Optimistic PAC Reinforcement Learning: the Instance-Dependent View

Andrea Tirinzoni, Aymen Al-Marjani, Emilie Kaufmann

Optimistic PAC Reinforcement Learning: the Instance-Dependent View

Andrea Tirinzoni tirinzoni@fb.com
Meta AI
Paris, France

Aymen Al-Marjani aymen.al_marjani@ens-lyon.fr
UMPA, ENS Lyon
Lyon, France

Emilie Kaufmann emilie.kaufmann@univ-lille.fr
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL
Lille, France

Abstract

Optimistic algorithms have been extensively studied for regret minimization in episodic tabular MDPs, both from a minimax and an instance-dependent view. However, for the PAC RL problem, where the goal is to identify a near-optimal policy with high probability, little is known about their instance-dependent sample complexity. A negative result of Wagenmaker et al. (2022) suggests that optimistic sampling rules cannot be used to attain the (still elusive) *optimal* instance-dependent sample complexity. On the positive side, we provide the first instance-dependent bound for an optimistic algorithm for PAC RL, BPI-UCRL, for which only minimax guarantees were available (Kaufmann et al., 2021). While our bound features some minimal visitation probabilities, it also features a refined notion of sub-optimality gap compared to the value gaps that appear in prior work. Moreover, in MDPs with deterministic transitions, we show that BPI-UCRL is actually near-optimal. On the technical side, our analysis is very simple thanks to a new "target trick" of independent interest. We complement these findings with a novel hardness result explaining why the instance-dependent complexity of PAC RL cannot be easily related to that of regret minimization, unlike in the minimax regime.

Keywords: Optimism, exploration, PAC reinforcement learning

1. Introduction

We are interested in the probably approximately correct (PAC) identification of the best policy in an episodic Markov Decision Process (MDP) with finite state space $\mathcal{S}$, action space $\mathcal{A}$, and horizon H. We denote by $\mathcal{M} := (\mathcal{S}, \mathcal{A}, (p_h, \nu_h)_{h\in[H]}, s_1, H)$ such an MDP. Each episode starts in the initial state $s_1 \in \mathcal{S}$ and lasts H steps (called stages). In each stage $h \in [H]$, the agent is in some state $s_h \in \mathcal{S}$, it takes an action $a_h \in \mathcal{A}$, it receives a random reward drawn from a distributions $\nu_h(s,a)$ with expectation $r_h(s,a)$, and it transitions to a next state $s_{h+1} \in \mathcal{S}$ with probability $p_h(\cdot|s_h, a_h)$. A (deterministic) policy $\pi = (\pi_h)_{h\in[H]}$ is a sequence of mappings $\pi_h : \mathcal{S} \to \mathcal{A}$. The action-value function $Q_h^\pi(s,a)$ quantifies the expected cumulative reward when starting in state s at stage h, taking action a and following policy π until the end of the episode. It satisfies the Bellman equations: for all $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$,

$$Q_h^\pi(s,a) = r_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}^\pi(s'),$$

where $V_h^\pi(s) := Q_h^\pi(s, \pi_h(s))$ is the corresponding value function (with $V_{H+1}^\pi = 0$). A policy $\pi^\star$ is optimal if $V_1^{\pi^\star}(s_1) = \max_\pi V_1^\pi(s_1)$. From the theory of MDPs (Puterman, 1994), a sufficient condition is that $\pi_h^\star(s) \in \arg\max_{a\in\mathcal{A}} Q_h^\star(s,a)$, where the optimal Q-function satisfies $Q_h^\star(s,a) = r_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}^\star(s')$, with $V_h^\star(s) = \max_{a\in\mathcal{A}} Q_h^\star(s,a)$ and $V_{H+1}^\star(s) = 0$. This condition implies that $\pi^\star$ maximizes the expected return at any state and stage simultaneously, while the (weaker) optimality condition only requires so at the initial state s_1.

In online episodic reinforcement learning (RL), the agent interacts with the MDP $\mathcal{M}$ by choosing, in each episode $t \in \mathbb{N}$, a policy π^t and collecting a trajectory in the MDP under this policy: $(s_h^t, a_h^t, r_h^t)_{h\in[H]}$ where $s_1^t = s_1$ and, for all $h \in [H]$, $a_h^t = \pi_h^t(s_h^t)$, $r_h^t \sim \nu_h(s_h^t, a_h^t)$, and $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$. The choice of π^t based on previously observed

trajectories is called the *sampling rule*. Several objectives have been studied in the literature. An agent seeking to maximize the total reward received in T episodes equivalently aims at minimizing the (pseudo) regret

$$\mathcal{R}_{\mathcal{M}}(T) := \sum_{t=1}^{T} \left(V_1^\star(s_1) - V_1^{\pi^t}(s_1) \right).$$

In PAC identification (or PAC RL), the agent's sampling rule is coupled with a (possibly adaptive) stopping rule τ after which the agent stops collecting trajectories and returns a guess for the optimal policy $\widehat{\pi}$. Given two parameters $\varepsilon, \delta > 0$ with $\delta \in (0,1)$, the algorithm $((\pi^t)_{t\in\mathbb{N}}, \tau, \widehat{\pi})$ is (ε,δ)-PAC if it returns an ε-optimal policy with high probability, i.e.,

$$\mathbb{P}_{\mathcal{M}} \left(V_1^{\widehat{\pi}}(s_1) \geq V_1^\star(s_1) - \varepsilon \right) \geq 1 - \delta.$$

The goal is to have (ε,δ)-PAC algorithms using a small number of exploration episodes τ (a.k.a. sample complexity).

The PAC RL framework was originally introduced by Fiechter (1994) and there exists algorithms attaining a sample complexity $O((SAH^3/\varepsilon^2)\log(1/\delta))$ (Dann and Brunskill, 2015; Ménard et al., 2021), which is optimal in a minimax sense in time-inhomogeneous MDPs (Domingues et al., 2021). These algorithms use an *optimistic* sampling rule coupled with a well-chosen stopping rule. Optimistic sampling rules, in which the policy π^t is the greedy policy with respect to an upper confidence bound on the optimal Q function, have been mostly proposed for regret minimization (see Neu and Pike-Burke (2020) for a survey). In particular, the UCBVI algorithm of Azar et al. (2017a) (with Bernstein bonuses) attains minimax optimal regret in episodic MDPs. Recent works have provided instance-dependent upper bounds on the regret for optimistic algorithms (Simchowitz and Jamieson, 2019; Xu et al., 2021; Dann et al., 2021). An instance-dependent bound features some complexity term which depends on the MDP instance, typically through some notion of sub-optimality gap. To the best of our knowledge, for PAC RL in episodic MDPs the only algorithms with instance-dependent upper bound on their sample complexity are MOCA (Wagenmaker et al., 2022) and EPRL (Tirinzoni et al., 2022), the latter being analyzed for MDPs with deterministic transitions. Neither of these algorithms are based on an optimistic sampling rule.

Notably, Wagenmaker et al. (2022) proved that no-regret sampling rules (including optimistic ones) cannot achieve the instance-optimal rate for PAC identification. The intuition is quite simple: an optimal algorithm for PAC RL must visit every state-action pair at least a certain amount of times, and this requires playing policies that cover the whole MDP in the minimum amount of episodes. On the other hand, a regret-minimizer focuses on playing high-reward policies which, depending on the MDP instance, might be arbitrarily bad at visiting hard-to-reach states.

Despite not being instance-optimal, optimistic sampling rules are simple (e.g., as opposed to the complex design of MOCA), computationally efficient, and do not require any sophisticated elimination rule (e.g., as opposed to the one proposed by Tirinzoni et al. (2022) to obtain the optimal gap dependence in deterministic MDPs). However, it remains an open question what instance-dependent complexity they can achieve.

Contributions Our main contribution is a new instance-dependent analysis for (a variant of) BPI-UCRL, a PAC RL algorithm based on an optimistic sampling rule proposed by Kaufmann et al. (2021) with only a worst-case sample complexity bound. In particular, in Theorem 2 we show that the sample complexity of BPI-UCRL can be bounded by

$$\tau \lesssim \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \frac{H^4 \log(1/\delta)}{p_h^{\min}(s,a) \max\{\widetilde{\Delta}_h(s,a), \varepsilon\}^2},$$

where $p_h^{\min}(s,a)$ is the minimum positive probability to reach (s,a) at stage h across all deterministic policies, while $\widetilde{\Delta}_h(s,a) := \min_{\pi: p_h^\pi(s,a)>0} \max_{\ell\in[H]} \max_{s': p_\ell^\pi(s')>0} (V_\ell^\star(s') - V_\ell^\pi(s'))$ is a new notion of sub-optimality gap that we call the *conditional return gap*.[1] Interestingly, we show that the gaps $\widetilde{\Delta}_h(s,a)$ are larger than both the value gaps of Wagenmaker et al. (2022) and of the (deterministic) return gaps of Tirinzoni et al. (2022). Notably, we prove this result with a remarkably simple analysis based on a new "target trick": instead of bounding the number of times each state-action-stage triplet (s,a,h) is visited (as it is common in the bandit literature), we control the number of times the played policy visits (s,a,h) with positive probability with (s,a,h) being the least visited triplet so far, an event that we refer to as (s,a,h) being "targeted".

1. We denote by $p_h^\pi(s,a)$ (resp. $p_h^\pi(s)$) the probability that π visits (s,a) (resp. s) at stage h.

Our second contribution is to prove that, unlike what happens in the minimax setting, there is no clear relationship between regret and sample complexity in the instance-dependent framework. Indeed, the "regret-to-PAC conversion" often proposed to turn a regret minimizer into an (ε, δ)-PAC algorithm for PAC RL (e.g., Jin et al., 2018; Ménard et al., 2021; Wagenmaker et al., 2022) cannot directly exploit an instance-dependent upper bound on the regret. In Theorem 4, we construct an MDP for which the sample complexity suggested by a regret-to-PAC conversion cannot be attained by any (ε, δ)-correct algorithm for PAC RL. In particular, this implies that one cannot take an instance-dependent regret bound for an optimistic algorithm (e.g., Simchowitz and Jamieson, 2019) and turn it into an instance-dependent sample complexity bound of the form above: a specific analysis for PAC RL, like the one proposed in this paper, is actually required.

2. The BPI-UCRL Algorithm

Let $n_h^t(s,a) := \sum_{j=1}^t \mathbb{1}\left(s_h^j = s, a_h^j = a\right)$ be the number of times the state-action pair (s,a) has been visited at stage h up to episode t. We introduce the maximum-likelihood estimators

$$\widehat{r}_h^t(s,a) := \frac{1}{n_h^t(s,a)} \sum_{j=1}^t \mathbb{1}\left(s_h^j = s, a_h^j = a\right) r_h^j \quad \text{and} \quad \widehat{p}_h^t(s'|s,a) := \frac{1}{n_h^t(s,a)} \sum_{j=1}^t \mathbb{1}\left(s_h^j = s, a_h^j = a, s_{h+1}^j = s'\right)$$

for $r_h(s,a)$ and $p_h(s'|s,a)$, respectively. As common, and without loss of generality, we shall assume that reward distributions are supported on $[0,1]$. We define inductively the following upper and lower bounds on the optimal value function. Letting $\overline{Q}_{H+1}^t = \underline{Q}_{H+1}^t = 0$, for all $h \in [H]$ we have

$$\overline{Q}_h^t(s,a) = \min\left(H-h+1,\ \widehat{r}_h^t(s,a) + b_h^t(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a)\overline{V}_{h+1}^t(s')\right), \qquad \overline{V}_h^t(s) = \max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a),$$

$$\underline{Q}_h^t(s,a) = \max\left(0,\ \widehat{r}_h^t(s,a) - b_h^t(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a)\underline{V}_{h+1}^t(s')\right), \qquad \underline{V}_h^t(s) = \max_{a\in\mathcal{A}} \underline{Q}_h^t(s,a),$$

where $b_h^t(s,a)$ is a confidence bonus defined as

$$b_h^t(s,a) := (H-h+1)\left(\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1\right)$$

for a suitable threshold β that we shall specify in the analysis. The BPI-UCRL algorithm (Kaufmann et al., 2021) can be described as follows:[2]

- the **sampling rule** prescribes $\pi_h^{t+1}(s) = \arg\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a)$ for each $t \in \mathbb{N}$;
- the **stopping rule** is $\tau = \inf\left\{t \in \mathbb{N} : \max_a \overline{Q}_1^t(s_1,a) - \max_a \underline{Q}_1^t(s_1,a) \le \varepsilon\right\}$;
- the **recommendation rule** is $\widehat{\pi}_h^\tau(s) = \arg\max_{a\in\mathcal{A}} \underline{Q}_h^\tau(s,a)$.

More precisely, in episode t, BPI-UCRL uses the policy π^t to generate a new trajectory $(s_h^t, a_h^t, r_h^t)_{h\in[H]}$. This trajectory is used to update the estimates of the dynamics to $\widehat{p}_h^t(s'|s,a)$ and $\widehat{r}_h^t(s,a)$ and the bounds on the optimal Q-values to $\underline{Q}_h^t$ and $\overline{Q}_h^t$. At the end of the t-th episode, BPI-UCRL checks for stopping and proceeds to the next episode if and only if $\max_a \overline{Q}_1^t(s_1,a) - \max_a \underline{Q}_1^t(s_1,a) > \varepsilon$. Upon stopping, it outputs as the guess for the optimal policy the greedy policy with respect to the lower confidence bounds on the optimal value, $\underline{Q}_h^\tau(s,a)$, where τ denotes the (random) number of episodes used before stopping.

2. The original BPI-UCRL algorithm uses slightly different Q-function bounds which do not feature $\widehat{r}_h^t(s,a)$ and $\widehat{p}_h^t(s'|s,a)$ explicitly but rather scale with KL confidence regions around them (see Appendix D of Kaufmann et al. (2021)). Here we write the explicit version obtained by appyling Pinsker's inequality, though our analysis also holds for the original confidence intervals.

Note that the sampling rule of BPI-UCRL is essentially the UCBVI algorithm with Hoeffding's bonuses proposed by Azar et al. (2017b) for regret minimization. Such bonuses can be improved using Bernstein's inequality, yielding either UCBVI with Bernstein's bonuses (Azar et al., 2017b) or EULER (Zanette and Brunskill, 2019). While this would likely reduce the dependence on the horizon from H^4 to H^3 in our final sample complexity bound, we focus on Hoeffding's bonuses for simplicity since the extension to Bernstein's bonuses is somewhat straightforward given existing analyses.

3. An Instance-dependent Analysis of BPI-UCRL

Before stating and proving our main result, we introduce our novel notion of sub-optimality gap. Formally, the *conditional return gap* of any state-action pair (s,a) at stage $h \in [H]$ is

$$\widetilde{\Delta}_h(s,a) := \min_{\pi \in \Pi : p_h^\pi(s,a) > 0} \max_{\ell \in [H]} \max_{s' \in \mathcal{S} : p_\ell^\pi(s') > 0} \left(V_\ell^\star(s') - V_\ell^\pi(s') \right), \tag{1}$$

where we recall that $p_h^\pi(s) := \mathbb{P}^\pi(s_h = s)$ and $p_h^\pi(s,a) = p_h^\pi(s)\mathbb{1}\left(\pi_h(s) = a\right)$. The intuition behind this definition is quite simple: in order to figure out whether (s,a) is sub-optimal at stage h, the agent must learn that all policies visiting (s,a) at stage h with positive probability are indeed sub-optimal. The complexity for detecting whether any of such policies (say, π) is sub-optimal is proportional to the maximum gap between the optimal value function and the one of π across all possible states visited by π itself. This is a gap between expected returns conditioned on different starting states and stages (hence the name conditional return gap). It turns out that these gaps are larger than both the value gaps $\Delta_h(s,a) := V_h^\star(s) - Q_h^\star(s,a)$ (Wagenmaker et al., 2022) and the variant of the return gaps $\overline{\Delta}_h(s,a) = V_1^\star(s_1) - \max_{\pi \in \Pi : p_h^\pi(s,a) > 0} V_1^\pi(s_1)$ introduced by Tirinzoni et al. (2022).[3]

Proposition 1 *For all $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, $\widetilde{\Delta}_h(s,a) \geq \Delta_h(s,a)$ and $\widetilde{\Delta}_h(s,a) \geq \overline{\Delta}_h(s,a)$. Moreover, if the MDP has deterministic transitions, $\widetilde{\Delta}_h(s,a) = \overline{\Delta}_h(s,a)$.*

Proof For the first inequality, we have

$$\widetilde{\Delta}_h(s,a) \geq V_h^\star(s) - \max_{\pi : p_h^\pi(s,a) > 0} V_h^\pi(s) = V_h^\star(s) - \max_{\pi : p_h^\pi(s,a) > 0} Q_h^\pi(s,a) = V_h^\star(s) - Q_h^\star(s,a) = \Delta_h(s,a).$$

The second one is trivial by lower bounding the maximum with $s' = s_1$ and $\ell = 1$. To see the equality, note that $V_h^\star(s) - V_h^\pi(s) = \mathbb{E}^\pi\left[\sum_{\ell=h}^H \Delta_\ell(s_\ell, \pi_\ell(s_\ell)) \mid s_h = s\right]$. In the deterministic case, this implies that $V_h^\star(s) - V_h^\pi(s)$ is a sum of $H - h + 1$ fixed (non-negative) value gaps. Therefore, the maximum in (1) must be attained at the initial stage and state, which implies the statement. ■

The first return gaps were actually introduced in the regret-minimization literature by Dann et al. (2021) as

$$\overline{\mathrm{gap}}_h(s,a) = \Delta_h(s,a) \vee \frac{1}{H} \min_{\pi \in \Pi : \mathbb{P}(\mathcal{B}_h^\pi(s,a)) > 0} \mathbb{E}^\pi\left[\sum_{\ell=1}^h \Delta_\ell(s_\ell, a_\ell) \mid \mathcal{B}_h(s,a)\right],$$

where $\mathcal{B}_h^\pi(s,a) = \{s_h = s, a_h = a, \exists \ell \leq h : \Delta_\ell(s_\ell, a_\ell) > 0\}$ is the event that policy π visits (s,a) at stage h after at least one sub-optimal action was played. We found no clear relationship between $\overline{\mathrm{gap}}_h(s,a)$ and $\widetilde{\Delta}_h(s,a)$ besides the fact that the former is also comparing returns (from stage 1), as for any policy playing optimally from stage $h+1$, $V^\star(s_1) - V^\pi(s_1) = \mathbb{E}^\pi\left[\sum_{\ell=1}^h \Delta_\ell(s_\ell, a_\ell)\right]$. We now state and prove our main result.

3. The return gaps were introduced by Tirinzoni et al. (2022) only for deterministic MDPs. Here we replace their maximum over policies visiting (s,a,h) with probability 1 by the one over policies visiting it with positive probability.

Theorem 2 *Let* $\beta(t,\delta) := (\sqrt{\beta^r(t,\delta)} + \sqrt{2\beta^p(t,\delta)})^2$, *where* $\beta^r(t,\delta) := \frac{1}{2}(\log(3SAH/\delta) + \log(e(1+t)))$ *and* $\beta^p(t,\delta) := \log(3SAH/\delta) + (S-1)\log(e(1+t/(S-1)))$. *With probability at least* $1-\delta$, *BPI-UCRL outputs a policy* $\widehat{\pi}^\tau$ *satisfying* $V_1^{\widehat{\pi}^\tau}(s_1) \geq V_1^\star(s_1) - \varepsilon$ *using a number of episodes upper bounded as*

$$\tau \leq H^4 \sum_{h=1}^{H} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \frac{720 \log \frac{3SAH}{\delta} + 1729 S \log\left(1152 \frac{S^2AH^5}{p_{\min}\varepsilon^2} \log \frac{3SAH}{\delta}\right)}{p_h^{\min}(s,a) \max\{\widetilde{\Delta}_h(s,a), \varepsilon\}^2},$$

where $p_h^{\min}(s,a) := \min_{\pi\in\Pi : p_h^\pi(s,a)>0} p_h^\pi(s,a)$ *and* $p_h^{\min}(s,a) = +\infty$ *when* (s,a,h) *is unreachable by any policy.*

Theorem 2 shows that the sample complexity of BPI-UCRL is upper bounded by a function that scales inversely with the conditional return gaps squared multipled by the minimum visitation probabilities of each triplet (s,a,h). We recall that BPI-UCRL also enjoys the worst-case sample complexity bound $\tau \leq \widetilde{O}(SAH^4 \log(1/\delta)/\varepsilon^2)$ proved by Kaufmann et al. (2021), which is minimax optimal up to a factor H. Thus, one can always take the minimum between this worst-case bound and the instance-dependent one of Theorem 2. Before proving our main theorem, we briefly discuss how it relates to existing results.

Comparison to Wagenmaker et al. (2022) The sample complexity upper bound achieved by the MOCA algorithm of Wagenmaker et al. (2022) is roughly

$$\tau \leq \widetilde{O}\left(H^2 \log(1/\delta) \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \min\left(\frac{1}{p_h^{\max}(s,a)\Delta_h(s,a)^2}, \frac{p_h^{\max}(s,a)}{\varepsilon^2}\right) + \frac{H^4 |\mathrm{OPT}(\varepsilon)| \log(1/\delta)}{\varepsilon^2}\right),$$

where $p_h^{\max}(s,a) := \max_{\pi\in\Pi : p_h^\pi(s,a)>0} p_h^\pi(s,a)$ and $\mathrm{OPT}(\varepsilon)$ is roughly the set of all ε-optimal triplets. In contrast to the bound we obtained for BPI-UCRL, this one scales with the maximum probabilities for reaching the different state-action pairs. This is obtained thanks to the clever exploration strategy of MOCA which focuses on efficiently covering the whole MDP. However, the bound of Wagenmaker et al. (2022) scales with value gaps which, from Proposition 1, are provably smaller than our conditional return gaps. Overall, the two bounds result non-comparable as there exist MDP instances where the one of BPI-UCRL is smaller, and viceversa for the one of MOCA. While we are able to show this improved gap dependence thanks to optimism alone, we are not sure how to achieve it with a suitable elimination rule that could be plugged into the MOCA exploration strategy to obtain the best of these two bounds.

The dependence on $p_h^{\min}(s,a)$ One might be wondering whether a better dependence than $p_h^{\min}(s,a)$ can be achieved with an optimistic rule like BPI-UCRL. We conjecture that this is not possible, at least in a worst-case sense. In fact, Wagenmaker et al. (2022) already proved that there exists an MDP instance in which any no-regret sampling rule (thus including optimistic ones) suffers a depence on the minimum visitation probabilities, while a "smart" PAC RL algorithm does not. The intuition is that a no-regret algorithm focuses on playing high-reward policies which, depending on the MDP instance, might be arbitrarily bad at exploring the state space. In our context, this means that, if the policy visiting (s,a,h) with largest reward is also the one that visits it with lowest probability, an optimistic sampling rule is likely to play such policy quite frequently and thus its sample complexity will scale inversely by $p_h^{\min}(s,a)$ as we show.

Deterministic MDPs (comparison to Tirinzoni et al. (2022)) If the MDP has deterministic transitions, we have $\widetilde{\Delta}_h(s,a) = \overline{\Delta}_h(s,a)$ (see Proposition 1) and $p_h^{\min}(s,a) = 1$ if state s is reachable by some policy at stage h, while $p_h^{\min}(s,a) = +\infty$ in the opposite case. Theorem 2 then implies that

$$\tau \leq \widetilde{O}\left(H^4 \sum_{h\in[H]} \sum_{s\in\mathcal{S}_h} \sum_{a\in\mathcal{A}} \frac{\log(1/\delta) + S\log\log(1/\delta)}{\max\{\overline{\Delta}_h(s,a), \varepsilon\}^2}\right),$$

where $\mathcal{S}_h$ is the subset of states reachable at stage h. Up to the extra multiplicative H^2 and $S\log\log(1/\delta)$ terms, this matches the bound obtained by Tirinzoni et al. (2022) for the EPRL algorithm with a maximum-diameter sampling rule that is informed a-priori about the MDP being deterministic. These extra terms arise because BPI-UCRL needs to concentrate the transition probabilities to work for general stochastic MDPs. If we knew that the MDP is deterministic, we could modify the bonuses as $b_h^t(s,a) := \sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \wedge 1$ and the thresholds as $\beta(t,\delta) := \beta^r(t,\delta)$. This would

yield sample complexity $\tau \leq \widetilde{O}\left(H^2 \sum_{h\in[H]} \sum_{s\in\mathcal{S}_h} \sum_{a\in\mathcal{A}} \log(1/\delta)/\max\{\overline{\Delta}_h(s,a),\varepsilon\}^2\right)$ which matches exactly the one of EPRL with maximum-diameter sampling and which is at most a factor of H^3 sub-optimal w.r.t. the instance-dependent lower bound of Tirinzoni et al. (2022). This is quite remarkable since EPRL obtains the "optimal" dependence on the deterministic return gaps $\overline{\Delta}_h(s,a)$ using a clever elimination rule, while here we show optimism alone is enough. We note, however, that reducing the sub-optimal dependence on H^3 still requires smarter exploration strategies than optimism, like the maximum-coverage one proposed by Tirinzoni et al. (2022).

3.1 Proof of Theorem 2

All lemmas and proofs not explicitly reported here can be found in Appendix A.

We carry out the proof under the "good event" $\mathcal{E} := \mathcal{E}^r \cap \mathcal{E}^p \cap \mathcal{E}^c$, where

$$\mathcal{E}^r := \left\{\forall t \in \mathbb{N}_{>0}, s\in\mathcal{S}, a\in\mathcal{A}, h\in[H] : \left|r_h(s,a) - \widehat{r}_h^t(s,a)\right| \leq \sqrt{\frac{\beta^r(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}}\right\},$$

$$\mathcal{E}^p := \left\{\forall t \in \mathbb{N}_{>0}, s\in\mathcal{S}, a\in\mathcal{A}, h\in[H] : \mathrm{KL}\left(\widehat{p}_h^t(\cdot|s,a), p_h(\cdot|s,a)\right) \leq \frac{\beta^p(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}\right\},$$

$$\mathcal{E}^c := \left\{\forall t \in \mathbb{N}_{>0}, s\in\mathcal{S}, a\in\mathcal{A}, h\in[H] : n_h^t(s,a) \geq \frac{1}{2}\overline{n}_h^t(s,a) - \log(3SAH/\delta)\right\}.$$

Note that event $\mathcal{E}^c$ relates the counts $n_h^t(s,a)$ to the pseudo-counts $\overline{n}_h^t(s,a) := \sum_{j=1}^t p_h^{\pi^j}(s,a)$. Thanks to Lemma 5, we have $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ and, thus, the final result will hold with the same probability.

This good event is identical to the one used in the original (minimax) analysis of BPI-UCRL (Kaufmann et al., 2021). On this good event, one can prove that our (slighlty different) bounds $\overline{Q}_h^t(s,a), \underline{Q}_h^t(s,a)$ are respectively upper and lower bounds on the optimal action value $Q_h^\star(s,a)$, for all (s,a,h) (see Lemma 6, which justifies the choice of threshold β). The correctness follows from this fact using the same arguments as Theorem 11 of Kaufmann et al. (2021). The original part of our proof is the way we upper bound the sample complexity on the good event $\mathcal{E}$.

Our proof is based on the following "target trick" which extends the one used by Tirinzoni et al. (2022) to MDPs with stochastic transitions. Fix any $s\in\mathcal{S}$, $a\in\mathcal{A}$, and $h\in[H]$. Let us introduce the event "(s,a,h) is targeted at time t" as

$$G_{s,a,h}^t := \left\{p_h^{\pi^t}(s,a) > 0, (s,a,h) \in \operatorname*{arg\,max}_{(s',a',\ell):p_\ell^{\pi^t}(s',a')>0} b_\ell^{t-1}(s',a')\right\}.$$

Intutively, (s,a,h) is targeted at time t if (1) it is visited with positive probability by π^t and (2) it maximizes the bonuses at time $t-1$ (i.e., the current uncertainty) across all triplets visited by π^t. Let $Z_h^\tau(s,a) := \sum_{t=1}^\tau \mathbb{1}\left(G_{s,a,h}^t\right)$ be the number of times (s,a,h) is targeted up the stopping time. Since at each time step at least one triplet is targeted,

$$\tau \leq \sum_{h=1}^H \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Z_h^{\tau-1}(s,a) + 1. \tag{2}$$

We shall now focus on bounding $Z_h^T(s,a)$ for some time $T>0$ at the end of which the algorithm did not stop. Thanks to (2), this will imply a bound on the final stopping time.

We first state the following crucial result which relates confidence intervals and conditional return gaps.

Lemma 3 *Under event $\mathcal{E}$, for any $t\in\mathbb{N}_{>0}, s\in\mathcal{S}, h\in[H]$,*

$$V_h^\star(s) - V_h^{\pi^{t+1}}(s) \leq 2\sum_{\ell=h}^H \sum_{s'\in\mathcal{S}} p_\ell^{\pi^{t+1}}(s'|s,h) b_\ell^t(s', \pi_\ell^{t+1}(s')),$$

where $p_\ell^\pi(s'|s,h) := \mathbb{P}^\pi(s_\ell = s'|s_h = s)$.

Let $(\tilde{s}_t, \tilde{h}_t) \in \arg\max_{(s',\ell):p_\ell^{\pi^t}(s')>0}(V_\ell^\star(s') - V_\ell^{\pi^t}(s'))$. Using Lemma 3 with this couple,

$$\max_{\ell\in[H]} \max_{s'\in\mathcal{S}:p_\ell^{\pi^t}(s')>0} \left(V_\ell^\star(s') - V_\ell^{\pi^t}(s')\right) \le 2\sum_{\ell=\tilde{h}_t}^{H}\sum_{s'\in\mathcal{S}} p_\ell^{\pi^t}(s'|\tilde{s}_t,\tilde{h}_t) b_\ell^{t-1}(s',\pi_\ell^t(s')).$$

Summing both sides for all episodes where (s,a,h) is targeted up to T and using that $p_h^{\pi^t}(s,a) > 0$ under $G_{s,a,h}^t$,

$$2\sum_{t=1}^{T} \mathbb{1}\left(G_{s,a,h}^t\right) \sum_{\ell=\tilde{h}_t}^{H}\sum_{s'\in\mathcal{S}} p_\ell^{\pi^t}(s'|\tilde{s}_t,\tilde{h}_t) b_\ell^{t-1}(s',\pi_\ell^t(s')) \ge Z_h^T(s,a)\widetilde{\Delta}_h(s,a). \tag{3}$$

Note that, for each time t, since $p_{\tilde{h}_t}^{\pi^t}(\tilde{s}_t) > 0$, then $p_\ell^{\pi^t}(s'|\tilde{s}_t,\tilde{h}_t) > 0$ implies that $p_\ell^{\pi^t}(s') > 0$. Using that, under $G_{s,a,h}^t$, (s,a,h) maximizes the bonuses at time $t-1$ over all triplets visited by π^t, we can upper bound the left-hand side as

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}\left(G_{s,a,h}^t\right) \sum_{\ell=\tilde{h}_t}^{H}\sum_{s'\in\mathcal{S}} p_\ell^{\pi^t}(s'|\tilde{s}_t,\tilde{h}_t) b_\ell^{t-1}(s',\pi_\ell^t(s')) &\le H\sum_{t=1}^{T} \mathbb{1}\left(G_{s,a,h}^t\right) b_h^{t-1}(s,a) \\
&\overset{(a)}{\le} 2H^2 \sum_{t=1}^{T} \mathbb{1}\left(G_{s,a,h}^t\right) \sqrt{\frac{\beta(\overline{n}_h^{t-1}(s,a),\delta)}{\overline{n}_h^{t-1}(s,a)\vee 1}} \\
&\overset{(b)}{\le} 2H^2 \sum_{t=1}^{T} \mathbb{1}\left(G_{s,a,h}^t\right) \sqrt{\frac{\beta(T,\delta)}{Z_h^{t-1}(s,a)p_h^{\min}(s,a)\vee 1}} \\
&\overset{(c)}{\le} 4H^2 \sqrt{\frac{\beta(T,\delta)Z_h^T(s,a)}{p_h^{\min}(s,a)}}.
\end{aligned}$$

where (a) uses Lemma 7 of Kaufmann et al. (2021) together with the definition of $b_h^{t-1}(s,a)$, (b) uses that $\overline{n}_h^{t-1}(s,a) \ge \sum_{j=1}^{t-1}\mathbb{1}\left(G_{s,a,h}^j\right) p_h^{\pi^j}(s,a) \ge Z_h^{t-1}(s,a)p_h^{\min}(s,a)$, and (c) uses the pigeon-hole principle (see Lemma 8). Plugging this into (3) and solving the resulting inequality in $Z_h^T(s,a)$, we obtain

$$Z_h^T(s,a) \le \frac{64H^4\beta(T,\delta)}{p_h^{\min}(s,a)\widetilde{\Delta}_h(s,a)^2}.$$

A similar derivation using the stopping rule definition together with the fact that the algorithm did not stop at T also allows us to prove that $Z_h^T(s,a) \le \frac{144H^4\beta(T,\delta)}{p_h^{\min}(s,a)\varepsilon^2}$ (see Lemma 9). Plugging these two bounds into (2) with $T = \tau - 1$,

$$\tau \le \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \frac{144H^4\beta(\tau-1,\delta)}{p_h^{\min}(s,a)\max\{\widetilde{\Delta}_h(s,a),\varepsilon\}^2} + 1. \tag{4}$$

The proof is concluded by noting that $\beta(\tau-1,\delta) \le 5\log\frac{3SAH}{\delta} + 4S + 4S\log(\tau)$ (see Lemma 10) and by using Lemma 11 to solve the resulting inequality in τ (see Appendix A.4). ■

4. On the Regret-to-PAC Conversion

In the minimax setting, the complexity of PAC RL and that of regret minimization are very related. Indeed, Jin et al. (2018) suggest the following regret-to-PAC conversion: one can take a regret minimizer, run it for T episodes, and output a policy $\widehat{\pi}$ uniformly drawn from the T played. Then, by Markov's inequality, $\mathbb{P}\left(V_1^{\widehat{\pi}}(s_1) < V_1^\star(s_1) - \varepsilon\right) \le \frac{1}{T\varepsilon}\sum_{t=1}^{T}\mathbb{E}[V_1^\star(s_1) - V_1^{\pi^t}(s_1)] = \frac{1}{\varepsilon}\mathbb{E}[\mathcal{R}_\mathcal{M}(T)/T]$. Thus, choosing T such that the expected average regret is smaller than $\varepsilon\delta$ yields an ε-optimal policy with probability $1-\delta$. This is why in the literature it is common to derive an upper bound $\overline{R}(T)$ on the expected average regret and then claim that the resulting sample complexity for PAC RL is $T_\varepsilon := \inf_{T\in\mathbb{N}}\left\{T : \overline{R}(T) \le \varepsilon\delta\right\}$. However, this claim can be misleading.

Applying this regret-to-PAC conversion to the UCBVI algorithm with Bernstein bonuses (Azar et al., 2017a), we get a sample complexity of order $O(SAH^3 \log(1/\delta)/(\varepsilon^2\delta^2))$, which is optimal in a minimax sense in all dependencies except δ.[4] However, this trick can only be perfomed when $\overline{R}(T)$ contains quantities known by the algorithm (e.g., it can be a worst-case bound but not an instance-dependent one). In fact, the regret minimizer is used as a *sampling rule* for PAC identification coupled with a *deterministic stopping rule* which simply stops after T_ε episodes. When T_ε is unknown, we need to use an *adaptive* stopping rule, in which case the claimed sample complexity T_ε might not be attainable. This is proved in the following theorem, where we show that there exist MDPs where T_ε can be exponentially (in S, A) smaller than the actual stopping time of any (ε, δ)-PAC algorithm.

Theorem 4 *For any $S \geq 4$, $A \geq 2$ and $H \geq \lceil \log_2(S) \rceil + 1$, there exists an MDP $\mathcal{M}$ with S states, A actions, and horizon H, and a regret minimization algorithm such that*

$$T_\varepsilon := \inf_{T \in \mathbb{N}} \left\{ T : \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{M}} \left[V_1^\star(s_1) - V_1^{\pi^t}(s_1) \right] \leq \varepsilon\delta \right\} \leq \frac{2}{\varepsilon^2\delta} \left(36 \log(2SAH) + 16 \log \frac{17}{\varepsilon^2\delta} + 9\varepsilon^2 \right) + 1.$$

Moreover, on the same instance any (ε, δ)-PAC identification algorithm must satisfy

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq \frac{SA \log(1/4\delta)}{16\varepsilon^2}.$$

Our proof (see Appendix B) essentially builds an MDP instance with many optimal actions. The intuition is that, in such MDP, it is relatively easy for a regret minimizer to start behaving near optimally (i.e., to have average regret below $\varepsilon\delta$). However, when this occurs the regret minimizer has still not enough confidence to produce an ε-optimal policy with probability at least $1 - \delta$. That is, a stopping rule for identification would not trigger, hence the separation between the two times.

The main implication is that the time T_ε at which the average regret goes below $\varepsilon\delta$ is not always a good proxy for the sample complexity that a regret minimizer would take for (ε, δ)-PAC identification. In particular, one cannot simply take an existing instance-dependent regret bound (e.g., Simchowitz and Jamieson, 2019; Dann et al., 2021; Xu et al., 2021) and turn it into a sample complexity bound by the regret-to-PAC conversion suggested above. A specific analysis for the PAC setting, like the one we propose in Section 3 or those of Wagenmaker et al. (2022); Tirinzoni et al. (2022), is actually needed.

Finally, we note that this result also solves an open question left by Wagenmaker et al. (2022) in their conclusion. First, it shows that the sample complexity stated in Equation (7.1) of Wagenmaker et al. (2022) for a regret-to-PAC conversion from an instance-dependent regret bound cannot always be attained by a PAC RL algorithm. Second, it shows that the extra term $|\mathrm{OPT}(\varepsilon)|/\varepsilon^2$ that appears in the complexity of MOCA is actually tight, at least in a worst-case sense, as our proof essentially builds an MDP where all ε-optimal state-action pairs must be visited $\Omega(1/\varepsilon^2)$ times.

5. Discussion

We derived the first instance-dependent sample complexity bound for an optimistic sampling rule (BPI-UCRL). It features a new notion of sub-optimality gap that we call "conditional return gap" and that is tighter than existing value gaps and (deterministic) return gaps. We proved this bound with a remarkably simple analysis based on a new "target trick" that could be of independent interest. We complemented this result by showing that one cannot directly leverage the standard regret-to-PAC conversion in the instance-dependent regime, thus making our novel analysis non-trivial.

In the bandit setting, it is known that optimism, when coupled with an appropriate stopping and recommendation rule, is near instance-optimal for best-arm identification with (sub)Gaussian distributions (Jamieson et al., 2014). In this work, we obtained a similar result for deterministic MDPs, where optimistic sampling rules are sub-optimal only by a factor H^3. This also explains the good empirical performance of BPI-UCRL observed by Tirinzoni et al. (2022) in such a setting. However, there seems to be a large gap for general stochastic MDPs, where our sample complexity scales with some minimal visitation probabilities that are avoided by algorithms like MOCA. This can be related to known results

4. The dependence on δ can be improved to $\log(1/\delta)^2$, see Appendix F of Kaufmann et al. (2021).

for structured bandits (Lattimore and Szepesvari, 2017), as a stochastic MDP presents a complex trade-off between collecting rewards and gathering information (i.e., exploring the state space) for which an optimistic algorithm can be arbitrarily sub-optimal.

Finding the right complexity (matching upper and lower bounds) for PAC RL in general stochastic MDPs remains the main open problem. In deterministic MDPs, upper and lower bounds are nearly matching and are expressed as (complex) functions of the (simple) deterministic return gaps (Tirinzoni et al., 2022). They were obtained by properly combining a coverage-based exploration strategy with a suitable elimination rule. We conjecture that a similar algorithmic design could be a good direction towards instance optimality in stochastic MDPs. This would involve the combination of (1) a coverage-based exploration strategy like MOCA (Wagenmaker et al., 2022) that ensures scaling with the "right" visitation probabilities, and (2) some elimination rule to avoid over-sampling that ensures scaling with the "right" notion of gap. Unfortunately, while there exist instance-dependent lower bounds for regret minimization (Tirinzoni et al., 2021; Dann et al., 2021), an analogous result for PAC RL is still unknown and thus it remains unclear what these "right" notions are. In this work, we take a step forward by proposing a novel and tighter gap definition, though it remains an open question whether our conditional return gaps can be related to an actual sample complexity lower bound.

References

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*, 2017a.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. volume 70 of *Proceedings of Machine Learning Research*, pages 263–272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017b. PMLR.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 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*, pages 5713–5723, 2017.

Christoph Dann, Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *CoRR*, abs/2107.01264, 2021. URL https://arxiv.org/abs/2107.01264.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory (ALT)*, 2021.

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

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*, 2014.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

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

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

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.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *NeurIPS*, 2020.

M.L. Puterman. *Markov Decision Processes. Discrete Stochastic. Dynamic Programming.* Wiley, 1994.

Max Simchowitz and Kevin G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *NeurIPS*, pages 1151–1160, 2019.

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. *arXiv:2203.09251*, 2022.

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

Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 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)*, 2019.

Appendix A. Proofs of Section 3

A.1 Additional notation

We define the following upper and lower confidence bounds over the value functions of each policy π. We initialize $\underline{V}^{t,\pi}_{H+1}(s) = \overline{V}^{t,\pi}_{H+1}(s) = 0$, then we define recursively

$$\overline{Q}^{t,\pi}_h(s,a) = \min\left(H-h+1,\ \widehat{r}^t_h(s,a) + b^t_h(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}^t_h(s'|s,a)\overline{V}^{t,\pi}_{h+1}(s')\right), \qquad \overline{V}^{t,\pi}_h(s) = \overline{Q}^{t,\pi}_h(s,\pi_h(s)),$$

$$\underline{Q}^{t,\pi}_h(s,a) = \max\left(0,\ \widehat{r}^t_h(s,a) - b^t_h(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}^t_h(s'|s,a)\underline{V}^{t,\pi}_{h+1}(s')\right), \qquad \underline{V}^{t,\pi}_h(s) = \underline{Q}^{t,\pi}_h(s,\pi_h(s)).$$

A.2 Proof of Lemma 3

Using event $\mathcal{E}$ and the fact that π^{t+1} is greedy w.r.t. $\overline{Q}^t_h(s,a)$,

$$\begin{aligned} V^\star_h(s) - V^{\pi^{t+1}}_h(s) &= \max_{a\in\mathcal{A}} Q^\star_h(s,a) - Q^{\pi^{t+1}}_h(s,\pi^{t+1}_h(s)) \le \max_{a\in\mathcal{A}} \overline{Q}^t_h(s,a) - Q^{\pi^{t+1}}_h(s,\pi^{t+1}_h(s)) \\ &= \overline{Q}^{t,\pi^{t+1}}_h(s,\pi^{t+1}_h(s)) - Q^{\pi^{t+1}}_h(s,\pi^{t+1}_h(s)). \end{aligned}$$

Let $a = \pi^{t+1}_h(s)$. Expanding the last quantity using the Bellman equations,

$$\begin{aligned} \overline{Q}^{t,\pi^{t+1}}_h(s,a) - Q^{\pi^{t+1}}_h(s,a) &\le \widehat{r}^t_h(s,a) - r_h(s,a) + \sum_{s'\in\mathcal{S}} (\widehat{p}^t_h(s'|s,a) - p_h(s'|s,a))\overline{V}^{t,\pi^{t+1}}_{h+1}(s') \\ &\quad + \sum_{s'\in\mathcal{S}} p_h(s'|s,a)\left(\overline{V}^{t,\pi^{t+1}}_{h+1}(s') - V^{\pi^{t+1}}_{h+1}(s')\right) + b^t_h(s,a) \\ &\le \sqrt{\frac{\beta^r(n^t_h(s,a),\delta)}{n^t_h(s,a)\vee 1}} \wedge 1 + (H-h)\sqrt{\frac{2\beta^p(n^t_h(s,a),\delta)}{n^t_h(s,a)\vee 1}} \wedge (H-h) \\ &\quad + \sum_{s'\in\mathcal{S}} p_h(s'|s,a)\left(\overline{V}^{t,\pi^{t+1}}_{h+1}(s') - V^{\pi^{t+1}}_{h+1}(s')\right) + b^t_h(s,a) \\ &\le 2b^t_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a)\left(\overline{V}^{t,\pi^{t+1}}_{h+1}(s') - V^{\pi^{t+1}}_{h+1}(s')\right), \end{aligned}$$

where in the second inequality we used event $\mathcal{E}$ as in the proof of Lemma 6. The statement follows by recursively applying this reasoning to $\overline{V}^{t,\pi^{t+1}}_{h+1}(s') - V^{\pi^{t+1}}_{h+1}(s') = \overline{Q}^{t,\pi^{t+1}}_{h+1}(s',\pi^{t+1}_{h+1}(s')) - Q^{\pi^{t+1}}_{h+1}(s',\pi^{t+1}_{h+1}(s'))$. ■

A.3 Other results

Lemma 5 *Using the threshold β defined in Theorem 2, $\mathbb{P}(\mathcal{E}) \ge 1-\delta$.*

Proof $\mathcal{E}^r$ and $\mathcal{E}^p$ hold with probability at least $1-\delta/3$ each by applying Proposition 1 and 2 of Jonsson et al. (2020) together with a union bound and Pinsker's inequality for the rewards. $\mathcal{E}^c$ holds with probability at least $1-3\delta$ by Lemma F.4 of Dann et al. (2017) and a union bound. Another union bound over the three events proves the statement. ■

Lemma 6 *Using the threshold β defined in Theorem 2, under event $\mathcal{E}$, for any $t \in \mathbb{N}_{>0}$, $s\in\mathcal{S}$, $a\in\mathcal{A}$, $h\in[H]$,*

$$\underline{Q}^{t,\pi}_h(s,a) \le Q^\pi_h(s,a) \le \overline{Q}^{t,\pi}_h(s,a),$$
$$\underline{Q}^t_h(s,a) \le Q^\star_h(s,a) \le \overline{Q}^t_h(s,a).$$

Proof Clearly, all inequalities hold at stage H since $Q_H^\pi(s,a) = Q_H^\star(s,a) = r_H(s,a)$ and, by event $\mathcal{E}^r$, together with the fact that rewards are bounded in $[0,1]$,

$$\left|r_H(s,a) - \widehat{r}_H^t(s,a)\right| \leq \sqrt{\frac{\beta^r(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \wedge 1 \leq b_H^t(s,a).$$

Now suppose the inequalities hold at stage $h+1 \leq H$. At stage h, we have

$$\begin{aligned}
Q_h^\pi(s,a) &= r_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) V_{h+1}^\pi(s') \\
&\overset{(a)}{\leq} r_h(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) \overline{V}_{h+1}^{t,\pi}(s') \\
&\overset{(b)}{\leq} r_h(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a) \overline{V}_{h+1}^{t,\pi}(s') + (H-h)\|p_h(s,a) - \widehat{p}_h^t(s,a)\|_1 \\
&\overset{(c)}{\leq} r_h(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a) \overline{V}_{h+1}^{t,\pi}(s') + (H-h)\sqrt{2\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a))} \\
&\overset{(d)}{\leq} \widehat{r}_h^t(s,a) + \sqrt{\frac{\beta^r(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a) \overline{V}_{h+1}^{t,\pi}(s') + (H-h)\sqrt{\frac{2\beta^p(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}},
\end{aligned}$$

where (a) is by assumption, (b) uses that $\overline{V}_{h+1}^{t,\pi}(s')$ is bounded by $H-h$, (c) is from Pinsker's inequality, and (d) uses the event $\mathcal{E}$. As before, since rewards are bounded in $[0,1]$ and $\sum_{s'\in\mathcal{S}}(p_h(s'|s,a) - \widehat{p}_h^t(s'|s,a))\overline{V}_{h+1}^{t,\pi}(s') \leq H-h$, we can clip the two bonuses above to 1 and $H-h$, respectively. This implies that,

$$\begin{aligned}
&\sqrt{\frac{\beta^r(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \wedge 1 + (H-h)\sqrt{\frac{2\beta^p(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \wedge (H-h) \\
&\qquad\leq \sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \wedge 1 + (H-h)\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \wedge (H-h) \\
&\qquad\leq (H-h+1)\sqrt{\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)\vee 1}} \wedge (H-h+1) \leq b_h^t(s,a).
\end{aligned}$$

This proves that $Q_h^\pi(s,a) \leq \overline{Q}_h^{t,\pi}(s,a)$. The proofs of all other inequalities follow analogously. ■

Lemma 7 *Under event $\mathcal{E}$, for all (s,a) and $h \leq H$,*

$$\max_a \overline{Q}_1^t(s_1,a) - \max_a \underline{Q}_1^t(s_1,a) \leq 3\sum_{h=1}^H \sum_{s,a} p_h^{\pi^{t+1}}(s,a) b_h^t(s,a),$$

Proof By definition of the optimistic rule, we first observe that

$$\max_a \overline{Q}_1^t(s_1,a) - \max_a \underline{Q}_1^t(s_1,a) = \overline{Q}_1^t(s_1,\pi_h^{t+1}(s_1)) - \max_a \underline{Q}_1^t(s_1,a) \leq D_1^t(s_1,\pi_1^{t+1}(s_1))$$

where we introduce the diameters

$$D_h^t(s,a) := \overline{Q}_h^t(s,a) - \underline{Q}_h^t(s,a).$$

Using the inductive definition of the confidence bounds, we get

$$\begin{aligned}
D_h(s,a) &\le 2b_h^t(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a)\left(\max_a \overline{Q}_{h+1}^t(s,a) - \max_a \underline{Q}_{h+1}^t(s,a)\right) \\
&\le 2b_h^t(s,a) + \sum_{s'\in\mathcal{S}} \widehat{p}_h^t(s'|s,a) D_{h+1}^t(s',\pi_{h+1}^{t+1}(s')) \\
&= 2b_h^t(s,a) + \sum_{s'\in\mathcal{S}} (\widehat{p}_h^t(s'|s,a) - p_h(s'|s,a)) D_{h+1}^t(s',\pi_{h+1}^{t+1}(s')) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a)) D_{h+1}^t(s',\pi_{h+1}^{t+1}(s')) \\
&\le 3b_h^t(s,a) + \sum_{s'\in\mathcal{S}} p_h(s'|s,a) D_{h+1}^t(s',\pi_{h+1}^{t+1}(s')),
\end{aligned}$$

and the result follows by induction. ■

Lemma 8 *For any $T > 0$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, $h \in [H]$,*

$$\sum_{t=1}^T \mathbb{1}\left(G_{s,a,h}^t\right) \sqrt{\frac{1}{Z_h^{t-1}(s,a) p_h^{\min}(s,a) \vee 1}} \le 2\sqrt{\frac{Z_h^T(s,a)}{p_h^{\min}(s,a)}}.$$

Proof Using the pigeon-hole principle together with the inequality $\sum_{i=1}^n 1/\sqrt{i} \le 2\sqrt{n} - 1$,

$$\begin{aligned}
\sum_{t=1}^T \mathbb{1}\left(G_{s,a,h}^t\right) \sqrt{\frac{1}{Z_h^{t-1}(s,a) p_h^{\min}(s,a) \vee 1}} &\le 1 + \frac{1}{\sqrt{p_h^{\min}(s,a)}} \sum_{j=2}^{Z_h^T(s,a)} \sqrt{\frac{1}{j-1}} \\
&\le 1 + \frac{2\sqrt{Z_h^T(s,a)} - 1}{\sqrt{p_h^{\min}(s,a)}} \le 2\sqrt{\frac{Z_h^T(s,a)}{p_h^{\min}(s,a)}}.
\end{aligned}$$

■

Lemma 9 *For any time $T > 0$ at the end of which the algorithm did not stop, for any $s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$,*

$$Z_h^T(s,a) \le \frac{144 H^4 \beta(T,\delta)}{p_h^{\min}(s,a)\varepsilon^2}.$$

Proof If the algorithm did not stop at the end of time T, by the definition of the stopping rule and Lemma 7, for all $t \le T$,

$$\varepsilon \le \max_a \overline{Q}_1^t(s_1,a) - \max_a \underline{Q}_1^t(s_1,a) \le 3\sum_{h=1}^H \sum_{s,a} p_h^{\pi^{t+1}}(s,a) b_h^t(s,a).$$

Summing both sides over times where (s,a,h) is targeted,

$$\begin{aligned}
\varepsilon Z_h^T(s,a) = \varepsilon \sum_{t=1}^T \mathbb{1}\left(G_{s,a,h}^t\right) &\le 3\sum_{t=1}^T \mathbb{1}\left(G_{s,a,h}^t\right) \sum_{\ell=1}^H \sum_{s',a'} p_\ell^{\pi^t}(s',a') b_\ell^{t-1}(s',a') \\
&\le 3H \sum_{t=1}^T \mathbb{1}\left(G_{s,a,h}^t\right) b_h^{t-1}(s,a) \le 12H^2 \sqrt{\frac{Z_h^T(s,a)\beta(T,\delta)}{p_h^{\min}(s,a)}},
\end{aligned}$$

where the last inequality was already derived in the proof of Theorem 2. The statement follows by solving the resulting inequality in $Z_h^T(s,a)$. ■

Lemma 10 *Let $S \geq 2$. For any time $t \geq 1$, $\beta(t-1,\delta) \leq 5\log\frac{3SAH}{\delta} + 4S + 4S\log(t)$.*

Proof Starting from the definition of β and using the inequality $(x+y)^2 \leq 2x^2 + 2y^2$,

$$\begin{aligned}
\beta(t-1,\delta) &= \left(\sqrt{\frac{1}{2}\left(\log\frac{3SAH}{\delta} + \log(et)\right)} + \sqrt{2\log\frac{3SAH}{\delta} + 2(S-1)\log\left(e\left(1+\frac{t-1}{S-1}\right)\right)}\right)^2 \\
&\leq 5\log\frac{3SAH}{\delta} + \log(et) + 4(S-1)\log\left(e\left(1+\frac{t-1}{S-1}\right)\right) \\
&\leq 5\log\frac{3SAH}{\delta} + \log(et) + 4(S-1)\log(et) \\
&\leq 5\log\frac{3SAH}{\delta} + 4S\log(et) \\
&= 5\log\frac{3SAH}{\delta} + 4S + 4S\log(t)\,.
\end{aligned}$$

■

Lemma 11 *Let $B, C \geq 1$. If $k \leq B\log(k) + C$, then*

$$k \leq B\log(B^2 + 2C) + C.$$

Proof Since $\log(k) \leq \sqrt{k}$ for any $k \geq 1$, we have that $k \leq B\sqrt{k} + C$. Solving this second-order inequality, we get the crude bound $\sqrt{k} \leq \frac{B}{2} + \sqrt{\frac{B^2}{4} + C}$, which in turns yields $k \leq B^2 + 2C$ using that $(x+y)^2 \leq 2(x^2+y^2)$ for $x, y \geq 0$. The statement follows by plugging this bound into the logarithm. ■

A.4 Explicit sample complexity bound

We show how to derive the sample complexity bound stated in Theorem 2 starting from the one derived in (4). Let $\mathcal{C}(\varepsilon) := \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{H^4}{p_h^{\min}(s,a)\max\{\widetilde{\Delta}_h(s,a),\varepsilon\}^2}$. Since $\beta(\tau-1,\delta) \leq 5\log\frac{3SAH}{\delta} + 4S + 4S\log(\tau)$ from Lemma 10, (4) implies that

$$\begin{aligned}
\tau &\leq 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 576\mathcal{C}(\varepsilon)S + 576\mathcal{C}(\varepsilon)S\log(\tau) + 1 \\
&\leq 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 577\mathcal{C}(\varepsilon)S + 576\mathcal{C}(\varepsilon)S\log(\tau)\,,
\end{aligned}$$

where the second inequality holds since $1 \leq \mathcal{C}(\varepsilon)S$. Using Lemma 11 with $B = 576\mathcal{C}(\varepsilon)S$ and $C = 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 577\mathcal{C}(\varepsilon)S$,

$$\begin{aligned}
\tau &\leq 576\mathcal{C}(\varepsilon)S\log\left(576^2\mathcal{C}(\varepsilon)^2S^2 + 1440\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 1154\mathcal{C}(\varepsilon)S\right) + 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 577\mathcal{C}(\varepsilon)S \\
&\leq 576\mathcal{C}(\varepsilon)S\log\left(4\cdot 576^2\mathcal{C}(\varepsilon)^2S^2\log\frac{3SAH}{\delta}\right) + 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 577\mathcal{C}(\varepsilon)S \\
&\leq 1152\mathcal{C}(\varepsilon)S\log\left(1152\mathcal{C}(\varepsilon)S\log\frac{3SAH}{\delta}\right) + 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta} + 577\mathcal{C}(\varepsilon)S \\
&\leq 1729\mathcal{C}(\varepsilon)S\log\left(1152\mathcal{C}(\varepsilon)S\log\frac{3SAH}{\delta}\right) + 720\mathcal{C}(\varepsilon)\log\frac{3SAH}{\delta},
\end{aligned}$$

where the inequalities use some trivial bounds to simplify the final expression. The result stated in Theorem 2 follows from here by noting that $\mathcal{C}(\varepsilon) \leq \frac{SAH^5}{p_{\min}\varepsilon^2}$.

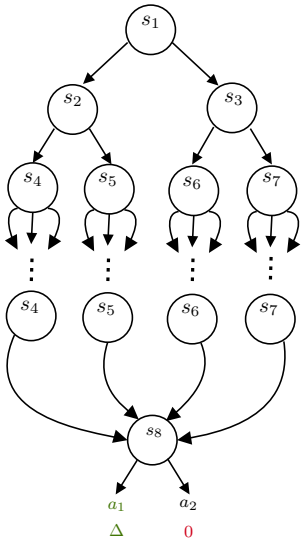

Figure 1: Example of the MDP instance for proving Theorem 4 where $S = 8$, $A = 3$, and $H \geq 4$. The reward is Gaussian with variance 1 and zero mean except for $r_H(s_8, a_1) = \Delta > 0$. The optimal policy takes action a_1 in s_8 at the last stage H, while any path of length $H - 1$ to reach that state is optimal.

Appendix B. Proof of Theorem 4

An example of the MDP instance we build to prove Theorem 4 is shown in Figure 1. Suppose $S \geq 4$, $A \geq 2$ and $H \geq \lceil \log_2(S) \rceil + 1$. We arrange the states in a binary tree, starting from the root and adding them one by one from left to right and top to bottom. The leaves of the binary tree have all A actions available, and so do the states at the second-last layer which have zero children. Such actions keep the agent in the same state up to layer $H - 1$. In layer $H - 1$, all A actions for all reachable states transition to state s_8. In the latter, only two actions are available, among which a_1 is the only in the whole MDP with a positive reward of $\Delta > 0$.

The intuition is that this is an extremely easy instance of regret minimization. In fact, it is essentially a two-armed bandit where the only thing that must be learned is the optimal action at stage H (i.e., a_1), while the agent can behave arbitrarily in all stages before and still suffer zero regret. On the other hand, this is an extremely hard instance for PAC identification since, in order to return an ε-optimal policy with enough confidence, any algorithm must explore all state-action pairs at stages from 1 to $H - 1$ up to an error below ε in order to assess that their rewards are all ε-close.

Proof of the sample-complexity lower bound Let us start by proving the lower bound on the sample complexity of any (ε, δ)-PAC algorithm. Let d be the depth of the tree, i.e., the first integer such that $\sum_{i=0}^{d-1} 2^i = 2^d - 1 \geq S - 1$. That is $d = \lceil \log_2(S) \rceil$. Note that, even if the last layer is not complete (i.e., it has less than 2^{d-1} states), the second last layer must be complete. Therefore, the are at least $2^{d-2} \geq S/4$ states with all A actions available that are reachable at stage $H - 1$. Call these states $\bar{s}_1, \dots, \bar{s}_m$ for m some integer with $m \geq S/4$. Moreover, $\overline{\Delta}_{H-1}(\bar{s}_i, a) = 0$ for all $i \in [m]$ and $a \in [A]$ since all paths are optimal up to stage $H - 1$, where $\overline{\Delta}$ denotes the deterministic return gaps of Tirinzoni et al. (2022). Note that the MDP is deterministic, so the lower bound of Theorem 2 by Tirinzoni et al. (2022) holds. By applying this result, we get

$$\forall i \in [m], a \in [A] : \mathbb{E}[n^{\tau}_{H-1}(\bar{s}_i, a)] \geq \frac{\log(1/4\delta)}{4\varepsilon^2}.$$

This directly implies that

$$\mathbb{E}[\tau] = \sum_{s \in \mathcal{S}_{H-1}} \sum_{a \in [A]} \mathbb{E}[n^{\tau}_{H-1}(s, a)] \geq \frac{SA \log(1/4\delta)}{16\varepsilon^2}.$$

Proof of the regret upper bound Let us now deal with regret minimization. Let us take the UCBVI algorithm (Azar et al., 2017a) with Hoeffding bonus (for general stochastic transitions) that we described in Section 2. We shall consider a slightly different *stage-dependent* definition of the bonuses $b_h^t(s,a)$. All we need is that, at any time $t \in \mathbb{N}$, they guarantee concentration for all (s,a,h) with probability at least $1 - \frac{1}{t^2}$. For our proof we only need to specify the specific form at stage H. Since at that stage we only need to concentrate rewards, using Hoeffding's inequality for sub-Gaussian distributions with $\sigma^2 = 1$, it is easy to see that

$$b_H^t(s,a) := \sqrt{\frac{2\log(2SAHt^2)}{n_H^t(s,a)}} \wedge 1$$

ensures $\overline{Q}_H^t(s,a) \geq Q_H^\star(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}, t \in \mathbb{N}$ with probability at least $1 - \frac{1}{t^2}$. For all stages $h = 1, \dots, H-1$, we can simply take the bonuses considered in the main paper with a decreasing schedule for δ, though their explicit expression is not really used in our proof.

Note that, in this particular MDP, the regret is zero whenever the agent plays action a_1 at stage H since all actions played from stage 1 to $H-1$ are optimal. In other words, this is equivalent to a bandit problem with two actions. Therefore, for any $T \geq 1$,

$$\sum_{t=1}^T \left(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\right) = \Delta \sum_{t=1}^T \mathbb{1}\left(a_H^t = a_2\right) = \Delta n_H^T(s_H, a_2).$$

Under the good event $\mathcal{G}_t$ in which the confidence intervals are valid at t, if $a_H^t = a_2$, then

$$b_H^{t-1}(s_H,a_2) \geq \frac{\Delta}{2} \implies n_H^{t-1}(s_H,a_2) \leq \frac{8}{\Delta^2}\log(2SAHt^2).$$

Therefore, the cumulative regret up to any time T in such good events can be bounded as

$$\sum_{t \leq T : \mathcal{G}_t} \left(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\right) \leq \Delta \sum_{t \leq T : \mathcal{G}_t} \mathbb{1}\left(a_H^t = a_2, n_H^{t-1}(s_H,a_2) \leq \frac{8}{\Delta^2}\log(2SAHT^2)\right) \leq \frac{8}{\Delta}\log(2SAHT^2).$$

On the other hand, the expected regret under the bad events is bounded as

$$\mathbb{E}\left[\sum_{t \leq T : \neg\mathcal{G}_t} \left(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\right)\right] \leq \Delta \sum_{t=1}^T \mathbb{P}(\neg\mathcal{G}_t) \leq \Delta \sum_{t=1}^T \frac{1}{t^2} \leq 2\Delta.$$

Combining these two we obtain the following bound on the expected cumulative regret:

$$\mathbb{E}\left[\sum_{t=1}^T \left(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\right)\right] \leq \frac{8}{\Delta}\log(2SAHT^2) + 2\Delta.$$

Finally,

$$T_\varepsilon := \inf_{T \in \mathbb{N}} \left\{ T : \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_1^\star(s_1) - V_1^{\pi^t}(s_1)\right] \leq \varepsilon\delta \right\} \leq \inf_{T \in \mathbb{N}} \left\{ T : \frac{8}{\Delta}\log(2SAHT^2) + 2\Delta \leq T\varepsilon\delta \right\}.$$

To bound T_ε, we need to solve the inequality on the right-hand side above. Using $\log(T) \leq \sqrt{T}$, it is easy to show that a crude bound is

$$T \leq \frac{260}{\Delta^2\varepsilon^2\delta^2}\left(4\log(2SAH) + \Delta^2\right).$$

Plugging this into the logarithm above yields

$$T_\varepsilon \leq \frac{2}{\Delta\varepsilon\delta}\left(4\log(2SAH) + 16\log\frac{17}{\Delta\varepsilon\delta} + 8\log\left(4\log(2SAH) + \Delta^2\right) + \Delta^2\right) + 1.$$

Setting $\Delta = \varepsilon$ and using $\log\left(4\log(2SAH) + \Delta^2\right) \leq 4\log(2SAH) + \Delta^2$ concludes the proof. ■