Towards Instance-Optimality in Online PAC Reinforcement Learning

Aymen Al-Marjani, Andrea Tirinzoni, Emilie Kaufmann

Towards Instance-Optimality in Online PAC Reinforcement Learning

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

[1]UMPA, ENS Lyon, Lyon, France
[2]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, Lille, France

Abstract

Several recent works have proposed instance-dependent upper bounds on the number of episodes needed to identify, with probability $1-\delta$, an ε-optimal policy in finite-horizon tabular Markov Decision Processes (MDPs). These upper bounds feature various complexity measures for the MDP, which are defined based on different notions of sub-optimality gaps. However, as of now, no lower bound has been established to assess the optimality of any of these complexity measures, except for the special case of MDPs with deterministic transitions. In this paper, we propose the first instance-dependent lower bound on the sample complexity required for the PAC identification of a near-optimal policy in any tabular episodic MDP. Additionally, we demonstrate that the sample complexity of the PEDEL algorithm of Wagenmaker and Jamieson (2022) closely approaches this lower bound. Considering the intractability of PEDEL, we formulate an open question regarding the possibility of achieving our lower bound using a computationally-efficient algorithm.

1 Introduction

We consider the online Probably Approximately Correct Reinforcement Learning (PAC RL) problem, in which an agent sequentially interacts with an environment modeled as a Markov Decision Process (MDP), with the goal of learning a near-optimal policy as quickly as possible. More precisely, given a precision $\varepsilon \geq 0$ and a risk parameter $\delta \in (0,1)$, the agent is required to return a policy $\widehat{\pi}$ whose value is within ε of the value of the optimal policy, with probability at least $1-\delta$. The agent's performance is evaluated through its *sample complexity*, defined as the number of interactions with the environment needed to output such a policy $\widehat{\pi}$.

Since its introduction by Fiechter (1994), this problem has been extensively investigated from a *minimax* point of view in two different settings: discounted MDPs (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020), in which the value of a policy is the expected (infinite) sum of rewards discounted by a factor $\gamma \in (0,1)$, and finite-horizon (or episodic) MDPs (Dann and Brunskill, 2015; Dann et al., 2019; Kaufmann et al., 2021; Ménard et al., 2021), in which the value is the expected sum of rewards up to a given horizon H. Notably, in the finite-horizon setting with S states, A actions, and horizon H, Dann and Brunskill (2015) proved that any PAC RL agent must play at least $\Omega(SAH^2 \log(1/\delta)/\varepsilon^2)$ episodes to identify an ε-optimal policy in the *worst-case*. Their lower bound was derived under the assumption of time-homogeneous rewards and transitions, while a lower bound of $\Omega(SAH^3 \log(1/\delta)/\varepsilon^2)$ episodes was later derived by Domingues et al. (2021) for the time-inhomogeneous case. There exist algorithms with sample complexity matching these lower bounds (Dann et al., 2019; Ménard et al., 2021).

Unfortunately, minimax optimality is not informative about the performance of an algorithm under different MDPs of the same size (H,S,A). For instance, let us imagine a first MDP with deterministic transitions and a tree structure which has a single optimal trajectory whose rewards are all considerably

higher than the rewards in any other trajectory. Let us also consider a second MDP in which all actions yield exactly the same reward, but this information is unknown to the agent beforehand. One would naturally expect the PAC RL task to be much easier in the first MDP where a few episodes should suffice to detect that the policy following the good trajectory is optimal. In the second one, however, no reasonable algorithm can *confidently state* that a policy is ε-optimal before having estimated uniformly well (with ε-precision) the value of all other policies.

This motivates a recent line of works focused on designing adaptive algorithms with *instance-dependent* guarantees, i.e., sample complexity bounds featuring some characteristics of the underlying MDP that go beyond its size as in minimax results. These characteristics have been expressed with different notions of *sub-optimality gaps*. The first algorithm of this kind is BESPOKE (Zanette et al., 2019), which was proposed for discounted MDPs. Its gap-based sample complexity is shown to be never worse than the minimax rate, while it can be significantly smaller in some MDPs. Taupin et al. (2022) later proposed GSS, a PAC RL algorithm for discounted linear MDPs (Jin et al., 2019) along with GSS-E, its counterpart for episodic linear MDPs. The problem of exact identification of the optimal policy ($\varepsilon = 0$) and the more complex identification of a Blackwell-optimal policy were treated by Marjani and Proutiere (2021) and Boone and Gaujal (2023), respectively. All these works assume that a generative model is available, i.e., that the learner can query a transition from any state at any time. In the more challenging setting where interaction is allowed only through trajectories, Al Marjani et al. (2021) studied exact best-policy identification in discounted MDPs, while a more recent line of works has considered approximate identification ($\varepsilon \geq 0$) in episodic MDPs (Wagenmaker et al., 2022a; Tirinzoni et al., 2022; Wagenmaker and Jamieson, 2022; Tirinzoni et al., 2023; Al-Marjani et al., 2023). All the proposed algorithms have sample complexity upper bounds of the form $\widetilde{\mathcal{O}}\big(\mathcal{C}(\mathcal{M}, \varepsilon) \log(1/\delta)\big)$, where $\mathcal{C}$ quantifies the hardness of learning an ε-optimal policy in the MDP $\mathcal{M}$, δ is the risk, while $\widetilde{\mathcal{O}}$ hides numerical constants and logarithmic factors of the relevant parameters. The expression of $\mathcal{C}$ is different for each algorithm but always dependent on some sub-optimality gaps (either values gaps or policy gaps, see Section 2 for a formal definition) and on state-visitation probabilities. We review these bounds in Section 4.

However, the lack of a general instance-dependent lower bound makes it difficult to assess the optimality of these approaches, i.e., how tight a complexity $\mathcal{C}(\mathcal{M}, \varepsilon)$ is compared to the best possible rate. Indeed, the only instance-dependent lower bounds for PAC RL without a generative model are either restricted to MDPs with deterministic transitions (Tirinzoni et al., 2022) or cover only the case of exact best-policy identification ($\varepsilon = 0$) under the assumption that the optimal policy is unique (Al Marjani et al., 2021). In this work, we fill this gap by answering the following question:

What is the best rate in $\log(1/\delta)$ *that a PAC RL algorithm can achieve on an episodic tabular MDP?*

Contributions We derive the first instance-dependent lower bound for PAC RL that holds for any $\varepsilon \geq 0$ and any tabular MDP (Theorem 1). As for bandit identification problems with many correct answers (Degenne and Koolen, 2019), our lower bound holds when $\delta \to 0$. Beyond the asymptotic regime, we strengthen this result with an additional lower bound that holds for all $\delta > 0$ in the special case of $\varepsilon = 0$ under the assumption that optimal policies share a unique state-action distribution (Theorem 2). Then, in Section 4, we review the complexity measures featured in existing upper bounds and show that the PEDEL algorithm of Wagenmaker and Jamieson (2022) matches our lower bound in tabular MDPs up to multiplicative H factors and an additive $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ term (Proposition 1). A shortcoming of PEDEL is that it is not computationally efficient as it explicitly enumerates all policies. We thus formulate an open question as to whether our bound can be attained by a computationally-efficient algorithm.

2 Preliminaries

We consider tabular finite-horizon *Markov decision processes* (MDPs). Formally, an MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{\nu_h\}_{h\in[H]}, s_1)$, where $\mathcal{S}$ is a finite set of S states, $\mathcal{A}$ is a finite set of A actions, H is the horizon, $p_h : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$ and $\nu_h : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathbb{R})$[1] respectively denote the transition kernel and the reward distribution at stage $h \in [H]$, while $s_1 \in \mathcal{S}$ is the initial state[2]. A learner interacts with $\mathcal{M}$ through episodes of length H. At the beginning of each episode, the learner starts in the initial state s_1. Then, for each stage $h \in [H]$, the learner plays an action $a_h \in \mathcal{A}$ and observes a stochastic transition to a new state $s_{h+1} \sim p_h(s_h, a_h)$ as well as a reward $R_h \sim \nu_h(s_h, a_h)$. The actions are usually chosen according to a Markovian (possibly stochastic) policy $\pi = \{\pi_h\}_{h\in[H]}$, i.e., a sequence of mappings $\pi_h : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, where $\pi_h(a|s)$ denotes the probability that the learner takes action a in state s at stage h. We denote by Π^{S} (resp. Π^{D}) the set of all Markovian stochastic (resp. deterministic) policies.

2.1 Policy gaps, value gaps, and state-action distributions

Denoting by $\mathbb{P}^\pi$ (resp. $\mathbb{E}^\pi$) the probability (resp. expectation) operator induced by the execution of a policy $\pi \in \Pi^{\mathrm{S}}$ for an episode on $\mathcal{M}$, we let $V_1^\pi := \mathbb{E}^\pi\big[\sum_{h=1}^H R_h \big| s_1\big]$ be the value of π at the initial state[3]. The *policy gap* of π is then defined as

$$\Delta(\pi) := V_1^\star - V_1^\pi,$$

where $V_1^\star := \max_{\pi\in\Pi^{\mathrm{D}}} V_1^\pi$ is the optimal value at s_1. We use $Q_h^\pi(s,a) := \mathbb{E}^\pi\big[\sum_{\ell=h}^H R_\ell \big| s_h = s, a_h = a\big]$ and $Q_h^\star(s,a) := \max_{\pi\in\Pi^{\mathrm{D}}} Q_h^\pi(s,a)$ to denote the action-value function of π and the optimal value function, respectively. The *value gap* of the triplet (h,s,a) is then defined as

$$\Delta_h(s,a) := \max_{b\in\mathcal{A}} Q_h^\star(s,b) - Q_h^\star(s,a)\ .$$

Moreover, we denote the visitation probability of (h,s,a) under π as $p_h^\pi(s,a) := \mathbb{P}^\pi(s_h = s, a_h = a)$ and $p_h^\pi(s) := \mathbb{P}^\pi(s_h = s)$. We let $\Omega := \big\{\big(p_h^\pi(s,a)\big)_{h,s,a} : \pi \in \Pi^{\mathrm{S}}\big\}$ denote the set of all valid state-action distributions. It is well known (e.g., Puterman, 1994) that Ω is a polytope defined by the linear constraints

$$\begin{aligned}
\forall \rho \in \Omega,\quad & \rho_h(s,a) \geq 0\ \forall (h,s,a),\\
& \sum_{a\in\mathcal{A}} \rho_1(s_1,a) = 1,\ \sum_{a\in\mathcal{A}} \rho_1(s,a) = 0\ \forall s \neq s_1,\\
& \sum_{a\in\mathcal{A}} \rho_h(s,a) = \sum_{s',a'} \rho_{h-1}(s',a') p_{h-1}(s|s',a')\ \forall (s,h) \in \mathcal{S}\times[|2,H|].
\end{aligned}$$

2.2 Learning problem

The learner interacts with an MDP $\mathcal{M}$ with unknown transition probabilities and reward distributions. Given a risk parameter $\delta \in (0,1)$ and a precision $\varepsilon \geq 0$ as input, the goal is to return a policy $\widehat{\pi} \in \Pi^{\mathrm{D}}$ with the guarantee that $\mathbb{P}_{\mathcal{M}}(\Delta(\widehat{\pi}) \leq \varepsilon) \geq 1-\delta$. To satisfy this requirement, the learner needs to gather samples from the transition and reward distributions of $\mathcal{M}$ by playing episodes in a sequential fashion. In each episode $t \in \mathbb{N}^*$, the learner selects a policy π^t (based on past observations) and collects a new trajectory $\mathcal{H}_t := \{(s_h^t, a_h^t, R_h^t)\}_{h\in[H]}$ under this policy, where $a_h^t \sim \pi^t(s_h^t)$. We let $\mathcal{F}_t := \sigma((\mathcal{H}_u)_{1\leq u\leq t})$ denote the sigma-algebra generated by trajectories up to episode t. The learner's performance is then evaluated by its *sample complexity* τ, which is a stopping time w.r.t, the filtration $(\mathcal{F}_t)_{t\geq 1}$ counting the (random) number of exploration episodes before termination.

[1] $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures over a set $\mathcal{X}$.
[2] This setting encompasses any initial state distribution by adding a transition from s_1 with the desired probabilities.
[3] Since the initial state s_1 is fixed, we drop it from the notation of value functions.

Definition 1 ((ε, δ)-PAC algorithm)**.** *Let $\mathfrak{M}$ be a set of MDPs. An algorithm is (ε, δ)-PAC on $\mathfrak{M}$ if for all MDPs $\mathcal{M} \in \mathfrak{M}$, with probability at least $1 - \delta$, it stops after playing $\tau < \infty$ episodes on $\mathcal{M}$ and returns a deterministic policy $\widehat{\pi} \in \Pi^{\mathrm{D}}$ satisfying $\Delta(\widehat{\pi}) \leq \varepsilon$.*

3 Lower Bounds

We consider the class $\mathfrak{M}_1$ of stochastic MDPs with *Gaussian rewards* of unit variance[4], in which $\nu_h(s,a) = \mathcal{N}(r_h(s,a), 1)$. While existing *upper* bounds commonly work under the stronger assumption that rewards lie in $[0,1]$ almost surely, we focus on this alternative setting since it has enabled the derivation of *closed-form lower bounds* that scale with intuitive quantities such as policy gaps (Dann et al., 2021; Tirinzoni et al., 2022). Moreover, the complexity of an MDP is mostly characterized by the expected rewards $r_h(s,a)$ rather than the full distributions $\nu_h(s,a)$, so that matching a lower bound for Gaussians while observing bounded rewards with the same mean is still very informative about the algorithm's adaptivity to the underlying problem.

3.1 General lower bound for approximate identification

Our first result is a general bound that holds for any $\varepsilon \geq 0$ in the asymptotic regime $\delta \to 0$. We use the notation $\Pi^{\varepsilon} := \{\pi \in \Pi^{D} : V_1^{\pi}(s_1) \geq V_1^{\star}(s_1) - \varepsilon\}$ for the set of all deterministic ε-optimal policies.

Theorem 1. *Any PAC RL algorithm that is (ε, δ)-PAC on $\mathfrak{M}_1$ satisfies, for any $\mathcal{M} \in \mathfrak{M}_1$,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau]}{\log(1/\delta)} \geq \mathcal{C}_{\mathrm{LB}}(\mathcal{M}, \varepsilon)$$

where

$$\mathcal{C}_{\mathrm{LB}}(\mathcal{M}, \varepsilon) := 2 \min_{\pi^{\varepsilon} \in \Pi^{\varepsilon}} \min_{\rho \in \Omega} \max_{\pi \in \Pi^{D}} \sum_{s,a,h} \frac{\left(p_h^{\pi}(s,a) - p_h^{\pi^{\varepsilon}}(s,a)\right)^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^{\varepsilon}) + \varepsilon)^2}.$$

Theorem 1 states that no matter how adaptive a PAC RL algorithm is, there is a minimal cost in terms of episodes that it must pay in order to learn an ε-optimal policy of $\mathcal{M}$. This cost is instance-dependent since it is a functional of $\mathcal{M}$, the MDP to be learned. The proof of Theorem 1 is deferred to Appendix A. It follows similar steps as the proof of the lower bound for ε-best arm identification (and other pure exploration problems) of Degenne and Koolen (2019).

3.2 Finite-δ bound for exact identification

In the case of exact identification (i.e. $\varepsilon = 0$), we further derive a lower bound which is valid for any $\delta \in (0,1)$ under the assumption that the optimal state-action distribution is unique. In particular, we assume that there exists $p^{\star} \in \Omega$ s.t. for any optimal policy $\pi^{\star}$ (i.e., with $V_1^{\pi^{\star}} = V_1^{\star}$) we have $p^{\pi^{\star}} = p^{\star}$. Note that this is a generalization of the assumption of "unique optimal trajectory" from Tirinzoni et al. (2022), under which we know that exact identification to be possible with a sample complexity that does not scale with ε. It is also the same assumption considered in Tirinzoni et al. (2021). As shown in that paper, it implies that there is a unique optimal action in states visited with positive probability by some optimal policy, but there can be arbitrary many optimal actions in all other states.[5]

[4] We trivially get results for Gaussian rewards with arbitrary variance σ^2 by multiplying our lower bounds by σ^2.

[5] Without a unique optimal state-action distribution, exact identification may not be even possible, as no algorithm may be able to stop in finite time and return an optimal policy w.h.p. while being $(0, \delta)$-PAC on the whole family $\mathfrak{M}_1$.

Theorem 2. *Fix any MDP $\mathcal{M} \in \mathfrak{M}_1$ s.t. the optimal state-action distribution $p^\star$ is unique. Then, for any PAC RL algorithm that is $(0,\delta)$-PAC on $\mathfrak{M}_1$,*

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq 2 \min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}: \Delta(\pi) > 0} \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\rho_h(s,a)\Delta(\pi)^2} \log\left(\frac{1}{2.4\delta}\right).$$

Remark 1. *When $S = H = 1$, this bound exactly coincides with the lower bound for best-arm identification in Gaussian multi-armed bandits (Garivier and Kaufmann, 2016).*

Proof. The idea of the proof is to explicitly compute the smallest KL divergence between the distribution of the observations under the MDP $\mathcal{M}$ and under any alternative $\widetilde{\mathcal{M}}$ that has the same transitions but a different mean reward function $\widetilde{r}_h$. Within the class $\mathfrak{M}_1$, the KL divergence of observations between $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ takes the simple form

$$\mathrm{KL}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = \sum_{h,s,a} \mathbb{E}_{\mathcal{M}}[n_h^\tau(s,a)] \frac{\big(r_h(s,a) - \widetilde{r}_h(s,a)\big)^2}{2}.$$

Note that, since $p^\star$ is unique, any $(0,\delta)$-PAC algorithm satisfies $\mathbb{P}_{\mathcal{M}}(V_1^{\widehat{\pi}} = V_1^\star) = \mathbb{P}_{\mathcal{M}}(p^{\widehat{\pi}} = p^\star) \geq 1-\delta$. Now fix a sub-optimal policy π for $\mathcal{M}$ (i.e., with $\Delta(\pi) > 0$). Note that $V_1^\pi = r^T p^\pi < V_1^\star = r^T p^\star$. We look for the closest alternative $\widetilde{\mathcal{M}}$ such that $\widetilde{r}^T p^\pi > \widetilde{r}^T p^\star$, i.e., where π becomes better than any optimal policy of $\mathcal{M}$. This can be computed by the quadratic program

$$\min_{\widetilde{r}: \widetilde{r}^T p^\pi > \widetilde{r}^T p^\star} \sum_{s,a,h} \mathbb{E}[n_h^\tau(s,a)] \frac{(r_h(s,a) - \widetilde{r}_h(s,a))^2}{2} = \frac{\Delta(\pi)^2}{2 \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\mathbb{E}[n_h^\tau(s,a)]}}.$$

By the $(0,\delta)$-PAC property, in such closest alternative we have $\mathbb{P}_{\widetilde{\mathcal{M}}}(p^{\widehat{\pi}} = p^\star) \leq \delta$. Then, Lemma 1 of Kaufmann et al. (2016) ensures that $\mathrm{KL}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) \geq \log\left(\frac{1}{2.4\delta}\right)$. Thus, for any π with $\Delta(\pi) > 0$,

$$1 \geq 2 \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\mathbb{E}[n_h^\tau(s,a)]\Delta(\pi)^2} \log\left(\frac{1}{2.4\delta}\right).$$

Multiplying both sides by $\mathbb{E}[\tau]$ and maximizing over sub-optimal policies, we obtain

$$\mathbb{E}[\tau] \geq 2 \max_{\pi \in \Pi^{\mathrm{D}}: \Delta(\pi) > 0} \sum_{s,a,h} \frac{\mathbb{E}[\tau]}{\mathbb{E}[n_h^\tau(s,a)]} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\Delta(\pi)^2} \log\left(\frac{1}{2.4\delta}\right).$$

Now it is easy to see that $\rho_h(s,a) := \mathbb{E}[n_h^\tau(s,a)]/\mathbb{E}[\tau]$ is a valid state-action distribution (i.e., $\rho \in \Omega$). Thus, minimizing the right-hand side over all $\rho \in \Omega$ concludes the proof. □

3.3 Interpreting the lower bound

While the expression of the lower bound might seem mysterious at a first glance, we provide an interpretation in terms of confidence intervals for the simpler setting of known transitions and unknown rewards. Our explanation hinges on the following concentration inequality, proved in Appendix B.

Lemma 1. *Assume the reward distribution $\nu_h(s,a)$ to be 1-subgaussian[6] with mean $r_h(s,a)$ for all (h,s,a). For any policy $\pi \in \Pi^{\mathrm{D}}$, define the estimator $\widehat{V}_1^{\pi,t} := \sum_{h,s,a} p_h^\pi(s,a)\widehat{r}_h^t(s,a)$, where $\widehat{r}_h^t(s,a)$ is the MLE of $r_h(s,a)$ using samples gathered until episode t. We have that*

$$\mathbb{P}\left(\forall t \geq t_0,\ \forall \pi, \pi' \in \Pi^{\mathrm{D}},\ \left|(\widehat{V}_1^{\pi,t} - \widehat{V}_1^{\pi',t}) - (V_1^\pi - V_1^{\pi'})\right| \leq \sqrt{\beta(t,\delta) \sum_{h,s,a} \frac{\big(p_h^\pi(s,a) - p_h^{\pi'}(s,a)\big)^2}{n_h^t(s,a)}}\right) \geq 1-\delta,$$

with $t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \text{ s.t. } \sup_\pi p_h^\pi(s) > 0\}$, and $\beta(t,\delta) := 4\log(1/\delta) + 12SH\log(A(1+t))$.

[6] A random variable X is σ^2-subgaussian if $\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\sigma^2\lambda^2/2}$ for any $\lambda \in \mathbb{R}$.

Imagine that a learner explores the MDP $\mathcal{M}$ using a fixed (stochastic) policy π^{exp}, whose state-action distribution is ρ^{exp}, and wants to figure out whether some policy π^ε is ε-optimal or not. Then, after playing π^{exp} for $K \geq 1$ episodes, $\mathbb{E}[n_h^K(s,a)] = K\rho_h^{\text{exp}}(s,a)$, so that the size of the confidence interval on $V_1^{\pi^\varepsilon} - V_1^\pi$ should roughly be $\sqrt{\beta(K,\delta)\sum_{h,s,a} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{K\rho_h^{\text{exp}}(s,a)}}$. Now, if the learner wishes to test whether π^ε is ε-optimal it has to determine the sign of $V_1^{\pi^\varepsilon} - V_1^\pi + \varepsilon$ for all other policies π. To do that, it is sufficient to shrink the size of the confidence interval on $V_1^{\pi^\varepsilon} - V_1^\pi$ below $\frac{1}{2}|V_1^{\pi^\varepsilon} - V_1^\pi + \varepsilon| = \frac{1}{2}|\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon|$ for all policies π. Solving for the minimal K that satisfies this condition, we see that playing roughly

$$K(\pi^{\text{exp}}, \pi^\varepsilon) \propto \log(1/\delta) \max_{\pi \in \Pi^D} \sum_{s,a,h} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{\rho_h^{\text{exp}}(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2}$$

episodes using the exploration policy π^{exp} is enough to determine whether π^ε is ε-optimal. Since the learner has the liberty to return *any* ε-optimal policy using *any* exploration policy, the lower bound corresponds to the minimum of $K(\pi^{\text{exp}}, \pi^\varepsilon)$ with respect to these two variables.

4 Towards a matching upper bound

4.1 Review of existing upper bounds

In this section, we review the main instance-dependent bounds within the PAC RL literature. We restrict our review to works on approximate identification (i.e., the general case with $\varepsilon \geq 0$).

PAC RL with a generative model Zanette et al. (2019) were the first to propose an instance-dependent PAC RL algorithm, called BESPOKE. In infinite-horizon tabular MDPs with a discount factor $\gamma \in [0,1)$ and when the agent has access to a simulator that can query observations from any state-action pair, BESPOKE finds an ε-optimal policy with a sample complexity of at most

$$\widetilde{\mathcal{O}}\left(\left[\sum_{s,a} \min\left(\frac{1}{(1-\gamma)^3\varepsilon^2}, \frac{\text{Var}[R(s,a)] + \gamma^2 \text{Var}_{s' \sim p(.|s,a)}[V^\star(s')]}{\max(\Delta_{sa}, (1-\gamma)\varepsilon)^2} + \frac{1}{(1-\gamma)\max(\Delta_{sa}, (1-\gamma)\varepsilon)} \right) \right] \log\left(\frac{1}{\delta}\right) \right),$$

where $\Delta_{sa} = V^\star(s) - Q^\star(s,a)$ is the value gap of state-action pair (s,a) and Var denotes the variance operator. A notable feature of this result is that the sample complexity of BESPOKE (i) scales as $\mathcal{O}(SA\log(1/\delta)/(1-\gamma)^3\varepsilon^2)$ in the worst-case, which is the conjectured minimax lower bound for the infinite-horizon discounted setting (Azar et al., 2012); (ii) it can be significantly smaller than minimax whenever the MDP is such that playing different actions yields very different total rewards, i.e., when the value gaps $(\Delta_{sa})_{s,a}$ are large compared to ε. For the setting of episodic linear MDPs (Jin et al., 2019), the GSS-E algorithm by Taupin et al. (2022) solves a G-optimal design to determine the sampling frequencies of each state-action pair. The sample complexity of GSS-E is upper bounded by

$$\widetilde{\mathcal{O}}\left(\frac{dH^4}{(\min_{h,s,a \neq \pi^\star(s)} \Delta_h(s,a) + \varepsilon)^2} (\log(1/\delta) + d) \right),$$

where d is the feature dimension. Up to H factors, this result improves upon the $\Omega(d^2H^2/\varepsilon^2)$ minimax bound for this setting (Wagenmaker et al., 2022b) whenever the minimum value gap in $\mathcal{M}$ is large.

PAC RL without a generative model On top of the sub-optimality gaps which characterize the bounds above, the instance-dependent complexities feature an additional component when a generative model is not available: visitation probabilities. These constitute the price that PAC RL algorithms have to

pay in order to navigate the MDP and collect observations from distant states. Existing high-probability bounds on the sample complexity are of the form[7]

$$\mathbb{P}\left(\tau = \widetilde{\mathcal{O}}\left(\mathcal{C}_{\mathrm{Alg}}(\mathcal{M},\varepsilon)\log\left(\frac{1}{\delta}\right)\right)\right) \geq 1-\delta,$$

where $\mathcal{C}_{\mathrm{Alg}}(\mathcal{M},\varepsilon)$ is a complexity measure corresponding to a given algorithm Alg. For example, for the MOCA algorithm Wagenmaker et al. (2022a) obtain

$$\mathcal{C}_{\mathrm{MOCA}}(\mathcal{M},\varepsilon) = H^2\sum_{h=1}^{H}\min_{\pi^{\mathrm{exp}}\in\Pi^{\mathrm{S}}}\max_{s,a}\frac{1}{p_h^{\pi^{\mathrm{exp}}}(s,a)}\min\left[\frac{1}{\Delta_h(s,a)^2},\ \frac{W_h(s)^2}{\varepsilon^2}\right] + \frac{H^4\left|\mathrm{OPT}(\mathcal{M},\varepsilon)\right|}{\varepsilon^2},$$

where $W_h(s) = \sup_\pi p_h^\pi(s)$ is the maximum reachability of state s at step $h\in[H]$ and $\mathrm{OPT}(\mathcal{M},\varepsilon)$ is a set of near-optimal triplets (h,s,a). In the above bound, the contribution of a triplet (h,s,a) to the total complexity will be small when either (i) its value gap $\Delta_h(s,a)$ is large or (ii) it is hard to reach by any policy, that is $W_h(s)\ll\varepsilon$. This "local complexity" of (h,s,a) is weighted by $1/p_h^{\pi^{\mathrm{exp}}}(s,a)$, which is the (expected) number of episodes that the agent needs to play in order to reach (h,s,a) when using π^{exp} as an exploration policy. Subsequent works have proposed alternative local complexity measure featuring policy gaps instead of value gaps (Tirinzoni et al., 2022; Wagenmaker and Jamieson, 2022; Al-Marjani et al., 2023). Policy gaps can be larger than value gaps. Notably, they always are in deterministic MDPs (Tirinzoni et al., 2022). For instance, for the PRINCIPLE algorithm, Al-Marjani et al. (2023) obtain

$$\mathcal{C}_{\mathrm{PRINCIPLE}}(\mathcal{M},\varepsilon) = H^3\min_{\pi^{\mathrm{exp}}\in\Pi^{\mathrm{S}}}\max_{h,s,a}\sup_{\pi\in\Pi^{\mathrm{S}}}\frac{p_h^\pi(s,a)}{p_h^{\pi^{\mathrm{exp}}}(s,a)\max(\varepsilon,\Delta(\pi))^2},$$

where we recall the definition of the policy gap $\Delta(\pi) := V_1^\star - V_1^\pi$. Compared to the bound of MOCA, here the contribution of (h,s,a) is small when all policies visiting it are largely sub-optimal. This can be the case even when a is an optimal action in state s, provided that no optimal policy reaches (h,s) with positive probability. We note that, while the lower bound of Theorem 1 only applies to algorithms that output a deterministic policy (see Definition 1), PRINCIPLE is allowed to return a stochastic policy.

4.2 PEDEL: A near-optimal algorithm

The PEDEL algorithm proposed by Wagenmaker and Jamieson (2022) has the sample complexity bound which resembles the most the complexity measure in our lower bound. To introduce it, we define the minimum policy gap $\Delta_{\min} := \min_{\pi\in\Pi^{\mathrm{D}}\setminus\{\pi^\star\}}\Delta(\pi)$, where $\pi^\star$ is an arbitrary optimal policy (i.e., $V_1^{\pi^\star} = V_1^\star$). Note that $\Delta_{\min} = 0$ whenever multiple optimal policies exist.

While PEDEL tackles the more general setting of identifying a near-optimal policy in linear MDPs, when instantiated for the special case of tabular MDPs, the leading term in its sample complexity bound is

$$\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) = H^4\sum_{h=1}^{H}\min_{\rho\in\Omega}\max_{\pi\in\Pi^{\mathrm{D}}}\sum_{s,a}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2},$$

where we ignore some additive lower-order term that is polynomial in $S, A, H, \log(1/\delta)$ and $\log(1/\varepsilon)$.

The next proposition, proved in Appendix C, compares this complexity measure to our lower bound.

Proposition 1. *For any MDP $\mathcal{M}$, it holds that*

$$\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) \leq 8H^5\mathcal{C}_{\mathrm{LB}}(\mathcal{M},\varepsilon) + \frac{4H^6}{(\varepsilon\vee\Delta_{\min})^2}.$$

[7]While we focus on the main complexity terms which scale with sub-optimality gaps, visitation probabilities, and $\log(1/\delta)$, it is worth noting that existing upper bounds all feature lower-order terms in either of these variables.

This shows that in MDPs in which the minimum policy gap is a constant w.r.t. other problem parameters, i.e., $\Delta_{\min} = \Omega(1)$, the complexity $\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M}, \varepsilon)$ is only a factor H^5 away from the instance-dependent lower bound. The same conclusion holds when we are interested in the regime $\varepsilon = \Omega(1)$.

More generally, the next proposition provides a sufficient condition on $\mathcal{M}$ for PEDEL to be instance-optimal up to polynomial multiplicative factors of the horizon, regardless of the values of ε and $\Delta_{\min}$. Let us define the following divergence measure between any pair of policies π, π':

$$d(\pi, \pi') := \sum_{h \in [H]} \mathrm{TV}(p_h^{\pi}, p_h^{\pi'})^2,$$

where $\mathrm{TV}(p_h^{\pi}, p_h^{\pi'}) := \frac{1}{2} \sum_{s,a} |p_h^{\pi}(s,a) - p_h^{\pi'}(s,a)|$ denotes the total variation distance.

Proposition 2. *Let $\varepsilon > 0$ and $\mathcal{M}$ be an MDP such that, for some constant $c > 0$,*

$$\min_{\pi^{\varepsilon} \in \Pi^{\varepsilon}} \max_{\pi \in \Pi^{\mathrm{D}} : \Delta(\pi) \le \varepsilon \vee \Delta_{\min}} d(\pi^{\varepsilon}, \pi) \ge c. \tag{1}$$

Then, $\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M}, \varepsilon) \le 2H^5 \big(4 + \frac{H}{c}\big) \mathcal{C}_{\mathrm{LB}}(\mathcal{M}, \varepsilon)$.

Proposition 2 essentially states that, for MDPs where near-optimal policies are sufficiently "diverse" (in the sense that for every ε-optimal policy there exists a sufficiently distant near-optimal policy), the complexity of PEDEL matches our lower bound up to only multiplicative factors of H. There are several classes of MDPs where the "diversity" condition (1) is satisfied. For instance, it is sufficient to find two near-optimal policies π^1, π^2 (i.e., such that $\Delta(\pi^1) \vee \Delta(\pi^2) \le \varepsilon \vee \Delta_{\min}$) with $\max_h \mathrm{TV}(p_h^{\pi^1}, p_h^{\pi^2}) = 1$ to guarantee that (1) holds with $c = 1/4$[8]. This happens in either of these cases:

- π^1 and π^2 deterministically visit some state s at some stage h (i.e., $p_h^{\pi^1}(s) = p_h^{\pi^2}(s) = 1$) in which they play different actions (i.e., $\pi_h^1(s) \neq \pi_h^2(s)$).
- π^1 and π^2 visit two disjoint sets of states at some stage h, i.e., $\{s : p_h^{\pi^1}(s) > 0\} \cap \{s : p_h^{\pi^2}(s) > 0\} = \emptyset$.
- π^1 and π^2 visit the same states with equal probabilities at some stage h (i.e., $p_h^{\pi^1}(s) = p_h^{\pi^2}(s)$ for any s) at which they play different actions (i.e., $\pi_h^1(s) \neq \pi_h^2(s)$ for all s). For instance, it is enough to have a constant reward at the last stage (i.e., for some α, $r_H(s,a) = \alpha$ for all s, a).

Remark 2. *Upon close inspection of its pseudocode, it seems that PEDEL was designed with the implicit assumption that $\varepsilon = \mathcal{O}(H/d^{3/2})$, where d is the dimension of the linear MDP* [9]*. When this assumption is not satisfied (e.g., when $\varepsilon = \Omega(1/d)$), the sample complexity of PEDEL can actually be d times larger than $\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M}, \varepsilon)$. We elaborate on this in Appendix C.3.*

5 Conclusion and perspective

We proposed the first general instance-dependent lower bound for online PAC RL and proved that it is nearly matched by PEDEL (Wagenmaker and Jamieson, 2022). Unfortunately, the algorithm is computationally intractable as it enumerates and stores the set of deterministic policies, which is of size A^{SH}, in order to eliminate suboptimal policies and solve an experimental design of the form

$$\min_{\rho \in \Omega} \max_{\pi \in \Pi_\ell} \sum_{s,a} \frac{\widehat{p}_h^{\pi,\ell}(s,a)^2}{\rho_h(s,a)}, \tag{2}$$

where $\Pi_\ell \subset \Pi^{\mathrm{D}}$ is the set of active policies at iteration ℓ (initialized as $\Pi_0 = \Pi^{\mathrm{D}}$) and $\widehat{p}_h^{\pi,\ell}(s,a)$ refers to the visitation probabilities of π under the empirical MDP $\widehat{\mathcal{M}}_\ell$. Therefore, we ask the following question

[8] This is because, due to the triangle inequality, $\max\big(\mathrm{TV}(p_h^{\pi^\varepsilon}, p_h^{\pi^1}), \mathrm{TV}(p_h^{\pi^\varepsilon}, p_h^{\pi^2})\big) \ge 1/2$ for any $\pi^\varepsilon \in \Pi^\varepsilon$.

[9] $d = SAH$ in our tabular setting.

Is there a PAC RL algorithm that can (nearly) match our lower bound while requiring a polynomial computational complexity in the size of the MDP?

We believe that answering this question would shed light on the (still elusive) problem of instance-optimality in PAC RL. Indeed, if the answer is negative then this would indicate a clear separation between MDPs and bandits, where we know that computationally-efficient instance-optimality is possible (Garivier and Kaufmann, 2016; Jedra and Proutiere, 2020).

As a starting point to answer the above question, it is natural to wonder whether it is possible to use the same policy-elimination approach as PEDEL while making it computationally efficient. This is precisely the idea of PRINCIPLE (Al-Marjani et al., 2023), which performs implicit policy elimination by adding linear constraints to the set of valid state-action distributions. However, while doing so, it only solves an upper bound on the "optimal" design (2) used by PEDEL of the form

$$\min_{\rho\in\Omega}\max_{\eta\in\Omega_\ell}\max_{s,a}\frac{\eta_h(s,a)}{\rho_h(s,a)},$$

where Ω_ℓ is the set of valid state-action distributions in $\widehat{\mathcal{M}}_\ell$ that satisfy certain near-optimality constraints. This makes the sample complexity of the polynomial-time algorithm PRINCIPLE strictly worse than that of PEDEL, thus not matching the lower bound. We leave as an open question whether an implicit policy elimination scheme can be made compatible with the optimal design (2), in which computing the objective itself seems to require enumerating all policies.

References

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

Alekh Agarwal, Sham M. Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Proceedings of the 33rd Conference On Learning Theory (COLT)*, 2020.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Aymen Al-Marjani, Andrea Tirinzoni, and Emilie Kaufmann. Active coverage for pac reinforcement learning. In *Proceedings of the 36th Conference On Learning Theory (COLT)*, 2023.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91:325–349, 2012.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Victor Boone and Bruno Gaujal. Identification of blackwell optimal policies for deterministic MDPs. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

V.V. Buldygin and Y.V. Kozachenko. Subgaussian random variables. *Ukrainian Mathematical Journal*, 1980.

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

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

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

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

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

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, 2016.

Aurélien Garivier and Emilie Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Annual Conference Computational Learning Theory*, 2019.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

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

Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes, 2021.

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.

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

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

Jerome Taupin, Yassir Jedra, and Alexandre Proutière. Best policy identification in linear mdps. *ArXiv*, abs/2208.05633, 2022.

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

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

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

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

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

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022b.

Andrea Zanette, Mykel J. Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5626–5635, 2019.

Appendix

Table of Contents

A Proof of Theorem 1 **13**

A.1 The change-of-measure argument 14

A.2 A max-min game formulation 15

A.3 Log-likelihood ratio for MDPs with the same transition kernel 16

A.4 Simplifying the expression of the characteristic time 19

B Concentration results **20**

B.1 Proof of Lemma 1 20

C PEDEL **21**

C.1 Proof of Proposition 1 21

C.2 Proof of Proposition 2 23

C.3 On the complexity of PEDEL in the moderate ε regime 23

A Proof of Theorem 1

As mentioned before, our proof is inspired by the one from Degenne and Koolen (2019). The key differences are in Lemma 6 which explicits the shape of the characteristic time for the PAC RL problem and Lemma 4 which relies on a slightly different martingale construction to concentrate the likelihood ratio. Indeed, our martingale involves the expected number of visits to state-action pairs instead of the actual number of visits as in Degenne and Koolen (2019), which is crucial to obtain the navigation constraints $\rho \in \Omega$ in the optimization program of the lower bound.

Notation For any $\pi^\varepsilon \in \Pi^\varepsilon$, we define the set of alternative MDPs that have the same transitions as $\mathcal{M}$ but in which π^ε is no longer ε-optimal:

$$\mathrm{Alt}\,(\pi^\varepsilon) := \Big\{ \widetilde{\mathcal{M}} \in \mathfrak{M}_1 :\ \forall (h,s,a),\ p_h(\cdot|s,a;\widetilde{\mathcal{M}}) = p_h(\cdot|s,a;\mathcal{M}) \text{ and } \exists \pi \in \Pi^D,\ V_1^{\widetilde{\mathcal{M}},\pi^\varepsilon} < V_1^{\widetilde{\mathcal{M}},\pi} - \varepsilon \Big\}.$$

Finally, we define the characteristic time to learn that π^ε is ε-optimal as

$$T(\mathcal{M},\pi^\varepsilon,\varepsilon) := \left(\sup_{\rho\in\Omega} \inf_{\widetilde{\mathcal{M}}\in \mathrm{Alt}(\pi^\varepsilon)} \sum_{h,s,a} \rho_h(s,a) \frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2} \right)^{-1}.$$

Further, for any set of MDPs $E \subset \mathfrak{M}_1$, we let $\overline{E}$ denote the closure of E where the limit points are defined w.r.t. the distance $d(\mathcal{M},\mathcal{M}') := \max_{h,s,a} |r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}'}(s,a)|$.

Proof. Let $\xi \in (0,1)$ and define $T := (1-\xi)\min_{\pi^\varepsilon\in\Pi^\varepsilon} T(\mathcal{M},\pi^\varepsilon,\varepsilon)\log(1/\delta)$[10]. Thanks to Markov's inequality we have that

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq T(1 - \mathbb{P}_{\mathcal{M}}(\tau < T)). \tag{3}$$

We will now upper bound the probability on the right-hand side above. Since the algorithm is (ε,δ)-PAC We have that

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\tau < T) &= \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} \notin \Pi^\varepsilon, \tau < T\big) + \sum_{\pi^\varepsilon\in\Pi^\varepsilon} \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) \\ &\leq \delta + \sum_{\pi^\varepsilon\in\Pi^\varepsilon} \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big).\end{aligned} \tag{4}$$

Now we fix $\pi^\varepsilon \in \Pi^\varepsilon$ and apply Lemma 2 for the event $\mathcal{C} = \big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) \in \mathcal{F}_T$, which yields that there exist $\widetilde{\mathcal{M}}_1,\dots,\widetilde{\mathcal{M}}_{SAH+1} \in \overline{\mathrm{Alt}\,(\pi^\varepsilon)}$ and $(\sigma_i)_{1\leq i\leq SAH+1} \in \mathbb{R}_+^{SAH+1}$ such that, for all $y > 0$,

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) &\leq \exp\Big(y + \frac{T}{T(\mathcal{M},\pi^\varepsilon,\varepsilon)}\Big) \max_{1\leq i\leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) \\ &\quad + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T\sigma_i^2}\Big) \\ &= \delta^{\xi-1}\exp(y) \max_{1\leq i\leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T\sigma_i^2}\Big).\end{aligned} \tag{5}$$

Now for any $i \in [|1, SAH+1|]$ since $\widetilde{\mathcal{M}}_i \in \overline{\mathrm{Alt}\,(\pi^\varepsilon)}$ there exists a sequence of MDPs $(\mathcal{M}'_n)_{n\geq 1}$ with values in $\mathrm{Alt}\,(\pi^\varepsilon)$ such that $\lim_{n\to\infty} \mathcal{M}'_n = \widetilde{\mathcal{M}}_i$[11].

[10] For simplicity, we assume the latter is an integer.

[11] Recall that the convergence was defined w.r.t. the distance $d(\mathcal{M},\mathcal{M}') := \max_{h,s,a} |r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}'}(s,a)|$

By definition of Alt (π^ε), we have that $\mathbb{P}_{\mathcal{M}'_n}(\widehat{\pi} = \pi^\varepsilon, \tau < T) \leq \mathbb{P}_{\mathcal{M}'_n}(\widehat{\pi} = \pi^\varepsilon) \leq \delta$ for all $n \geq 1$. Therefore

$$\begin{aligned}\mathbb{P}_{\widetilde{\mathcal{M}}_i}(\widehat{\pi} = \pi^\varepsilon, \tau < T) &\leq \mathbb{P}_{\widetilde{\mathcal{M}}_i}(\widehat{\pi} = \pi^\varepsilon) \\ &\overset{(a)}{\leq} \liminf_{n\to\infty} \mathbb{P}_{\mathcal{M}'_n}(\widehat{\pi} = \pi^\varepsilon) \leq \delta, \end{aligned} \tag{6}$$

where (a) uses Fatou's lemma. Combining (5) with (6) for the value $y = \xi \log(1/\delta)/2$ yields

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) &\leq \delta^\xi \exp(y) + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T\sigma_i^2}\Big) \\ &\overset{(a)}{=} \delta^{\xi/2} + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{\xi^2 \log(1/\delta)}{4(1-\xi)\min_{\pi^\varepsilon \in \Pi^\varepsilon} T(\mathcal{M}, \pi^\varepsilon, \varepsilon)\sigma_i^2}\Big), \end{aligned} \tag{7}$$

where (a) uses the definition of T. Therefore $\lim_{\delta\to 0} \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^\varepsilon, \tau < T\big) = 0$. This, combined with (4) gives that $\lim_{\delta\to 0} \mathbb{P}_{\mathcal{M}}(\tau < T) = 0$. Plugging this back into (3) and using the definition of T yields

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau]}{\log(1/\delta)} \geq (1-\xi) \min_{\pi^\varepsilon \in \Pi^\varepsilon} T(\mathcal{M}, \pi^\varepsilon, \varepsilon).$$

To finish the proof of Theorem 1, we take the limit when ξ goes to zero and use the simplified expression of the characteristic time given in Lemma 6. □

A.1 The change-of-measure argument

Lemma 2. *Consider* $(\widetilde{\mathcal{M}}_i)_{1\leq i\leq SAH+1} \in \overline{\text{Alt}\,(\pi^\varepsilon)}^{SAH+1}$ *given by Lemma 3 and let* $T \geq 1$. *Then for any event* $\mathcal{C} \in \mathcal{F}_T$ *and any* $y > 0$ *we have*

$$\mathbb{P}_{\mathcal{M}}(C) \leq \exp\big(y + \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)}\big) \max_{1\leq i\leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}(C) + \sum_{i=1}^{SAH+1} \exp\big(-\frac{y^2}{2T\sigma_i^2}\big),$$

where $\sigma_i^2 := \frac{H^2}{4} d(\mathcal{M}, \widetilde{\mathcal{M}}_i)^2 (1 + d(\mathcal{M}, \widetilde{\mathcal{M}}_i))^2$.

Proof. Consider the simplex vector $\lambda^\star \in \Delta_{SAH+1}$ given by Lemma 3. We define the mixture distribution $\mathbb{Q} = \sum_{i=1}^{SAH+1} \lambda_i^\star \mathbb{P}_{\widetilde{\mathcal{M}}_i}$ and the corresponding log-likelihood ratio

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{Q}) := \log \frac{d\mathbb{P}_{\mathcal{M}}}{d\mathbb{Q}}(\mathcal{H}_T).$$

Using Lemma 3.1 from (Garivier and Kaufmann, 2021) we have that for any event $C \in \mathcal{F}_T$ and any $x > 0$,

$$\mathbb{P}_{\mathcal{M}}(C) \leq e^x \mathbb{Q}(C) + \mathbb{P}_{\mathcal{M}}(L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{Q}) > x). \tag{8}$$

We bound each term in the right-hand side separately. Since $\lambda^\star \in \Delta_{SAH+1}$, for any event C,

$$\mathbb{Q}(C) = \sum_{i=1}^{SAH+1} \lambda_i^\star \mathbb{P}_{\widetilde{\mathcal{M}}_i}(C) \leq \max_{1\leq i\leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}(C) \tag{9}$$

On the other hand, we have that

$$
\begin{aligned}
L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{Q}) &\overset{(a)}{\leq} \sum_{i=1}^{SAH+1} \lambda_i^\star \log \frac{d\mathbb{P}_{\mathcal{M}}}{d\mathbb{P}_{\widetilde{\mathcal{M}}_i}}\Big((s_1^t, a_1^t, R_1^t, \ldots, s_H^t, a_H^t, R_H^t)_{1\leq t\leq T}\Big) \\
&= \sum_{i=1}^{SAH+1} \lambda_i^\star L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) \\
&\overset{(b)}{=} \sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) + \sum_{i=1}^{SAH+1} \lambda_i^\star \sum_{h,s,a} \mathbb{E}_{\mathcal{M}}[n_h^T(s,a)] \frac{\big(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2} \\
&= \sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) + T \sum_{i=1}^{SAH+1} \lambda_i^\star \sum_{h,s,a} \frac{\mathbb{E}_{\mathcal{M}}[n_h^T(s,a)]}{T} \frac{\big(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2} \\
&\overset{(c)}{\leq} \sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) + \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)},
\end{aligned}
$$

where (a) uses the convexity of $x \mapsto \log(1/x)$ and Jensen's inequality, (b) uses Lemma 4 and (c) uses the second statement of Lemma 3 and the fact that the vector $\big[\frac{\mathbb{E}_{\mathcal{M}}[n_h^T(s,a)]}{T}\big]_{h,s,a}$ belongs to $\Omega(\mathcal{M})$. Therefore for any $y > 0$, we have that

$$
\begin{aligned}
\mathbb{P}_{\mathcal{M}}\Big(L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{Q}) > \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)} + y\Big) &\leq \mathbb{P}_{\mathcal{M}}\Big(\sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) > y\Big) \\
&\leq \sum_{i=1}^{SAH+1} \mathbb{P}_{\mathcal{M}}\Big(M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) > y\Big) \\
&\leq \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T\sigma_i^2}\Big), \qquad (10)
\end{aligned}
$$

where in the last line we defined $\sigma_i^2 := \frac{H^2}{4} d(\mathcal{M}, \widetilde{\mathcal{M}}_i)^2 (1 + d(\mathcal{M}, \widetilde{\mathcal{M}}_i))^2$ and used Azuma-Hoeffding inequality along with Lemma 4. Combining (9) and (10) with (8) for $x = \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)} + y$ gives the result. □

A.2 A max-min game formulation

We define $\Delta_{SAH+1} := \{\lambda \in \mathbb{R}_+^{SAH+1} : \sum_{i=1}^{SAH+1} \lambda_i = 1\}$ to be the simplex of dimension SAH. Further, for any set of MDPs $E \subset \mathfrak{M}_1$, we let $\overline{E}$ denote the closure of E where the convergence is defined w.r.t. the distance $d(\mathcal{M}, \mathcal{M}') := \max_{h,s,a} |r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}'}(s,a)|$. $\mathrm{Conv}(E)$ refers to the convex hull of E. Finally, we define the set of KL-divergence vectors generated by alternative instances in $\mathrm{Alt}(\pi^\varepsilon)$,

$$
\mathcal{D}(\pi^\varepsilon) := \Big\{ \Big[\frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2} \Big]_{h,s,a} \in \mathbb{R}^{SAH} \text{ s.t. } \widetilde{\mathcal{M}} \in \mathrm{Alt}(\pi^\varepsilon) \Big\}.
$$

Lemma 3. *Fix $\pi^\varepsilon \in \Pi^\varepsilon$.*

Then there exists $\rho^\star \in \Omega, \lambda^\star \in \Delta_{SAH+1}$ and $\widetilde{\mathcal{M}}_1, \ldots, \widetilde{\mathcal{M}}_{SAH+1} \in \overline{\mathrm{Alt}(\pi^\varepsilon)}$ such that

$$
T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} = \sum_{i=1}^{SAH+1} \lambda_i^\star \Big[\sum_{h,s,a} \rho_h^\star(s,a) \frac{\big(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2} \Big].
$$

Furthermore, for any $\rho \in \Omega$ we have that

$$\sum_{i=1}^{SAH+1} \lambda_i^\star \left[\sum_{h,s,a} \rho_h(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \right] \leq T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1}.$$

Proof. Observe that we can rewrite the expression of the characteristic time $T(\mathcal{M}, \pi^\varepsilon, \varepsilon)$ as follows,

$$\begin{aligned}
T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} &= \sup_{\rho \in \Omega} \inf_{\widetilde{\mathcal{M}} \in \mathrm{Alt}(\pi^\varepsilon)} \sum_{h,s,a} \rho_h(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \\
&= \sup_{\rho \in \Omega} \inf_{\widetilde{d} \in \mathcal{D}(\pi^\varepsilon)} \rho^\top \widetilde{d} \\
&= \sup_{\rho \in \Omega} \inf_{\widetilde{d} \in \overline{\mathcal{D}(\pi^\varepsilon)}} \rho^\top \widetilde{d} \\
&= \sup_{\rho \in \Omega} \inf_{\widetilde{d} \in \mathrm{Conv}(\overline{\mathcal{D}(\pi^\varepsilon)})} \rho^\top \widetilde{d}, \qquad (11)
\end{aligned}$$

where $\mathrm{Conv}(\mathcal{D}(\pi^\varepsilon))$ denotes the convex hull of $\mathcal{D}(\pi^\varepsilon)$. Now let $(\rho^\star, d^\star)$ be an optimal solution to (11). Since $\mathcal{D}(\pi^\varepsilon) \subset \mathbb{R}^{SAH}$, by Carathéodory's extension theorem we have that there exists $\lambda^\star \in \Delta_{SAH+1}$ and $d_1, \ldots, d_{SAH+1} \in \overline{\mathcal{D}(\pi^\varepsilon)}$ such that $d^\star = \sum_{i=1}^{SAH+1} \lambda_i^\star d_i$. This means that there exists $\rho^\star \in \Omega$ and $\widetilde{\mathcal{M}}_1, \ldots, \widetilde{\mathcal{M}}_{SAH+1} \in \overline{\mathrm{Alt}\,(\pi^\varepsilon)}$ such that

$$\begin{aligned}
T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} &= (\rho^\star)^\top d^\star \\
&= \sum_{i=1}^{SAH+1} \lambda_i^\star (\rho^\star)^\top d_i \\
&= \sum_{i=1}^{SAH+1} \lambda_i^\star \left[\sum_{h,s,a} \rho_h^\star(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \right].
\end{aligned}$$

This proves the first statement. Now for the second statement, using Sion's minimax theorem (Sion (1958), Theorem 3.4) we know that

$$(\rho^\star)^\top d^\star = \sup_{\rho \in \Omega} \inf_{\widetilde{d} \in \mathrm{Conv}(\overline{\mathcal{D}(\pi^\varepsilon)})} \rho^\top \widetilde{d} = \inf_{\widetilde{d} \in \mathrm{Conv}(\overline{\mathcal{D}(\pi^\varepsilon)})} \sup_{\rho \in \Omega} \rho^\top \widetilde{d},$$

i.e $(\rho^\star, d^\star)$ is a saddle point of (11). This means that for all $\rho \in \Omega$

$$\rho^\top d^\star \leq (\rho^\star)^\top d^\star = T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1}.$$

Expanding the left-hand side proves the second statement. □

A.3 Log-likelihood ratio for MDPs with the same transition kernel

In the following we fix an algorithm $\mathfrak{A}$. For $T \geq 1$ we define the history up to the end of episode T as $\mathcal{H}_T := (s_1^t, a_1^t, R_1^t, \ldots, s_H^t, a_H^t, R_H^t, \mathbb{1}(t \leq \tau_\delta))_{1 \leq t \leq T}$. For any MDP $\mathcal{M}$, we write $\mathbb{P}_\mathcal{M}$ to denote the probability distribution over possible histories when $\mathfrak{A}$ interacts with $\mathcal{M}$[12]. Further $(\mathcal{F}_T)_{T \geq 1}$ will denote the sigma algebra generated by $(\mathcal{H}_T)_{T \geq 1}$. Finally, for a pair of MDPs $\mathcal{M}, \widetilde{\mathcal{M}}$, we define the log-likelihood

[12]Since we will be considering the same algorithm $\mathfrak{A}$ interacting with different MDPs, we do not index the probability distributions by $\mathfrak{A}$.

ratio of observations at the end of any episode T[13]

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) := \log \frac{d\mathbb{P}_{\mathcal{M}}}{d\mathbb{P}_{\widetilde{\mathcal{M}}}}(\mathcal{H}_T)$$
$$= \log\Bigg(\prod_{t=1}^{T}\prod_{h=1}^{H} \frac{\exp\big(-[R_h^t - r_h^{\mathcal{M}}(s_h^t, a_h^t)]^2/2\big) p_{h-1}^{\mathcal{M}}(s_h^t | s_{h-1}^t, a_{h-1}^t)}{\exp\big(-[R_h^t - r_h^{\widetilde{\mathcal{M}}}(s_h^t, a_h^t)]^2/2\big) p_{h-1}^{\widetilde{\mathcal{M}}}(s_h^t | s_{h-1}^t, a_{h-1}^t)}\Bigg).$$

Lemma 4. *For any pair of MDPs $\mathcal{M}, \widetilde{\mathcal{M}} \in \mathfrak{M}_1$, there exists a martingale (under $\mathbb{E}_{\mathcal{M}}$) $\big(M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}})\big)_{T\geq 1}$ whose increments are $\frac{H^2}{4} d(\mathcal{M}, \widetilde{\mathcal{M}})^2 (1 + d(\mathcal{M}, \widetilde{\mathcal{M}}))^2$-subgaussian and such that the likelihood ratio at the end of episode T satisfies*

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) + \sum_{h,s,a} \mathbb{E}_{\mathcal{M}}[n_h^T(s,a)] \frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2}.$$

Proof. Using that the MDPs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ share the same transition kernels and have Gausssian reward distributions with unit variance, we can simplify their log-likelihood ratio as follows,

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{h=1}^{H}\Big[\big(R_h^t - r_h^{\mathcal{M}}(s_h^t, a_h^t)\big)^2 - \big(R_h^t - r_h^{\widetilde{\mathcal{M}}}(s_h^t, a_h^t)\big)^2\Big]$$
$$= \frac{1}{2}\sum_{h,s,a}\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\Big[\big(R_h^t - r_h^{\widetilde{\mathcal{M}}}(s,a)\big)^2 - \big(R_h^t - r_h^{\mathcal{M}}(s,a)\big)^2\Big]. \tag{12}$$

Now for any fixed (h,s,a) we can define $\widehat{r}_h^T(s,a) := \frac{\sum_{t=1}^T \mathbb{1}(s_h^t = s, a_h^t = a) R_h^t}{n_h^T(s,a)}$ if $n_h^T(s,a) > 0$ and $\widehat{r}_h^T(s,a) := 0$ otherwise. Then we can write that

$$\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\big(R_h^t - r_h^{\mathcal{M}}(s_h, a_h)\big)^2$$
$$= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\Big[\big(R_h^t - \widehat{r}_h^T(s,a)\big) + \big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big)\Big]^2$$
$$= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\Big[\big(R_h^t - \widehat{r}_h^T(s,a)\big)^2 + \big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2\Big]$$
$$+ 2\big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big) \underbrace{\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\big(R_h^t - \widehat{r}_h^T(s,a)\big)}_{=0}$$
$$= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\Big[\big(R_h^t - \widehat{r}_h^T(s,a)\big)^2 + \big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2\Big]. \tag{13}$$

Similarly, one can show that

$$\sum_{h,s,a}\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\big(R_h^t - r_h^{\widetilde{\mathcal{M}}}(s_h^t, a_h^t)\big)^2$$
$$= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\Big[\big(R_h^t - \widehat{r}_h^T(s,a)\big)^2 + \big(\widehat{r}_h^T(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\big)^2\Big]. \tag{14}$$

[13] With the convention that $p_0(\cdot|s_0,a_0) = \mathbb{1}(s_1 = \cdot)$ for all (s_0,a_0). Also note that we have simplified the probabilities of choosing actions $\pi^t(a_h^t | s_h^t, a_{h-1}^t, \ldots, s_1^t, \mathcal{H}_{t-1})$ and of stopping $\pi^t(\tau_\delta = t | \mathcal{H}_t)$ as they only depend on the history, therefore having the same value for $\mathcal{M}$ and $\widetilde{\mathcal{M}}$.

Combining equations (12), (13) and (14) we get that

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = \frac{1}{2}\sum_{h,s,a}\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a)\left[\left(\widehat{r}_h^T(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)^2 - \left(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2\right]$$
$$= \frac{1}{2}\sum_{h,s,a} n_h^T(s,a)\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)\left(2\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right). \tag{15}$$

Next we define the sequences

$$M_T(h,s,a) := \frac{1}{2}\Big[n_h^T(s,a)\big(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\big)\big(2\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\big)$$
$$- \mathbb{E}_{\mathcal{M}}[n_h^T(s,a)]\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2\Big].$$
$$M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) := \sum_{h,s,a} M_T(h,s,a).$$

Using (15) one can check that

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) + \sum_{h,s,a}\mathbb{E}_{\mathcal{M}}[n_h^T(s,a)]\frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2}.$$

This proves the second statement. Now for the first statement we note that for $T \geq 2$,

$$M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) - M_{T-1}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}})$$
$$= \frac{1}{2}\sum_{h,s,a}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)\mathbb{1}(s_h^T = s, a_h^T = a)\left(2R_h^T - r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)$$
$$- \mathbb{P}_{\mathcal{M}}(s_h^T = s, a_h^T = a)\left(r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2$$
$$= \underbrace{\frac{1}{2}\sum_{h,s,a}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)\mathbb{1}\left(s_h^T = s, a_h^T = a\right)\left(R_h^T - r_h^{\mathcal{M}}(s,a)\right)}_{:=X_T}$$
$$+ \underbrace{\frac{1}{2}\sum_{h,s,a}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)^2\left(\mathbb{1}\left(s_h^T = s, a_h^T = a\right) - \mathbb{P}_{\mathcal{M}}\left(s_h^t = s, a_h^t = a\right)\right).}_{:=Y_T}$$

X_T satisfies

$$\mathbb{E}[X_T|\mathcal{F}_{T-1}] = \mathbb{E}\left[\frac{1}{2}\sum_{h,s,a}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)\underbrace{\mathbb{E}\left[\left(R_h^T - r_h^{\mathcal{M}}(s,a)\right)|S_h^T, A_h^T\right]}_{=0}\middle|\mathcal{F}_{T-1}\right] = 0$$

and $X_T = \sum_{h=1}^{H} X_{T,h}$ where

$$X_{T,h} = \frac{\left(r_h^{\mathcal{M}}(s_h^T, a_h^T) - r_h^{\widetilde{\mathcal{M}}}(s_h^T, a_h^T)\right)}{2}\left(R_h^T - r_h^{\mathcal{M}}(s_h^T, a_h^T)\right)$$

is subgaussian with variance $\frac{d(\mathcal{M},\widetilde{\mathcal{M}})^2}{4}$ conditionally to $\mathcal{F}_{T-1}$ (using that $R_h^T - r_h^{\mathcal{M}}(s_h^T, a_h^T)$ is 1-subgaussian). Therefore, by Lemma 5 stated below, X_T is subgaussian with $\sigma_X^2 = \frac{H^2 d(\mathcal{M},\widetilde{\mathcal{M}})^2}{4}$.

Y_T satisfies

$$\mathbb{E}[Y_T|\mathcal{F}_{T-1}] = \frac{1}{2}\sum_{h,s,a}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)^2 \mathbb{E}\left[\mathbb{1}(s_h^T = s, a_h^T = a) - \mathbb{P}^{\mathcal{M}}(s_h^T = s, a_h^T = a)|\mathcal{F}_{T-1}\right] = 0$$

and $|Y_T| \leq \frac{Hd(\mathcal{M},\widetilde{\mathcal{M}})^2}{2}$. Therefore Y_T is subgaussian with $\sigma_Y^2 = \frac{H^2 d(\mathcal{M},\widetilde{\mathcal{M}})^4}{4}$.

By Lemma 5, $M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) - M_{T-1}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}})$ is conditionally subgaussian with variance

$$\frac{H^2 d(\mathcal{M},\widetilde{\mathcal{M}})^2(1 + d(\mathcal{M},\widetilde{\mathcal{M}}))^2}{4}.$$

□

Lemma 5 (sum of subgaussian random variables, e.g. Buldygin and Kozachenko (1980))**.** *Let X an Y be two random variables that are σ_X^2 and σ_Y^2 subgaussian respectively. Then $X+Y$ is $(\sigma_X+\sigma_Y)^2$-subgaussian.*

Proof. Using Hölder inequality and the definition of subgaussian variables, we can write, for any $p \geq 1, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$

$$\begin{aligned}
\mathbb{E}[\exp\left(t(X+Y)\right)] &= \mathbb{E}[\exp(tX)\exp(tY)] \\
&\leq \mathbb{E}[\exp(ptX)]^{1/p}\mathbb{E}[\exp(qtY)]^{1/q} \\
&\leq \exp\left(\frac{p^2t^2\sigma_X^2}{2}\right)^{1/p} \exp\left(\frac{q^2t^2\sigma_Y^2}{2}\right)^{1/q} \\
&= \exp\left(\frac{t^2(p\sigma_X^2 + q\sigma_Y^2)}{2}\right).
\end{aligned}$$

The conclusion follows by choosing $p = \frac{\sigma_X+\sigma_Y}{\sigma_X}$ and $q = \frac{\sigma_X+\sigma_Y}{\sigma_Y}$ for which $p\sigma_X^2 + q\sigma_Y^2 = (\sigma_X + \sigma_Y)^2$.

□

A.4 Simplifying the expression of the characteristic time

Lemma 6. *For any $\mathcal{M} \in \mathfrak{M}_1$ and $\pi^\varepsilon \in \Pi^\varepsilon$ we have*

$$T(\mathcal{M}, \pi^\varepsilon, \varepsilon) = 2 \inf_{\rho\in\Omega} \max_{\pi\in\Pi^D} \sum_{s,a,h} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2}.$$

Proof. Let us first solve the inner minimization program in the definition of $T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1}$. Using the definition of $\mathrm{Alt}(\pi^\varepsilon)$, we have that

$$\inf_{\widetilde{\mathcal{M}}\in\mathrm{Alt}(\pi^\varepsilon)} \sum_{h,s,a} \rho_h(s,a)\frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} = \min_{\pi\in\Pi^D} \inf_{\widetilde{\mathcal{M}}: V_1^{\widetilde{\mathcal{M}},\pi^\varepsilon} < V_1^{\widetilde{\mathcal{M}},\pi} - \varepsilon} \sum_{h,s,a} \rho_h(s,a)\frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2}. \tag{16}$$

Now observe that we can rewrite $V_1^{\widetilde{\mathcal{M}},\pi^\varepsilon} < V_1^{\widetilde{\mathcal{M}},\pi} - \varepsilon$ as linear constraint in the rewards of $\widetilde{\mathcal{M}}$:

$$\begin{aligned}
&\sum_{h,s,a}(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))r_h^{\widetilde{\mathcal{M}}}(s,a) > \varepsilon, \\
\iff &\sum_{h,s,a}(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right) > V_1^{\pi^\varepsilon} - V_1^\pi + \varepsilon, \\
\iff &\sum_{h,s,a}(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right) > \Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon
\end{aligned}$$

Therefore, letting $u_h(s,a) = r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)$, the program in (16) is equivalent to

$$\min_{\pi \in \Pi^{\mathrm{D}}} \inf_{u \text{ s.t: } \sum_{h,s,a} (p_h^{\pi}(s,a) - p_h^{\pi^\varepsilon}(s,a)) u_h(s,a) > \Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon} \sum_{h,s,a} \rho_h(s,a) \frac{u_h(s,a)^2}{2}. \tag{17}$$

Solving the KKT conditions of the previous program, we get that

$$\inf_{u \text{ s.t: } \sum_{h,s,a} (p_h^{\pi}(s,a) - p_h^{\pi^\varepsilon}(s,a)) u_h(s,a) > \Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon} \sum_{h,s,a} \rho_h(s,a) \frac{u_h(s,a)^2}{2} = \left(\sum_{h,s,a} \frac{(p_h^{\pi}(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2} \right)^{-1}.$$

Summing up all the inequalities, we conclude that

$$T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} = \frac{1}{2} \sup_{\rho \in \Omega} \min_{\pi \in \Pi^{\mathrm{D}}} \left(\sum_{h,s,a} \frac{(p_h^{\pi}(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2} \right)^{-1}.$$

□

B Concentration results

We report here useful concentration results from previous literature.

Proposition 3. *(Lemma 26, Al-Marjani et al. (2023)[14]) Let the reward distribution $\nu_h(s,a)$ be 1-subgaussian with mean $r_h(s,a)$ for all (h,s,a), and let $\widehat{r}_h^t(s,a)$ be the MLE of $r_h(s,a)$ using samples gathered until episode t. Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$ and $Z := |\mathcal{Z}|$. With probability at least $1-\delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$,*

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

Proposition 4. *(Lemma 30, Al-Marjani et al. (2023)) Let $n \in \mathbb{N}$, $q, b \in \mathbb{R}^n$ with b having strictly positive entries, and $c \in \mathbb{R}_{\geq 0}$. Then,*

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

B.1 Proof of Lemma 1

Proof. Fix any pair of policies π, π'. We write

$$\begin{aligned}
(\widehat{V}_1^{\pi,t} - \widehat{V}_1^{\pi',t}) - (V_1^{\pi} - V_1^{\pi'}) &= (p^{\pi} - p^{\pi'})^{\top} (\widehat{r}^t - r) \\
&= \sum_{h,s,a} (p_h^{\pi}(s,a) - p_h^{\pi'}(s,a)) (\widehat{r}_h^t(s,a) - r_h(s,a)) \\
&= \sum_{h,s,a} \mathbb{1}\big(a \in \{\pi_h(s), \pi'_h(s)\}\big) (p_h^{\pi}(s,a) - p_h^{\pi'}(s,a)) (\widehat{r}_h^t(s,a) - r_h(s,a)),
\end{aligned}$$

[14] Note that, while Al-Marjani et al. (2023) state this lemma for rewards bounded in $[0,1]$, they actually prove it for any 1-subgaussian distribution. Indeed, their proof simply combines the concentration result of Abbasi-Yadkori et al. (2011), which holds for any subgaussian distribution, with a trick from Réda et al. (2021).

where we used vector notation $p^\pi = [p_h^\pi(s,a)]_{h,s,a}$. Now applying Proposition 3 with $\delta' = \delta/(A^{2SH})$ and the set $\mathcal{Z} = \big\{(h,s,a) \mid (h,s) \in [H] \times \mathcal{S}, a \in \{\pi_h(s), \pi'_h(s)\} \text{ s.t. } \sup_\pi p_h^\pi(s) > 0 \big\}$ we get that with probability at least $1-\delta'$,

$$\forall t \geq t_0,\ \sum_{h,s,a} \mathbb{1}\big(a \in \{\pi_h(s), \pi'_h(s)\}\big) n_h^t(s,a) \big(\widehat{r}_h^t(s,a) - r_h(s,a)\big)^2 \leq 4\log(1/\delta') + 4SH\log(A(1+t))$$
$$:= \widetilde{\beta}(t,\delta'),$$

where we used that $|\mathcal{Z}| \leq 2SH$. Next, for each pair of policies (π, π') we use Proposition 4 with $q = p^\pi - p^{\pi'}$ which yields that

$$\big|(\widehat{V}_1^{\pi,t} - \widehat{V}_1^{\pi',t}) - (V_1^\pi - V_1^{\pi'})\big| \leq \sqrt{\widetilde{\beta}(t,\delta') \sum_{h,s,a} \mathbb{1}\big(a \in \{\pi_h(s), \pi'_h(s)\}\big) \frac{\big(p_h^\pi(s,a) - p_h^{\pi'}(s,a)\big)^2}{n_h^t(s,a)}}$$
$$= \sqrt{\widetilde{\beta}(t,\delta') \sum_{h,s,a} \frac{\big(p_h^\pi(s,a) - p_h^{\pi'}(s,a)\big)^2}{n_h^t(s,a)}},$$

with probability at least $1 - \delta/(A^{2SH})$. We conclude the proof with a union bound over pairs of policies $(\pi, \pi') \in \Pi^{\mathrm{D}} \times \Pi^{\mathrm{D}}$ and remarking that

$$\widetilde{\beta}(t,\delta') = 4\log(1/\delta) + 12SH\log(A) + 4SH\log(1+t) \leq \beta(t,\delta).$$

□

C PEDEL

C.1 Proof of Proposition 1

First, let us introduce the intermediate complexity measure

$$C(\mathcal{M}, \varepsilon) := \min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi) \vee \varepsilon \vee \Delta_{\min})^2}.$$

We start by showing that $H^3 C(\mathcal{M},\varepsilon) \leq \mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) \leq H^5 C(\mathcal{M},\varepsilon)$. For $h \in [H]$ consider any $\rho^{\star,h} \in \arg\min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2}$. Now, letting $\widetilde{\rho} := \frac{1}{H}\sum_{h=1}^H \rho^{\star,h}$, we see that since Ω is a convex set, $\widetilde{\rho} \in \Omega$. Furthermore,

$$\begin{aligned}
C(\mathcal{M},\varepsilon) &= \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\leq \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\widetilde{\rho}_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\overset{(a)}{\leq} \sum_{h=1}^H \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\widetilde{\rho}_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\overset{(b)}{\leq} H \sum_{h=1}^H \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h^{\star,h}(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&= H \sum_{h=1}^H \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&= H^{-3}\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon),
\end{aligned}$$

where (a) uses the fact that $\max_\pi \sum_h f(\pi,h) \le \sum_h \max_\pi f(\pi,h)$ and (b) uses the crude bound $\widetilde{\rho}_h(s,a) \ge \rho_h^{\star,h}(s,a)/H$. On the other hand we have

$$\begin{aligned}
\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) &= H^4 \sum_{h=1}^{H} \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\le H^4 \sum_{\ell=1}^{H} \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&= H^5 C(\mathcal{M},\varepsilon).
\end{aligned}$$

Therefore, we just proved that

$$H^3 C(\mathcal{M},\varepsilon) \le \mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) \le H^5 C(\mathcal{M},\varepsilon). \tag{18}$$

Now we compare $C(\mathcal{M},\varepsilon)$ and $\mathcal{C}_{\mathrm{LB}}(\mathcal{M},\varepsilon)$. Using that $a^2 \le 2(a-b)^2 + 2b^2$, we note that for any $\rho \in \Omega$ and any $\pi^\varepsilon \in \Pi^\varepsilon$,

$$\begin{aligned}
&\max_{\pi\in\Pi^{\mathrm{D}}} \frac{\sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)}}{(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\quad\le \max_{\pi\in\Pi^{\mathrm{D}}} \left[\sum_{s,a,h} \frac{2(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \sum_{s,a,h} \frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \right] \\
&\quad\le \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{2(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\quad= \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{2(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \sum_{s,a,h} \frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\rho_h(s,a)(\varepsilon\vee\Delta_{\min})^2}.
\end{aligned} \tag{19}$$

Now let us define $\rho^0 := \arg\min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2}$ and $\widetilde{\rho}^1 := \frac{\rho^0+p^{\pi^\varepsilon}}{2}$. Then we have that

$$\begin{aligned}
C(\mathcal{M},\varepsilon) &= \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\le \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\widetilde{\rho}_h^1(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} \\
&\overset{(a)}{\le} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{2(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\widetilde{\rho}_h^1(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \sum_{s,a,h} \frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\widetilde{\rho}_h^1(s,a)(\varepsilon\vee\Delta_{\min})^2} \\
&\overset{(b)}{\le} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{4(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h^0(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \sum_{s,a,h} \frac{4p_h^{\pi^\varepsilon}(s,a)^2}{p_h^{\pi^\varepsilon}(s,a)(\varepsilon\vee\Delta_{\min})^2} \\
&= 4 \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \frac{4H}{(\varepsilon\vee\Delta_{\min})^2},
\end{aligned}$$

where (a) uses (19) and (b) uses the fact that for all (h,s,a), $\widetilde{\rho}_h^1(s,a) \ge \max(\rho_h^0(s,a), p_h^{\pi^\varepsilon}(s,a))/2$. Since this holds for any π^ε,

$$\begin{aligned}
C(\mathcal{M},\varepsilon) &\le 4 \min_{\pi\in\Pi^\varepsilon} \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2} + \frac{4H}{(\varepsilon\vee\Delta_{\min})^2} \\
&\le 16 \min_{\pi\in\Pi^\varepsilon} \min_{\rho\in\Omega} \max_{\pi\in\Pi^{\mathrm{D}}} \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)+\varepsilon-\Delta(\pi^\varepsilon))^2} + \frac{4H}{(\varepsilon\vee\Delta_{\min})^2} \\
&= 8\mathcal{C}_{\mathrm{LB}}(\mathcal{M},\varepsilon) + \frac{4H}{(\varepsilon\vee\Delta_{\min})^2},
\end{aligned} \tag{20}$$

where in the second inequality we used that $\Delta(\pi)+\varepsilon-\Delta(\pi^\varepsilon)\le 2(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})$. Combining (18) and (20) proves the first inequality.

C.2 Proof of Proposition 2

Combining the first inequality in the sequence (20) with (18), we have that

$$\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon)\le 4H^5\underbrace{\min_{\pi\in\Pi^\varepsilon}\min_{\rho\in\Omega}\max_{\pi\in\Pi^{\mathrm{D}}}\sum_{s,a,h}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min})^2}}_{(\star)}+\frac{4H^6}{(\varepsilon\vee\Delta_{\min})^2}. \tag{21}$$

We now lower bound $(\star)$ as a function of $1/(\varepsilon\vee\Delta_{\min})^2$. We have

$$\begin{aligned}
(\star)&\ge \min_{\pi\in\Pi^\varepsilon}\min_{\rho\in\Omega}\max_{\pi\in\Pi^{\mathrm{D}}:\Delta(\pi)\le\varepsilon\vee\Delta_{\min}}\sum_{s,a,h}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\varepsilon\vee\Delta_{\min})^2}\\
&\ge \min_{\pi\in\Pi^\varepsilon}\max_{\pi\in\Pi^{\mathrm{D}}:\Delta(\pi)\le\varepsilon\vee\Delta_{\min}}\sum_{h\in[H]}\min_{\rho\in\Omega}\sum_{s,a}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\varepsilon\vee\Delta_{\min})^2}\\
&\ge \min_{\pi\in\Pi^\varepsilon}\max_{\pi\in\Pi^{\mathrm{D}}:\Delta(\pi)\le\varepsilon\vee\Delta_{\min}}\sum_{h\in[H]}\min_{\rho\in\mathcal{P}(\mathcal{S}\times\mathcal{A})}\sum_{s,a}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\varepsilon\vee\Delta_{\min})^2}\\
&= \frac{1}{(\varepsilon\vee\Delta_{\min})^2}\min_{\pi\in\Pi^\varepsilon}\max_{\pi\in\Pi^{\mathrm{D}}:\Delta(\pi)\le\varepsilon\vee\Delta_{\min}}\sum_{h\in[H]}\left(\sum_{s,a}|p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a)|\right)^2\\
&= \frac{4}{(\varepsilon\vee\Delta_{\min})^2}\min_{\pi\in\Pi^\varepsilon}\max_{\pi\in\Pi^{\mathrm{D}}:\Delta(\pi)\le\varepsilon\vee\Delta_{\min}} d(\pi^\varepsilon,\pi)\ge\frac{4c}{(\varepsilon\vee\Delta_{\min})^2},
\end{aligned}$$

where the first equality uses that $\min_{\rho\in\mathcal{P}(\mathcal{X})}\sum_{x\in\mathcal{X}}\frac{f(x)}{\rho(x)}=(\sum_{x\in\mathcal{X}}\sqrt{f(x)})^2$ for any non-negative function f. This implies that

$$\frac{4H^6}{(\varepsilon\vee\Delta_{\min})^2}\le\frac{H^6}{c}(\star).$$

Plugging this into (21) and using that $(\star)\le 2\mathcal{C}_{\mathrm{LB}}(\mathcal{M},\varepsilon)$ as in (20) concludes the proof.

C.3 On the complexity of PEDEL in the moderate ε regime

PEDEL has a loop structure where at each iteration it seeks to halve the precision of its estimate of the value for all the policies that are still active. Taking a closer look into the design of PEDEL, we notice that it starts the first iteration with the parameter $\ell_0=\lceil\log_2\frac{d^{3/2}}{H}\rceil$ and ends at $\lceil\log\frac{4}{\varepsilon}\rceil$. From Theorem 7 in Wagenmaker and Jamieson (2022), we get that the number of episodes played during the initial iteration is

$$\mathcal{O}\left(H^4\sum_{h=1}^{H}\frac{\inf_{\Lambda_{exp}\in\Omega_h}\max_{\varphi\in\Phi}\|\varphi\|_{\Lambda_{exp}^{-1}}}{\varepsilon_{exp}}\right),\ \text{where}\quad \varepsilon_{exp}:=\frac{\varepsilon_{\ell_0}^2}{\beta_{\ell_0}},$$
$$\varepsilon_{\ell_0}:=2^{-\ell_0}=\frac{H}{d^{3/2}},\ \beta_{\ell_0}:=64H^2\log(\frac{4H^2|\Pi|\ell_0^2}{\delta}).$$

As a consequence, running just the initial iteration of PEDEL requires the number of episodes

$$\mathcal{C}_0:=\mathcal{O}\left(d^3H^4\log(|\Pi|/\delta)\min_{\rho\in\Omega}\max_{\pi\in\Pi^{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)}\right).$$

When $\varepsilon = \Omega(1/d)$, we have that $d^2 = \Omega(\frac{1}{(\varepsilon \vee \Delta(\pi) \vee \Delta_{\min})^2})$ for all policies π so that

$$\mathcal{C}_0 = \Omega\left(dH^4 \log(|\Pi|/\delta) \min_{\rho \in \Omega} \max_{\pi \in \Pi^{\mathrm{D}}} \frac{\sum_{s,a,h} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)}}{(\varepsilon \vee \Delta(\pi) \vee \Delta_{\min})^2} \right).$$

Therefore when $\varepsilon = \Omega(1/SAH)$, we get that the sample complexity of PEDEL for tabular MDPs satisfies

$$\tau = \Omega\left(SAH \times \mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M}, \varepsilon) \log(1/\delta) \right),$$

almost surely.