mathcal{E}}$ denote the complementary event of $\mathcal{E}$. We start by the following decomposition

$$\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{good}^{RF}})\leq\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{vis}})+\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}})+\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_p^{RF}}\cap\mathcal{E}_{vis}\cap\mathcal{E}_{cov}).$$

Now we bound each term separately. First observe that applying Theorem 51 with parameter $\varepsilon_0 = \varepsilon/4SH^2$ yields $\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{vis}}) \leq \delta/3$. Second, using Corollary 4 we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) &\leq \sum_{k=0}^{\infty} \mathbb{P}_{\mathcal{M}}(\text{CovGame with inputs } (c^k, \delta/6(k+1)^2) \text{ fails}) \\ &\leq \sum_{k=0}^{\infty} \frac{\delta}{6(k+1)^2} = \frac{\delta\pi^2}{36} \leq \delta/3.\end{aligned}$$

Next, note that by design of PCE $n_h^0(s,a) = n_h(s,a;\widetilde{\mathcal{D}}_0)$ and $c^0 = \mathbb{1}_{\widehat{\mathcal{X}}}$ so that $\mathcal{E}_{cov} \subset \big(\forall (h,s,a) \in \widehat{\mathcal{X}},\ n_h^0(s,a) \geq 1\big)$. Therefore we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_p^{RF}} \cap \mathcal{E}_{vis} \cap \mathcal{E}_{cov}) &\leq \mathbb{P}_{\mathcal{M}}\big(\overline{\mathcal{E}_p^{RF}},\ \big\{(h,s) : \sup_{\pi} p_h^{\pi}(s) \geq \frac{\varepsilon}{4SH^2}\big\} \subseteq \widehat{\mathcal{X}},\ \forall (h,s,a) \in \widehat{\mathcal{X}}\ n_h^0(s,a) \geq 1\big) \\ &= \mathbb{P}_{\mathcal{M}}\Bigg(\big\{(h,s) : \sup_{\pi} p_h^{\pi}(s) \geq \frac{\varepsilon}{4SH^2}\big\} \subseteq \widehat{\mathcal{X}},\ \exists k \geq 0\ \exists \pi \in \Pi^D\ \exists r \in [0,1]^{SAH} : \\ &\qquad \Big|\sum_{s,a,h} \big(\widehat{p}_h^{\pi,k}(s,a) - p_h^{\pi}(s,a)\big) r_h(s,a)\Big| > \sqrt{\beta^{RF}(t_k, \delta/3) \sum_{(s,a,h)\in\widehat{\mathcal{X}}} \frac{p_h^{\pi}(s,a)^2}{n_h^k(s,a)}} + \frac{\varepsilon}{4}\Bigg) \\ &\overset{(a)}{\leq} \mathbb{P}_{\mathcal{M}}\Bigg(\big\{(h,s) : \sup_{\pi} p_h^{\pi}(s) \geq \frac{\varepsilon}{4SH^2}\big\} \subseteq \widehat{\mathcal{X}},\ \exists t \geq t_0\ \exists \pi \in \Pi^D\ \exists r \in [0,1]^{SAH} : \\ &\qquad \Big|\sum_{s,a,h} \big(\widehat{p}_h^{\pi,t}(s,a) - p_h^{\pi}(s,a)\big) r_h(s,a)\Big| > \sqrt{\beta^{RF}(t, \delta/3) \sum_{(s,a,h)\in\widehat{\mathcal{X}}} \frac{p_h^{\pi}(s,a)^2}{n_h^t(s,a)}} + \frac{\varepsilon}{4}\Bigg) \\ &\overset{(b)}{\leq} \delta/3,\end{aligned}$$

where in (a) we introduced $t_0 = \inf\{t \geq 1 : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \widehat{\mathcal{X}}\}$ and switched back to indexing counts and estimates by the episode number (instead of the phase) in order to apply Theorem 27 in (b) with $\mathcal{Z} = \{(h,s,a) : (h,s) \in \widehat{\mathcal{X}}\}$ and $\varepsilon_0 = \varepsilon/4SH^2$. Combining the four inequalities above yields the desired result. ■

E.2. Low concentrability / Good coverage of all policies

The next lemma shows that PCE achieves proportional coverage.

Lemma 32 *Under the good event, for all phases $k \geq 0$, we have that*

$$n_h^k(s,a) \geq 2^k \sup_{\pi} p_h^{\pi}(s,a) \quad \forall (h,s,a) \in \widehat{\mathcal{X}}.$$

Proof First of all, note that for any triplet $(h,s,a) \in \widehat{\mathcal{X}}$, $\sup_{\pi} p_h^{\pi}(s,a)$ is always attained by some deterministic policy. Therefore, it is sufficient to prove that, given a fixed deterministic policy $\pi \in \Pi^D$,

$$\forall k \geq 0, \forall (h,s,a) \in \widehat{\mathcal{X}}, \quad n_h^k(s,a) \geq 2^k p_h^{\pi}(s,a)\ .$$

We do this by induction over k. For $k = 0$ the result is trivial since, under the good event, we have that for all $(h, s, a) \in \widehat{\mathcal{X}}$, $n_h^0(s, a) \geq c_h^0(s, a) = 1 \geq 2^0 p_h^\pi(s, a)$. Now suppose that the property holds for phase k. Then under the good event we know that for all (h, s, a), $n_h^{k+1}(s, a) - n_h^k(s, a) = n_h(s, a, \mathcal{D}_{k+1}) \geq c_h^{k+1}(s, a)$. Plugging the definition of c^{k+1} (Line 9 of Algorithm 2) we get that for any $(h, s, a) \in \widehat{\mathcal{X}}$,

$$\begin{aligned} n_h^{k+1}(s, a) &\geq c_h^{k+1}(s, a) \\ &= 2^{k+1} \overline{W}_h(s) \\ &\geq 2^{k+1} \sup_\pi p_h^\pi(s) \\ &= 2^{k+1} \sup_\pi p_h^\pi(s, a), \end{aligned} \tag{18}$$

where the second inequality uses the event $\mathcal{E}_{vis}$. ■

E.3. Correctness

Lemma 33 *Let $\widehat{p}$ be the estimate of the transition probabilities that* PCE *outputs. For any reward function r, let $\hat{\pi}_r$ be an optimal policy in the MDP $(\widehat{p}, r)$. Then*

$$\mathbb{P}\left(\forall r \in [0, 1]^{SAH}, V_1^{\hat{\pi}_r}(s_1; r) \geq V_1^\star(s_1; r) - \varepsilon\right) \geq 1 - \delta.$$

In other words, PCE *is (ε, δ)-PAC for reward-free exploration.*

Proof Assume that PCE stops as phase k and let $\widehat{p}^k$ denote the empirical transition estimates that it returns. Fix any reward function $r = [r_h(s, a)]_{h,s,a} \in [0, 1]^{SAH}$ and let $\widehat{\pi} \in \arg\max_{\pi \in \Pi^D} (\widehat{p}^{\pi,k})^\top r$ be the policy obtained when planning for reward function r under the transition model $\widehat{p}^k$. Further define $\pi^\star \in \arg\max_{\pi \in \Pi^D} (p^\pi)^\top r$, $V_1^\star := (p^{\pi^\star})^\top r$, and $V_1^{\widehat{\pi}} := (p^{\widehat{\pi}})^\top r$. Note that both $\widehat{\pi}$ and $\pi^\star$ are deterministic. Therefore under the good event $\mathcal{E}_{good}^{RF}$ we have

$$\begin{aligned} V_1^{\widehat{\pi}} &= (p^{\widehat{\pi}})^\top r \\ &\overset{(a)}{\geq} (\widehat{p}^{\widehat{\pi},k})^\top r - \sqrt{\beta^{RF}(t_k, \delta/3) \sum_{(s,a,h) \in \widehat{\mathcal{X}}} \frac{p_h^{\widehat{\pi}}(s, a)^2}{n_h^k(s, a)}} - \frac{\varepsilon}{4} \\ &\overset{(b)}{\geq} (\widehat{p}^{\pi^\star,k})^\top r - \sqrt{\beta^{RF}(t_k, \delta/3) \sum_{(s,a,h) \in \widehat{\mathcal{X}}} \frac{p_h^{\widehat{\pi}}(s, a)^2}{n_h^k(s, a)}} - \frac{\varepsilon}{4} \\ &\overset{(c)}{\geq} (p^{\pi^\star})^\top r - \sqrt{\beta^{RF}(t_k, \delta/3) \sum_{(s,a,h) \in \widehat{\mathcal{X}}} \frac{p_h^{\pi^\star}(s, a)^2}{n_h^k(s, a)}} - \sqrt{\beta^{RF}(t_k, \delta/3) \sum_{(s,a,h) \in \widehat{\mathcal{X}}} \frac{p_h^{\widehat{\pi}}(s, a)^2}{n_h^k(s, a)}} - \frac{\varepsilon}{2} \\ &\overset{(d)}{\geq} V_1^\star - 2\sqrt{H \beta^{RF}(t_k, \delta/3) 2^{-k}} - \frac{\varepsilon}{2} \\ &\overset{(e)}{\geq} V_1^\star - \varepsilon, \end{aligned}$$

where (a) and (c) use the good event $\mathcal{E}_p^{RF}$ for policies $\widehat{\pi}$ and $\pi^\star$ respectively, (b) uses the definition of $\widehat{\pi}$, (d) uses Lemma 32 and (e) uses the stopping condition of PCE (Line 10 in Algorithm **??**). Note that the inequality above holds, under the good event $\mathcal{E}_{good}$, jointly for all reward functions r. Since $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_{good}) \geq 1-\delta$, we have just proved that PCE is (ε, δ)-PAC for reward-free exploration. ■

E.4. Upper bound on the number of phases

Lemma 34 *Define the index of the final phase of PCE,* $\kappa_f := \inf\left\{k \in \mathbb{N}_+ : \sqrt{H\beta^{RF}(t_k, \delta/3)2^{4-k}} \leq \varepsilon\right\}$. *Further let* τ *denote the number of episodes played by the algorithm. Then under the good event, it holds that* $\kappa_f < \infty$ *and*

$$2^{\kappa_f} \leq \frac{32H\beta^{RF}(\tau, \delta/3)}{\varepsilon^2}.$$

Proof First we prove that κ_f is finite. Under the good event we have

$$\begin{aligned} t_k &= \sum_{j=0}^{k} d_j \\ &\leq \sum_{j=0}^{k} \left[64 m_j \varphi^\star(c^j) + \widetilde{\mathcal{O}}\big(m_j \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2(\log(6(j+1)^2/\delta) + S)\big)\right], \end{aligned}$$

where we recall that $m_j = \log_2\left(\frac{\max_{s,a,h} c_h^j(s,a)}{\min_{s,a,h} c_h^j(s,a) \vee 1}\right) \vee 1$. Now using the fact that $c_h^j(s,a) \leq 2^j \mathbb{1}((h,s,a) \in \widehat{\mathcal{X}})$ for $j \geq 0$ we deduce that $m_0 = 1$ and $m_j \leq j \ \forall j \geq 1$ so that

$$\begin{aligned} t_k &\leq \sum_{j=0}^{k} \left[8(j+1)2^j \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) + \widetilde{\mathcal{O}}\big((j+1)\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2(\log(4(j+1)^2/\delta) + S)\big)\right] \\ &= \mathcal{O}_{k\to\infty}\big(k^2 2^k\big). \end{aligned} \tag{19}$$

Now recall that the threshold β^{RF} was defined in Appendix D as

$$\beta^{RF}(t, \delta) = 4H^2 \log(1/\delta) + 24SH^3 \log(A(1+t)) \tag{20}$$

Combining (19) and (20) gives that

$$\beta^{RF}(t_k, \delta/3) = o_{k\to\infty}\big(2^k\big).$$

Therefore $\kappa_f = \inf\left\{k \in \mathbb{N}_+ : \sqrt{H\beta^{RF}(t_k, \delta/3)2^{4-k}} \leq \varepsilon\right\}$ is indeed finite. The proof of the second statement is straightforward by noting that $\kappa_f - 1$ does not satisfy the stopping condition (Line 12 in Algorithm 2) and using the (crude) upper bound $t_{\kappa_f - 1} \leq \tau$. ■

E.5. Upper bound on the phase length

Lemma 35 *Let $k \geq 1$ be such that* PCE *did not stop before phase k. Under the good event, the number of episodes played by* PCE *during phase k satisfies*

$$d_k \leq c_1 k H \beta^{RF}(\tau, \delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) + \widetilde{\mathcal{O}}\left(k\frac{S^3A^2H^5(\log(6(k+1)^2/\delta)+S)}{\varepsilon}\right),$$

where $c_1 = 73728$. Furthermore, the duration of the initial phase is upper bounded as

$$d_0 \leq \widetilde{\mathcal{O}}\left(\frac{S^3A^2H^5(\log(6/\delta)+S)}{\varepsilon}\right).$$

Proof Using the good event and the definition of c^k we write

$$\begin{aligned} d_k &\leq 64 m_k \varphi^\star\left(\left[2^k \overline{W}_h(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) + \widetilde{\mathcal{O}}\big(m_k\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6(k+1)^2/\delta)+S)\big) \\ &\overset{(a)}{\leq} 64k\varphi^\star\left(\left[2^k\overline{W}_h(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) + \widetilde{\mathcal{O}}\big(k\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6(k+1)^2/\delta)+S)\big), \end{aligned} \tag{21}$$

where (a) uses that $m_k = \log_2\left(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a)\vee 1}\right) \vee 1 \leq k$. Now by definition of the good event we have that for any triplet $(h,s,a) \in \widehat{\mathcal{X}}$, $\overline{W}_h(s) \leq 36 \sup_\pi p_h^\pi(s)$. Therefore

$$\begin{aligned} \varphi^\star\left(\left[2^k\overline{W}_h(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) &\overset{(a)}{\leq} \varphi^\star\left(\left[36\times 2^k \sup_\pi p_h^\pi(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) \\ &\overset{(b)}{\leq} \varphi^\star\left(\left[\frac{1152H\beta^{RF}(\tau,\delta/3)\sup_\pi p_h^\pi(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\overset{(c)}{\leq} 1152H\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s)\geq\frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right), \end{aligned} \tag{22}$$

where (a) uses that $\varphi^\star(c) \leq \varphi^\star(c')$ if $\forall(h,s,a)\ c_h(s,a) \leq c'_h(s,a)$, (b) uses Lemma 34 and the fact that $k \leq \kappa_f$ since PCE did not stop before phase k and (c) uses Lemma 10 and the fact that $\widehat{\mathcal{X}} \subseteq \left\{(h,s,a) : \sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right\}$ on the good event. Using again this last property yields

$$\begin{aligned} \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) &\leq \sum_{h,s,a} \frac{\mathbb{1}\left((h,s,a)\in\hat{\mathcal{X}}\right)}{\sup_\pi p_h^\pi(s,a)} \\ &= \sum_{(h,s,a)\in\widehat{\mathcal{X}}} \frac{1}{\sup_\pi p_h^\pi(s)} \leq \frac{32H^3S^2A}{\varepsilon}, \end{aligned} \tag{23}$$

where the first inequality uses Lemma 12. Combining (21), (22) and (23) proves the statement for $k \geq 1$. Now it remains to upper bound the duration of the burn-in phase. To that end, we write that by definition of the good event

$$d_0 \leq 64 m_0 \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) + \widetilde{\mathcal{O}}\big(\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2(\log(6/\delta) + S)\big),$$

where $m_0 = \log_2\left(\frac{\max_{s,a,h} c_h^0(s,a)}{\min_{s,a,h} c_h^0(s,a) \vee 1}\right) \vee 1 = 1$. Therefore

$$\begin{aligned} d_0 &\leq \widetilde{\mathcal{O}}\big(\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2(\log(6/\delta) + S)\big) \\ &\leq \widetilde{\mathcal{O}}\left(\frac{S^3A^2H^5(\log(6/\delta) + S)}{\varepsilon}\right), \end{aligned}$$

where the last inequality uses (23). ■

E.6. Total sample complexity

Theorem 36 *With probability at least $1-\delta$, the total sample complexity of* PCE *satisfies*

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

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in S, A, H, ε and $\log(1/\delta)$.

Proof Denoting by T_{vis} the number of episodes used by the ESTIMATEREACHABILITY sub-routine in line 2 of the algorithm, we write

$$\begin{aligned} \tau &= T_{vis} + \sum_{k=0}^{\kappa_f} d_k \\ &\leq T_{vis} + \widetilde{\mathcal{O}}\left(\frac{S^3A^2H^5(\log(6/\delta)+S)}{\varepsilon}\right) \\ &\quad + \sum_{k=1}^{\kappa_f}\left[c_1 kH\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s,a)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right)\right. \\ &\quad \left. + \widetilde{\mathcal{O}}\left(k\frac{S^3A^2H^5(\log(6(k+1)^2/\delta)+S)}{\varepsilon}\right)\right] \\ &\leq T_{vis} + c_1\kappa_f^2 H\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right), \end{aligned} \tag{24}$$

where we used Lemma 35 to upper bound $(d_k)_{k\geq 0}$. From Theorem 51, we know that T_{vis} is deterministic and satisfies

$$T_{vis} = \widetilde{\mathcal{O}}\left(\frac{S^3AH^4\left(\log\left(\frac{SAH}{\delta}\right)+S\right)}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right). \tag{25}$$

Combining inequalities (24) and (25) with the definition of the threshold $\beta^{RF}(t,\delta) = 4H^2\log(1/\delta) + 24SH^3\log(A(1+t))$ we get

$$\begin{aligned}\tau &\le c_1\kappa_f^2 H\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \ge \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right) \\ &\le c_2\kappa_f^2\left(H^3\log(1/\delta) + SH^4\log(A(1+\tau))\right)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \ge \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right), \end{aligned} \tag{26}$$

where $c_2 = 24c_1$ On the other hand, thanks to Lemma 34 and the definition of the threshold β^{RF} we have that

$$\kappa_f \le \log_2\left(\frac{128H^3\log(1/\delta) + 768SH^4\log(A(1+\tau))}{\varepsilon^2}\right) \tag{27}$$

Combining (26) with (27) and solving for τ we get that

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

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in S, A, H, ε and $\log(1/\delta)$. ■

E.7. Benign instances for PCE

In this section we propose some MDP instances in which the quantity

$$\mathcal{C}(\text{PCE},\varepsilon) := \varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a})H^3/\varepsilon^2, \tag{28}$$

which is (an upper bound on) the leading term in the small (δ,ε) regime in our sample complexity bound for PCE, can be smaller than the minimax rate SAH^3/ε^2.

E.7.1. DISGUISED CONTEXTUAL BANDITS

Lemma 37 *Suppose that $\mathcal{M}$ is a "disguised" contextual bandit, i.e.,*

$$\forall(h,s,a,s'),\ p_h(s'|s,a) = p_h(s'|s).$$

Then $\mathcal{C}(PCE,\varepsilon) = AH^3/\varepsilon^2$.

Proof In this case for any (h,s) and any policy π, $p_h^\pi(s) = p_h(s)$ is independent of the policy. Thanks to Lemma 9 we have

$$\begin{aligned}
\varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a}) &= \inf_{\pi^{exp}\in\Pi^S} \max_{s,a,h} \frac{\sup_\pi p_h^\pi(s,a)}{p_h^{\pi^{exp}}(s,a)} \\
&= \inf_{\pi^{exp}\in\Pi^S} \max_{s,a,h} \frac{p_h(s)\sup_\pi \pi_h(a|s)}{p_h(s)\pi_h^{exp}(a|s)} \\
&= \inf_{\pi^{exp}\in\Pi^S} \max_{s,h} \frac{1}{\min_a \pi_h^{exp}(a|s)} \\
&= A,
\end{aligned}$$

where the last equality is because $(\min_a \pi_h^{exp}(a|s))^{-1} \geq A$ and the infimum over Π^S is achieved by the uniform policy. ■

E.7.2. ERGODIC MDPS

Let $\alpha, \beta \in (0,1)$ such that $\alpha > \beta$. Further define the set of probability vectors such that

$$\mathcal{P}_{\alpha,\beta} = \left\{ q \in \mathbb{R}_+^S : \sum_{i=1}^S q_i = 1,\ \max_i q_i \leq S^{\alpha-1}, \min_i q_i \geq \frac{1-S^{\beta-1}}{S-1} \right\}.$$

Note that such set is never empty since the vector $(S^{\beta-1}, \frac{1-S^{\beta-1}}{S-1}, \dots, \frac{1-S^{\beta-1}}{S-1})$ always satisfies the inequalities in its definition. We define the class of MDPs $\mathfrak{M}_{erg}$ such that their transition kernel satisfies

$$\forall (h,s,a),\ p_h(.|s,a) \in \mathcal{P}_{\alpha,\beta}.$$

Lemma 38 *Assume that $\mathcal{M} \in \mathfrak{M}_{erg}$, then $\mathcal{C}(PCE,\varepsilon) \leq S^\alpha AH^4/\varepsilon^2$.*

Remark 39 *Note that the "ergodicity" of MDPs in $\mathfrak{M}_{erg}$ can be as small as one wishes: by taking the limit $\beta \to 1$, the constraint $\min_{s'} p_h(s'|s,a) \geq \frac{1-S^{\beta-1}}{S-1}$ becomes vacuous so the MDP can be non-ergodic. In that regime, $\alpha = 1$ and we recover the minimax sample complexity (up to an H factor) SAH^3/ε^2.*

Proof First of all we note that

$$\begin{aligned}
\forall \pi \in \Pi\ \forall s \in \mathcal{S},\ p_h^\pi(s) &= \sum_{s'\in\mathcal{S}} p_{h-1}^\pi(s) p_h(s|s', \pi_{h-1}(s')) \\
&\leq \sum_{s'\in\mathcal{S}} p_{h-1}^\pi(s) S^{\alpha-1} = S^{\alpha-1}. \qquad (29)
\end{aligned}$$

Similarly

$$\forall \pi \in \Pi\ \forall s \in \mathcal{S},\ p_h^\pi(s) \geq \frac{1-S^{\beta-1}}{S-1}. \qquad (30)$$

Now using Lemma 13 we have that

$$\begin{aligned}\varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a}) &\le \sum_{h=1}^H \inf_{\pi^{exp}\in\Pi^S} \max_s \frac{1}{p_h^{\pi^{exp}}(s)} \sum_a \sup_\pi p_h^\pi(s,a) \\ &= \sum_{h=1}^H \inf_{\pi^{exp}\in\Pi^S} \max_s \frac{A\sup_\pi p_h^\pi(s)}{p_h^{\pi^{exp}}(s)} = A\sum_{h=1}^H \underbrace{\inf_{\pi^{exp}\in\Pi^S} \max_s \frac{\sup_\pi p_h^\pi(s)}{p_h^{\pi^{exp}}(s)}}_{:=\mathcal{C}_h},\end{aligned} \tag{31}$$

Now fix $h \in [H]$ and denote by π^s any policy in $\arg\max_{\pi\in\Pi} p_h^\pi(s)$. Further define the stochastic policy $\widetilde{\pi}$ such that

$$p^{\widetilde{\pi}} = \frac{\sum_{s'\in\mathcal{S}} p^{\pi^s}}{S}.$$

Using (30) we have that

$$\begin{aligned}\forall s\in\mathcal{S},\ p_h^{\widetilde{\pi}}(s) &= \frac{\sum_{s'\in\mathcal{S}} p_h^{\pi^{s'}}(s)}{S} \\ &\ge \frac{\sup_{\pi\in\Pi} p_h^\pi(s) + (S-1)\frac{1-S^{\beta-1}}{S-1}}{S} \\ &= \frac{\sup_{\pi\in\Pi} p_h^\pi(s) + 1 - S^{\beta-1}}{S}.\end{aligned} \tag{32}$$

Therefore

$$\begin{aligned}\mathcal{C}_h &= \inf_{\pi^{exp}\in\Pi^S} \max_s \frac{\sup_\pi p_h^\pi(s)}{p_h^{\pi^{exp}}(s)} \\ &\le \max_s \frac{\sup_\pi p_h^\pi(s)}{p_h^{\widetilde{\pi}}(s)} \\ &\overset{(a)}{\le} \max_s \frac{S\sup_\pi p_h^\pi(s)}{\sup_{\pi\in\Pi} p_h^\pi(s) + 1 - S^{\beta-1}} \\ &= \max_s \frac{S}{1 + \frac{1-S^{\beta-1}}{\sup_\pi p_h^\pi(s)}} \\ &\overset{(b)}{\le} \max_s \frac{S}{1 + S^{1-\alpha}(1-S^{\beta-1})} \\ &= \frac{S}{1 + S^{1-\alpha} - S^{\beta-\alpha}} \le S^\alpha,\end{aligned}$$

where (a) uses (32) and (b) uses (29). Combining (31) with the previous inequality yields that $\varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a}) \le S^\alpha AH$. ■

Appendix F. PRINCIPLE and its Analysis

F.1. Pseudo-code of PRINCIPLE

The pseudo code of PRINCIPLE is detailed in Algorithm 3.

Algorithm 3 PRINCIPLE (PRoportIoNal Coverage with Implicit PoLicy Elimination)

Input: Precision ε, Confidence δ, set of reachable states $\mathcal{S}$
Output: A policy $\widehat{\pi}$ that is ε-optimal w.p larger than $1-\delta$
Define target function $c_h^0(s,a) = 1$ for all (h,s,a)
Execute CovGame$\big(c^0,\ \delta/4\big)$ to get dataset $\mathcal{D}_0$ and number of episodes d_0 // BURN-IN PHASE
Initialize episode count $t_0 \leftarrow d_0$ and statistics $n_h^0(s,a), \widehat{r}_h^0(s,a), \widehat{p}_h^0(.|s,a)$ using $\widetilde{\mathcal{D}}_0$
Initialize the set of active distributions $\Omega^0 \leftarrow \Omega(\widehat{p}^0)$
for $k = 1, \dots$ **do**
  // PROPORTIONAL COVERAGE
  Compute $c_h^k(s,a) := 2^k \min\big(\sup_{\widehat{\rho}\in\Omega^{k-1}} \widehat{\rho}_h(s,a) + 2\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^k,\delta/2)2^{1-k}},\ 1\big)$ for all (h,s,a)
  Execute CovGame$\big(c^k,\ \delta/4(k+1)^2\big)$ to get dataset $\widetilde{\mathcal{D}}_k$ and number of episodes T_k
  if $T_k > SAH2^k$ **then**
    Run PruneDataset$(\widetilde{\mathcal{D}}_k, c^k)$ to get *effective* dataset $\mathcal{D}_k$ and *effective phase length* d_k
  else
    Set $d_k \leftarrow T_k$ and $\mathcal{D}_k \leftarrow \widetilde{\mathcal{D}}_k$
  end if
  Update *effective episode count* $t_k \leftarrow t_{k-1} + d_k$ and statistics $n_h^k(s,a), \hat{r}_h^k(s,a), \widehat{p}_h^k(.|s,a)$ using $\mathcal{D}_k$
  // STATE-ACTION-DISTRIBUTION ELIMINATION
  Compute the lower confidence bound

$$\underline{V}_1^k := \sup_{\substack{\widehat{\rho}\in\Omega(\widehat{p}^k),\\ \max_{h,s,a} \widehat{\rho}_h(s,a)/n_h^k(s,a)\le 2^{-k}}} \widehat{\rho}^\top \widehat{r}^k - \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)}$$

  Update the set of active state-action distributions

$$\Omega^k \leftarrow \left\{\widehat{\rho}\in\Omega(\widehat{p}^k):\ \widehat{\rho}^\top\widehat{r}^k \ge \underline{V}_1^k \text{ and } \max_{h,s,a} \widehat{\rho}_h(s,a)/n_h^k(s,a) \le 2^{-k}\right\}$$

  if $\sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \le \varepsilon$ **then**
    Compute any $\widehat{\rho}^\star \in \arg\max_{\widehat{\rho}\in\Omega^k} \widehat{\rho}^\top\widehat{r}^k$ and extract the corresponding policy $\widehat{\pi}$
    return $\widehat{\pi}$
  end if
end for

F.2. Analysis of PRINCIPLE

To simplify the presentation of the algorithm and the analysis, we index the counts as well as the empirical estimates of transitions and rewards by their phase number. Hence, for each triplet (h,s,a), $n_h^k(s,a), \widehat{p}_h^k(.|s,a)$ and $\widehat{r}_h^k(s,a)$ will refer to the number of visits, the empirical transition kernel and the empirical mean reward respectively after t_k episodes, i.e. at the end of the k-th phase. For a transition kernel $\widetilde{p}$, we define the corresponding set of state-action distributions as

Algorithm 4 PruneDataset

Input: Target counts c, Dataset $\widetilde{\mathcal{D}}$ such that $n_h(s,a;\widetilde{\mathcal{D}}) \geq c_h(s,a)$ for all (h,s,a)
Output: A dataset $\mathcal{D}$ of $d \leq SAH2^k$ episodes satisfying $n_h(s,a;\mathcal{D}) \geq c_h(s,a)$ for all (h,s,a)

Initialize dataset $\mathcal{D} \leftarrow \emptyset$, episode number $d \leftarrow 0$ and dataset-counts $n_h(s,a;\mathcal{D}) \leftarrow 0$ for all (h,s,a)
for episode $e = (s_\ell^e, a_\ell^e, R_\ell^e)_{1\leq \ell \leq H}$ in $\widetilde{\mathcal{D}}$ **do**
  if $\exists \ell \in [H]$ such that $n_\ell(s_\ell^e, a_\ell^e; \mathcal{D}) < c_\ell(s_\ell^e, a_\ell^e)$ **then**
    Update dataset-counts $n_h(s_h^e, a_h^e; \mathcal{D}) \leftarrow n_h(s_h^e, a_h^e; \mathcal{D}) + 1$ for all $h \in [H]$
    Update dataset $\mathcal{D} \leftarrow \mathcal{D} \cup \{e\}$ and episode number $d \leftarrow d+1$
    if $n_h(s,a;\mathcal{D}) \geq c_h(s,a)$ for all (h,s,a) **then**
      return $(\mathcal{D}, d)$
    end if
  end if
end for

$\Omega(\widetilde{p}) = \{\widetilde{p}^\pi : \pi \in \Pi^S\}$. Finally, for a dataset of episodes $\mathcal{D}$, $n_h(s,a;\mathcal{D})$ denotes the number of visits of (h,s,a) in the episodes stored in $\mathcal{D}$.

F.2.1. GOOD EVENT

We introduce the following events

$$\mathcal{E}_{bpi} := \Bigg(\forall k \in \mathbb{N}^\star, \forall \pi \in \Pi^S,\ \big|\widehat{V}_1^{\pi,k} - V_1^\pi\big| \leq \sqrt{\beta^{bpi}(t_k,\delta/2) \min\Big(\sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}, \sum_{s,a,h} \frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)} \Big)}$$

$$\text{and } \Big| \sum_{s,a,h} \big(\widehat{p}_h^{\pi,k}(s,a) - p_h^\pi(s,a)\big)\widetilde{r}_h(s,a) \Big| \leq \sqrt{\beta^{bpi}(t_k,\delta/2) \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}} \text{ for all } \widetilde{r} \in [0,1]^{SAH} \Bigg),$$

$$\mathcal{E}_{cov} := \Bigg(\forall k \in \mathbb{N},\ \text{CovGame run with inputs } (c^k, \delta/4(k+1)^2) \text{ terminates after at most}$$

$$64 m_k \varphi^\star(c^k) + \widetilde{\mathcal{O}}\big(m_k \varphi^\star(\mathbb{1}) SAH^2 (\log(4(k+1)^2/\delta) + S)\big) \text{ episodes and returns a dataset } \widetilde{\mathcal{D}}_k$$

$$\text{such that for all } (h,s,a), n_h(s,a;\widetilde{\mathcal{D}}_k) \geq c_h^k(s,a) \Bigg),$$

where $m_k = \log_2\big(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a) \vee 1}\big) \vee 1$ and $\beta^{bpi}(t,\delta) = 16H^2 \log(2/\delta) + 96SAH^3 \log(1+t)$ is defined in Appendix D. Then our good event is defined as the intersection

$$\mathcal{E}_{good} := \mathcal{E}_{bpi} \cap \mathcal{E}_{cov}.$$

Lemma 40 *We have that* $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_{good}) \geq 1 - \delta$.

Proof Let $\overline{\mathcal{E}}$ denote the complementary event of $\mathcal{E}$. We start by the following decomposition

$$\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{good}}) \leq \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) + \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{bpi}} \cap \mathcal{E}_{cov}).$$

Now we bound each term separately. First observe that using Corollary 4 we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) &\leq \sum_{k=0}^{\infty} \mathbb{P}_{\mathcal{M}}(\text{CovGame with inputs } (c^k, \delta/4(k+1)^2) \text{ fails}) \\ &\leq \sum_{k=0}^{\infty} \frac{\delta}{4(k+1)^2} = \frac{\delta\pi^2}{24} \leq \delta/2.\end{aligned}$$

Next, note that by design of PRINCIPLE $n_h^0(s,a) = n_h(s,a;\widetilde{\mathcal{D}}_0)$ and $c^0 = \mathbb{1}$ so that $\mathcal{E}_{cov} \subset \big(\forall(h,s,a),\ n_h^0(s,a) \geq 1\big)$. Therefore we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{bpi}} \cap \mathcal{E}_{cov}) &\leq \mathbb{P}_{\mathcal{M}}\big(\overline{\mathcal{E}_{bpi}} \text{ and } \forall(h,s,a)\ n_h^0(s,a) \geq 1\big) \\ &\leq \delta/2,\end{aligned}$$

where we applied Theorem 28 and used the fact that $\beta^p(t,\delta) \leq \beta^{bpi}(t,\delta)$. Combining the two inequalities above yields the desired result. ■

F.2.2. LOW CONCENTRABILITY / GOOD COVERAGE OF OPTIMAL POLICIES

Lemma 41 *Under the good event, for all $k \geq 1$ such that PRINCIPLE did not stop before phase k, it holds that $n_h(s,a,\mathcal{D}_k) \geq c_h^k(s,a)$ for all (h,s,a) and $d_k \leq SAH2^k$.*

Proof Fix $k \geq 1$ such that PRINCIPLE did not stop before phase k. By definition of the good event we know that at the end of CovGame, $n_h(s,a;\widetilde{\mathcal{D}}_k) \geq c_h^k(s,a)$ for all (h,s,a). Now we distinguish two cases. **If $T_k \leq SAH2^k$:** then the result follows immediately since in this case, by design of PRINCIPLE (line 13 in Algorithm 3), $\mathcal{D}_k = \widetilde{\mathcal{D}}_k$ and $d_k = T_k$.
If $T_k > SAH2^k$: the first statement is a direct consequence of the stopping condition of PRUNEDATASET run with parameters $(\widetilde{\mathcal{D}}_k, c^k)$ (lines 7-8 in Algorithm 4). Now for the second statement, observe that each new episode e added by PRUNEDATASET to $\mathcal{D}_k$ increments the dataset-count of at least one triplet (h,s,a) that is not yet covered, i.e. $n_h(s,a;\mathcal{D}_k) < c_h^k(s,a)$. By the pigeon-hole principle it takes at most $\sum_{h,s,a} c_h^k(s,a)$ episodes to ensure that $n_h(s,a,\mathcal{D}_k) \geq c_h^k(s,a)$ for all (h,s,a). Therefore

$$d_k \leq \sum_{h,s,a} c_h^k(s,a) \leq SAH2^k,$$

where we used that $c_h^k(s,a) \leq 2^k$ due to the clipping. ■

The next lemma shows that the set of active state-action distributions always contains the distributions induced by optimal policies.

Lemma 42 *Under the good event, for all optimal policies $\pi^\star \in \Pi^\star$ and all phases $k \geq 0$, we have that*

$$\widehat{p}^{\pi^\star,k} \in \Omega^k \quad \text{and } n_h^k(s,a) \geq 2^k p_h^{\pi^\star}(s,a) \quad \forall(h,s,a).$$

Proof We fix an optimal policy $\pi^\star$ and prove the statement by induction. For $k = 0$, the fact that $\widehat{p}^{\pi^\star,0} \in \Omega^0$ is trivial since $\Omega^0 = \Omega(\widehat{p}^0)$ consists of all possible state-action distributions induced in the MDP whose transition kernel is $\widehat{p}^0$. Furthermore, under the good event we have that, for all (h,s,a), $n_h^0(s,a) \geq c_h^0(s,a) = 1 \geq 2^0 \max\big(p_h^{\pi^\star}(s,a), \widehat{p}_h^{\pi^\star,0}(s,a)\big)$. Now suppose that the property holds for phase k. Then we know that for any (h,s,a)

$$
\begin{aligned}
\big|\widehat{p}_h^{\pi^\star,k+1}(s,a) - \widehat{p}_h^{\pi^\star,k}(s,a)\big| &\leq \big|\widehat{p}_h^{\pi^\star,k+1}(s,a) - p_h^{\pi^\star}(s,a)\big| + \big|p_h^{\pi^\star}(s,a) - \widehat{p}_h^{\pi^\star,k}(s,a)\big| \\
&\overset{(a)}{\leq} \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} + \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k}(s,a)}} \\
&\overset{(b)}{\leq} 2\sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k}(s,a)}} \\
&\overset{(c)}{\leq} 2\sqrt{\beta^{bpi}(t_{k+1},\delta/2)H2^{-k}} \\
&= 2\sqrt{\beta^{bpi}(t_k + d_{k+1},\delta/2)H2^{-k}} \\
&\overset{(d)}{\leq} 2\sqrt{\beta^{bpi}(t_k + SAH2^{k+1},\delta/2)H2^{-k}},
\end{aligned} \tag{33}
$$

where (a) uses the event $\mathcal{E}_{bpi}$ for the reward $\widetilde{r}_\ell(s',a') = \mathbb{1}\big((\ell,s',a') = (h,s,a)\big)$, (b) uses the facts that $t \mapsto \beta(t,\delta)$ is non-decreasing and $n_h^{k+1}(s,a) \geq n_h^k(s,a)$, (c) uses the induction hypothesis which yields that $n_h^k(s,a) \geq 2^k p_h^{\pi^\star}(s,a)$ and (d) uses Lemma 41. Similarly we have that

$$
\big|p_h^{\pi^\star}(s,a) - \widehat{p}_h^{\pi^\star,k}(s,a)\big| \leq \sqrt{\beta^{bpi}(t_k + SAH2^{k+1},\delta/2)H2^{-k}} \tag{34}
$$

Now thanks to Lemma 41, we know that for all (h,s,a), $n_h^{k+1}(s,a) - n_h^k(s,a) = n_h(s,a,\mathcal{D}_{k+1}) \geq c_h^{k+1}(s,a)$. Plugging the definition of c^{k+1} (Line 8 of Algorithm 3) we get that,

$$
\begin{aligned}
n_h^{k+1}(s,a) &\geq 2^{k+1}\min\Big(\sup_{\widehat{\rho}\in\Omega^k}\widehat{\rho}_h(s,a) + 2\sqrt{H\beta^{bpi}(t_k + SAH2^{k+1},\delta/2)2^{-k}},\ 1\Big) \\
&\overset{(a)}{\geq} 2^{k+1}\min\Big(\widehat{p}_h^{\pi^\star,k}(s,a) + 2\sqrt{H\beta^{bpi}(t_k + SAH2^{k+1},\delta/2)2^{-k}},\ 1\Big) \\
&\overset{(b)}{\geq} 2^{k+1}\max\big(\widehat{p}_h^{\pi^\star,k+1}(s,a),\ p_h^{\pi^\star}(s,a)\big),
\end{aligned} \tag{35}
$$

where (a) uses that, by the induction hypothesis, $\widehat{p}^{\pi^\star,k} \in \Omega^k$ and (b) uses (33) along with (34). In particular we have proved that $\max_{h,s,a}\widehat{p}_h^{\pi^\star,k+1}(s,a)/n_h^{k+1}(s,a) \leq 2^{-(k+1)}$. Now it remains to show that $(\widehat{p}^{\pi^\star,k+1})^\top \widehat{r}^{k+1} \geq \underline{V}_1^{k+1}$. Let us consider $\widetilde{\rho}$ achieving the supremum in the definition of $\underline{V}_1^{k+1}$, i.e.

$$
\widetilde{\rho} \in \operatorname*{arg\,max}_{\substack{\widehat{\rho}\in\Omega(\widehat{p}^{k+1}),\\ \max_{h,s,a}\widehat{\rho}_h(s,a)/n_h^{k+1}(s,a)\leq 2^{-(k+1)}}} \widehat{\rho}^\top \widehat{r}^{k+1},
$$

and let $\widetilde{\pi}$ be a policy corresponding to $\widetilde{\rho}$[4]. Then we have that

$$\begin{aligned}
(\widehat{p}^{\pi^\star,k+1})^\top \widehat{r}^{k+1} &\overset{(a)}{\geq} V_1^\star - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} \\
&\geq V_1^{\widetilde{\pi}} - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} \\
&\overset{(b)}{\geq} \widetilde{\rho}^\top \widehat{r}^{k+1} - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{\widetilde{\rho}_h(s,a)^2}{n_h^{k+1}(s,a)}} - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} \\
&\overset{(c)}{\geq} \widetilde{\rho}^\top \widehat{r}^{k+1} - 2\sqrt{2^{-(k+1)}H\beta^{bpi}(t_{k+1},\delta/2)} \\
&= \underline{V}_1^{k+1}
\end{aligned} \tag{36}$$

where (a) uses the event $\mathcal{E}_{bpi}$ for policy $\pi^\star$, (b) uses the same event combined with the fact that $\widetilde{\rho} = \widehat{p}^{\widetilde{\pi},k+1}$, and (c) uses (35) and the fact that by definition of $\widetilde{\rho}$, $\max_{h,s,a}\widetilde{\rho}_h(s,a)/n_h^{k+1}(s,a) \leq 2^{-(k+1)}$.

Now combining (35) with (36) gives that $\widehat{p}^{\pi^\star,k+1} \in \Omega^{k+1}$. This finishes the proof. ■

F.2.3. CORRECTNESS

Lemma 43 *Under the good event, if PRINCIPLE stops then the recommended policy satisfies* $V_1^{\widehat{\pi}} \geq V_1^\star - \varepsilon$.

Proof Suppose that PRINCIPLE stops at phase $k \geq 1$. Let $\pi^\star$ be any optimal policy and recall the definition $\widehat{\rho}^\star = \arg\max_{\widehat{\rho}\in\Omega^k} \widehat{\rho}^\top \widehat{r}^k$ with ties broken arbitrarily. We have that

$$\begin{aligned}
V_1^{\widehat{\pi}} &\overset{(a)}{\geq} (\widehat{\rho}^\star)^\top \widehat{r}^k - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{\rho}_h^\star(s,a)^2}{n_h^k(s,a)}} \\
&\overset{(b)}{\geq} (\widehat{p}^{\pi^\star,k})^\top \widehat{r}^k - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{\rho}_h^\star(s,a)^2}{n_h^k(s,a)}} \\
&\overset{(c)}{\geq} V_1^\star - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi^\star,k}(s,a)^2}{n_h^k(s,a)}} - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{\rho}_h^\star(s,a)^2}{n_h^k(s,a)}} \\
&\overset{(d)}{\geq} V_1^\star - 2\sqrt{2^{-k}H\beta^{bpi}(t_k,\delta/2)} \overset{(e)}{\geq} V_1^\star - \varepsilon,
\end{aligned}$$

where (a) uses the event $\mathcal{E}_{bpi}$ for policy $\widehat{\pi}$ and the fact that $\widehat{\rho}^\star = \widehat{p}^{\widehat{\pi},k}$, (b) uses the definition of $\widehat{\rho}^\star$ and the fact that, by Lemma 42, $\widehat{p}^{\pi^\star,k} \in \Omega^k$, (c) uses the event $\mathcal{E}_{bpi}$ for the policy $\pi^\star$, and (d) uses that for all $\rho \in \Omega^k, \max_{h,s,a}\rho_h(s,a)/n_h^k(s,a) \leq 2^{-k}$ and (e) uses the stopping condition of PRINCIPLE (Line 20 of Algorithm 3). ■

4. i.e. $\widetilde{\pi}$ is the policy obtained by renormalization of $\widetilde{\rho}$.

F.2.4. UPPER BOUND ON THE NUMBER OF PHASES

Lemma 44 *Define the index of the final phase of PRINCIPLE,* $\kappa_f := \inf\big\{k \in \mathbb{N}_+ : \sqrt{2^{2-k} H \beta^{bpi}(t_k, \delta/2)} \leq \varepsilon\big\}$. *Further let* τ *denote the number of episodes played by the algorithm. Then under the good event, it holds that* $\kappa_f < \infty$ *and*

$$2^{\kappa_f} \leq \frac{8H\beta^{bpi}(\tau, \delta/2)}{\varepsilon^2}.$$

Proof To prove that κ_f is finite we write

$$\begin{aligned} t_k &= \sum_{j=0}^{k} d_j \\ &\leq d_0 + SAH \sum_{j=1}^{k} 2^j \\ &\leq \widetilde{\mathcal{O}}\Big(\varphi^\star(\mathbb{1}) SAH^2 \big(\log(4/\delta) + S\big)\Big) + SAH 2^{k+1}, \end{aligned} \tag{37}$$

where we have used the coverage event $\mathcal{E}_{cov}$ and Lemma 41 to upper bound d_0 and $(d_k)_{1 \leq j \leq k}$ respectively. This means that $t_k = \mathcal{O}_{k\to\infty}\big(2^k\big)$. Now recall that

$$\beta^{bpi}(t, \delta) := 16H^2 \log(1/\delta) + 96SAH^3 \log(1 + t). \tag{38}$$

Combining (37) and (38) gives that

$$\beta^{bpi}(t_k, \delta/2) = o_{k\to\infty}\big(2^k\big).$$

Therefore $\kappa_f = \inf\big\{k \in \mathbb{N}_+ : \sqrt{2^{2-k} H \beta^{bpi}(t_k, \delta/2)} \leq \varepsilon\big\}$ is indeed finite. The proof of the second statement is straightforward by noting that $\kappa_f - 1$ does not satisfy the stopping condition (Line 12 in Algorithm 3) and using the (crude) upper bound $t_{\kappa_f - 1} \leq \tau$. ■

Lemma 45 (UPPER BOUND ON PHASES WHERE A SUBOPTIMAL POLICY IS ACTIVE) *Let* π *be any suboptimal policy and* k *such that PRINCIPLE did not stop at phase* k *and* $\widehat{p}^{\pi,k} \in \Omega^k$. *Further let* τ *denote the number of episodes played by the algorithm. Then under the good event, we have the inequality*

$$2^k \leq \frac{16H\beta^{bpi}(\tau, \delta/2)}{\max(\varepsilon, \Delta(\pi))^2},$$

where $\Delta(\pi) := V_1^\star(s_1; r) - V_1^\pi(s_1; r)$ *denotes the policy gap of* π.

Proof Let $\pi^\star$ be any optimal policy. Then we have

$$\begin{aligned}
V_1^\star - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi^\star,k}(s,a)^2}{n_h^k(s,a)}} &\overset{(a)}{\leq} (\widehat{p}^{\pi^\star,k})^\top \widehat{r}^k \\
&\overset{(b)}{\leq} \sup_{\substack{\widehat{\rho}\in\Omega(\widehat{p}^k),\\ \max_{h,s,a}\widehat{\rho}_h(s,a)/n_h^k(s,a)\leq 2^{-k}}} \widehat{\rho}^\top \widehat{r}^k \\
&= \underline{V}_1^{\star,k} + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \\
&\overset{(c)}{\leq} (\widehat{p}^{\pi,k})^\top \widehat{r}^k + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \\
&\overset{(d)}{\leq} V_1^\pi + \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}} + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)},
\end{aligned}$$

where (a) uses the event $\mathcal{E}_{bpi}$ for $\pi^\star$, (b) uses the definition of Ω^k along with Lemma 42 which gives that $\widehat{p}^{\pi^\star,k} \in \Omega^k$, (c) uses our assumption that $\widehat{p}^{\pi,k} \in \Omega^k$ and (d) uses the event $\mathcal{E}_{bpi}$ for policy π. Rewriting the inequality above we get that

$$\begin{aligned}
\Delta(\pi) &= V_1^\star - V_1^\pi \\
&\leq \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi^\star,k}(s,a)^2}{n_h^k(s,a)}} + \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}} + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \\
&\leq 2\sqrt{2^{-k}H\beta^{bpi}(t_k,\delta/2)} + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} = 4\sqrt{2^{-k}H\beta^{bpi}(t_k,\delta/2)}, \qquad (39)
\end{aligned}$$

where the last inequality uses the fact that $\widehat{p}^{\pi^\star,k} \in \Omega^k$ by Lemma 42 and that $\widehat{p}^{\pi,k} \in \Omega^k$ by assumption. Therefore, using a crude bound $t_k \leq \tau$ we get that

$$2^k \leq \frac{16H\beta^{bpi}(\tau,\delta/2)}{\Delta(\pi)^2}.$$

Combining the result above with Lemma 44 and the fact that $k \leq \kappa_f$ yields the final result. ■

F.2.5. UPPER BOUND ON THE PHASE LENGTH

Lemma 46 *Let T_k denote the number of episodes played by PRINCIPLE during phase $k \geq 1$. Then we have*

$$\begin{aligned}
T_k \leq\ & 256H\beta^{bpi}(\tau,\delta/2)k\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right) \\
&+ 48k\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^k}\varphi^\star(\mathbb{1}) \\
&+ \widetilde{\mathcal{O}}\left(k\varphi^\star(\mathbb{1})SAH^2\big(\log(4(k+1)^2/\delta)+S\big)\right).
\end{aligned}$$

Proof Define $m_k = \log_2\left(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a)\vee 1}\right) \vee 1$. Under the good event we have

$$\begin{aligned} T_k &\le 64 m_k \varphi^\star(c^k) + \widetilde{\mathcal{O}}\Big(m_k \varphi^\star(\mathbb{1}) SAH^2\big(\log(4(k+1)^2/\delta) + S\big)\Big) \\ &\le 64 k \varphi^\star(c^k) + \widetilde{\mathcal{O}}\Big(k \varphi^\star(\mathbb{1}) SAH^2\big(\log(4(k+1)^2/\delta) + S\big)\Big), \end{aligned} \tag{40}$$

where the last inequality uses the fact that for all (h,s,a), $c_h^k(s,a) \le 2^k$. Now we simplify the expression of $\varphi^\star(c^k)$ as follows

$$\begin{aligned} \varphi^\star(c^k) &= \varphi^\star\left(\left[2^k \min\big(\sup_{\widehat{\rho}\in\Omega^{k-1}} \widehat{\rho}_h(s,a) + 2\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{1-k}},\ 1\big)\right]_{h,s,a}\right) \\ &\le \varphi^\star\left(\left[\sup_{\substack{\pi\in\Pi^S:\\ \widehat{p}^{\pi,k-1}\in\Omega^{k-1}}} 2^k \widehat{p}_h^{\pi,k-1}(s,a) + 2\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{k+1}}\right]_{h,s,a}\right), \end{aligned} \tag{41}$$

where we have used that $\varphi^\star(c) \le \varphi^\star(c')$ if $\forall (h,s,a)\ c_h(s,a) \le c'_h(s,a)$. Now fix a policy π in the set $\{\pi \in \Pi^S : \widehat{p}^{\pi,k-1} \in \Omega^{k-1}\}$. Using the event $\mathcal{E}_{bpi}$ for the rewards $\widetilde{r}_\ell(s',a') = \mathbb{1}\big((\ell,s',a') = (h,s,a)\big)$ we have that for all (h,s,a)

$$\begin{aligned} 2^k \widehat{p}_h^{\pi,k-1}(s,a) &\le 2^k p_h^\pi(s,a) + 2^k \sqrt{\beta^{bpi}(t_{k-1},\delta/2) \sum_{s',a',\ell} \frac{\widehat{p}_\ell^{\pi,k-1}(s',a')^2}{n_\ell^{k-1}(s',a')}} \\ &\overset{(a)}{\le} 2^k p_h^\pi(s,a) + 2^k \sqrt{\beta^{bpi}(t_{k-1},\delta/2) H 2^{1-k}} \\ &\le 2^k p_h^\pi(s,a) + \sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{k+1}} \\ &\overset{(b)}{\le} \frac{32 H \beta^{bpi}(\tau,\delta/2) p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2} + \sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{k+1}}, \end{aligned}$$

where (a) uses that $\max_{s',a',\ell} \frac{\widehat{p}_\ell^{\pi,k-1}(s',a')}{n_\ell^{k-1}(s',a')} \le 2^{1-k}$ since $\widehat{p}^{\pi,k-1} \in \Omega^{k-1}$ and (b) uses Lemma 45. Plugging the inequality above into (41) we get that

$$\begin{aligned} \varphi^\star(c^k) &\le \varphi^\star\left(\left[\sup_{\pi\in\Pi^S} \frac{32 H \beta^{bpi}(\tau,\delta/2) p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2} + 3\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{k+1}}\right]_{h,s,a}\right) \\ &\le 32 H \beta^{bpi}(\tau,\delta/2) \varphi^\star\left(\left[\sup_{\pi\in\Pi^S} \frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right) \\ &\quad + 3\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{k+1}}\varphi^\star(\mathbb{1}), \end{aligned} \tag{42}$$

where we used Lemma 10 in the last step. Combining (40) and (42) finishes the proof. ■

F.2.6. TOTAL SAMPLE COMPLEXITY

Theorem 47 *With probability at least $1-\delta$, the total sample complexity of PRINCIPLE satisfies*

$$\tau \leq \widetilde{\mathcal{O}}\left((H^3\log(1/\delta)+SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)+\frac{\varphi^\star(\mathbb{1})}{\varepsilon}+\varphi^\star(\mathbb{1})\right]\right),$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, H, \varepsilon, \log(1/\delta)$ and $\varphi^\star(\mathbb{1})$ and $\Delta(\pi) := V_1^\star(s_1;r) - V_1^\pi(s_1;r)$ denotes the policy gap of π.

Proof We write

$$\begin{aligned}
\tau &= \sum_{k=0}^{\kappa_f} T_k \\
&\leq \widetilde{\mathcal{O}}\Big(\varphi^\star(\mathbb{1})^2 SAH^2\big(\log(4/\delta)+S\big)\Big) + \sum_{k=1}^{\kappa_f} T_k \\
&\leq \underbrace{\sum_{k=1}^{\kappa_f} 256H\beta^{bpi}(\tau,\delta/2)k\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)}_{:=A} + \underbrace{\sum_{k=1}^{\kappa_f} 48k\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^k}\varphi^\star(\mathbb{1})}_{:=B} \\
&\quad + \underbrace{\widetilde{\mathcal{O}}\Big(\sum_{k=1}^{\kappa_f} k\varphi^\star(\mathbb{1})SAH^2\big(\log(4(k+1)^2/\delta)+S\big)\Big)}_{:=C},
\end{aligned}$$

where we have used Lemma 46. Now we bound each term separately. First note that

$$\begin{aligned}
A &\leq 256H\beta^{bpi}(\tau,\delta/2)\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\kappa_f^2 \\
&\overset{(a)}{\leq} 256H\beta^{bpi}(\tau,\delta/2)\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right) \\
&\overset{(b)}{\leq} \mathcal{O}\left([H^3\log(1/\delta)+SAH^4\log(1+\tau)]\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\right),
\end{aligned}$$

where (a) uses Lemma 44 and (b) uses the definition of β^{bpi}. Similarly

$$\begin{aligned}
B &\leq 48\sqrt{H\beta^{bpi}(\tau+SAH2^{\kappa_f-1},\delta/2)2^{\kappa_f}}\varphi^\star(\mathbb{1})\kappa_f^2 \\
&\overset{(a)}{\leq} 48\sqrt{\frac{4H^2\beta^{bpi}(\tau+SAH2^{\kappa_f-1},\delta/2)\beta^{bpi}(\tau,\delta/2)}{\varepsilon^2}}\varphi^\star(\mathbb{1})\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right) \\
&\leq \frac{48H}{\varepsilon}\beta^{bpi}(\tau+SAH2^{\kappa_f-1},\delta/2)\varphi^\star(\mathbb{1})\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right) \\
&\overset{(b)}{\leq} \mathcal{O}\left(\frac{\varphi^\star(\mathbb{1})}{\varepsilon}\left[H^3\log(1/\delta)+SAH^4\log\left(1+\tau+\frac{4SAH^2\beta^{bpi}(\tau,\delta/2)}{\varepsilon^2}\right)\right]\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\right),
\end{aligned}$$

where (a) and (b) use Lemma 44. Finally

$$C \le \widetilde{\mathcal{O}}\Big(\varphi^\star(\mathbb{1})SAH^2\big(\log(4(\kappa_f+1)^2/\delta)+S\big)\kappa_f^2\Big)$$
$$\le \widetilde{\mathcal{O}}\Big(\varphi^\star(\mathbb{1})SAH^2\big[\log\big(\frac{4\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)}{\delta}\big)+S\big]\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\Big),$$

where we have used Lemma 44 again. Combining the three inequalities with the definition of β^{bpi} we get that

$$\tau \le \mathcal{O}\Bigg((H^3\log(1/\delta)+SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)+\frac{\varphi^\star(\mathbb{1})}{\varepsilon}+\varphi^\star(\mathbb{1})\right]$$
$$\times\,\mathrm{polylog}(\tau,S,A,H,\varphi^\star(\mathbb{1}),\varepsilon,\log(1/\delta))\Bigg).$$

Solving for τ yields

$$\tau \le \widetilde{\mathcal{O}}\Bigg((H^3\log(1/\delta)+SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)+\frac{\varphi^\star(\mathbb{1})}{\varepsilon}+\varphi^\star(\mathbb{1})\right]\Bigg),$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, H, \varepsilon, \log(1/\delta)$ and $\varphi^\star(\mathbb{1})$. ■

Remark 48 (Reachability) *While for the PCE algorithm we were able to reduce the sample complexity by ignoring states that are hard to reach (which also allows using PCE when Assumption 1 is violated), we did not manage to propose a similar improvement for PRINCIPLE. This is because in reward-free exploration it is sufficient to guarantee that the* true confidence intervals *that depend on the visitation probabilities* under the true MDP *are small, i.e.,* $\sqrt{\beta^{\mathrm{RF}}(t_k,\delta)\sum_{(h,s,a)}\frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}} \le 2^k$. *This allows us to filter out all* (h,s,a) *for which* $\sup_\pi p_h^\pi(s,a) \le \mathcal{O}(\varepsilon/SH^2)$*, by arguing that their contribution to the true confidence interval is negligible. In contrast, the analysis of PRINCIPLE crucially relies on concentrating the values of policies by minimizing* their empirical confidence intervals*, i.e.,* $\sqrt{\beta^{bpi}(t_k,\delta)\sum_{(h,s,a)}\frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}} \le 2^k$. *We do not see a straightforward way to ignore the contribution of hard-to-reach states to these empirical confidence intervals.*

F.3. Comparison with other BPI-algorithms

In this section we compare PRINCIPLE with other algorithms for Best-Policy Identification algorithms that enjoy problem-dependent guarantees, namely PEDEL (Wagenmaker and Jamieson, 2022) and MOCA (Wagenmaker et al., 2022). Recalling that $\Delta(\pi) = V_1^\star(s_1) - V_1^\pi(s_1)$ denotes the policy gap of π, we first note that by Theorem 7, the leading term in the sample complexity of PRINCIPLE in the small (ε,δ) regime is $\mathrm{PRINCIPLE}(\mathcal{M},\varepsilon)\log(1/\delta)$ where

$$\mathrm{PRINCIPLE}(\mathcal{M},\varepsilon) := H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi^{\mathrm{s}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right).$$

We will now compare this term with the leading terms in the sample complexities of PEDEL and MOCA respectively, in the same asymptotic regime.

F.3.1. COMPARISON WITH PEDEL

Define the minimum policy gap

$$\Delta_{\min}(\Pi^{\mathrm{D}}) := \begin{cases} \min_{\pi \neq \pi^\star} \Delta(\pi), & \text{if the optimal policy } \pi^\star \text{ is unique} \\ 0, & \text{otherwise.} \end{cases}$$

Then instantiating Theorem 1 from Wagenmaker and Jamieson (2022) for our setting of tabular MDPs (i.e. with $d = SAH$ and $\Pi = \Pi^{\mathrm{D}}$), we see that the sample complexity achieved by PEDEL satisfies

$$\tau \leq \widetilde{O}\Big(\mathrm{PEDEL}(\mathcal{M}, \varepsilon)(\log(1/\delta) + SH) + \mathrm{poly}(SAH, \log(1/\varepsilon), \log(1/\delta))\Big)$$

$$\text{where} \quad \mathrm{PEDEL}(\mathcal{M}, \varepsilon) := H^4 \sum_{h=1}^{H} \min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi^{\mathrm{D}}))^2}.$$

Therefore the leading term PEDEL's complexity in the small (ε, δ) regime is $\mathrm{PEDEL}(\mathcal{M}, \varepsilon)\log(1/\delta)$. The next lemma shows that, up to H factors, this rate is always better than the complexity measure achieved by PRINCIPLE.

Lemma 49 *For any MDP $\mathcal{M}$, it holds that $\mathrm{PEDEL}(\mathcal{M}, \varepsilon) \leq H^2 \mathrm{PRINCIPLE}(\mathcal{M}, \varepsilon)$.*

Proof Fix any $h \in [H], \rho \in \Omega, \pi \in \Pi^{\mathrm{D}}$. Then we have

$$\sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)} \leq \left(\max_{s,a,h} \frac{p_h^\pi(s,a)}{\rho_h(s,a)}\right) \sum_{s,a} p_h^\pi(s,a) = \max_{s,a,h} \frac{p_h^\pi(s,a)}{\rho_h(s,a)}.$$

Therefore for all $h \in [H]$, using that $\Pi^{\mathrm{D}} \subset \Pi^{\mathrm{S}}$ we have

$$\begin{aligned}
\min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi^{\mathrm{D}}))^2} &\leq \min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}} \max_{s,a,h} \frac{p_h^\pi(s,a)/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi^{\mathrm{D}}))^2} \\
&= \min_{\rho \in \Omega} \max_{s,a,h} \max_{\pi \in \Pi^{\mathrm{D}}} \frac{p_h^\pi(s,a)/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi^{\mathrm{D}}))^2} \\
&\leq \min_{\rho \in \Omega} \max_{s,a,h} \sup_{\pi \in \Pi^{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\rho_h(s,a)\max(\varepsilon, \Delta(\pi))^2} \\
&= \varphi^\star\left(\left[\sup_{\pi \in \Pi^{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\max(\varepsilon, \Delta(\pi))^2}\right]_{h,s,a}\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{PEDEL}(\mathcal{M}, \varepsilon) &:= H^4 \sum_{h=1}^{H} \min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi^{\mathrm{D}}))^2} \\
&\leq H^5 \varphi^\star\left(\left[\sup_{\pi \in \Pi^{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\max(\varepsilon, \Delta(\pi))^2}\right]_{h,s,a}\right) \\
&= H^2 \mathrm{PRINCIPLE}(\mathcal{M}, \varepsilon).
\end{aligned}$$

■

F.3.2. COMPARISON WITH MOCA

Let us define the complexity functional,

$$\text{MOCA}(\mathcal{M}, \varepsilon) := H^2 \sum_{h=1}^{H} \min_{\rho \in \Omega} \max_{s,a} \frac{1}{\rho_h(s,a)} \min\Big(\frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\varepsilon^2}\Big) + \frac{H^4 \big|(h,s,a) : \widetilde{\Delta}_h(s,a) \le 3\varepsilon / W_h(s)\big|}{\varepsilon^2},$$

where $W_h(s) := \sup_\pi p_h^\pi(s)$ is the reachability of (h, s) and

$$\widetilde{\Delta}_h(s,a) := \begin{cases} \min_{b \neq a} V_h^\star(s) - Q_h^\star(s,b) & \text{if } a \text{ is the unique optimal action at } (h,s), \\ V_h^\star(s) - Q_h^\star(s,a) & \text{otherwise} \end{cases}$$

is the value gap of (h, s, a). Theorem 1 together with Proposition 2 of Wagenmaker et al. (2022) yield that the stopping time of MOCA satisfies

$$\tau \le \widetilde{\mathcal{O}}\bigg(\text{MOCA}(\mathcal{M}, \varepsilon) \log(1/\delta) + \frac{\text{poly}\big(SAH, \log(1/\varepsilon), \log(1/\delta)\big)}{\varepsilon}\bigg).$$

Therefore we see that $\text{MOCA}(\mathcal{M}, \varepsilon) \log(1/\delta)$ is the dominating term in the sample complexity of MOCA in the regime of small ε and small δ. On the other hand, as stated earlier, the leading term in PRINCIPLE's complexity in that regime is $\text{PRINCIPLE}(\mathcal{M}, \varepsilon) \log(1/\delta)$. Therefore we compare $\text{MOCA}(\mathcal{M}, \varepsilon)$ with $\text{PRINCIPLE}(\mathcal{M}, \varepsilon)$ to assess which algorithm is better in this regime.

Lemma 50 *Fix any $\Delta \in (0, 1]$. There exists an MDP $\mathcal{M}$ where*

$$\textit{MOCA}(\mathcal{M}, \varepsilon) = \Omega\bigg(\frac{H^5 SA}{\varepsilon^2}\bigg) \quad \textit{while PRINCIPLE}(\mathcal{M}, \varepsilon) = \mathcal{O}\bigg(\frac{H^4 SA}{\varepsilon\Delta} + \frac{H^4 \log(S) \log(A)}{\varepsilon^2}\bigg).$$

Proof Consider the MDP in figure F.3.2 which consists of an initial state s_1 and two sub-MDPs depending on the action taken at step $h = 1$. If the learner takes action a_1 it receives a reward $\Delta > 0$ and makes a transition to a sub-MDP $\mathcal{M}_1$ for which $|\mathcal{S}_1| = \log(S), |\mathcal{A}_1| = \log(A), H_1 = H - 1$ and where the rewards can be anything. On the other hand, if it takes action a_2 the learner will receive zero reward and make a transition to a sub-MDP $\mathcal{M}_2$ for which $|\mathcal{S}_2| = S - \log(S), |\mathcal{A}_2| = A, H_2 = H - 1$, the rewards are equal to zero everywhere and the transitions are deterministic, i.e. $p(s'|s,a) \in \{0, 1\}$ for all $(s, a) \in \mathcal{S}_2 \times \mathcal{A}_2$.

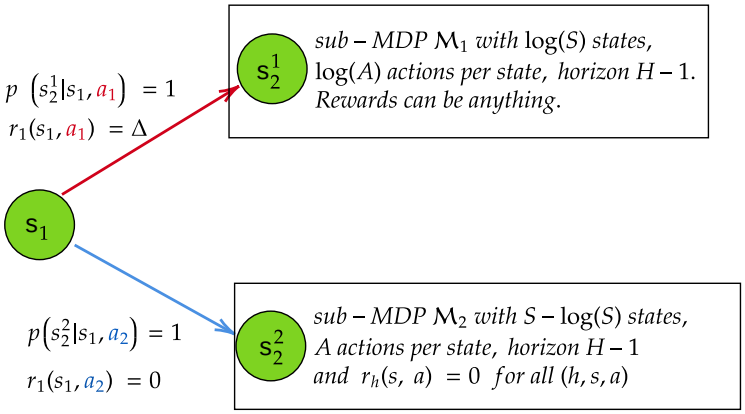

Figure 1: MDP instance with large policy gaps and small value gaps.

Note that in this example $\widetilde{\Delta}_h(s,a)=0$ for all $(h,s,a)\in\mathcal{M}_2$. Therefore

$$\begin{aligned}\text{MOCA}(\mathcal{M},\varepsilon) &\geq \frac{H^4\left|(h,s,a):\ \widetilde{\Delta}_h(s,a)\leq 3\varepsilon/W_h(s)\right|}{\varepsilon^2},\\ &\geq \frac{H^4(H-1)(S-\log(S))A}{\varepsilon^2}.\end{aligned} \tag{43}$$

On the other hand for all triplets (h,s,a) in the sub-MDP $\mathcal{M}_2$ we have

$$\sup_{\pi\in\Pi^{\text{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\leq\sup_{\pi\in\Pi^{\text{S}}}\frac{4\pi_1(a_2|s_1)}{(\varepsilon+\Delta(\pi))^2}, \tag{44}$$

where we used that $p_h^\pi(s,a)\leq\pi_1(a_2|s_1)$ (since the only path to reach (h,s,a) is by playing action a_2 at s_1) and that $\max(a,b)\geq(a+b)/2$. Now, by the performance-difference lemma we have

$$\begin{aligned}\Delta(\pi)&=\sum_{h,s,a}p_h^\pi(s,a)[V_h^\star(s)-Q_h^\star(s,a)]\\ &\geq p_1^\pi(s_1,a_2)[V_1^\star(s_1)-Q_1^\star(s_1,a_2)]=\pi_1(a_2|s_1)\Delta.\end{aligned}$$

Plugging this back into (44), we get

$$\begin{aligned}\sup_{\pi\in\Pi^{\text{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}&\leq\sup_{\pi\in\Pi^{\text{S}}}\frac{4\pi_1(a_2|s_1)}{(\varepsilon+\pi_1(a_2|s_1)\Delta)^2}\\ &=\sup_{x\in[0,1]}\frac{4x}{(\varepsilon+x\Delta)^2}=\frac{1}{\varepsilon\Delta}\end{aligned}$$

For triplets (h,s,a) outside of $\mathcal{M}_2$ (i.e. either at s_1 or in the sub-MDP $\mathcal{M}_1$) we use the crude bound

$$\sup_{\pi\in\Pi^{\text{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\leq\frac{\sup_{\pi\in\Pi^{\text{S}}}p_h^\pi(s,a)}{\varepsilon^2}.$$

Therefore

$$
\begin{aligned}
\text{PRINCIPLE}(\mathcal{M},\varepsilon) &= H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi^{\mathrm{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\\
&= H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi^{\mathrm{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}(\mathbb{1}\left((h,s,a)\in\mathcal{M}_2\right)+\mathbb{1}\left((h,s,a)\notin\mathcal{M}_2\right))\right]_{h,s,a}\right)\\
&\overset{(a)}{\leq} H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi^{\mathrm{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\mathbb{1}\left((h,s,a)\in\mathcal{M}_2\right)\right]_{h,s,a}\right)\\
&\quad + H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi^{\mathrm{S}}}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\mathbb{1}\left((h,s,a)\notin\mathcal{M}_2\right)\right]_{h,s,a}\right)\\
&\leq H^3\varphi^\star\left(\left[\frac{\mathbb{1}\left((h,s,a)\in\mathcal{M}_2\right)}{\varepsilon\Delta}\right]_{h,s,a}\right)+H^3\varphi^\star\left(\left[\frac{\mathbb{1}\left((h,s,a)\notin\mathcal{M}_2\right)\sup_{\pi\in\Pi^{\mathrm{S}}}p_h^\pi(s,a)}{\varepsilon^2}\right]_{h,s,a}\right)\\
&\overset{(b)}{\leq} H^3\sum_{(h,s,a)\in\mathcal{M}_2}\frac{1}{\varepsilon\Delta\sup_{\pi\in\Pi^{\mathrm{S}}}p_h^\pi(s,a)}+H^3\sum_{(h,s,a)\notin\mathcal{M}_2}\frac{1}{\varepsilon^2}\\
&\overset{(c)}{=}\frac{H^3(H-1)(S-\log(S))A}{\varepsilon\Delta}+\frac{H^3(H-1)\log(S)\log(A)}{\varepsilon^2}
\end{aligned}
\tag{45}
$$

where (a) uses the sub-linearity of the flow from Lemma 10, (b) uses the bound on $\varphi^\star$ from Lemma 12 and (c) uses that the sub-MDP $\mathcal{M}_2$ has deterministic transitions. Combining (43) and (45) finishes the proof. ■

Appendix G. Estimating State Reachability

Let $\mathcal{A}^\Pi$ be a regret minimizer that has a small regret for a (fixed) reward function r. If we set this reward function to $r_{h'}^{(h,s)}(s',a') = \mathbb{1}((s'=s,h'=h))$ for a target pair (h,s) intuitively the regret minimizer will visit as much as possible state s in step h and the total reward collected by the algorithm, $n_h^t(s)=\sum_{a\in\mathcal{A}}n_h^t(s,a)$, will be close to $t\times W_h(s)$, where the maximum visitation probability $W_h(s)=\max_\pi p_h^\pi(s)$ is actually the optimal value function in the MDP with reward function $r^{(h,s)}$. The empirical number of visitations can thus be used to estimate the unknown visitation probability.

This idea is already at the heart of the initialization phase of the MOCA algorithm, which relies on repeatedly running the Euler algorithm. We propose a slightly simpler version below, that doesn't need any restart and relies on a generic algorithm $\mathcal{A}^\Pi$ satisfying some first-order regret bound scaling with a quantity $\mathcal{R}_\delta^\Pi(T)$, as specified in the following theorem. ESTIMATEREACHABILITY $((h,s);\varepsilon_0,\delta)$ outputs a valid confidence interval $[\underline{W}_h(s),\overline{W}_h(s)]$ on the value of $W_h(s)$, which can be further used to eliminate all (h,s) whose maximum visitation probability is (slightly) smaller than a target ε_0.

Algorithm 5 ESTIMATEREACHABILITY $((h,s);\varepsilon_0,\delta)$

Input: Step h, state s, threshold $\varepsilon_0 > 0$, failure probability $\delta \in (0,1)$, regret minimizer $\mathcal{A}^{\Pi}$
Output: An interval $[\underline{W}_h(s), \overline{W}_h(s)]$
Compute $T = T(\varepsilon_0,\delta) = \inf\left\{T \in \mathbb{N} : 4\mathcal{R}^{\Pi}_{\delta/2}(T) + 6\log\left(\frac{4}{\delta}\right) \le \frac{\varepsilon_0}{4}T\right\}$
Collect T episodes $\{(s_1^t, a_1^t, \dots, s_H^t, a_H^t)\}_{t\le T}$ using $\mathcal{A}^{\Pi}$ with reward $\widetilde{r}_{h'}(s',a') = \mathbb{1}((s' = s, h' = h))$ and confidence $1-\delta/2$
Let $n_h^T(s) = \sum_{t=1}^T \mathbb{1}(s_h^t = s)$ be the number of visits of (h,s)
Define $\underline{W}_h(s) = \left(\frac{n_h^T(s)}{2T} - \frac{\varepsilon_0}{16}\right) \vee 0$ and $\overline{W}_h(s) = \left(\frac{2n_h^T(s)}{T} + \frac{\varepsilon_0}{4}\right) \wedge 1$

Theorem 51 *Assume that, for all (h,s), when $\mathcal{A}^{\Pi}$ is run for the reward function $r = r^{(h,s)}$ and confidence $1-\delta$ up to some horizon $T \in \mathbb{N}$, with probability larger than $1-\delta$,*

$$\sum_{t=1}^{T} V_1^{\star}(s_1;r) - \sum_{t=1}^{T} V_1^{\pi^t}(s_1;r) \le \sqrt{\mathcal{R}^{\Pi}_{\delta}(T) T V^{\star}(s_1;r)} + \mathcal{R}^{\Pi}_{\delta}(T). \tag{46}$$

For all (h,s), let $[\underline{W}_h(s), \overline{W}_h(s)]$ be the output of ESTIMATEREACHABILITY *$(h,s;\varepsilon_0,\delta/(SH))$ and define*

$$\widehat{\mathcal{X}} = \left\{(h,s) : \underline{W}_h(s) \ge \frac{\varepsilon_0}{8}\right\}.$$

With probability $1-\delta$, the following holds:

- *For all (h,s), $W_h(s) \in \left[\underline{W}_h(s), \overline{W}_h(s)\right]$*
- $\{(h,s) : W_h(s) \ge \varepsilon_0\} \subseteq \widehat{\mathcal{X}} \subseteq \left\{(h,s) : W_h(s) \ge \frac{\varepsilon_0}{8}\right\}$
- *For all $(h,s) \in \widehat{\mathcal{X}}$, $\overline{W}_h(s) \le 36 W_h(s)$.*

Moreover, the (deterministic) sample complexity necessary to construct $\widehat{\mathcal{X}}$ is

$$T_{\varepsilon_0}(\delta) := SH \times \inf\left\{T \in \mathbb{N}^{\star} : T \in \mathbb{N} : 4\mathcal{R}^{\Pi}_{\delta/(2SH)}(T) + 6\log\left(\frac{4}{\delta}\right) \le \frac{\varepsilon_0}{4}T\right\}.$$

In particular, using UCBVI as the regret minimizer, we have $T_{\varepsilon_0}(\delta) = \widetilde{\mathcal{O}}\left(\frac{S^2AH^2\left(\log\left(\frac{SAH}{\delta}\right)+S\right)}{\varepsilon_0}\right)$.

Proof Let $T = T(\varepsilon_0,\delta)$ be the (deterministic) number of episodes of ESTIMATEREACHABILITY $((h,s);\varepsilon_0,\delta)$, which satisfies

$$4\mathcal{R}^{\Pi}_{\delta/2}(T) + 6\log\left(\frac{4}{\delta}\right) \le \alpha\varepsilon_0 T \quad \text{for} \quad \alpha := \frac{1}{4}. \tag{47}$$

The analysis relies on the first-order bound on the regret of $\mathcal{A}^{\Pi}$ assumed in (46) and on a tight control of the martingale

$$M_T = \sum_{t=1}^{T} \left[\mathbb{1}(s_h^t = s) - p_h^{\pi_t}(s)\right]$$

where $p_h^\pi(s) = p_h^\pi(s, \pi(s))$ is the probability to reach s under policy π. Observing that the increment of this martingale is bounded in $[-1, 1]$ and that its variance is upper bounded by $W_h(s)$, we can use Bernstein's inequality to get that

$$\mathbb{P}\left(|M_T| \leq \sqrt{2TW_h(s)\log\left(\frac{4}{\delta}\right)} + \frac{2}{3}\log\left(\frac{4}{\delta}\right)\right) \geq 1 - \frac{\delta}{2}.$$

Remarking that the regret of $\mathcal{A}^\Pi$ for the reward function $r = r^{(h,s)}$ can be written

$$\sum_{t=1}^{T} V_1^\star(s_1; r) - \sum_{t=1}^{T} V_1^{\pi^t}(s_1; r) = TW_h(s) - \sum_{t=1}^{T} p_h^{\pi^t}(s) = TW_h(s) - n_h^T(s) + M_T$$

and that $n_h^T(s) \leq TW_h(s) + M_T$, we obtain that with probability larger than $1 - \delta$, the following two inequalities hold:

$$\begin{aligned} n_h^T(s) &\geq TW_h(s) - \left[\sqrt{\mathcal{R}_{\delta/2}(T)TW_h(s)} + \mathcal{R}_{\delta/2}(T) + \sqrt{2\log\left(\frac{4}{\delta}\right)TW_h(s)} + \frac{2}{3}\log\left(\frac{4}{\delta}\right)\right] \\ TW_h(s) &\geq n_h^T(s) - \left[\sqrt{2\log\left(\frac{4}{\delta}\right)TW_h(s)} + \frac{2}{3}\log\left(\frac{4}{\delta}\right)\right] \end{aligned}$$

Using the AM-GM inequality above, this first yields

$$n_h^T(s)/2 - g(\delta) \leq TW_h(s) \leq 2n_h^T(s) + f(T, \delta),$$

where $f(T, \delta) := 4\mathcal{R}_{\delta/2}(T) + \frac{16}{3}\log\left(\frac{4}{\delta}\right)$ and $g(\delta) := \frac{7}{6}\log\left(\frac{4}{\delta}\right)$. Observing that $g(\delta) \leq \frac{1}{4}f(T, \delta)$ and $f(T, \delta) \leq \alpha\varepsilon_0 T$ by inequality (47), we get

$$\frac{n_h^T(s)}{2T} - \frac{\alpha\varepsilon_0}{4} \leq W_h(s) \leq \frac{2n_h^T(s)}{T} + \alpha\varepsilon_0,$$

which also implies

$$\frac{W_h(s)}{2} - \frac{\alpha\varepsilon_0}{2} \leq \frac{n_h^T(s)}{T} \leq 2W_h(s) + \frac{\alpha\varepsilon_0}{2}.$$

As the output of ESTIMATEREACHABILITY $((h, s); \varepsilon_0, \delta)$ can be written

$$\left[\underline{W}_h(s) = \left(\frac{n_h^T(s)}{2T} - \frac{\alpha\varepsilon_0}{4}\right) \vee 0, \overline{W}_h(s) = \left(\frac{2n_h^T(s)}{T} + \alpha\varepsilon_0\right) \wedge 1\right]$$

and we get that with probability larger than $1 - \delta$:

1. For any value of $W_h(s)$,

$$\frac{W_h(s)}{4} - \frac{\alpha\varepsilon_0}{2} \leq \underline{W}_h(s) \leq W_h(s) \leq \overline{W}_h(s) \leq 4W_h(s) + 2\alpha\varepsilon_0.$$

2. If $W_h(s) \geq \varepsilon_0$, then $W_h(s) \in [\underline{W}_h(s), \overline{W}_h(s)] \in [\frac{1-2\alpha}{4}W_h(s), (4 + 2\alpha)\, W_h(s)]$.

3. If $W_h(s) < \varepsilon_0$, then $W_h(s) \in [\underline{W}_h(s), \overline{W}_h(s)] \in [0, (4+2\alpha)\,\varepsilon_0]$.

Now if $[\underline{W}_h(s), \overline{W}_h(s)]$ is the output of ESTIMATEREACHABILITY $((h,s); \varepsilon, \delta/SH)$ and

$$\widehat{\mathcal{X}} = \left\{(h,s) : \underline{W}_h(s) \geq \frac{1-2\alpha}{4}\varepsilon_0\right\}$$

we deduce that, with probability $1-\delta$:

- (h,s) with $W_h(s) \geq \varepsilon_0$ are all in $\widehat{\mathcal{X}}$.
- Since $\underline{W}_h(s) \leq W_h(s)$, any (h,s) with $W_h(s) < \frac{1-2\alpha}{4}\varepsilon_0$ does not belong to $\widehat{\mathcal{X}}$.

This proves that $\{(h,s) : W_h(s) \geq \varepsilon_0\} \subseteq \hat{\mathcal{X}} \subseteq \{(h,s) : W_h(s) \geq \frac{1-2\alpha}{4}\varepsilon_0\}$. To prove the last statement we remark that for $(h,s) \in \widehat{\mathcal{X}}$, if $W_h(s) \geq \varepsilon_0$, we have by 2. that $\overline{W}_h(s) \leq (4+2\alpha)\,W_h(s)$ while if $W_h(s) \in \left[\frac{1-2\alpha}{4}\varepsilon_0, \varepsilon_0\right)$ we have by 3. that

$$\overline{W}_h(s) \leq (4+2\alpha)\,\varepsilon_0 \leq 4\frac{4+2\alpha}{1-2\alpha}W_h(s)$$

Plugging the value $\alpha = 1/4$ yields $\overline{W}_h(s) \leq 36 W_h(s)$ in both cases.

To get an upper bound on the number of episodes used by an instance of ESTIMATEREACHABILITY, we need to find a T that satisfies

$$T - 1 \leq \frac{16}{\varepsilon_0}\mathcal{R}^{\Pi}_{\delta/(2SH)}(T) + \frac{24}{\varepsilon_0}\log\left(\frac{SAH}{\delta}\right). \tag{48}$$

For UCBVI, Theorem 19 yields $\mathcal{R}_\delta(T) = 256^2 SAH\left(\log\left(\frac{2SAH}{\delta}\right) + 6S\right)\log^2(T+1)$. Using the bound $\log^2(x) \leq 4\sqrt{x}$ we get a first crude upper bound on T by solving a quadratic equation which gives the final scaling by plugging back this crude bound in (48). ■