Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, Michal Valko

Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited

Omar Darwiche Domingues OMAR.DARWICHE-DOMINGUES@INRIA.FR
Inria Lille

Pierre Ménard* PIERRE.MENARD@OVGU.DE
Otto von Guericke University Magdeburg

Emilie Kaufmann EMILIE.KAUFMANN@UNIV-LILLE.FR
CNRS & ULille (CRIStAL), Inria Lille

Michal Valko VALKOM@DEEPMIND.COM
DeepMind Paris & Inria Lille

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

In this paper, we propose new problem-independent lower bounds on the sample complexity and regret in episodic MDPs, with a particular focus on the *non-stationary case* in which the transition kernel is allowed to change in each stage of the episode. Our main contribution is a lower bound of $\Omega((H^3SA/\varepsilon^2)\log(1/\delta))$ on the sample complexity of an (ε, δ)-PAC algorithm for best policy identification in a non-stationary MDP, relying on a construction of "hard MDPs" which is different from the ones previously used in the literature. Using this same class of MDPs, we also provide a rigorous proof of the $\Omega(\sqrt{H^3SAT})$ regret bound for non-stationary MDPs. Finally, we discuss connections to PAC-MDP lower bounds.

Keywords: reinforcement learning, episodic, lower bounds

1. Introduction

In episodic reinforcement learning (RL), an agent interacts with an environment in episodes of length H. In each stage $h \in \{1, \dots, H\}$, the agent is in a state s_h, takes an action a_h then observes the next state s_{h+1} sampled according to a transition kernel $p_h(\cdot|s_h, a_h)$, and receives a reward $r_h(s_h, a_h)$. The quality of a RL algorithm, which adaptively selects the next action to perform based on past observation, can be measured with different performance metrics.

On the one hand, the sample complexity quantifies the number of episodes in which an algorithm makes mistakes (in the PAC-MDP setting) or the number of episodes needed before outputing a near optimal policy (in the best policy identification setting). On the other hand, the regret quantifies the difference between the total reward gathered by an optimal policy and that of the algorithm. Minimax upper bounds on the sample complexity or the regret of episodic RL algorithms in finite MDPs have been given in the prior work, for instance in the work of Dann and Brunskill (2015); Dann et al. (2017); Azar et al. (2017); Jin et al. (2018), and Zanette and Brunskill (2019). Deriving *lower bounds* is also helpful to assess the quality of these upper bounds, in particular in terms of their scaling in the horizon H, the number of states S and the number of actions A.

* This work was done while Pierre Ménard was at Inria Lille.

Sample complexity lower bounds Sample complexity has mostly been studied in the γ-discounted setting for PAC-MDP algorithms (Kakade, 2003), for which the number of time steps in which an algorithm acts ε-sub-optimaly (called the *sample complexity*) has to be upper bounded, with probability larger than $1-\delta$. State-of-the art lower bounds are a $\Omega\big(\frac{SA}{\varepsilon^2}\log\big(\frac{S}{\delta}\big)\big)$ bound by Strehl et al. (2009) and a $\Omega\Big(\frac{SA}{(1-\gamma)^3\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$ bound by Lattimore and Hutter (2012). A lower bound of the same order is provided by Azar et al. (2012) for the number of steps algorithms that have access to a generative model need to identify an ε-optimal policy.

PAC-MDP algorithms in the episodic setting were later studied by Dann and Brunskill (2015), who also provide a lower bound. Unlike the previous ones, they *do not* lower bound the number of ε-mistakes of the algorithm, but rather state that any algorithm that outputs a deterministic policy $\widehat{\pi}$ that is ε-optimal with probability at least $1-\delta$, there exists an MDP where the expected number of episodes before $\widehat{\pi}$ is returned must be at least $\Omega\Big(\frac{SAH^2}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$. This lower bound therefore applies to the sample complexity of best-policy identification (see Section 2 for a formal definition), which is our main focus in this paper. The "hard MDP" instances used to prove this worse-case bound are inspired by the the ones of Strehl et al. (2009) and consist of S multi-armed bandit (MAB) problems played in parallel. Jiang et al. (2017); Dann et al. (2017); Yin et al. (2020) show that the PAC lower bound has an extra factor H when the transition kernels are allowed to depend on the stage h of the episode, and also rely on a construction of hard instances based on parallel MAB instances. In this paper, we aim for a result that applies to a more general class of strategies than what has been previously shown. Unlike the prior lower bound constructions with parallel MAB instances, we design a class of MDPs where each of them has stage-dependent transitions and behaves as *single* bandit instance with $\Theta(HSA)$ arms. In Theorem 7, we prove that in this class there exists an MDP for which the expected number of samples needed to identify an ε-optimal policy with probability $1-\delta$ is at least $\Omega\Big(\frac{SAH^3}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$. Our construction avoids unnecessary assumptions without which prior analyses would not work.

Regret lower bounds In the average-reward setting, Jaksch et al. (2010) prove a regret lower bound of $\Omega(\sqrt{DSAT})$ where D is the diameter of the MDP and T is the total number of actions taken in the environment. In the episodic setting, the total number of actions taken is HT, where T is now the number of episodes, and H is roughly the equivalent of the diameter D.[1] Hence, intuitively, the lower bound of Jaksch et al. (2010) should be translated to $\Omega(\sqrt{H^2SAT})$ for episodic MDPs after T episodes. Yet, to the best of our knowledge, a precise proof of this claim has not been given in the literature. The proof of Jaksch et al. (2010) relies on building a set of hard MDPs with "bad" states (with zero reward) and "good" states (with reward 1), and can be adapted to episodic MDPs by making the good states absorbing. However, this construction does not include MDPs whose transitions are allowed to change at every stage h. In the case of stage-dependent transitions, Jin et al. (2018) claim that the lower bound becomes $\Omega(\sqrt{H^3SAT})$, by using the construction of Jaksch et al. (2010) and a mixing-time argument, but they do not provide a complete proof. In Theorem 9, we provide a detailed proof of their statement, by relying on the same class of hard MDPs given for our sample complexity lower bound.

[1] The diameter D is the minimum average time to go from one state to another. In an episodic MDP, if the agent can come back to the same initial state s_1 after H steps, the average time between any pair of states is bounded by $2H$, if we restrict the state set to the states that are reachable from s_1 in H steps.

Algorithm	Setting	h-dependent optimal	h-independent optimal
UCBVI (Azar et al., 2017)	Regret	Yes[3]	Yes
Q-Learning+UCB (Jin et al., 2018)	Regret	Yes	No
BPI-UCBVI (Ménard et al., 2020)	BPI	Yes	No
ORLC (Dann et al., 2019)[4]	BPI, PAC	Yes[3]	Yes

Table 1: Algorithms matching the lower bounds in different settings.[2]

Our contributions Our main contribution are unified, simple, and complete proofs of minimax lower bounds for different episodic RL settings. In particular, using a single class of hard MDPs and the same information-theoretic tools, we provide regret and sample-complexity lower bounds for episodic reinforcement learning algorithms for stage-dependent transitions. For T episodes, the regret bound is $\Omega(\sqrt{H^3SAT})$, which is the same as the one sketched by Jin et al. (2018), but we provide a detailed proof with a different construction. This lower bound is matched by the optimistic Q-learning algorithm of Jin et al. (2018). For the sample complexity of best-policy identification (BPI), we prove the first lower bound for MDPs with stage-dependent transitions for algorithms that may output randomized policies after a random stopping time. This bound is of order $\Omega\Big(\frac{SAH^3}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$ and is matched by the BPI-UCBVI algorithm of Ménard et al. (2020). As a corollary of the BPI lower bound, we also obtain a lower bound of $\Omega\Big(\frac{SAH^3}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$ in the PAC-MDP setting. Finally, note that our proof technique also provides rigorous proofs of the bounds $\Omega(\sqrt{H^2SAT})$ and $\Omega\Big(\frac{SAH^2}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$ for regret and best-policy identification with stage-independent transitions. Table 1 shows algorithms whose upper bounds match the lower bounds[2] presented in this paper for the regret, BPI, and PAC settings, both for stage-dependent and stage-independent transitions.[3] [4]

2. Setting and Performance Measures

Markov decision process We consider an episodic Markov decision process (MDP) defined as a tuple $\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, H, \mu, p, r)$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, H is the number of steps in one episode, μ is the initial state distribution, $p = \{p_h\}_h$ and $r = \{r_h\}_h$ are sets of transitions and rewards for $h \in [H]$ such that taking an action a in state s results in a reward $r_h(s,a) \in [0,1]$ and a transition to $s' \sim p_h(\cdot|s,a)$. We assume that the cardinalities of $\mathcal{S}$ and $\mathcal{A}$ are finite, and denote them by S and A, respectively.

[2] Up to constants and logarithmic terms, and assuming that either T is large enough (for the regret) or ε is small enough (for the sample complexity).

[3] UCBVI and ORLC have been proposed for MDPs with h-independent transitions, but they can readily be used for MDPs with h-dependent transitions by viewing them as MDPs with HS states and h-independent transitions.

[4] Dann et al. (2019) analyze the ORLC algorithm in a slightly different setting, proving that it outputs "Individual POlicy Certificates" (IPOC). ORLC can be converted to an (ε, δ)-PAC algorithm for BPI (see Definition 3) by setting the stopping rule to be the first time the optimality certificate is smaller than ε. Sample complexity guarantees for both BPI and PAC-MDP setting can be deduced from their analysis.

Markov and history-dependent policies Let $\Delta(\mathcal{A})$ be the set of probability distributions over the action set and let

$$\mathcal{I}_h^t = \big((\mathcal{S}\times\mathcal{A})^{H-1}\times\mathcal{S}\big)^{t-1}\times(\mathcal{S}\times\mathcal{A})^{h-1}\times\mathcal{S}$$

be the set of possible histories up to step h of episode t, that is, the set of tuples of the form

$$\big(s_1^1, a_1^1, s_2^1, a_2^1, \dots, s_H^1, \dots, s_1^t, a_1^t, s_2^t, a_2^t, \dots, s_h^t\big) \in \mathcal{I}_h^t.$$

A *Markov policy* is a function $\pi : \mathcal{S}\times[H]\to\Delta(\mathcal{A})$ such that $\pi(a|s,h)$ denotes the probability of taking action a in state s at step h. A *history-dependent policy* is a family of functions denoted by $\boldsymbol{\pi} \triangleq (\pi_h^t)_{t\geq 1, h\in[H]}$, where $\pi_h^t : \mathcal{I}_h^t \to \Delta(\mathcal{A})$ such that $\pi_h^t(a \,|\, i_h^t)$ denotes the probability of taking action a at time (t,h) after observing the history $i_h^t \in \mathcal{I}_h^t$. We denote by Π_{Markov} and Π_{Hist} the sets of Markov and history-dependent policies, respectively.

Probabilistic model A policy $\boldsymbol{\pi}$ interacting with an MDP defines a stochastic process denote by $(S_h^t, A_h^t)_{t\geq 1, h\in[H]}$, where S_h^t and A_h^t are the random variables representing the state and the action at time (t,h). As explained by (Lattimore and Szepesvári, 2020), the Ionescu-Tulcea theorem ensures the existence of probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\mathcal{M}})$ such that

$$\mathbb{P}_{\mathcal{M}}\big[S_1^t = s\big] = \mu(s),\ \mathbb{P}_{\mathcal{M}}\big[S_{h+1}^t = s | A_h^t, I_h^t\big] = p_h(s|S_h^t, A_h^t),\ \text{and}\quad \mathbb{P}_{\mathcal{M}}\big[A_h^t = a | I_h^t\big] = \pi_h^t(a|I_h^t),$$

where $\boldsymbol{\pi} = (\pi_h^t)_{t\geq 1, h\in[H]}$ and for any (t,h),

$$I_h^t \triangleq \big(S_1^1, A_1^1, S_2^1, A_2^1, \dots, S_H^1, \dots, S_1^t, A_1^t, S_2^t, A_2^t, \dots, S_h^t\big)$$

is the random vector taking values in $\mathcal{I}_h^t$ containing all state-action pairs observed up to step h of episode t, but not including A_h^t. We denote by $\mathcal{F}_h^t$ the σ-algebra generated by I_h^t. Next, we denote by $\mathbb{P}_{\mathcal{M}}^{I_H^T}$ the pushforward measure of I_H^T under $\mathbb{P}_{\mathcal{M}}$,

$$\mathbb{P}_{\mathcal{M}}^{I_H^T}\big[i_H^T\big] \triangleq \mathbb{P}_{\mathcal{M}}\big[I_H^T = i_H^T\big] = \prod_{t=1}^{T}\mu(s_1^t)\prod_{h=1}^{H-1}\pi_h^t(a_h^t|i_h^t)p_h\big(s_{h+1}^t|s_h^t, a_h^t\big),$$

where $i_h^t \triangleq (s_1^1, a_1^1, \dots, s_H^1, \dots, s_1^t, a_1^t, \dots, s_h^t) \in \mathcal{I}_h^t$. Moreover, let $\mathbb{E}_{\mathcal{M}}$ be the expectation under $\mathbb{P}_{\mathcal{M}}$. Notice that the dependence of $\mathbb{P}_{\mathcal{M}}$ and $\mathbb{E}_{\mathcal{M}}$ on the policy $\boldsymbol{\pi}$ is denoted implicitly, and we denote them explicitly as $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}$ and $\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}$ when it is relevant.

Value function In an episode t, the value of a policy $\boldsymbol{\pi}$ in the MDP $\mathcal{M}$ is defined as

$$V^{\boldsymbol{\pi},t}(i_H^{t-1}, s) \triangleq \mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}\left[\sum_{h=1}^{H} r_h(S_h^t, A_h^t)\Big| I_H^{t-1} = i_H^{t-1}, S_1^t = s\right],$$

where i_H^{t-1} are the states and actions observed before episode t and $\boldsymbol{\pi}$ can be history-dependent. In particular, for a Markov policy π, the value does not depend on i_H^{t-1} and we have

$$V^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathcal{M}}\left[\sum_{h=1}^{H} r_h(S_h^1, A_h^1)\Big| S_1^1 = s\right].$$

The optimal value function V^* is defined as $V^*(s) \triangleq \max_{\pi \in \Pi_{\text{Markov}}} V^\pi(s)$ which is achieved by an optimal policy π^* that satisfies $V^{\pi^*}(s) = V^*(s)$ for all $s \in \mathcal{S}$. As a consequence of Theorem 5.5.1 of Puterman (1994), we have $V^*(s) \geq V^{\boldsymbol{\pi},t}(i_H^{t-1}, s)$, which shows that Markov policies are sufficient to achieve an optimal value function. We also define $\rho^* \triangleq \rho^{\pi^*}$ and the average value functions over the initial state as

$$\rho^{\boldsymbol{\pi},t}(i_H^{t-1}) \triangleq \mathbb{E}_{s\sim\mu}\big[V^{\boldsymbol{\pi},t}(i_H^{t-1}, s)\big], \quad \rho^\pi \triangleq \mathbb{E}_{s\sim\mu}[V^\pi(s)].$$

Algorithm We define a reinforcement-learning algorithm as a history-dependent policy $\boldsymbol{\pi}$ used to interact with the environment. In the BPI setting, where we eventually stop and recommend a policy, an algorithm is defined as a triple $(\boldsymbol{\pi}, \tau, \widehat{\pi}_\tau)$ where τ is a stopping time with respect to the filtration $(\mathcal{F}_H^t)_{t\geq 1}$, and $\widehat{\pi}_\tau$ is a Markov policy recommended after τ episodes.

Performance criteria The performance of RL algorithms has been commonly measured according to its regret or under a Probably Approximately Correct (PAC) framework, as defined below.

Definition 1 *The expected regret of an algorithm $\boldsymbol{\pi}$ in an MDP $\mathcal{M}$ after T episodes is defined as*

$$\mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}) \triangleq \mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}\left[\sum_{t=1}^{T}\big(\rho^* - \rho^{\boldsymbol{\pi},t}(I_H^{t-1})\big)\right].$$

Definition 2 *An algorithm $\boldsymbol{\pi}$ is (ε, δ)-PAC for exploration in an MDP $\mathcal{M}$ (or PAC-MDP) if there exists a polynomial function $F_{\text{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta))$ such that its* sample complexity

$$\mathcal{N}_\varepsilon^{\text{PAC}} \triangleq \sum_{t=1}^{\infty} \mathbb{1}\big\{\rho^* - \rho^{\boldsymbol{\pi},t}(I_H^{t-1}) > \varepsilon\big\}$$

satisfies $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}\big[\mathcal{N}_\varepsilon^{\text{PAC}} > F_{\text{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta))\big] \leq \delta$.

Definition 3 *An algorithm $(\boldsymbol{\pi}, \tau, \widehat{\pi}_\tau)$ is (ε, δ)-PAC for best-policy identification in an MDP $\mathcal{M}$ if the policy $\widehat{\pi}_\tau$ returned after τ episodes satisfies*

$$\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}\Big[\rho^{\widehat{\pi}_\tau} \leq \rho^* - \varepsilon\Big] \leq \delta.$$

The sample complexity *is defined as the number of episodes τ required for stopping.*

3. Lower Bound Recipe

In this section, we present the two main ingredients for the proof of our minimax lower bounds. These lower bounds consider a class $\mathcal{C}$ of hard MDPs instances (on which the optimal policy is difficult to identify), that are typically close to each other, but for which the behavior of an algorithm is expected to be different (because they do not share the same optimal policy). The class $\mathcal{C}$ used to derive all our results is presented in Section 3.1. Then, lower bound proofs use a change of distribution between two well-chosen MDPs in $\mathcal{C}$ in order to obtain inequalities on the expected number of visits of certain state-action pairs in one of them. The information-theoretic tools that we use for these changes of distributions are gathered in Section 3.2.

3.1. Hard MDP instances

From a high-level perspective, the family of MDPs that we use for our proofs behave like multi-armed bandits with $\Theta(HSA)$ arms. To gain some intuition about the construction, assume that $S = 4$ and consider the MDP in Figure 1. The agent starts in a *waiting state* s_{w} where it can take an action a_{w} to stay in s_{w} up to a stage $\overline{H} < H$, after which the agent has to leave s_{w}. From s_{w}, the agent can only transition to a state s_1, from which it can reach two absorbing states, a "good" state s_g and a "bad" state s_b. The state s_g is the only state where the agent can obtain a reward, which starts to be 1 at stage $\overline{H}+2$. There is a single action a^* in state s_1 that increases by ε the probability of arriving to the good state, and this action must be taken at a specific stage h^*. The intuition is that, in order to maximize the rewards, the agent must choose the right moment $h \in \{1, \dots, \overline{H}\}$ to leave s_{w}, and then choose the good action $a^* \in \{1, \dots, A\}$ in s_1. This results in a total of $\overline{H}A$ possible choices, or "arms" and the maximal reward is $\Theta(\overline{H})$. By analogy with the existing minimax regret bound for multi-armed bandits (Auer et al., 2002; Lattimore and Szepesvári, 2020), the regret lower bound should be $\Omega(H\sqrt{HAT})$, by taking $\overline{H} = \Theta(H)$.

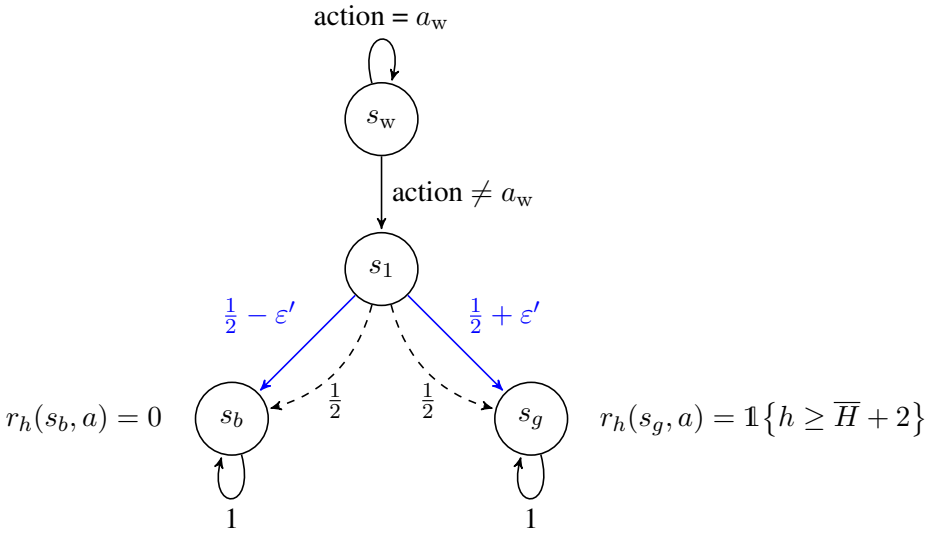

Figure 1: Illustration of the class of hard MDPs for $S = 4$.

Inspired by the tree construction of Lattimore and Szepesvári (2020) for the lower bound in the average-reward setting, we now generalize these MDPs to $S > 4$. Consider a family of MDPs described as follows and illustrated in Figure 2. First, we state the following assumption, which we relax in Appendix D.

Assumption 1 *The number of states and actions satisfy $S \geq 6$, $A \geq 2$, and there exists an integer d such that $S = 3 + (A^d - 1)/(A - 1)$, which implies $d = \Theta(\log_A S)$. We further assume that $H \geq 3d$.*

As in the previous case, there are three special states: a "waiting" state s_{w} where the agent starts and can choose to stay up to a stage $\overline{H}$, a "good" state s_g that is absorbing and is the only state where the agent obtains rewards, and a "bad" state s_b that is absorbing and gives no reward. The other $S - 3$

states are arranged in a full A-ary tree of depth $d-1$, which can be done since we assume there exists an integer d such that $S-3=\sum_{i=0}^{d-1} A^i$. The root of the tree is denoted by s_{root}, which can only be reached from s_{w}, and the states s_g and s_b can only be reached from the leaves of the tree.

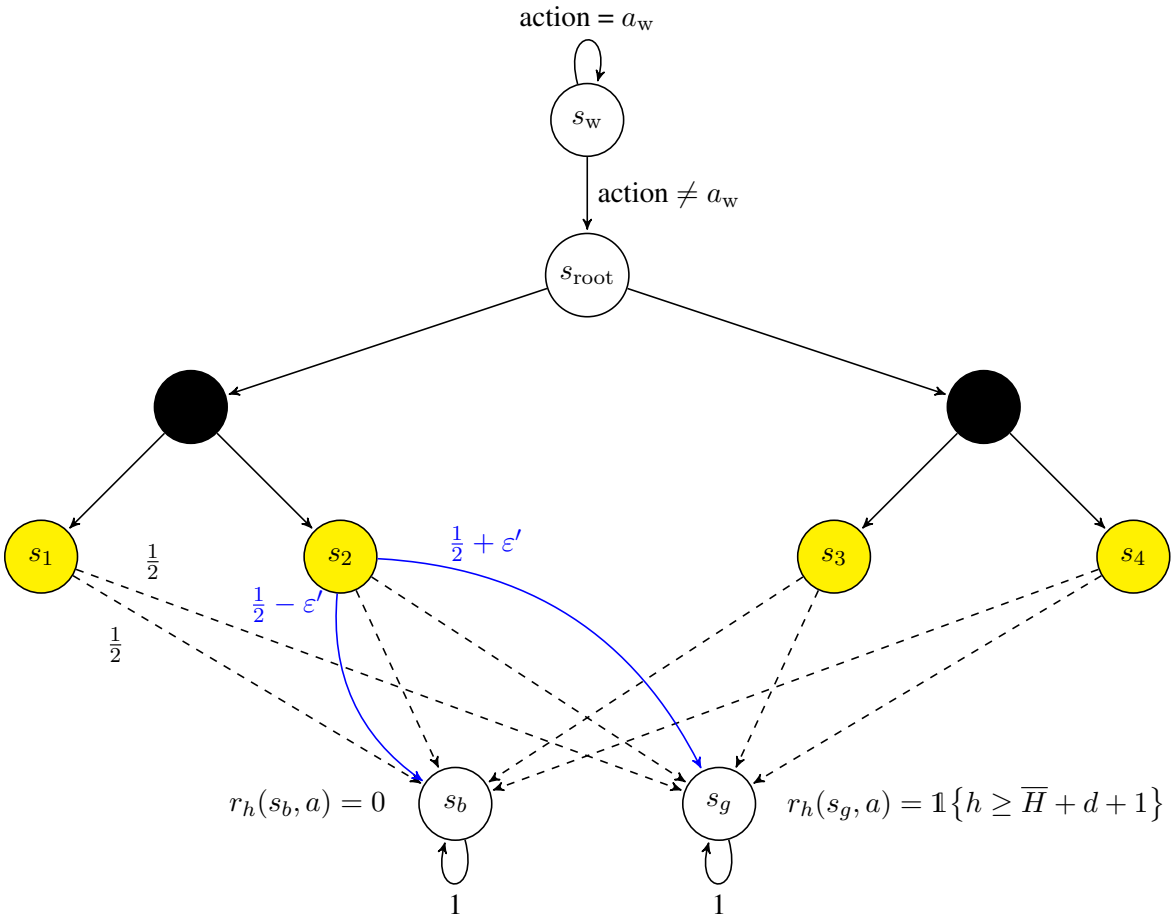

Figure 2: Illustration of the class of hard MDPs used in the proofs of Theorems 7 and 9.

Let $\overline{H} \leq H-d$ be an integer that will be a parameter of the class of MDPs. Letting $\mathcal{L}=\{s_1,s_2,\ldots,s_L\}$ be the set of L leaves of the tree, we define for each

$$(h^*,\ell^*,a^*)\in\{1+d,\ldots,\overline{H}+d\}\times\mathcal{L}\times\mathcal{A},$$

an MDP $\mathcal{M}_{(h^*,\ell^*,a^*)}$ as follows. For any state in the tree, the transitions are deterministic: the a-th action in a node leads to the a-th child of that node. The transitions from s_{w} are given by

$$p_h(s_{\text{w}}|s_{\text{w}},a)\triangleq\mathbb{1}\{a=a_{\text{w}},h\leq\overline{H}\}\quad\text{and}\quad p_h(s_{\text{root}}|s_{\text{w}},a)\triangleq 1-p_h(s_{\text{w}}|s_{\text{w}},a).$$

That is, there is an action a_{w} that allows the agent to stay at s_{w} up to a stage $\overline{H}$. After stage $\overline{H}$, the agent has to traverse the tree down to the leaves. The transitions from any leaf $s_i \in \mathcal{L}$ are given by

$$p_h(s_g|s_i,a) \triangleq \frac{1}{2} + \Delta_{(h^*,\ell^*,a^*)}(h,s_i,a) \quad \text{and} \quad p_h(s_b|s_i,a) \triangleq \frac{1}{2} - \Delta_{(h^*,\ell^*,a^*)}(h,s_i,a), \tag{1}$$

where $\Delta_{(h^*,\ell^*,a^*)}(h,s_i,a) \triangleq \mathbb{1}\{(h,s_i,a) = (h^*,s_{\ell^*},a^*)\} \cdot \varepsilon'$, for some $\varepsilon' \in [0,1/2]$ that is the second parameter of the class. This means that there is a single leaf ℓ^* where the agent can choose an action a^* at stage h^* that increases the probability of arriving to the good state s_g. Finally, the states s_g and s_b are absorbing, that is, for any action a, we have $p_h(s_b|s_b,a) \triangleq p_h(s_g|s_g,a) \triangleq 1$. The reward function depends only on the state and is defined as

$$\forall a \in \mathcal{A}, \quad r_h(s,a) \triangleq \mathbb{1}\big\{s = s_g, h \geq \overline{H} + d + 1\big\}$$

so that the agent does not miss any reward if it chooses to stay at s_{w} until stage $\overline{H}$.

We further define a reference MDP $\mathcal{M}_0$ which is an MDP of the above type but for which $\Delta_0(h,s_i,a) \triangleq 0$ for all (h,s_i,a). For every ε' and $\overline{H}$, we define the class $\mathcal{C}_{\overline{H},\varepsilon'}$ to be the set

$$\mathcal{C}_{\overline{H},\varepsilon'} \triangleq \{\mathcal{M}_0\} \bigcup \big\{\mathcal{M}_{(h^*,\ell^*,a^*)}\big\}_{(h^*,\ell^*,a^*) \in \{1+d,\dots,\overline{H}+d\} \times \mathcal{L} \times \mathcal{A}}.$$

3.2. Change of Distribution Tools

Definition 4 *The Kullback-Leibler divergence between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on a measurable space $(\Omega,\mathcal{G})$ is defined as*

$$\mathrm{KL}(\mathbb{P}_1,\mathbb{P}_2) \triangleq \int_\Omega \log\left(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}(\omega)\right)\mathrm{d}\mathbb{P}_1(\omega),$$

if $\mathbb{P}_1 \ll \mathbb{P}_2$ and $+\infty$ otherwise. For Bernoulli distributions, we define $\forall (p,q) \in [0,1]^2$,

$$\mathrm{kl}(p,q) \triangleq \mathrm{KL}\left(\mathcal{B}(p),\mathcal{B}(q)\right) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

Lemma 5 (proof in Appendix A) *Let $\mathcal{M}$ and $\mathcal{M}'$ be two MDPs that are identical except for their transition probabilities, denoted by p_h and p'_h, respectively. Assume that we have $\forall (s,a)$, $p_h(\cdot|s,a) \ll p'_h(\cdot|s,a)$. Then, for any stopping time τ with respect to $(\mathcal{F}_H^t)_{t\geq 1}$ that satisfies $\mathbb{P}_{\mathcal{M}}[\tau < \infty] = 1$,*

$$\mathrm{KL}\Big(\mathbb{P}_{\mathcal{M}}^{I_H^\tau}, \mathbb{P}_{\mathcal{M}'}^{I_H^\tau}\Big) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h\in[H-1]} \mathbb{E}_{\mathcal{M}}\left[N_{h,s,a}^\tau\right] \mathrm{KL}\big(p_h(\cdot|s,a), p'_h(\cdot|s,a)\big), \tag{2}$$

where $N_{h,s,a}^\tau \triangleq \sum_{t=1}^\tau \mathbb{1}\big\{(S_h^t,A_h^t) = (s,a)\big\}$ and $I_H^\tau : \Omega \to \bigcup_{t\geq 1}\mathcal{I}_H^t : \omega \mapsto I_H^{\tau(\omega)}(\omega)$ is the random vector representing the history up to episode τ.

Lemma 6 (Lemma 1, Garivier et al., 2019) *Consider a measurable space $(\Omega,\mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$-measurable function $Z : \Omega \to [0,1]$, we have*

$$\mathrm{KL}(\mathbb{P}_1,\mathbb{P}_2) \geq \mathrm{kl}(\mathbb{E}_1[Z],\mathbb{E}_2[Z]),$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively.

4. Sample Complexity Lower Bounds

We are ready to state a new minimax lower bound on the sample complexity of best policy identification (see Definition 3), in an MDP with stage-dependent transitions. We note that unlike existing sample complexity lower bounds which also construct "bandit-like" hard instances (Strehl et al., 2009; Lattimore and Hutter, 2012; Dann and Brunskill, 2015), we do not refer to the bandit lower bound of Mannor and Tsitsiklis (2004), but instead use explicit change of distribution arguments based on the tools given in Section 3.2. This allows us to provide BPI lower bounds for algorithms that output randomized policies and to have a self-contained proof. As a consequence of this result, we then easily derive a PAC-MDP (see Definition 2) lower bound in Corollary 8, which is proved in Appendix B.

Theorem 7 *Let $(\boldsymbol{\pi}, \tau, \widehat{\pi}_\tau)$ be an algorithm that is (ε, δ)-PAC for best policy identification in any finite episodic MDP. Then, under Assumption 1, there exists an MDP $\mathcal{M}$ with stage-dependent transitions such that for $\varepsilon \leq H/24$, $H \geq 4$ and $\delta \leq 1/16$,*

$$\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}[\tau] \geq \frac{1}{3456}\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right).$$

Corollary 8 *Let $\boldsymbol{\pi}$ be an algorithm that is (ε, δ)-PAC for exploration according to Definition 2 and that, in each episode t, plays a deterministic policy π_t. Then, under the assumptions of Theorem 7, there exists an MDP $\mathcal{M}$ such that*

$$\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}\left[\mathcal{N}_\varepsilon^{\mathrm{PAC}} > \frac{1}{6912}\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right) - 1\right] > \delta.$$

Proof of Theorem 7 Without loss of generality, we assume that for any $\mathcal{M}$, the algorithm satisfies $\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}[\tau < \infty] = 1$. Otherwise, there exists an MDP with $\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}[\tau] = +\infty$ and the lower bound is trivial.

We will prove that the lower bound holds for the reference MDP $\mathcal{M}_0$ defined in Section 3.1, that has no optimal action. To do so, we will consider changes of distributions with other MDPs in the class $\mathcal{C}_{\overline{H},\widetilde{\varepsilon}}$ for $\overline{H}$ to be chosen later and $\widetilde{\varepsilon} \triangleq 2\varepsilon/(H - \overline{H} - d)$. These MDPs are of the form $\mathcal{M}_{(h^*,\ell^*,a^*)}$ with $(h^*, \ell^*, a^*) \in \{1 + d, \dots, \overline{H} + d\} \times \mathcal{L} \times \mathcal{A}$, for which

$$\Delta_{(h^*,\ell^*,a^*)}(h, s_i, a) = \mathbb{1}\{h = h^*, s_i = s_{\ell^*}, a = a^*\}\,\widetilde{\varepsilon},$$

We recall that $d - 1$ is the depth of the tree. We denote by $\mathbb{P}_{(h^*,\ell^*,a^*)} \triangleq \mathbb{P}_{\boldsymbol{\pi},\mathcal{M}_{(h^*,\ell^*,a^*)}}$ and $\mathbb{E}_{(h^*,\ell^*,a^*)} \triangleq \mathbb{E}_{\boldsymbol{\pi},\mathcal{M}_{(h^*,\ell^*,a^*)}}$ the probability measure and expectation in the MDP $\mathcal{M}_{(h^*,\ell^*,a^*)}$ by following $\boldsymbol{\pi}$ and by $\mathbb{P}_0$ and $\mathbb{E}_0$ the corresponding operators in the MDP $\mathcal{M}_0$.

Suboptimality gap of $\widehat{\pi}_\tau$ We can show that the value of the optimal policy in any of the MDPs $\mathcal{M}_{(h^*,\ell^*a^*)}$ is $\rho^* = (H - \overline{H} - d)\left(\frac{1}{2} + \widetilde{\varepsilon}\right)$ and the value of the recommended policy $\widehat{\pi}_\tau$ is

$$\rho^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)} = (H - \overline{H} - d)\left(\frac{1}{2} + \widetilde{\varepsilon}\,\mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}[S_{h^*} = s_{\ell^*}, A_{h^*} = a^*]\right)$$

where $\mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}$ is the probability distribution over states and actions $(S_h, A_h)_{h\in[H]}$ following the Markov policy $\widehat{\pi}_\tau$ in the MDP $\mathcal{M}_{(h^*,\ell^*a^*)}$. Notice that $\rho^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}$ is a random variable and

$\mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}$ is a random measure that are $\mathcal{F}_H^\tau$-measurable. Hence,

$$\rho^* - \rho^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)} = 2\varepsilon\Big(1 - \mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}[S_{h^*} = s_{\ell^*}, A_{h^*} = a^*]\Big)$$

and

$$\rho^* - \rho^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)} < \varepsilon \iff \mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}[S_{h^*} = s_{\ell^*}, A_{h^*} = a^*] > \frac{1}{2}\,.$$

Definition of a "good" event $\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}$ for $\mathcal{M}_{(h^*,\ell^*a^*)}$ The transitions of all MDPs are the same up to the stopping time $\eta = \min\{h \in [H] : S_h \in \mathcal{L}\}$ when a leaf is reached. Hence, η depends only on the policy that is followed, and not on the parameters of the MDP, which allows us to define the random measure $\mathbf{P}^{\widehat{\pi}_\tau}$ as

$$\begin{aligned}\mathbf{P}^{\widehat{\pi}_\tau}[S_{h^*} = s_{\ell^*}, A_{h^*} = a^*] &\triangleq \mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}[S_\eta = s_{\ell^*}, A_\eta = a^*, \eta = h^*] \\ &= \mathbf{P}^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)}[S_{h^*} = s_{\ell^*}, A_{h^*} = a^*]\end{aligned} \tag{3}$$

since the probability distribution of (S_η, A_η, η) on the RHS of (3) does not depend on the parameters of the MDP (h^*, ℓ^*, a^*), given $\eta = h^*$. We define the event

$$\mathcal{E}^\tau_{(h^*,\ell^*,a^*)} \triangleq \left\{\mathbf{P}^{\widehat{\pi}_\tau}[S_{h^*} = s_{\ell^*}, A_{h^*} = a^*] > \frac{1}{2}\right\},$$

which is said to be "good" due to the fact that $\mathcal{E}^\tau_{(h^*,\ell^*,a^*)} = \left\{\rho^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)} > \rho^* - \varepsilon\right\}$. Since the algorithm is assumed to be (ε, δ)-PAC for any MDP, we have

$$\mathbb{P}_{(h^*,\ell^*,a^*)}\left[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right] = \mathbb{P}_{(h^*,\ell^*,a^*)}\left[\rho^{\widehat{\pi}_\tau}_{(h^*,\ell^*,a^*)} > \rho^* - \varepsilon\right] \geq 1 - \delta.$$

Lower bound on the expectation of τ in the reference MDP $\mathcal{M}_0$ Recall that

$$N^\tau_{(h^*,\ell^*,a^*)} = \sum_{t=1}^{\tau} \mathbb{1}\left\{S^t_{h^*} = s_{\ell^*}, A^t_{h^*} = a^*\right\},$$

such that $\sum_{(h^*,\ell^*,a^*)} N^\tau_{(h^*,\ell^*,a^*)} = \tau$. For any $\mathcal{F}_H^\tau$-measurable random variable Z taking values in $[0,1]$, we have

$$\begin{aligned}&\mathbb{E}_0\left[N^\tau_{(h^*,\ell^*,a^*)}\right]\frac{16\varepsilon^2}{(H - \overline{H} - d)^2} \geq \mathbb{E}_0\left[N^\tau_{(h^*,\ell^*,a^*)}\right]\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + \widetilde{\varepsilon}\right) \quad \text{by Lemma 14} \\ &= \mathrm{KL}\left(\mathbb{P}_0^{I_H^\tau}, \mathbb{P}^{I_H^\tau}_{(h^*,\ell^*,a^*)}\right) \quad \text{by Lemma 5} \\ &\geq \mathrm{kl}\big(\mathbb{E}_0[Z], \mathbb{E}_{(h^*,\ell^*,a^*)}[Z]\big) \quad \text{by Lemma 6}\end{aligned}$$

for any (h^*, ℓ^*, a^*), provided that $\widetilde{\varepsilon} \leq 1/4$. Letting $Z = \mathbb{1}\left\{\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right\}$ yields

$$\begin{aligned}&\mathrm{kl}\big(\mathbb{E}_0[Z], \mathbb{E}_{(h^*,\ell^*,a^*)}[Z]\big) = \mathrm{kl}\left(\mathbb{P}_0\left[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right], \mathbb{P}_{(h^*,\ell^*,a^*)}\left[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right]\right) \\ &\geq \left(1 - \mathbb{P}_0\left[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right]\right)\log\left(\frac{1}{1 - \mathbb{P}_{(h^*,\ell^*,a^*)}\left[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right]}\right) - \log(2) \quad \text{by Lemma 15} \\ &\geq \left(1 - \mathbb{P}_0\left[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\right]\right)\log\left(\frac{1}{\delta}\right) - \log(2).\end{aligned}$$

Consequently,

$$\mathbb{E}_0\Big[N^\tau_{(h^*,\ell^*,a^*)}\Big] \geq \frac{(H-\overline{H}-d)^2}{16\varepsilon^2}\left[\left(1-\mathbb{P}_0\Big[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\Big]\right)\log\left(\frac{1}{\delta}\right)-\log(2)\right].$$

Summing over all MDP instances, we obtain

$$\begin{aligned}\mathbb{E}_0[\tau] &\geq \sum_{(h^*,\ell^*,a^*)} \mathbb{E}_0\Big[N^\tau_{(h^*,\ell^*,a^*)}\Big] \\ &\geq \frac{(H-\overline{H}-d)^2}{16\varepsilon^2}\left[\left(\overline{H}LA-\sum_{(h^*,\ell^*,a^*)}\mathbb{P}_0\Big[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\Big]\right)\log\left(\frac{1}{\delta}\right)-\overline{H}LA\log(2)\right]. \qquad (4)\end{aligned}$$

Now, we have

$$\sum_{(h^*,\ell^*,a^*)}\mathbb{P}_0\Big[\mathcal{E}^\tau_{(h^*,\ell^*,a^*)}\Big] = \mathbb{E}_0\left[\sum_{(h^*,\ell^*,a^*)}\mathbb{1}\left\{\mathbf{P}^{\widehat{\pi}_\tau}[S_{h^*}=s_{\ell^*},A_{h^*}=a^*]>\frac{1}{2}\right\}\right]\leq 1. \qquad (5)$$

Above we used the fact that

$$\sum_{(h^*,\ell^*,a^*)}\mathbf{P}^{\widehat{\pi}_\tau}[S_{h^*}=s_{\ell^*},A_{h^*}=a^*] = \sum_{h^*}\mathbf{P}^{\widehat{\pi}_\tau}[S_{h^*}\in\mathcal{L}]=1$$

since, at a single stage $h^* \in \big\{1+d,\overline{H}+d\big\}$, a leaf state will be reached almost surely. This implies that, if there exists (h^*,ℓ^*,a^*) such that $\mathbf{P}^{\widehat{\pi}_\tau}[S_{h^*}=s_{\ell^*},A_{h^*}=a^*]>\frac{1}{2}$, then, for any other $(h',\ell',a')\neq(h^*,\ell^*,a^*)$, we have $\mathbf{P}^{\widehat{\pi}_\tau}[S_{h'}=s_{\ell'},A_{h'}=a']<\frac{1}{2}$, which proves (5).

Plugging (5) in (4) yields

$$\begin{aligned}\mathbb{E}_0[\tau] &\geq \frac{(H-\overline{H}-d)^2}{16\varepsilon^2}\left[\big(\overline{H}LA-1\big)\log\left(\frac{1}{\delta}\right)-\overline{H}LA\log(2)\right] \\ &\geq \overline{H}LA\frac{(H-\overline{H}-d)^2}{32\varepsilon^2}\log\left(\frac{1}{\delta}\right) \qquad (6)\end{aligned}$$

where we used the assumption that $\delta\leq 1/16$. The number of leaves $L=(1-1/A)(S-3)+1/A$ satisfies $L\geq S/4$, since we assume $A\geq 2$, $S\geq 6$. Taking $\overline{H}=H/3$ and with the assumption $d\leq H/3$, we obtain

$$\mathbb{E}_0[\tau]\geq\frac{H^3SA}{3456\varepsilon^2}\log\left(\frac{1}{\delta}\right).$$

Finally, the condition $\varepsilon\leq H/24$ implies that $\widetilde{\varepsilon}\leq 1/4$, as required above. ■

5. Regret Lower Bound

Using again change of distributions between MDPs in a class $\mathcal{C}_{\overline{H},\varepsilon}$, we prove the following result.

Theorem 9 *Under Assumption 1, for any algorithm $\boldsymbol{\pi}$, there exists an MDP $\mathcal{M}_{\boldsymbol{\pi}}$ whose transitions depend on the stage h, such that, for $T \geq HSA$*

$$\mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\pi}}) \geq \frac{1}{48\sqrt{6}}\sqrt{H^3SAT}\,.$$

Proof Consider the class of MDPs $\mathcal{C}_{\overline{H},\varepsilon}$ introduced in Section 3.1, with $\overline{H}$ and ε to be chosen later. This class contains a reference MDP $\mathcal{M}_0$ and MDPs of the form $\mathcal{M}_{(h^*,\ell^*,a^*)}$ parameterized by

$$(h^*, \ell^*, a^*) \in \{1+d, \dots, \overline{H}+d\} \times \mathcal{L} \times \mathcal{A}$$

in which

$$\Delta_{(h^*,\ell^*,a^*)}(h, s_i, a) \triangleq \mathbb{1}\{(h, s_i, a) = (h^*, s_{\ell^*}, a^*)\}\varepsilon.$$

As already mentioned, this family of MDPs behave like bandits, hence our proof follows the one for minimax lower bound in bandits (see, e.g., Bubeck and Cesa-Bianchi 2012).

Regret of $\boldsymbol{\pi}$ in $\boldsymbol{\mathcal{M}_{(h^*,\ell^*,a^*)}}$ The mean reward gathered by $\boldsymbol{\pi}$ in $\mathcal{M}_{(h^*,\ell^*,a^*)}$ is given by

$$\begin{aligned}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[\sum_{t=1}^{T}\sum_{h=1}^{H} r_h(S_h^t, A_h^t)\right] &= \sum_{t=1}^{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[\sum_{h=\overline{H}+d+1}^{H}\mathbb{1}\{S_h^t = s_g\}\right] \\ &= (H - \overline{H} - d)\sum_{t=1}^{T}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S^t_{\overline{H}+d+1} = s_g\right].\end{aligned}$$

For any $h \in \{1+d, \dots, \overline{H}+d\}$,

$$\begin{aligned}&\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{h+1}^t = s_g\right] \\ &= \mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_g\right] + \frac{1}{2}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] + \mathbb{1}\{h = h^*\}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_{\ell^*}, A_h^t = a^*\right]\varepsilon. \quad (7)\end{aligned}$$

Indeed, if $S_{h+1}^t = s_g$, we have either $S_h^t = s_g$ or $S_{h+1}^t \in \mathcal{L}$. In the latter case, the agent has $1/2$ probability of arriving at s_g, plus ε if the stage is h^*, the leaf is s_{ℓ^*} and the action is a^*.

Using the facts that $\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{1+d}^t = s_g\right] = 0$ because the agent needs first to traverse the tree and $\sum_{h=1+d}^{\overline{H}+d}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] = 1$ because the agent traverses the tree only once in one episode, we obtain from (7) that

$$\begin{aligned}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S^t_{\overline{H}+d+1} = s_g\right] &= \sum_{h=1+d}^{\overline{H}+d}\frac{1}{2}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] + \mathbb{1}\{h = h^*\}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_{\ell^*}, A_h^t = a^*\right]\varepsilon \\ &= \frac{1}{2} + \varepsilon\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{h^*}^t = s_{\ell^*}, A_{h^*}^t = a^*\right].\end{aligned}$$

Hence, the optimal value in any of the MDPs is $\rho^* = (H - \overline{H} - d)(1/2 + \varepsilon)$, which is obtained by the policy that starts to traverse the tree at step $h^* - d$ then chooses to go to the leaf s_{ℓ^*} and performs action a^*. The regret of $\boldsymbol{\pi}$ in $\mathcal{M}_{(h^*,\ell^*,a^*)}$ is then

$$\mathcal{R}_T\big(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}\big) = T(H - \overline{H} - d)\varepsilon\left(1 - \frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]\right)$$

where $N^T_{(h^*,\ell^*,a^*)} = \sum_{t=1}^{T}\mathbb{1}\{S_{h^*}^t = s_{\ell^*}, A_{h^*}^t = a^*\}$.

Maximum regret of π over all possible $\mathcal{M}_{(h^*,\ell^*,a^*)}$ We first lower bound the maximum of the regret by the mean over all instances

$$\begin{aligned}\max_{(h^*,\ell^*,a^*)} \mathcal{R}_T\big(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}\big) &\geq \frac{1}{\overline{H}LA} \sum_{(h^*,\ell^*,a^*)} \mathcal{R}_T\big(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}\big) \\ &\geq T(H-\overline{H}-d)\varepsilon\left(1 - \frac{1}{\overline{H}LAT} \sum_{(h^*,\ell^*,a^*)} \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]\right), \end{aligned} \tag{8}$$

so that, in order to lower bound the regret, we need an upper bound on the sum of $\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]$ over all MDP instances (h^*, ℓ^*, a^*). For this purpose, we will relate each expectation to the expectation of the same quantity under the reference MDP $\mathcal{M}_0$.

Upper bound on $\sum \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]$ Since $N^T_{(h^*,\ell^*,a^*)}/T \in [0,1]$, Lemma 6 gives us

$$\mathrm{kl}\left(\frac{1}{T}\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right], \frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]\right) \leq \mathrm{KL}\left(\mathbb{P}_0^{I_H^T}, \mathbb{P}_{(h^*,\ell^*,a^*)}^{I_H^T}\right).$$

By Pinsker's inequality, $(p-q)^2 \leq (1/2)\,\mathrm{kl}(p,q)$, it implies

$$\frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right] \leq \frac{1}{T}\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right] + \sqrt{\frac{1}{2}\,\mathrm{KL}\left(\mathbb{P}_0^{I_H^T}, \mathbb{P}_{(h^*,\ell^*,a^*)}^{I_H^T}\right)}$$

and, by Lemma 5, we know that

$$\mathrm{KL}\left(\mathbb{P}_0^{I_H^T}, \mathbb{P}_{(h^*,\ell^*,a^*)}^{I_H^T}\right) = \mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right] \mathrm{kl}(1/2, 1/2+\varepsilon)$$

since $\mathcal{M}_0$ and $\mathcal{M}_{(h^*,\ell^*,a^*)}$ only differ at stage h^* when $(s,a) = (s_{\ell^*}, a^*)$. Assuming that $\varepsilon \leq 1/4$, we have $\mathrm{kl}(1/2, 1/2+\varepsilon) \leq 4\varepsilon^2$ by Lemma 14, and, consequently

$$\frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right] \leq \frac{1}{T}\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right] + \sqrt{2}\varepsilon\sqrt{\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right]}. \tag{9}$$

The sum of $N^T_{(h^*,\ell^*,a^*)}$ over all instances $(h^*,\ell^*,a^*) \in \{1+d,\dots,\overline{H}+d\} \times \mathcal{L} \times \mathcal{A}$ is

$$\sum_{(h^*,\ell^*,a^*)} N^T_{(h^*,\ell^*,a^*)} = \sum_{t=1}^{T} \sum_{h^*=1+d}^{\overline{H}+d} \mathbb{1}\{S^t_{h^*} \in \mathcal{L}\} = T \tag{10}$$

since for a single stage $h^* \in \{1+d,\dots,\overline{H}+d\}$, we have $S^t_{h^*} \in \mathcal{L}$ almost surely.

Summing (9) over all instances (h^*,ℓ^*,a^*) and using (10), we obtain using the Cauchy-Schwartz inequality that

$$\begin{aligned}\frac{1}{T}\sum_{(h^*,\ell^*,a^*)} \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right] &\leq 1 + \sqrt{2}\varepsilon \sum_{(h^*,\ell^*,a^*)} \sqrt{\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right]} \\ &\leq 1 + \sqrt{2}\varepsilon\sqrt{\overline{H}LAT}\,. \end{aligned} \tag{11}$$

Optimizing ε and choosing $\overline{H}$ Plugging (11) in (8), we obtain

$$\max_{(h^*,\ell^*,a^*)} \mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}) \geq T(H - \overline{H} - d)\varepsilon\left(1 - \frac{1}{\overline{H}LA} - \frac{\sqrt{2}\varepsilon\sqrt{\overline{H}LAT}}{\overline{H}LA}\right).$$

The value of ε which maximizes the lower bound is $\varepsilon = \frac{1}{2\sqrt{2}}\left(1 - \frac{1}{\overline{H}LA}\right)\sqrt{\frac{\overline{H}LA}{T}}$ which yields

$$\max_{(h^*,\ell^*,a^*)} \mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}) \geq \frac{1}{4\sqrt{2}}\left(1 - \frac{1}{\overline{H}LA}\right)(H - \overline{H} - d)\sqrt{\overline{H}LAT}. \tag{12}$$

The number of leaves is $L = (1 - 1/A)(S - 3) + 1/A \geq S/4$, since $A \geq 2$ and $S \geq 6$. We choose $\overline{H} = H/3$ and use the assumptions that $A \geq 2$ and $d \leq H/3$ to obtain

$$\max_{(h^*,\ell^*,a^*)} \mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}) \geq \frac{1}{48\sqrt{6}} H\sqrt{HSAT}.$$

Finally, the assumption that $\varepsilon \leq 1/4$ is satisfied if $T \geq HSA$. ■

6. Discussion

The lower bounds presented in Theorems 7 and 9 hold for MDPs with stage-dependent transitions. As explained in Appendix C, their proof can be easily adapted to the case where the transitions $p_h(\cdot|s,a)$ do not depend on h and the bounds become $\Omega\left(\frac{SAH^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ and $\Omega(\sqrt{H^2SAT})$, respectively.

Our proofs require us to be able to build a full A-ary tree containing roughly S nodes whose depth d is small when compared to the horizon H, that is $d \leq H/3$ (Assumption 1). In Appendix D, we explain how to obtain the same bounds if we cannot build a full tree, and how the bounds become exponential in H if $d > H/3$.

Acknowledgments

The research presented was supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, French National Research Agency project BOLD (ANR19-CE23-0026-04) and the SFI Sachsen-Anhalt for the project RE-BCI.

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 2002.

Mohammad Gheshlaghi Azar, Ré Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

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

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 2012.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 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*, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2019.

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Sham Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

Tor Lattimore and Marcus Hutter. PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*, 2012.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 2004.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. *arXiv:2007.13442*, 2020.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 2009.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv:2007.03760*, 2020.

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*, 2019.

Appendix A. Change of Distribution: Proof of Lemma 5

The pushforward measure of $\mathbb{P}_{\mathcal{M}}$ under I_H^τ is given by

$$\forall T,\ \forall i_H^T \in \mathcal{I}_H^T, \quad \mathbb{P}_{\mathcal{M}}^{I_H^\tau}[i_H^T] = \mathbb{P}_{\mathcal{M}}^{I_H^T}[\tau = T, i_H^T] = \mathbb{P}_{\mathcal{M}}\left[\tau = T \middle| I_H^T = i_H^T\right] \mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T].$$

If $\mathbb{P}_{\mathcal{M}'}\left[\tau = T \middle| I_H^T = i_H^T\right] > 0$ and $\mathbb{P}_{\mathcal{M}'}^{I_H^T}[i_H^T] > 0$, we have

$$\frac{\mathbb{P}_{\mathcal{M}}^{I_H^\tau}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}^{I_H^\tau}[i_H^T]} = \frac{\mathbb{P}_{\mathcal{M}}\left[\tau = T \middle| I_H^T = i_H^T\right] \mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}\left[\tau = T \middle| I_H^T = i_H^T\right] \mathbb{P}_{\mathcal{M}'}^{I_H^T}[i_H^T]} = \frac{\mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}^{I_H^T}[i_H^T]}$$

where we use the fact that $\mathbb{P}_{\mathcal{M}}\left[\tau = T \middle| I_H^T = i_H^T\right] = \mathbb{P}_{\mathcal{M}'}\left[\tau = T \middle| I_H^T = i_H^T\right]$ since the event $\{\tau = T\}$ depends only on I_H^T. This implies that

$$\mathbb{P}_{\mathcal{M}}^{I_H^\tau}[i_H^T] \log\left(\frac{\mathbb{P}_{\mathcal{M}}^{I_H^\tau}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}^{I_H^\tau}[i_H^T]}\right) = \mathbb{P}_{\mathcal{M}}\left[\tau = T \middle| I_H^T = i_H^T\right] \mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T] \log\left(\frac{\mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}^{I_H^T}[i_H^T]}\right)$$

under the convention that $0 \log(0/0) = 0$. Hence,

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\mathcal{M}}^{I_H^\tau}, \mathbb{P}_{\mathcal{M}'}^{I_H^\tau}\right) &= \sum_{T=1}^{\infty} \sum_{i_H^T \in \mathcal{I}_H^T} \mathbb{P}_{\mathcal{M}}^{I_H^\tau}[i_H^T] \log\left(\frac{\mathbb{P}_{\mathcal{M}}^{I_H^\tau}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}^{I_H^\tau}[i_H^T]}\right) \\
&= \sum_{T=1}^{\infty} \sum_{i_H^T} \mathbb{P}_{\mathcal{M}}\left[\tau = T \middle| I_H^T = i_H^T\right] \mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T] \log\left(\frac{\mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T]}{\mathbb{P}_{\mathcal{M}'}^{I_H^T}[i_H^T]}\right) \\
&= \sum_{T=1}^{\infty} \sum_{i_H^T} \mathbb{P}_{\mathcal{M}}\left[\tau = T \middle| I_H^T = i_H^T\right] \mathbb{P}_{\mathcal{M}}^{I_H^T}[i_H^T] \sum_{t=1}^{T} \sum_{h=1}^{H-1} \log\left(\frac{p_h(s_{h+1}^t | s_h^t, a_h^t)}{p'_h(s_{h+1}^t | s_h^t, a_h^t)}\right) \\
&= \sum_{T=1}^{\infty} \mathbb{E}_{\mathcal{M}}\left[\mathbb{1}\{\tau = T\} \sum_{t=1}^{T} \sum_{h=1}^{H-1} \log\left(\frac{p_h(S_{h+1}^t | S_h^t, A_h^t)}{p'_h(S_{h+1}^t | S_h^t, A_h^t)}\right)\right] \\
&= \mathbb{E}_{\mathcal{M}}\left[\sum_{t=1}^{\tau} \sum_{h=1}^{H-1} \log\left(\frac{p_h(S_{h+1}^t | S_h^t, A_h^t)}{p'_h(S_{h+1}^t | S_h^t, A_h^t)}\right)\right].
\end{aligned}$$

Now, we apply Lemma 13 by taking $X_t = \sum_{h=1}^{H-1} \log\left(\frac{p_h(S_{h+1}^t | S_h^t, A_h^t)}{p'_h(S_{h+1}^t | S_h^t, A_h^t)}\right)$ and $\mathcal{F}_t = \mathcal{F}_H^t$. Notice that X_t is bounded almost surely, since when $p_h(S_{h+1}^t | S_h^t, A_h^t) = p'_h(S_{h+1}^t | S_h^t, A_h^t) = 0$, the trajectory containing $(S_h^t, A_h^t, S_{h+1}^t)$ has zero probability. Lemma 13 and the Markov property give us

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}_{\mathcal{M}}^{I_H^\tau}, \mathbb{P}_{\mathcal{M}'}^{I_H^\tau}\right) &= \mathbb{E}_{\mathcal{M}}\left[\sum_{t=1}^{\tau} \sum_{h=1}^{H-1} \mathbb{E}_{\mathcal{M}}\left[\log\left(\frac{p_h(S_{h+1}^t | S_h^t, A_h^t)}{p'_h(S_{h+1}^t | S_h^t, A_h^t)}\right) \middle| S_h^t, A_h^t\right]\right] \\
&= \mathbb{E}_{\mathcal{M}}\left[\sum_{t=1}^{\tau} \sum_{h=1}^{H-1} \mathrm{KL}\big(p_h(\cdot | S_h^t, A_h^t), p'_h(\cdot | S_h^t, A_h^t)\big)\right] = \sum_{s,a,h} \mathbb{E}_{\mathcal{M}}\big[N_{h,s,a}^\tau\big] \,\mathrm{KL}\big(p_h(\cdot | s, a), p'_h(\cdot | s, a)\big).
\end{aligned}$$

Appendix B. PAC-MDP Lower Bound: Proof of Corollary 8

Recall that $\mathcal{N}_\varepsilon^{\mathrm{PAC}} = \sum_{t=1}^{\infty} \mathbb{1}\{\rho^* - \rho^{\pi_t} > \varepsilon\}$ and let

$$T(\varepsilon, \delta) \triangleq \frac{1}{6912} \frac{H^3 SA}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) - 1.$$

We proceed by contradiction and assume that the claim in Corollary 8 is false. Then we have

$$\text{for all MDP } \mathcal{M}, \quad \mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}\left[\mathcal{N}_\varepsilon^{\mathrm{PAC}} \le T(\varepsilon,\delta)\right] \ge 1 - \delta. \tag{13}$$

that is, the algorithm satisfies Definition 2 with $F_{\mathrm{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta)) = T(\varepsilon, \delta)$. In particular, (13) holds for any MDP in the class $\mathcal{C}_{\overline{H},\widetilde{\varepsilon}}$ used to prove Theorem 7, for which $\overline{H} = H/3$ and $\widetilde{\varepsilon} = 2\varepsilon/(H - \overline{H} - d)$.

This allows us to build from $\boldsymbol{\pi}$ a best policy identification algorithm that outputs an ε-optimal policy with probability larger than $1 - \delta$ for every MDP in $\mathcal{C}_{H/3,\widetilde{\varepsilon}}$. We proceed as follows: the sampling rule is that of the algorithm $\boldsymbol{\pi}$ while the stopping rule is deterministic and set to $\tau \triangleq 2T(\varepsilon, \delta) + 1$. Letting $N_t(\pi)$ be the number of times that the algorithm plays a deterministic policy π up to episode t, we let the recommendation rule be $\widehat{\pi}_\tau = \arg\max_\pi N_\tau(\pi)$.

For every $\mathcal{M} \in \mathcal{C}_{H/3,\widetilde{\varepsilon}}$, the event $\left\{\mathcal{N}_\varepsilon^{\mathrm{PAC}} \le T(\varepsilon,\delta)\right\}$ implies $\widehat{\pi}_\tau = \pi^*$. This is trivial for $\mathcal{M}_0$, where any policy is optimal, and this holds for any other $\mathcal{M}_{(h^*,\ell^* a^*)} \in \mathcal{C}_{H/3,\widetilde{\varepsilon}}$ since there is a unique optimal policy π^* and it satisfies $(\rho^{\pi^*} - \rho^{\pi}) = 2\varepsilon > \varepsilon$ in $\mathcal{M}_{(h^*,\ell^* a^*)}$ for any other deterministic policy π. Hence, if $\widehat{\pi}_\tau \neq \pi^*$, the number of mistakes $\mathcal{N}_\varepsilon^{\mathrm{PAC}}$ would be larger than $T(\varepsilon, \delta)$. Thus we proved that the BPI algorithm that we defined satisfies

$$\forall \mathcal{M} \in \mathcal{C}_{H/3,\widetilde{\varepsilon}}, \quad \mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}[\widehat{\pi}_\tau = \pi^*] \ge \mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}\left[\mathcal{N}_\varepsilon^{\mathrm{PAC}} \le T(\varepsilon,\delta)\right] \ge 1 - \delta.$$

Under these conditions, we established in the proof of Theorem 7 that, for $\mathcal{M}_0 \in \mathcal{C}_{H/3,\widetilde{\varepsilon}}$,

$$\tau = \mathbb{E}_{\mathcal{M}_0}[\tau] \ge \frac{1}{3456} \frac{H^3 SA}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)$$

which yields

$$2T(\varepsilon, \delta) + 1 \ge \frac{1}{3456} \frac{H^3 SA}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)$$

and contradicts the definition of $T(\varepsilon, \delta)$.

Appendix C. Recovering the lower bounds for stage-independent transitions

The proofs of Theorem 7 and Theorem 9 can be adapted to the case where the transitions $p_h(\cdot|s, a)$ do not depend on h. To do so, we need to have a set of hard MDPs with stage-independent transitions. For that, we remove the waiting state s_{w} and the agent starts at s_{root}, which roughly corresponds to setting $\overline{H} = 1$ in the proofs, and we take

$$\Delta_{(h^*,\ell^*,a^*)}(h, s_i, a) \triangleq \mathbb{1}\{(s_i, a) = (s_{\ell^*}, a^*)\}\varepsilon'$$

to be independent of h. We also take h-independent rewards as

$$\forall a \in \mathcal{A}, \quad r_h(s, a) = \mathbb{1}\{s = s_g\}.$$

Since $\overline{H} = 1$ and no longer $H/3$, the regret bound becomes $\Omega(\sqrt{H^2 SAT})$ and the BPI bound becomes $\Omega\left(\frac{SAH^2}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$.

Appendix D. Relaxing Assumption 1

In the proofs of Theorems 7 and 9, we use Assumption 1 stating that

(i) there exists an integer d such that $S = 3 + (A^d - 1)/(A - 1)$, and

(ii) $H \geq 3d$,

which we discuss below.

D.1. Relaxing (i)

Assumption (i) makes the proof simpler by allowing us to consider a *full* A-ary tree with $S - 3$ nodes, which implies that all the leaves are at the same level $d - 1$ in the tree. The proof can be generalized to any $S \geq 6$ by arranging the states in a balanced, but not necessarily full, A-ary tree. In this case, there might be subset of the leaves at a level $d - 1$ and another subset at a level $d - 2$, which creates an asymmetry in the leaf nodes. To handle this, we proceed as follows:

- First, using $(S - 3)/2$ states, we build a balanced A-ary tree of depth $d - 1$;
- For each leaf at depth $d - 2$, we add another state (taken among the remaining $(S - 3)/2$) as its child.
- Any remaining state that was not added to the tree (and is not s_{w}, s_g or s_b), can be merged to the absorbing states s_g or s_b.

This construction ensures that we have a tree with at least $(S - 3)/2$ and at most $(S - 3)$ nodes, where all the leaves are at the same depth $d - 1$, for

$$d = \lceil \log_A((S - 3)(A - 1) + 1) \rceil \in [\log_A S - 1, \log_A S + 2]. \tag{14}$$

Lemma 16 shows that the number of leaves L in this tree satisfies $S \geq L \geq (S - 3)/8$. Hence, in the proofs of Theorem 7 (Eq. 6) and Theorem 9 (Eq. 12) , we take $L \geq (S - 3)/8$ and obtain lower bounds of the same order.

D.2. Relaxing (ii)

Equation (14) implies that there exists a constant $c \in [-1, 2]$ such that $d = \log_A S + c$. Assumption (ii), stating that $H \geq 3d = 3 \log_A S + 3c$ ensures that the horizon is large enough with respect to the size of the MDP for the agent to be able to traverse the tree down to the rewarding state. If this condition is not satisfied, that is, if $H < 3 \log_A S + 3c$, we have $S \geq A^{\frac{H}{3} - 2}$. In this case, we can build a tree using a subset of the state space containing $\left\lceil A^{\frac{H}{3} - 2} \right\rceil$ states, and merge the remaining $S - \left\lceil A^{\frac{H}{3} - 2} \right\rceil$ states to the absorbing states s_b or s_g. In this case, the resulting bounds will replace S by $\left\lceil A^{\frac{H}{3} - 2} \right\rceil$, and become exponential in the horizon H,

$$\Omega\left(\frac{\left\lceil A^{\frac{H}{3} - 2} \right\rceil A H^3}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right) \quad \text{and} \quad \Omega\left(\sqrt{H^3 \left\lceil A^{\frac{H}{3} - 2} \right\rceil T} \right)$$

for BPI and regret, respectively.

The arguments above give us Theorem 10, Corollary 11 and Theorem 12 below, which state BPI, PAC-MDP and regret lower bounds, respectively, without requiring Assumption 1.

Theorem 10 *Let $(\boldsymbol{\pi}, \tau, \widehat{\pi}_\tau)$ be an algorithm that is (ε, δ)-PAC for best policy identification in any finite episodic MDP. Then, if $S \geq 11$, $A \geq 4$ and $H \geq 6$, there exists an MDP $\mathcal{M}$ with stage-dependent transitions such that for $\varepsilon \leq H/24$ and $\delta \leq 1/16$,*

$$\mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}[\tau] \geq c_1 \min\left(S, A^{\frac{H}{3}-2}\right) \frac{H^3 A}{\varepsilon^2} \log\left(\frac{1}{\delta}\right).$$

where c_1 is an absolute constant.

Proof If $S \leq A^{\frac{H}{3}-2}$, then $H \geq 3d$, where d is given in Equation 14. In this case, we follow the proof of Theorem 7 up to Equation 6, where we take $L \geq (S-3)/8$ according to the arguments in Section D.1. If $S > A^{\frac{H}{3}-2}$, then $H < 3d$ and we follow the arguments in Section D.2. ■

Corollary 11 *Let $\boldsymbol{\pi}$ be an algorithm that is (ε, δ)-PAC for exploration according to Definition 2 and that, in each episode t, plays a deterministic policy π_t. Then, under the conditions of Theorem 10, there exists an MDP $\mathcal{M}$ such that*

$$\mathbb{P}_{\boldsymbol{\pi},\mathcal{M}}\left[\mathcal{N}_\varepsilon^{\mathrm{PAC}} > c_2 \min\left(S, A^{\frac{H}{3}-2}\right) \frac{H^3 A}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) - 1\right] > \delta.$$

where c_2 is an absolute constant.

Proof Analogous to the proof of Corollary 8, using Theorem 10 instead of Theorem 7. ■

Theorem 12 *If $S \geq 11$, $A \geq 4$ and $H \geq 6$, for any algorithm $\boldsymbol{\pi}$, there exists an MDP $\mathcal{M}_{\boldsymbol{\pi}}$ whose transitions depend on the stage h, such that, for $T \geq HSA$*

$$\mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\pi}}) \geq c_3 \sqrt{\min\left(S, A^{\frac{H}{3}-2}\right)} \sqrt{H^3 A T}\,.$$

where c_3 is an absolute constant.

Proof If $S \leq A^{\frac{H}{3}-2}$, then $H \geq 3d$, where d is given in Equation 14. In this case, we follow the proof of Theorem 9 up to Equation 12, where we take $L \geq (S-3)/8$ according to the arguments in Section D.1. If $S > A^{\frac{H}{3}-2}$, then $H < 3d$ and we follow the arguments in Section D.2. ■

Appendix E. Technical Lemmas

Lemma 13 *Let $(X_t)_{t\geq 1}$ be a stochastic process adapted to the filtration $(\mathcal{F}_t)_{t\geq 1}$. Let τ be a stopping time with respect to $(\mathcal{F}_t)_{t\geq 1}$ such that $\tau < \infty$ with probability 1. If there exists a constant*

c such that $\sup_t |X_t| \le c$ *almost surely, then*

$$\mathbb{E}\left[\sum_{t=1}^{\tau} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{E}[X_t|\mathcal{F}_{t-1}]\right].$$

Proof Let $M_n \triangleq \sum_{t=1}^{n}(X_t - \mathbb{E}[X_t|\mathcal{F}_{t-1}])$. Then, M_n is a martingale and, by Doob's optional stopping theorem, $\mathbb{E}[M_\tau] = \mathbb{E}[M_0] = 0$. ■

Lemma 14 *If* $\varepsilon \in [0, 1/4]$, *then* $\mathrm{kl}(1/2, 1/2+\varepsilon) \le 4\varepsilon^2$.

Proof Using the inequality $-\log(1-x) \le 1/(1-x) - 1$ for any $0 < x < 1$, we obtain

$$\mathrm{kl}(1/2, 1/2+\varepsilon) = -\frac{1}{2}\log(1-4\varepsilon^2) \le \frac{1}{2}\left(\frac{1}{1-4\varepsilon^2} - 1\right) = \frac{2\varepsilon^2}{1-4\varepsilon^2}.$$

If $\varepsilon \le 1/4$, then $1 - 4\varepsilon^2 \ge 3/4 > 1/2$, which implies the result. ■

Lemma 15 *For any* $p, q \in [0, 1]$,

$$\mathrm{kl}(p, q) \ge (1-p)\log\left(\frac{1}{1-q}\right) - \log(2).$$

Proof If follows from the definition of $\mathrm{kl}(p, q)$ and the fact that the entropy $H(p) \triangleq p\log(1/p) + (1-p)\log(1/(1-p))$ satisfies $H(p) \le \log(2)$:

$$\begin{aligned}
\mathrm{kl}(p, q) &= p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \\
&= (1-p)\log\left(\frac{1}{1-q}\right) + (1-p)\log\left(\frac{1}{1-q}\right) + p\log\left(\frac{1}{q}\right) - H(p) \\
&\ge (1-p)\log\left(\frac{1}{1-q}\right) - \log(2).
\end{aligned}$$

■

Lemma 16 *Let L be the number of leaves in a balanced A-ary tree with S nodes and $A \ge 2$. Then, $L \ge S/4$.*

Proof Let d be the depth of the tree. There exists an integer R such that $0 < R \le A^d$ such that

$$S = \frac{A^d - 1}{A - 1} + R.$$

The number of leaves is given by $L = R + A^{d-1} - \left\lceil \frac{R}{A} \right\rceil$. We consider two cases: either $\frac{A^d-1}{A-1} \le \frac{S}{2}$ or $\frac{A^d-1}{A-1} > \frac{S}{2}$. If $\frac{A^d-1}{A-1} \le \frac{S}{2}$, we have $R \ge S/2$ which implies $L \ge S/2 > S/4$. If $\frac{A^d-1}{A-1} > \frac{S}{2}$, we have $L \ge A^{d-1} > \frac{1}{A} + \frac{S}{2}\left(1 - \frac{1}{A}\right) \ge S/4$. ■