Proceedings of Machine Learning Research vol 178:1–61, 2022 35th Annual Conference on Learning Theory

Beyond No Regret: Instance-Dependent PAC Reinforcement Learning

Andrew Wagenmaker AJWAGEN@CS.WASHINGTON.EDU
University of Washington, Seattle, WA

Max Simchowitz MSIMCHOW@MIT.EDU
MIT, Cambridge, MA

Kevin Jamieson JAMIESON@CS.WASHINGTON.EDU
University of Washington, Seattle, WA

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

The theory of reinforcement learning has focused on two fundamental problems: achieving low regret, and identifying ϵ-optimal policies. While a simple reduction allows one to apply a low-regret algorithm to obtain an ϵ-optimal policy and achieve the worst-case optimal rate, it is unknown whether low-regret algorithms can obtain the instance-optimal rate for policy identification. We show this is not possible—there exists a fundamental tradeoff between achieving low regret and identifying an ϵ-optimal policy at the instance-optimal rate.

Motivated by our negative finding, we propose a new measure of instance-dependent sample complexity for PAC tabular reinforcement learning which explicitly accounts for the attainable state visitation distributions in the underlying MDP. We then propose and analyze a novel, planning-based algorithm which attains this sample complexity—yielding a complexity which scales with the suboptimality gaps and the "reachability" of a state. We show our algorithm is nearly minimax optimal, and on several examples that our instance-dependent sample complexity offers significant improvements over worst-case bounds.

1. Introduction

Two of the most fundamental problems in Reinforcement Learning (RL) are regret minimization, and PAC (Probably Approximately Correct) policy identification. In the former setting, the goal of the agent is simply to play actions that collect sufficient reward in an online fashion, while in the latter, the goal of the agent is to explore their environment in order to identify an ϵ-optimal policy with probability $1 - \delta$.

These objectives are intimately related: for an agent to achieve low-regret they must play "good" policies, and therefore can solve the PAC problem as well. Indeed, in the worst case, optimal performance can be achieved by the "online-to-batch" reduction: running a worst-case optimal regret algorithm for K episodes, and averaging its chosen policies (or choosing one at random) to make a recommendation. In this paper, we ask if online-to-batch is all there is to PAC learning. Focusing on the non-generative tabular setting, we ask

> *Does the online-to-batch reduction yield tight instance-dependent guarantees in non-generative, tabular PAC reinforcement learning? Or, are there other algorithmic principles and measures of sample complexity that emerge in the PAC setting but are absent when studying regret?*

Mirroring recent developments in the regret setting which obtain instance-dependent regret guarantees, we approach this question from an instance-dependent perspective, and seek to develop instance-dependent PAC guarantees.

Our focus on the non-generative setting brings to light the role of exploration in learning good policies. The majority of low-regret algorithms rely on playing actions they believe will lead to large reward (the principle of *optimism*) and only explore enough to ensure they do not overcommit to suboptimal actions. While this is sufficient to balance the exploration-exploitation tradeoff and induce enough exploration to obtain low regret, as we will see, when the goal is simply *exploration* and no concern is given for the online reward obtained, much more aggressive exploration can be used to efficiently traverse the MDP and learn a good policy. Hence, in addressing our question above, we aim to understand more broadly what are the most effective exploration strategies for traversing an unknown MDP when the goal is to learn a good policy.

1.1. Our Contributions

We demonstrate the importance of non-optimistic planning via three main contributions:

- *New measure of instance-dependent complexity.* We propose a novel, fully instance-dependent measure of complexity for MDPs, the *gap-visitation complexity*:

$$\mathcal{C}(\mathcal{M},\epsilon) := \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$$

  where here $w_h^\pi(s,a)$ is the probability of visiting (s,a) at step h under policy π, $\Delta_h(s,a)$ is a measure of the suboptimality of choosing action a at state s and step h, $W_h(s)$ is the *maximum reachability* of state s at step h, and $\mathrm{OPT}(\epsilon)$ is the set of all "near-optimal" state-action tuples. We show that $\mathcal{C}(\mathcal{M},\epsilon)$ is no larger than the minimax optimal PAC rate, and that in some cases, $\mathcal{C}(\mathcal{M},\epsilon)$ is equivalent to the instance-optimal complexity.

- *A novel planning-based algorithm.* We propose and analyze a computationally efficient planning-based algorithm, Moca, which returns an ϵ-optimal policy with probability at least $1-\delta$ after $\widetilde{\mathcal{O}}(\mathcal{C}(\mathcal{M},\epsilon)\cdot\log 1/\delta)$ episodes, for finite $\delta > 0$ and $\epsilon > 0$. Rather than relying on optimism to guarantee exploration, it employs an aggressive exploration strategy which seeks to reach states of interest as quickly as possible, coupling this with a Monte Carlo estimator and action-elimination procedure to identify suboptimal actions.

- *Insufficiency of online-to-batch.* We show, through several explicit instances, that low-regret algorithms cannot achieve our proposed measure of complexity, and indeed can do arbitrarily worse. This shows that optimistic planning does not suffice to attain sharp instance-dependent PAC guarantees in tabular reinforcement learning.

A Motivating Example. Consider the MDP in Figure 1. In state s_0, action a_1 is optimal and transitions to state s_1 with probability $1-p$ and state s_2 with probability p. Action a_2 is suboptimal and transitions to state s_2 with probability 1. To learn a good policy, we need to identify the optimal action in both s_1 and s_2. An optimistic or low-regret algorithm

will primarily play a_1 in s_0, as this action is optimal, and it will therefore only reach s_2 approximately $\mathcal{O}(pK)$ times. It follows that a low-regret algorithm will take at least $\Omega(\frac{1}{p\Delta_2^2})$ episodes to learn the optimal action in s_2. In contrast, we could instead play a_2 in s_0, collecting less reward but learning the optimal action in s_2 in only $\Omega(\frac{1}{\Delta_2^2})$ episodes. For small p, this could be arbitrarily better. The following result makes this formal, illustrating that for identifying good policies in MDPs, existing low-regret and optimistic approaches can be highly suboptimal, and more intentional exploration procedures are needed.

Proposition 1 (Informal) *On the example in Figure 1, any low-regret algorithm must run for at least* $K \geq \Omega\big(\frac{\log 1/\delta}{\Delta_1^2} + \frac{\log 1/\delta}{p\Delta_2^2}\big)$ *episodes to identify the optimal policy, while* MOCA *will terminate and output the optimal policy after only* $K \leq \mathcal{O}\big(\frac{\log 1/\delta}{\Delta_1^2} + \frac{\log 1/\delta}{\Delta_2^2}\big)$ *episodes.*

We stress that our goal in this work is *not* to match the $\delta \to 0$ scaling of the optimal instance-dependent lower bound for (ϵ, δ)-PAC, but rather to obtain an instance-dependent complexity that captures the *finite-time* difficulty of learning an ϵ-optimal policy, and scales with an intuitive notion of MDP explorability, as in the example above. Even in the much simpler bandits setting, hitting the instance-optimal rate usually requires algorithms that "track" the optimal allocation, which can typically only be accomplished in the aforementioned $\delta \to 0$ limit, making such algorithms impractical in practice (Garivier and Kaufmann, 2016). In contrast to this approach, we focus on the non-asymptotic regime, avoiding mixing-time and tracking arguments, and seeking to instead obtain "practical" instance-dependence.

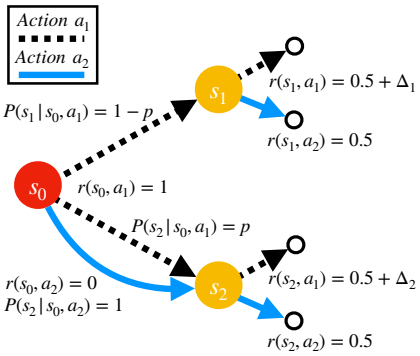

Figure 1: A motivating example

2. Related Work

The literature on PAC RL is vast and dates back at least two decades (Kearns and Singh, 2002; Kakade, 2003). We cannot do it justice here so we review only the most relevant works.

Minimax (ϵ, δ)-PAC Bounds. The vast majority of work has focused on minimax sample complexities that hold for *any* tabular MDP with bounded rewards (Lattimore and Hutter, 2012; Dann and Brunskill, 2015; Azar et al., 2017; Dann et al., 2017, 2019; Ménard et al., 2020). In addition, a PAC guarantee can be obtained from any low-regret algorithm using an online-to-batch conversion. If an algorithm has a regret bound of $\mathcal{O}(\sqrt{CK})$, one can obtain an ϵ-optimal policy with probability $1-\delta$ after $K \geq \frac{C}{\epsilon^2\delta^2}$ episodes, allowing low-regret algorithms such as (Jin et al., 2018; Zhang et al., 2020b) to solve the PAC problem as well. See (Jin et al., 2018; Ménard et al., 2020) for a more in-depth discussion of this approach.

Instance-Dependent Regret Bounds for Episodic MDPs. Optimistic planning algorithms have been shown to obtain gap-dependent regret bounds that scale as $\log(K) \cdot \sum_{s,a,h} \frac{1}{\Delta_h(s,a)}$ (Simchowitz and Jamieson, 2019; Xu et al., 2021; Dann et al., 2021). Using

the online-to-batch conversion, this gives a PAC guarantee scaling as $\sum_{s,a,h} \frac{1}{\Delta_h(s,a)\cdot\epsilon} \cdot \frac{1}{\delta^2}$. Ok et al. (2018) propose an algorithm that achieves instance-optimal regret, though it is not computationally efficient and asymptotic, $T \to \infty$. The algorithm of Zanette and Brunskill (2019), Euler, achieves a first-order style regret bound of $\sqrt{SAK \min\{\mathbb{Q}_\star H, \mathcal{G}^2\}}$, where $\mathbb{Q}_\star$ and $\mathcal{G}^2$ are problem-dependent quantities.

Towards Instance-Dependent PAC Learning. To date, only several works have derived instance-dependent PAC bounds in the non-generative setting. Jonsson et al. (2020) obtains a complexity that scales as the Q-value gap for the first time step but exponentially in H. Marjani et al. (2021) study the problem of best-policy identification ($\epsilon = 0$), and obtain an instance-dependent complexity, yet their results are asymptotic ($\delta \to 0$). Wagenmaker et al. (2021) provide an instance-optimal (ϵ, δ)-PAC algorithm, to our knowledge the only such result, yet their result holds only in certain classes of continuous state MDPs.

Generative Model Setting. In the simpler generative model setting, the agent can query any s and a and observe the next state and reward. Many minimax-style guarantees have been developed in this setting (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020; Li et al., 2020). Recently, several instance-dependent results have been shown (Zanette et al., 2019; Marjani and Proutiere, 2020; Khamaru et al., 2020, 2021). Most relevant is the work of Zanette et al. (2019) which proposes the Bespoke algorithm and achieves a sample complexity of $\sum_{s,a} \frac{\log(1/\delta)}{\max\{\epsilon^2, \Delta(s,a)^2\}}$. Note that this always scales at least as $\Omega(S/\epsilon^2)$.

Lower Bounds. We are unaware of any instance-dependent lower bound for (ϵ, δ)-PAC for MDPs. However, it is straightforward to obtain lower bounds for exact best policy identification ($\epsilon = 0$) (Marjani and Proutiere, 2020; Marjani et al., 2021), though such lower bounds are uninterpretable solutions to non-convex optimization problems. Furthermore, at present no algorithm is known to hit the best policy identification lower bound.

3. Preliminaries

Notation. We let $[N] = \{1, 2, \dots, N\}$. $\triangle(\mathcal{X})$ denotes the set of probability distributions over a set $\mathcal{X}$. $\mathbb{E}_\pi[\cdot]$ denotes the expectation over the trajectories induced by policy π and $\mathbb{P}_\pi[\cdot]$ denotes the probability measure induced by π. We let $\gtrsim$ refer to inequality up to absolute constants, and let $\mathcal{O}(\cdot)$ hide absolute constants, and $\widetilde{\mathcal{O}}(\cdot)$ hide absolute constants as well as poly log terms. In general, we use log to denote the base 2 logarithm.

Markov Decision Processes. We study finite-horizon, time inhomogeneous Markov Decision Processes (MDPs) given by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, P_0, \{R_h\}_{h=1}^H)$. Here $\mathcal{S}$ is the set of states ($S := |\mathcal{S}|$), $\mathcal{A}$ the set of actions ($A := |\mathcal{A}|$), H the horizon, $P_h : \mathcal{S} \times \mathcal{A} \to \triangle(\mathcal{S})$ the transition kernel at step h, $P_0 \in \triangle(\mathcal{S})$ the initial state distribution, and $R_h : \mathcal{S} \times \mathcal{A} \to \triangle([0,1])$ the reward distribution, with $r_h(s,a) = \mathbb{E}[R_h(s,a)]$. We assume that $\{P_h\}_{h=1}^H, P_0$, and $\{R_h\}_{h=1}^H$ are all initially unknown to the learner.

An *episode* is a trajectory $\{(s_h, a_h, R_h)\}_{h=1}^H$ where $s_1 \sim P_0$, $s_{h+1} \sim P_h(\cdot|s_h, a_h)$, and $R_h \sim R_h(s_h, a_h)$. After H steps, the MDP restarts and the process repeats. A *policy* π is a mapping from states to actions: $\pi : \mathcal{S} \times [H] \to \triangle(\mathcal{A})$. $\pi_h(a|s)$ denotes the probability that π

chooses a at (s,h). If for all (s,h), $\pi_h(a|s) = 1$ for some a, we say π is a *deterministic policy* and denote $\pi_h(s)$ the action it chooses at (s,h). Otherwise we say π is a *stochastic policy*.

Given a policy π, the Q-value function, $Q^\pi : \mathcal{S} \times \mathcal{A} \times [H] \to [0,H]$, denotes the expected reward obtained by playing action a in state s at time h, and then playing π for all subsequent time. Formally, it is defined as

$$Q_h^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{h'=h}^H R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a\right].$$

We also define the value function, $V^\pi : \mathcal{S} \times [H] \to [0,H]$, as $V_h^\pi(s) := Q_h^\pi(s, \pi_h(s))$. The Q-function satisfies the Bellman equation:

$$Q_h^\pi(s,a) = r_h(s,a) + \sum_{s'} P_h(s'|s,a) V_{h+1}^\pi(s').$$

We let $V_{H+1}^\pi(s) = 0$ and $Q_{H+1}^\pi(s,a) = 0$. We define the optimal Q-function as $Q_h^\star(s,a) := \sup_\pi Q_h^\pi(s,a)$, $V_h^\star(s) := \sup_\pi V_h^\pi(s)$, and let $\pi^\star$ denote an optimal policy. $V_0^\pi := \sum_s P_0(s) V_1^\pi(s)$ denotes the *value* of a policy, the expected reward it will obtain, and $V_0^\star := \sup_\pi V_0^\pi$.

Suboptimality Gaps. Critical to our analysis is the concept of a *suboptimality gap*:

$$\Delta_h(s,a) := V_h^\star(s) - Q_h^\star(s,a).$$

In words, $\Delta_h(s,a)$ denotes the suboptimality of taking action a in (s,h), and then playing the optimal policy henceforth. We let $\Delta_h^\pi(s,a) := \max_{a'} Q_h^\pi(s,a') - Q_h^\pi(s,a)$ and $\Delta_{\min}(s,h) := \min_{a:\Delta_h(s,a)>0} \Delta_h(s,a)$. For action a with $Q_h^\star(s,a) = \max_{a'} Q_h^\star(s,a')$, we define $\Delta_h(s,a) := \Delta_{\min}(s,h)$, so that $\Delta_h(s,a)$ is always non-zero. Throughout the remainder of the body, we make the following assumption:

Assumption 3.1 (Unique Optimal Actions) *For each (s,h), there exists a unique action, a, such that $Q_h^\star(s,a) = \max_{a'} Q_h^\star(s,a')$—each state has a unique optimal action.*

This assumption is purely for notational convenience—all our results hold for MDPs with multiple optimal actions at each state, as we show in the appendix. Finally, we introduce the idea of a *state-action visitation distribution*. We define

$$w_h^\pi(s,a) := \mathbb{P}_\pi[s_h = s, a_h = a], \quad w_h^\pi(s) := \mathbb{P}_\pi[s_h = s].$$

Note that $w_h^\pi(s,a) = \pi_h(a|s) w_h^\pi(s)$. We denote the *maximum reachability* of (s,h) by $W_h(s) := \sup_\pi w_h^\pi(s)$, the maximum probability with which we could hope to reach (s,h).

PAC Reinforcement Learning Problem. In this work we study PAC RL. Formally, in PAC RL, the goal is to, with probability $1-\delta$, identify a policy $\widehat{\pi}$ such that

$$V_0^\star - V_0^{\widehat{\pi}} \le \epsilon \tag{3.1}$$

using as few episodes as possible. We say that a policy satisfying (3.1) is *ϵ-optimal* and that an algorithm which returns a policy satisfying (3.1) with probability at least $1-\delta$ is (ϵ,δ)-PAC. Note that our goal is to find a single policy not a distribution over policies[1].

1. That is, we want to find some policy $\widehat{\pi}$ such that $V_0^\star - V_0^{\widehat{\pi}} \le \epsilon$, not a distribution over policies $\lambda \in \triangle(\Pi)$ such that $V_0^\star - \sum_{\pi\in\Pi} \lambda_\pi V_0^\pi \le \epsilon$. Note that returning a single policy is the standard goal of PAC RL found in the literature.

4. Instance-Dependent PAC Policy Identification

Before stating our main result, we introduce our new notion of sample complexity for MDPs.

Definition 4.1 (Gap-Visitation Complexity) *For a given MDP $\mathcal{M}$, we define the* gap-visitation complexity *as:*

$$\mathcal{C}(\mathcal{M}, \epsilon) := \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min \left\{ \frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

where the infimum is over all policies, both deterministic and stochastic, and:

$$\mathrm{OPT}(\epsilon) := \big\{(s,a,h) \ : \ \epsilon \geq W_h(s)\Delta_h(s,a)/3\big\}.$$

We also define the best-policy gap-visitation complexity *as:*

$$\mathcal{C}^\star(\mathcal{M}) := \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}.$$

Since $w_h^\pi(s,a) = \pi_h(a|s)w_h^\pi(s)$, as long as $w_h^\pi(s) > 0$ for some π, we can always choose our policy such that all actions are supported and $w_h^\pi(s,a) > 0$ for all a[2]. Recall that we have defined $\Delta_h(s,a)$ so that $\Delta_h(s,a) > 0$ for all (s,a,h), implying that as $\epsilon \to 0$, $|\mathrm{OPT}(\epsilon)| \to 0$. Given this new notion of sample complexity, we are now ready to state our main result.

Theorem 2 *There exists an (ϵ,δ)-PAC algorithm,* Moca*, which with probability at least $1-\delta$ terminates after running for at most*

$$\mathcal{C}(\mathcal{M},\epsilon) \cdot H^2 c_\epsilon \log\tfrac{1}{\delta} + \tfrac{C_{\mathrm{LOT}}(\epsilon)}{\epsilon}$$

episodes and returns an ϵ-optimal policy, for lower-order term $C_{\mathrm{LOT}}(\epsilon) = \mathrm{poly}(S,A,H,\log\frac{1}{\epsilon},\log\frac{1}{\delta})$ and $c_\epsilon = \mathrm{poly}\log(SAH/\epsilon)$. Furthermore, if $\epsilon < \epsilon^\star := \min\{\min_{s,a,h} W_h(s)\Delta_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$, Moca *terminates after at most*

$$\mathcal{C}^\star(\mathcal{M}) \cdot H^2 c_{\epsilon^\star} \log\tfrac{1}{\delta} + \tfrac{C_{\mathrm{LOT}}(\epsilon^\star)}{\epsilon^\star}$$

episodes and returns $\pi^\star$, the optimal policy, with probability $1-\delta$.

In addition, Moca is computationally efficient with computational cost scaling polynomially in problem parameters. In Section 6, we provide a sketch of the proof of Theorem 2 and state the definition Moca. The full proof is deferred to Appendix C.

2. Here, we adopt the convention that, in the trivial case $W_h(s) = 0$ (and thus $w_h^\pi(s,a) = 0$), $\frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}$ evaluates to 0.

4.1. Interpreting the Complexity

Intuitively, the first term in the gap-visitation complexity quantifies how quickly we can eliminate all actions at least $\epsilon/W_h(s)$-suboptimal for all s and h, given that we must explore in our particular MDP. For a given s and h, if we play policy π for K episodes, we will reach (s,h) on average $w_h^\pi(s)K$ times. Thus, if we imagine that there is a bandit at (s,h), to eliminate action a will require that we run for at least $\frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}$ episodes. The following result makes this rigorous—up to H factors, a complexity of $\mathcal{O}(\mathcal{C}^\star(\mathcal{M})\cdot\log 1/\delta)$, which MOCA achieves, cannot be improved on in general for best-policy identification.

Proposition 3 *Fix some $S>1, A>1, H>1$, $\bar{h}\in[H]$, transition kernels $\{P_h\}_{h=1}^{\bar{h}-1}$, and gaps $\{\mathsf{gap}(s,a)\}_{s\in[S],a\in[A-1]}\subseteq(0,1/2)^{SA}$. Then there exists some MDP $\mathcal{M}$ with S states, A actions, horizon H, transition kernel P_h for $h\le\bar{h}-1$, and gaps*

$$\Delta_{\bar{h}}(s,a)=\mathsf{gap}(s,a),\quad \forall s\in\mathcal{S},a\in\mathcal{A},a\neq\pi^\star_{\bar{h}}(s),\qquad \Delta_h(s,a)\ge 1,\quad \forall s\in\mathcal{S},a\in\mathcal{A},h\neq\bar{h},$$

such that any $(0,\delta)$-PAC algorithm with stopping time K_δ requires:

$$\mathbb{E}_{\mathcal{M}}[K_\delta]\gtrsim \inf_\pi \max_{s,a}\frac{1}{w^\pi_{\bar{h}}(s,a)\Delta_{\bar{h}}(s,a)^2}\cdot\log\frac{1}{2.4\delta}.$$

In this instance, as $\Delta_h(s,a)\ge 1$ for $h\neq\bar{h}$, assuming $\{P_h\}_{h=1}^{\bar{h}-1}$ is chosen such that $W_h(s)$ is not too small for each s and $h\le\bar{h}$, we will have that $\mathcal{C}^\star(\mathcal{M})=\mathcal{O}(\inf_\pi\max_{s,a}\frac{1}{w^\pi_{\bar{h}}(s,a)\Delta_{\bar{h}}(s,a)^2})$, so Proposition 3 implies that we must have $\mathbb{E}_{\mathcal{M}}[K_\delta]\ge\Omega(\mathcal{C}^\star(\mathcal{M})\cdot\log 1/\delta)$, matching the upper bound given in Theorem 2 up to H factors.

The second term in $\mathcal{C}(\mathcal{M},\epsilon)$, $H^2|\mathrm{OPT}(\epsilon)|/\epsilon^2$, captures the complexity of ensuring that, after eliminating $\epsilon/W_h(s)$-suboptimal actions, sufficient exploration is performed to guarantee the returned policy is ϵ-optimal. While this will be no worse than H^3SA/ϵ^2, it could be much better, if in our MDP the number of (s,a,h) with $\Delta_h(s,a)\lesssim\epsilon/W_h(s)$ is small (note that since $\Delta_h(s,a)\ge\Delta_{\min}(s,h)$ by definition, $\mathrm{OPT}(\epsilon)$ will only contain states for which the minimum *non-zero* gap is less than $\epsilon/W_h(s)$). We next obtain the following bounds on $\mathcal{C}(\mathcal{M},\epsilon)$, providing an interpretation of $\mathcal{C}(\mathcal{M},\epsilon)$ in terms of the maximum reachability, and illustrating $\mathcal{C}(\mathcal{M},\epsilon)$ is no larger than the minimax optimal complexity. This implies MOCA is nearly worst-case optimal, matching the lower bound of $\Omega(\frac{SAH^2}{\epsilon^2}\cdot\log 1/\delta)$ from Dann and Brunskill (2015) up to H and log factors[3].

Proposition 4 *The following bounds hold:*

1. $\mathcal{C}(\mathcal{M},\epsilon)\le\frac{H^3SA}{\epsilon^2}$
2. $\mathcal{C}(\mathcal{M},\epsilon)\le\sum_{h=1}^H\sum_{s,a}\min\{\frac{1}{W_h(s)\Delta_h(s,a)^2},\frac{W_h(s)}{\epsilon^2}\}+\frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$
3. $\mathcal{C}(\mathcal{M},\epsilon)\le\sum_{h=1}^H\sum_{s,a}\frac{1}{\epsilon\max\{\Delta_h(s,a),\epsilon\}}+\frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$.

3. This lower bound is for the stationary setting. As noted in Ménard et al. (2020), one would expect a lower bound of $\Omega(\frac{SAH^3}{\epsilon^2}\cdot\log 1/\delta)$ in the non-stationary setting, implying MOCA is H^2 off the lower bound.

In the special case of multi-armed and contextual bandits, the gap-visitation complexity simplifies considerably.

Proposition 5 *If $\mathcal{M}$ is a multi-armed bandit, then*

$$\mathcal{C}(\mathcal{M}, \epsilon) = \sum_a \min\left\{\frac{1}{\Delta(a)^2}, \frac{1}{\epsilon^2}\right\}, \quad \mathcal{C}^\star(\mathcal{M}) = \sum_a \frac{1}{\Delta(a)^2}.$$

Furthermore, if $\mathcal{M}$ is a contextual bandit, then

$$\mathcal{C}^\star(\mathcal{M}) = \max_s \frac{1}{W(s)} \sum_a \frac{1}{\Delta(s,a)^2}.$$

The values given here are known to be the optimal problem-dependent constants for both best arm identification and (ϵ, δ)-PAC for multi-armed bandits (Kaufmann et al., 2016; Degenne and Koolen, 2019). To our knowledge, the lower bound for best-policy identification in contextual bandits has never been formally stated, yet it is obvious it will take the form of $\mathcal{C}^\star(\mathcal{M})$ given here. It follows that in the special cases of multi-armed bandits and contextual bandits, MOCA is instance-optimal, up to logarithmic factors and lower-order terms.

Several additional interpretations of the gap-visitation complexity are given in Appendix A.2. The above results show that the gap-visitation complexity cleanly interpolates between the worst-case optimal rate for (ϵ, δ)-PAC, and, in certain MDPs, the instance-optimal rate for best-policy identification. In between these extremes, it captures an intuitive sense of instance-dependence. As we will show in the following section, this instance-dependence can offer significant improvements over worst-case optimal approaches.

Remark 4.1 (Comparison to Marjani et al. (2021)) *Our notion of best-policy gap-visitation complexity is closely related to the measure of complexity introduced in Marjani et al. (2021), though they study the infinite-horizon, discounted case. Notably, however, their analysis only considers best-policy identification ($\epsilon = 0$) and is purely asymptotic ($\delta \to 0$), while ours holds for $\delta > 0$ and $\epsilon > 0$. Further, our best-policy gap-visitation complexity offers a non-trivial improvement over their complexity, scaling as $(\min_s w_h^\pi(s,a)\Delta_{\min}(s,h)^2)^{-1}$ instead of $(\min_s \Delta_{\min}(s)^2 \cdot \min_s w^\pi(s, \pi^\star(s)))^{-1}$ which Marjani et al. (2021) obtains.*

Remark 4.2 (Dependence on $\log 1/\delta$) *While the leading term in the sample complexity of* MOCA *only scales as $\log 1/\delta$, the lower order term scales as a suboptimal $\log^3 1/\delta$. These additional factors of $\log 1/\delta$ are due to the regret-minimization algorithm used in the exploration procedure we employ. We show in Remark D.2 that it can be improved to $\log 1/\delta \cdot \log\log 1/\delta$ and leave completely removing the suboptimal δ scaling for future work.*

Remark 4.3 (Improving H Dependence) *As noted above,* MOCA *attains a worst-case H dependence that is a factor of H^2 worse than the lower bound. Our analysis relies on Hoeffding's inequality to argue about the concentration of our estimate of $Q_h^{\widehat{\pi}}(s,a)$. Rather than depending on the variance of the next-state value function, our confidence interval therefore depends on H^2, an upper bound on the variance. If desired, we could instead employ an empirical Bernstein-style inequality (Maurer and Pontil, 2009), which would allow us to replace this H^2 scaling with the variance of the reward obtained from playing a at (s,h) and then playing $\widehat{\pi}$. We believe that this modification may allow us to refine the H dependence of* MOCA. *As the focus of this work is obtaining an instance-dependent complexity, we leave the details of this for future work.*

5. Low-Regret Algorithms are Suboptimal for PAC

Using our instance-dependence complexity, we next show that running a low-regret algorithm and applying an online-to-batch conversion can be very suboptimal for PAC RL. We first define a low-regret algorithm and our learning protocol:

Definition 5.1 (Low-Regret Algorithm) *We say an algorithm $\mathcal{R}$ is a* low-regret algorithm *if it has expected regret bounded as* $\mathrm{Regret}(K) = \sum_{k=1}^{K} \mathbb{E}_{\mathcal{R}}[V_0^\star - V_0^{\pi_k}] \leq C_1 K^\alpha + C_2$, *for some constants* C_1, C_2, $\alpha \in (0,1)$, *and where* π_k *is the policy* $\mathcal{R}$ *plays at episode* k.

Protocol 5.1 (Low-Regret to PAC) *We consider the following procedure:*

1. *Learner runs* $\mathcal{R}$ *satisfying Definition 5.1 for* K *episodes, collects data* $\mathfrak{D}_{\mathcal{R}}(K)$.
2. *Using* $\mathfrak{D}_{\mathcal{R}}(K)$ *any way it wishes, the learner proposes a (possibly stochastic) policy* $\widehat{\pi}$.

Note that the setting considered in Proposition 1 is precisely that considered here. We now present an additional instance class where any learner following Protocol 5.1 with a low regret algorithm $\mathcal{R}$ is provably suboptimal.

Instance Class 5.1 *Given a number of states* $S \in \mathbb{N}$, *consider an MDP with horizon* $H = 2$, S *states, and* $S+1$ *actions, defined as in Figure 2.*

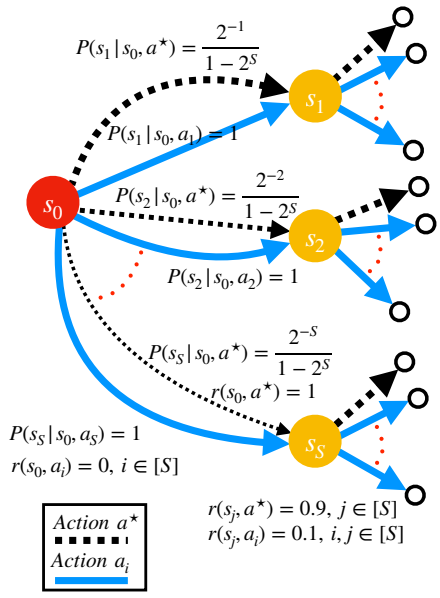

Figure 2: MDP from Instance Class 5.1

Similar to the example considered in Proposition 1, here $a^\star$ is the optimal action in every state, yet in state s_0, taking action a_i is much more informative. The following result shows that this structure results in poor performance for low-regret algorithms.

Proposition 6 (Informal) *For the MDP in Instance Class 5.1 with* S *states and small enough* ϵ, *to find an* ϵ*-optimal policy with probability* $1-\delta$ *any learner executing Protocol 5.1 with a low-regret algorithm satisfying Definition 5.1 must collect at least* $\Omega(\frac{S \log 1/\delta}{\epsilon})$ *episodes. In contrast, on this example* $\mathcal{C}^\star(\mathcal{M}) = \mathcal{O}(S^2)$ *and* $\epsilon^\star = 1/3$, *so, for* $\epsilon \leq 1/3$, *with probability* $1-\delta$, MOCA *terminates and output* $\pi^\star$ *in* $\widetilde{\mathcal{O}}(\mathrm{poly}(S))$ *episodes.*

In particular, this example shows that there is an exponential separation between low-regret algorithms and MOCA. For exponentially small ϵ, learning the optimal policy following Protocol 5.1 takes $\widetilde{\Omega}(2^S)$ samples, yet MOCA finds the optimal policy in $\widetilde{\mathcal{O}}(\mathrm{poly}(S))$ samples.

Proposition 6 implies that the true complexity of finding a good policy is often much smaller than the complexity of finding a good policy *given that we explore to minimize regret*. The key piece in this example, and the example of Proposition 1, is that the optimal

action in the initial state is very uninformative—if we want to learn the optimal action in a *subsequent* state, we should not take the optimal action in the initial state, but should instead take an action that leads us to the subsequent state with high probability. Nearly all existing works rely on algorithms which play policies which *converge to a good policy*. For instance-dependent PAC RL, instead of *playing* good policies, our examples show that an algorithm ought to explore efficiently, possibly taking very suboptimal actions in the process, ultimately *recommending* a good policy. This shortcoming of greedy algorithms motivates our design of MOCA, where we seek to incorporate this insight.

While it is known that low-regret algorithms are minimax optimal for PAC RL, these instances show that running a low-regret algorithm and then an online-to-batch procedure is suboptimal by an arbitrarily large factor for PAC RL. We conclude that minimax optimality is far from being the complete story for PAC RL, and that if our goal is to simply identify a good policy, we can do much better than running a low-regret algorithm.

Remark 5.1 (Performance of Optimistic Algorithms) *Optimistic algorithms that rely on standard bonuses will also achieve low regret. This implies that recent works specifically targeting PAC bounds such as (Dann et al., 2019; Ménard et al., 2020), which rely on optimism, will also fail to hit the optimal instance-dependent rate, or a rate of $\mathcal{O}(\mathcal{C}(\mathcal{M}, \epsilon))$. In addition, even works such as Xu et al. (2021) which do not explicitly rely on the principle of optimism and do not have known $\mathcal{O}(T^{\alpha})$-style regret bounds can also be shown to fail on our examples as they only take actions which may be optimal.*

6. Algorithm and Techniques

We turn now to the definition of our algorithm, MOCA, and sketch out the proof of Theorem 2. A full algorithm definition is given in Appendix A.3 and detailed proof in Appendix C.

6.1. Compounding Errors

In a standard multi-armed bandit, the value of a particular action is determined solely by the environment. However, in an MDP, the value of an action depends not only on the environment, but also on the actions the learner chooses to play in subsequent steps. If we run a policy $\widehat{\pi}$ after reaching some (s, h), though we may be able to identify the optimal action to play at (s, h) *given that we then play* $\widehat{\pi}$, if $\widehat{\pi}$ is suboptimal, this action may also be suboptimal. The following result, a refinement of the celebrated performance-difference lemma (Kakade, 2003), is a key piece in our analysis, allowing us to effectively handle this compounding nature of errors, and may be of independent interest.

Proposition 7 *Assume that for each h and s, $\widehat{\pi}$ plays an action which is $\epsilon_h(s)$-suboptimal with respect to $\widehat{\pi}$. That is, $\max_a Q_h^{\widehat{\pi}}(s, a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \le \epsilon_h(s)$. Then the suboptimality of $\widehat{\pi}$ is bounded as: $V_0^\star - V_0^{\widehat{\pi}} \le \sum_{h=1}^H \sup_\pi \sum_s w_h^\pi(s)\epsilon_h(s)$.*

Proposition 7 shows it is sufficient to learn actions that perform well *as compared to the best actions one could take given that $\widehat{\pi}$ is played in subsequent steps*. This observation motivates the basic premise of our algorithm: we treat every state as an individual bandit, and run an action elimination-style algorithm at each state (Even-Dar et al., 2006) to shrink $\epsilon_h(s)$.

6.2. Efficient Exploration

By Proposition 7, (s,h) adds at most $\sup_\pi w_h^\pi(s)\epsilon_h(s) = W_h(s)\epsilon_h(s)$ to the total suboptimality. If we play the policy achieving $w_h^\pi(s) = W_h(s)$ for K episodes, we will reach (s,h) $\mathcal{O}(W_h(s)K)$ times. It follows that we need $K \gtrsim \Omega(\frac{1}{W_h(s)\epsilon_h(s)^2})$ to learn an $\epsilon_h(s)$-optimal action at (s,h). If we set $\epsilon_h(s) \sim \epsilon/W_h(s)$, the suboptimality of our policy will be proportional to ϵ and we only need $K \gtrsim \Omega(\frac{W_h(s)}{\epsilon^2})$: the difficulty of reaching a state is balanced by the fact that hard-to-reach states do not contribute significantly to the suboptimality.

Navigating the MDP by Grouping States. Naively performing the above strategy results in a worst-case sample complexity suboptimal in its dependence on S. To overcome this, we propose an exploration procedure which *groups states*—rather than exploring states individually, it seeks to reach any number of states which are "nearby", in the sense that a single policy may reach any of them with similar probability.

To make this practical, we take inspiration from the "reward-free" algorithm of Zhang et al. (2020a), itself inspired by the classical Rmax algorithm (Brafman and Tennenholtz, 2002). We create an augmented reward function which gives a reward of "1" to any (s,a) we wish to visit, and "0" otherwise. We then run a (variance-sensitive) regret minimization algorithm, Euler (Zanette and Brunskill, 2019), on this modified reward function to generate a set of policies that can effectively traverse the MDP to visit the desired states. The resulting algorithm, `Learn2Explore`, takes as input some $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, the (s,a) pairs we wish to visit, and returns a partition $\{\mathcal{X}_j\}_j$ of $\mathcal{X}$, a collection of policies $\{\Pi_j\}_j$, and values $\{N_j\}_j$ satisfying the following guarantee:

Theorem 8 (Performance of `Learn2Explore`, informal) *With high probability, the partition $\{\mathcal{X}_j\}_j$ returned by* `Learn2Explore` *satisfies $\sup_\pi \sum_{(s,a)\in\mathcal{X}_j} w_h^\pi(s,a) \le 2^{-j+1}$. Moreover, the policy classes Π_j are such that, by executing a single trajectory of each $\pi \in \Pi_j$ once, we visit every $(s,a) \in \mathcal{X}_j$ at least $\frac{1}{2}N_j$ times with high probability, where*

$$N_j = \mathcal{O}\big(S^3A^2H^4\log^3\tfrac{1}{\delta}/|\mathcal{X}\setminus\cup_{j'=1}^{j-1}\mathcal{X}_{j'}|\big), \quad |\Pi_j| = \mathcal{O}(2^jS^3A^2H^4\log^3 1/\delta)$$

Furthermore, `Learn2Explore` *terminates after at most* $\mathrm{poly}(S,A,H,\log\frac{1}{\delta\epsilon})\cdot\frac{1}{\epsilon}$ *episodes.*

In other words, the sets $\mathcal{X}_j$ are groupings of "nearby" states that are increasingly difficult to reach, and the sets Π_j give a policy cover which navigates to each $(s,a) \in \mathcal{X}_j$. If we wish to collect n samples from each $(s,a) \in \mathcal{X}_j$, it will only require running for $\mathcal{O}(|\Pi_j|n/N_j) = \mathcal{O}(2^j|\mathcal{X}\setminus\cup_{j'=1}^{j-1}\mathcal{X}_{j'}|\cdot n) \le \mathcal{O}(2^jSAn)$ episodes.

6.3. Eliminating Suboptimal Actions

`Moca-SE`, our primary subroutine, combines both of these insights to eliminate suboptimal actions and determine an ϵ-optimal $\widehat{\pi}$. `Moca-SE` proceeds backwards in h, first learning near-optimal actions in every (s,H), which gives us $\widehat{\pi}_H$. More generally, at step h, given some policy $\{\widehat{\pi}_{h'}\}_{h'=h+1}^H$, we take an action a and then play $\{\widehat{\pi}_{h'}\}_{h'=h+1}^H$. The total reward obtained, a *Monte Carlo* rollout of $\{\widehat{\pi}_{h'}\}_{h'=h+1}^H$, is an unbiased estimate of $Q_h^{\widehat{\pi}}(s,a)$. Using these rollouts to estimate $Q_h^{\widehat{\pi}}(s,a)$, we eliminate actions suboptimal with respect to $\widehat{\pi}$.

Algorithm 1 Monte Carlo Action Elimination - Single Epoch (`MoCA-SE`(ϵ, δ, `FinalRound`))

input: tolerance ϵ, confidence δ, final round flag `FinalRound`
for each (s,h) **do** // loop over all s,h to learn maximum reachability
  Attempt to reach (s,h), set $\widehat{W}_h(s)$ to estimate of $W_h(s)$
set $\ell_\epsilon \leftarrow \lceil \log \frac{H}{\epsilon} \rceil$, $\widehat{\pi}_h(s) \leftarrow$ arbitrary action, $\mathcal{A}_h^0(s) \leftarrow \mathcal{A}$, $\mathcal{Z}_h \leftarrow$ reachable states
for $h = H, H-1, \dots, 1$ **do** // loop over horizon
  for $i = 1, 2, \dots, \lceil \log \frac{64}{H^2 S \epsilon} \rceil$ **do** // loop over estimated maximum reachability
    $\mathcal{Z}_{hi} \leftarrow \{s \in \mathcal{Z}_h \ : \ \widehat{W}_h(s) \in [2^{-i}, 2^{-i+1}]\}$
    for $\ell = 1, \dots, \ell_\epsilon$ **do** // loop over tolerance ϵ_ℓ
      $\epsilon_\ell \leftarrow H 2^{-\ell}$, $\mathcal{Z}_{hi}^\ell \leftarrow \{(s,a) \ : \ s \in \mathcal{Z}_{hi}, a \in \mathcal{A}_h^{\ell-1}(s), |\mathcal{A}_h^{\ell-1}(s)| > 1\}$
      Run `Learn2Explore` to collect $\widetilde{\mathcal{O}}(H^2 \widehat{W}_h(s)^2/\epsilon_\ell^2)$ samples from $\forall (s,a) \in \mathcal{Z}_{hi}^\ell$
      For all $s \in \mathcal{Z}_{hi}$, remove actions from $\mathcal{A}_h^{\ell-1}(s)$ that are $\mathcal{O}(\epsilon_\ell / \widehat{W}_h(s))$-suboptimal
  if `FinalRound` is `true` **then** // ensure $\widehat{\pi}$ ϵ-optimal
    $\mathcal{Z}_h^{\ell_\epsilon+1} \leftarrow \{(s,a) \ : \ s \in \mathcal{Z}_h, a \in \mathcal{A}_h^{\ell_\epsilon}(s), |\mathcal{A}_h^{\ell_\epsilon}(s)| > 1\}$
    Run `Learn2Explore` to collect $\widetilde{\mathcal{O}}(H^4 2^{-2j(s)}/\epsilon^2)$ samples from $\forall (s,a) \in \mathcal{Z}_h^{\ell_\epsilon+1}$,
      where $2^{-j(s)}$ is the "group reachability" of s
    For all $s \in \mathcal{Z}_h$, remove actions from $\mathcal{A}_h^{\ell_\epsilon}(s)$ that are $\widetilde{\mathcal{O}}(2^j \epsilon / H)$-suboptimal.
  else
    $\mathcal{A}_h^{\ell_\epsilon+1}(s) \leftarrow \mathcal{A}_h^{\ell_\epsilon}(s)$ for all $s \in \mathcal{Z}_h$
  Set $\widehat{\pi}_h(s)$ to any action in $\mathcal{A}_h^{\ell_\epsilon+1}(s)$ for all $s \in \mathcal{Z}_h$
return $\widehat{\pi}$, $\max_{s,h} |\mathcal{A}_h^{\ell_\epsilon+1}(s)|$

Proposition 7 then allows us to relate the local suboptimality of an action, $\max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s))$, to the global suboptimality of $\widehat{\pi}$. Critical to making this procedure efficient, we apply `Learn2Explore` to reach states for which we have not identified the optimal action. The following results show that this procedure is able to efficiently eliminate $\epsilon_\ell / W_h(s) = H 2^{-\ell} / W_h(s)$-suboptimal actions and that its complexity is bounded by a quantity reminiscent of $\mathcal{C}(\mathcal{M}, \epsilon)$.

Lemma 6.1 (Informal) *With high probability, any $a \in \mathcal{A}_h^\ell(s)$ satisfies $\Delta_h(s,a) \lesssim \frac{\epsilon_\ell}{W_h(s)}$.*

Lemma 6.2 (Informal) *With high probability, for a given value of h and i, the number of episodes the loop over ℓ on Line 8 collects is at most*

$$\widetilde{\mathcal{O}}\left(H^2 \inf_\pi \max_{s \in \mathcal{Z}_{hi}} \max_a \min \left\{ \frac{1}{w_h^\pi(s,a) \widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a) \epsilon^2} \right\} \right). \tag{6.1}$$

Proof Sketch We first call `Learn2Explore` with input set $\mathcal{X} = \mathcal{Z}_{hi}^\ell$. Let $\mathcal{X}_{hij}^\ell$ denote the partition of $\mathcal{Z}_{hi}^\ell$ returned. To collect $\widetilde{\mathcal{O}}(H^2 \widehat{W}_h(s)^2/\epsilon_\ell^2)$ samples from each $(s,a) \in \mathcal{X}_{hij}^\ell$, Theorem 8 implies that it suffices to run for approximately $\widetilde{\mathcal{O}}(2^j |\mathcal{X}_{hij}^\ell| \cdot H^2 \widehat{W}_h(s)^2/\epsilon_\ell^2)$ episodes; implying a total complexity of $\widetilde{\mathcal{O}}(\sum_j 2^j |\mathcal{X}_{hij}^\ell| \cdot H^2 \widehat{W}_h(s)^2/\epsilon_\ell^2)$. Theorem 8 also gives

$$\sup_\pi \min_{(s,a) \in \mathcal{X}_{hij}^\ell} |\mathcal{X}_{hij}^\ell| w_h^\pi(s,a) \le \sup_\pi \textstyle\sum_{(s,a) \in \mathcal{X}_{hij}^\ell} w_h^\pi(s,a) \le 2^{-j+1}.$$

Rearranging this gives that

$$2^j|\mathcal{X}^\ell_{hij}| \lesssim \inf_\pi \max_{(s,a)\in\mathcal{X}^\ell_{hij}} \frac{1}{w^\pi_h(s,a)}.$$

Using that $\widehat{W}_h(s) \approx W_h(s)$, and since actions in stage ℓ are only active if $\Delta_h(s,a) \lesssim \frac{W_h(s)}{\epsilon_\ell}$, we see that $\widetilde{\mathcal{O}}(\sum_j 2^j|\mathcal{X}^\ell_{hij}| \cdot H^2\widehat{W}_h(s)^2/\epsilon_\ell^2)$ can be bounded by (6.1).

■

6.4. Putting Everything Together: Moca

Algorithm 2 **MO**nte **C**arlo **A**ction Elimination (Moca)

1: **input:** tolerance ϵ, confidence δ
2: **for** $m = 1, \dots, \lceil\log(H/\epsilon)\rceil - 1$ **do**
3: $\quad\widehat{\pi}^m$, `MaxOpt` $\leftarrow$ `Moca-SE`$(H2^{-m}, \frac{\delta}{36m^2}$, `false`$)$
4: $\quad$**if** `MaxOpt` $= 1$ **then** $\widehat{\pi}^m$ is optimal, **return** $\widehat{\pi}^m$
5: **return** `Moca-SE`$(\epsilon, \frac{\delta}{36\lceil\log(H/\epsilon)\rceil^2}$, `true`$)$

Our main algorithm, Moca, calls `Moca-SE` multiple times with geometrically decreasing tolerance ϵ'. When run with $\epsilon' < \epsilon$ it sets `FinalRound` = `false`. If `Moca-SE` is able to identify the optimal action in each (s,h), thereby identifying $\pi^\star$, Moca simply terminates and output $\pi^\star$. However, if this does not occur, on the final call to `Moca-SE`, when $\epsilon' \leftarrow \epsilon$, we set `FinalRound` = `true`, which triggers an additional round of exploration necessary to guarantee $\widehat{\pi}$ is ϵ-optimal. Critically, while in the first stage we only sample (s,a) in proportion to the maximum reachability of s, in this stage we sample each (s,a) in proportion with the *reachability of the partition containing* (s,a). Combining Theorem 8 with our choice for the number of samples taken in this final round, we obtain the following guarantees.

Lemma 6.3 (Informal) *Any $a \in \mathcal{A}_h^{\ell_\epsilon+1}(s)$ satisfies $\Delta^{\widehat{\pi}}_h(s,a) \le \mathcal{O}(\frac{\epsilon}{H\cdot 2^{-j(s)+1}})$, where $j(s)$ denotes the index of the partition returned by* Learn2Explore *which contains s.*

Lemma 6.4 (Informal) *If* `Moca-SE` *is run with* ***`FinalRound`*** $=$ ***`true`****, the procedure within the if statement on Line 12 terminates after at most $\widetilde{\mathcal{O}}(H^4|\mathcal{Z}_h^{\ell_\epsilon+1}|/\epsilon^2)$ episodes.*

Using Lemma 6.1, one can show that $\cup_h \mathcal{Z}_h^{\ell_\epsilon+1} \subseteq \mathrm{OPT}(\epsilon)$, so $\sum_h |\mathcal{Z}_h^{\ell_\epsilon+1}| \le |\mathrm{OPT}(\epsilon)|$. A simple calculation combining Lemma 6.3 and Proposition 7 gives that this exploration is sufficient to guarantee $\widehat{\pi}$ is ϵ-optimal.

Lemma 6.5 (Informal) *With high probability, if* `Moca-SE` *is run with* ***`FinalRound`*** $=$ ***`true`****, it returns a policy $\widehat{\pi}$ which is ϵ-optimal.*

Proof Sketch By Proposition 7, we can bound the suboptimality of $\widehat{\pi}$ as:

$$V_0^\star - V_0^{\widehat{\pi}} \le \sum_{h=1}^H \sup_\pi \sum_s w^\pi_h(s)\epsilon_h(s) \le \sum_{h=1}^H \sum_j \sup_\pi \sum_{s\in\mathcal{X}^{\ell_\epsilon+1}_{hj}} w^\pi_h(s)\epsilon_h(s).$$

By Lemma 6.3, $\epsilon_h(s) \le \frac{\epsilon}{H \cdot 2^{-j+1}}$ for all $s \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$. By Theorem 8, $\sup_\pi \sum_{s \in \mathcal{X}_{hj}^{\ell_\epsilon+1}} w_h^\pi(s) \le 2^{-j}$. Thus, the above is bounded as $\sum_{h=1}^H \sum_j \frac{\epsilon}{H \cdot 2^{-j+1}} \cdot 2^{-j} \lesssim \epsilon$, which proves the result. ■

7. Conclusion

In this work, we proposed a new instance-dependent measure of complexity for PAC RL, the gap-visitation complexity, showed that our algorithm, MOCA, hits this complexity, and, through several examples, showed that running a low-regret procedure cannot be instance-optimal for PAC RL. Our work opens several interesting directions for future work.

- While the gap-visitation complexity takes into account the maximum reachability of a given state, it does not take into account how easily a given state may be reached by a near-optimal policy. One could imagine an MDP where some state, s, is easily reached by a suboptimal policy but is never visited by near-optimal policies. In this case, a PAC algorithm need not learn a good action in this state to return an ϵ-optimal policy, yet MOCA currently would do so. We believe that this idea—weighting states during exploration not by their maximum visitation but by their visitation from near-optimal policies—could be incorporated into our current framework, but leave the details of this to future work.
- Neither this work nor Marjani et al. (2021) hit the true instance-optimal lower bound which, as shown in Marjani et al. (2021), is the solution to a non-convex optimization problem even for best-policy identification. The above discussion suggests that $\mathcal{C}(\mathcal{M}, \epsilon)$ is not in general the instance-dependent lower bound, though Proposition 3 and Proposition 5 show that in certain cases it does match the instance-dependent lower bound. Relating $\mathcal{C}(\mathcal{M}, \epsilon)$ to the true lower bound in general and developing algorithms that hit the lower bound would both be interesting directions for future work.
- By running an algorithm that achieves gap-dependent logarithmic regret (such as Simchowitz and Jamieson (2019)) and performing an online-to-batch conversion, one can obtain a PAC sample complexity of

$$\mathcal{O}\Bigg(\sum_{s,a,h:\Delta_h(s,a)>0} \frac{1}{\Delta_h(s,a)\epsilon} \cdot \frac{1}{\delta^2} \Bigg). \tag{7.1}$$

  While Proposition 4 shows that MOCA achieves a similar complexity, albeit with a $\log 1/\delta$ scaling, it must also pay for the $\frac{|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$ term, which could dominate the $\frac{1}{\Delta_h(s,a)\epsilon}$ term. We believe removing this term (or showing it is necessary) and obtaining a sample complexity of the form (7.1) but that scales instead with $\log 1/\delta$ is an important step in understanding the true complexity of PAC reinforcement learning.

Acknowledgements

The work of AW is supported by an NSF GFRP Fellowship DGE-1762114. MS is generously supported by an Open Philanthropy AI Fellowship. The work of KJ was funded in part by the AFRL and NSF TRIPODS 2023166.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

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.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *arXiv preprint arXiv:1510.08906*, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *arXiv preprint arXiv:1902.03475*, 2019.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *arXiv preprint arXiv:2006.05879*, 2020.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

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.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

Koulik Khamaru, Eric Xia, Martin J Wainwright, and Michael I Jordan. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced q-learning. *arXiv preprint arXiv:2106.14352*, 2021.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33, 2020.

Aymen Al Marjani and Alexandre Proutiere. Best policy identification in discounted mdps: Problem-specific sample complexity. *arXiv preprint arXiv:2009.13405*, 2020.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *arXiv preprint arXiv:2106.02847*, 2021.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *arXiv preprint arXiv:1806.00775*, 2018.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018.

Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*, 2019.

Alexandre B Tsybakov. Introduction to nonparametric estimation., 2009.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. *arXiv preprint arXiv:2102.05214*, 2021.

Haike Xu, Tengyu Ma, and Simon S Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. *arXiv preprint arXiv:2102.04692*, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

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*, volume 32. Curran Associates, Inc., 2019.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020a.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020b.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems 26*, 2013.

Contents

Appendix A. Full Algorithm Description and Complexity

We turn now to providing several additional interpretations of the gap-visitation complexity in Appendix A.2, and a full description of Moca in Appendix A.3. First, however, we relax the assumption that each state has a unique optimal action, Assumption 3.1, in Appendix A.1.

A.1. Non-Unique Optimal Actions

Towards relaxing Assumption 3.1, we construct an effective gap, $\widetilde{\Delta}_h(s,a)$, which coincides with $\Delta_h(s,a)$ for states where the optimal action is unique, but could be 0 for states where the optimal action is non-unique. Formally, the effective gap is defined as follows:

$$\widetilde{\Delta}_h(s,a) := \begin{cases} \Delta_h(s,a) & a \text{ is a suboptimal action} \\ \Delta_{\min}(s,h) & a \text{ is the unique action at } s,h \text{ for which } \Delta_h(s,a) = 0 \\ 0 & a \text{ is a non-unique action at } s,h \text{ for which } \Delta_h(s,a) = 0. \end{cases}$$

We can then define the gap-visitation complexity in terms of the effective gap:

$$\mathcal{C}(\mathcal{M},\epsilon) := \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

for:

$$\mathrm{OPT}(\epsilon) := \left\{(s,a,h) \;:\; \epsilon \geq W_h(s)\widetilde{\Delta}_h(s,a)/3\right\}.$$

Note that this definition coincides with the definition of $\mathcal{C}(\mathcal{M},\epsilon)$ given in Definition 4.1 in the case when optimal actions are unique in each state. Theorem 2 holds identically with this modified definition of the gap-visitation complexity, as do Proposition 4 and Proposition 5.

Note that the best-policy gap-visitation complexity does not have a natural analogue in the case when some state has a non-unique optimal action. As the best-policy gap-visitation complexity corresponds to the complexity of finding the optimal policy, and as it is not possible to guarantee the optimal action has been found if there are multiple optimal actions, in the case of best-policy identification, we still assume that the MDP has unique optimal actions in each state.

For the remainder of the appendix, we will consider MDPs that may not have unique optimal actions, and as such, will use the effective gap throughout.

A.2. Interpreting $\mathcal{C}(\mathcal{M},\epsilon)$

Proposition 9 *The gap-visitation complexity, $\mathcal{C}(\mathcal{M},\epsilon)$, satisfies*

$$\mathcal{C}(\mathcal{M},\epsilon) = \sum_{h=1}^{H} \inf_{\pi} \max_{s} \frac{1}{w_h^\pi(s)} \sum_{a} \min\left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

Furthermore, when $\mathcal{M}$ has unique optimal actions, the best-policy gap-visitation complexity, $\mathcal{C}^\star(\mathcal{M})$, satisfies

$$\mathcal{C}^\star(\mathcal{M}) = \sum_{h=1}^{H} \inf_{\pi} \max_{s} \frac{1}{w_h^\pi(s)} \sum_{a:\Delta_h(s,a)>0} \frac{1}{\Delta_h(s,a)^2}.$$

Proof Consider the optimization

$$\min_{\lambda \in \triangle(X)} \max_{x \in X} a_x/\lambda_x.$$

It is easy to see that

$$\sum_{x \in X} a_x = \min_{\lambda \in \triangle(X)} \max_{x \in X} a_x/\lambda_x$$

and the optimal λ is

$$\lambda_x^* = \frac{a_x}{\sum_{x' \in X} a_{x'}}.$$

For any policy π, we will have that $\sum_a \pi_h(a|s) = 1$, and $\pi_h(a|s)$ must be a valid distribution over a. This implies that $w_h^\pi(s,a) = w_h^\pi(s)\pi_h(a|s)$. Now fix π for steps $h' = 1, \ldots, h-1$, then it follows that

$$\inf_{\pi_h} \max_{s,a} \min \left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} = \inf_{\pi_h} \max_{s} \frac{1}{w_h^\pi(s)} \max_{a} \frac{1}{\pi_h(a|s)} \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

Now for a given s, we can use that $w_h^\pi(s)$ is independent of π_h and apply our above calculation to get that

$$\inf_{\pi_h} \frac{1}{w_h^\pi(s)} \max_{a} \frac{1}{\pi_h(a|s)} \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} = \frac{1}{w_h^\pi(s)} \sum_a \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

As the maximum over s is over a finite set and $\pi_h(\cdot|s)$ can be chosen independently of $\pi_h(\cdot|s')$ for any $s \neq s'$, we have that

$$\inf_{\pi_h} \max_{s} \frac{1}{w_h^\pi(s)} \max_{a} \frac{1}{\pi_h(a|s)} \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} = \max_{s} \frac{1}{w_h^\pi(s)} \sum_a \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

Since taking an inf over π is equivalent to taking an inf over $\{\pi_{h'}\}_{h'=1}^{h-1}$ and π_h, we can take the inf of this over $\{\pi_{h'}\}_{h'=1}^{h-1}$ to get

$$\inf_{\pi} \max_{s,a} \min \left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} = \inf_{\pi} \max_{s} \frac{1}{w_h^\pi(s)} \sum_a \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

The same line of reasoning can be used to obtain the expression for $\mathcal{C}^\star(\mathcal{M})$. ■

Proposition 10 *We can bound*

$$\mathcal{C}(\mathcal{M},\epsilon) \le \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \frac{4}{w_h^\pi(s,a)\widetilde{\Delta}_h^\epsilon(s,a)^2 + \frac{\epsilon^2}{SA}}$$

where

$$\widetilde{\Delta}_h^\epsilon(s,a) := \begin{cases} \widetilde{\Delta}_h(s,a) & \frac{\epsilon}{W_h(s)} < \frac{\widetilde{\Delta}_h(s,a)}{3} \\ \epsilon/H & \frac{\epsilon}{W_h(s)} \ge \frac{\widetilde{\Delta}_h(s,a)}{3} \end{cases}.$$

Proof Let $\mathrm{OPT}_h(\epsilon) = \{(s,a) \,:\, \widetilde{\Delta}_h(s,a)W_h(s)/3 \le \epsilon\}$ so that $\mathrm{OPT}(\epsilon) = \cup_h \mathrm{OPT}_h(\epsilon)$. We can always bound $|\mathrm{OPT}_h(\epsilon)| \le SA$, and furthermore,

$$\begin{aligned}
\frac{H^2|\mathrm{OPT}_h(\epsilon)|}{\epsilon^2} &= \min\left\{\frac{H^2}{1/|\mathrm{OPT}_h(\epsilon)|\cdot\epsilon^2}, \frac{H^2SA}{\epsilon^2}\right\} \\
&\overset{(a)}{=} \inf_{\lambda\in\triangle(\mathrm{OPT}_h(\epsilon))} \max_{(s,a)\in\mathrm{OPT}_h(\epsilon)} \min\left\{\frac{H^2}{\lambda_{sa}\epsilon^2}, \frac{H^2SA}{\epsilon^2}\right\} \\
&\le \inf_{\pi} \max_{(s,a)\in\mathrm{OPT}_h(\epsilon)} \min\left\{\frac{H^2}{w_h^\pi(s,a)\epsilon^2}, \frac{H^2SA}{\epsilon^2}\right\} \\
&\overset{(b)}{\le} \inf_{\pi} \max_{(s,a)\in\mathrm{OPT}_h(\epsilon)} \frac{2H^2}{w_h^\pi(s,a)\epsilon^2 + \frac{\epsilon^2}{SA}}
\end{aligned}$$

where (a) follows since the optimal distribution will simply place a mass of $1/|\mathrm{OPT}_h(\epsilon)|$ on each $(s,a) \in \mathrm{OPT}_h(\epsilon)$, and (b) follows since $\min\{\frac{1}{a},\frac{1}{b}\} = \frac{1}{\max\{a,b\}} \le \frac{1}{a/2+b/2}$.

Consider the distribution π' which is a mixture of distribution π $1/2$ of the time, and the distribution π^{sh} $1/(2SA)$ of the time, where π^{sh} is the distribution which achieves $w_h^{\pi^{sh}}(s) = W_h(s)$. In other words, we will have $w_h^{\pi'}(s,a) \ge w_h^\pi(s,a)/2 + W_h(s)/(2SA)$. Given this, we can bound

$$\begin{aligned}
&\inf_{\pi}\max_{s,a}\min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}\right\} \\
&\quad\le \inf_{\pi}\max_{s,a}\min\left\{\frac{2}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2 + W_h(s)\widetilde{\Delta}_h(s,a)^2/SA}, \frac{2W_h(s)^2}{w_h^\pi(s,a)\epsilon^2 + W_h(s)\epsilon^2/SA}\right\} \\
&\quad\le \inf_{\pi}\left[\max_{(s,a)\in\mathrm{OPT}_h(\epsilon)^c} \frac{2}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2 + W_h(s)\widetilde{\Delta}_h(s,a)^2/SA} + \max_{(s,a)\in\mathrm{OPT}_h(\epsilon)} \frac{2}{w_h^\pi(s,a)\epsilon^2 + \epsilon^2/SA}\right].
\end{aligned}$$

If $(s,a) \in \mathrm{OPT}_h(\epsilon)^c$, then $\widetilde{\Delta}_h(s,a)W_h(s) > 3\epsilon$, so $W_h(s)\widetilde{\Delta}_h(s,a)^2 \ge 3\widetilde{\Delta}_h(s,a)\epsilon \ge \epsilon^2$. Thus, we can bound the above as

$$\le \inf_{\pi}\left[\max_{(s,a)\in\mathrm{OPT}_h(\epsilon)^c} \frac{2}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2 + \epsilon^2/SA} + \max_{(s,a)\in\mathrm{OPT}_h(\epsilon)} \frac{2}{w_h^\pi(s,a)\epsilon^2 + \epsilon^2/SA}\right].$$

The result then follows combining this with the bound on $\frac{H^2|\mathrm{OPT}_h(\epsilon)|}{\epsilon^2}$ given above, and using the definition of $\widetilde{\Delta}_h^\epsilon(s,a)$. ■

Proposition 11 *We can bound*

$$\mathcal{C}(\mathcal{M}, \epsilon) \leq \sum_{s,a,h} \frac{1}{\epsilon \max\{\widetilde{\Delta}_h(s,a), \epsilon\}} + \frac{H^2 |\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

Proof This follows from Proposition 9 and noting that

$$\begin{aligned}\min\left\{\frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2}\right\} &\leq \min\left\{\frac{1}{\sqrt{W_h(s)}\widetilde{\Delta}_h(s,a)}, \frac{\sqrt{W_h(s)}}{\epsilon}\right\} \cdot \frac{\sqrt{W_h(s)}}{\epsilon} \\ &\leq \min\left\{\frac{1}{\widetilde{\Delta}_h(s,a)\epsilon}, \frac{1}{\epsilon^2}\right\}.\end{aligned}$$

■

Proof [Proof of Proposition 4] Let π^{sh} denote the policy that achieves $w_h^{\pi^{sh}}(s) = W_h(s)$. Consider the state visitation distribution:

$$w'_h(s) = \frac{\sum_{s'} w_h^{\pi^{s'h}}(s) \cdot \sum_a \min\left\{\frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2}\right\}}{\sum_{s',a} \min\left\{\frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2}\right\}}.$$

Since the set of state visitations realizable on a given MDP is convex and for any realizable state distribution there exists a policy with that state distribution by Proposition 12, and since w'_h is a convex combination of state visitation distributions, it follows that there exists some policy $\widetilde{\pi}$ such that $w'_h(s) = w_h^{\widetilde{\pi}}(s)$. Furthermore, by definition,

$$w_h^{\widetilde{\pi}}(s) \geq \frac{w_h^{\pi^{sh}}(s) \cdot \sum_a \min\left\{\frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2}\right\}}{\sum_{s',a} \min\left\{\frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2}\right\}} = W_h(s) \cdot \frac{\sum_a \min\left\{\frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2}\right\}}{\sum_{s',a} \min\left\{\frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2}\right\}}.$$

Thus, since $\widetilde{\pi}$ is a feasible policy, using the expression for $\mathcal{C}(\mathcal{M}, \epsilon)$ given in Proposition 9, it follows that

$$\begin{aligned}\mathcal{C}(\mathcal{M}, \epsilon) &= \sum_{h=1}^{H} \inf_{\pi} \max_{s} \frac{1}{w_h^{\pi}(s)} \sum_a \min\left\{\frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2}\right\} + \frac{H^2 |\mathrm{OPT}(\epsilon)|}{\epsilon^2} \\ &\leq \sum_{h=1}^{H} \sum_{s,a} \min\left\{\frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2}\right\} + \frac{H^2 |\mathrm{OPT}(\epsilon)|}{\epsilon^2}.\end{aligned}$$

To obtain the first bound, we use the second bound to get

$$\mathcal{C}(\mathcal{M}, \epsilon) \leq \sum_{s,a,h} \frac{H^2 W_h(s)}{\epsilon^2} \leq \frac{H^3 SA}{\epsilon^2}$$

and use that $|\mathrm{OPT}(\epsilon_{\mathrm{tol}})| \leq SAH$. ■

Proof [Proof of Proposition 5] This follows directly from Proposition 9. ■

A.3. Full Algorithm Description

We turn now to the full definition of our algorithm, Moca.

A.3.1. Learn2Explore Overview

Algorithm 3 Learn2Explore

function Learn2Explore(active set $\mathcal{X} \subseteq \mathcal{S}\times\mathcal{A}$, step h, confidence δ, sampling confidence δ_{samp}, tolerance ϵ_{L2E})
  if $|\mathcal{X}| = 0$ **then return** $\{(\emptyset, \emptyset, 0, 0)\}_{j=1}^{\lceil \log(1/\epsilon_{\text{L2E}})\rceil}$
  for $j = 1, \dots, \lceil \log(1/\epsilon_{\text{L2E}})\rceil$ **do**
    $K_j \leftarrow K_j(\delta/\lceil \log(1/\epsilon_{\text{L2E}})\rceil, \delta_{\text{samp}})$ as defined in (D.1), $M_j \leftarrow |\mathcal{X}|$, $N_j \leftarrow K_j/(4|\mathcal{X}|\cdot 2^j)$
    $\mathcal{X}_j, \Pi_j \leftarrow$ FindExplorableSets$(\mathcal{X}, h, \delta, K_j, N_j)$
    $\mathcal{X} \leftarrow \mathcal{X}\backslash\mathcal{X}_j$
  return $\{(\mathcal{X}_j, \Pi_j, N_j, M_j)\}_{j=1}^{\lceil \log(1/\epsilon_{\text{L2E}})\rceil}$

function FindExplorableSets(active set $\mathcal{X} \subseteq \mathcal{S}\times\mathcal{A}$, step h, confidence δ, epochs to run K, samples to collect N)
  Set $r_h^1(s,a) \leftarrow 1$ for $(s,a) \in \mathcal{X}$ and 0 otherwise, $N(s,a,h) \leftarrow 0$, $\mathcal{Y} \leftarrow \emptyset$, $\Pi \leftarrow \emptyset$, $j \leftarrow 1$
  for $k = 1, 2, \dots, K$ **do**
    // Euler is as defined in Zanette and Brunskill (2019)
    Run Euler on reward function r_h^j, get trajectory $\{(s_h^k, a_h^k, h)\}_{h=1}^H$ and policy π_k
    $N(s_h^k, a_h^k) \leftarrow N(s_h^k, a_h^k) + 1$, $\Pi \leftarrow \Pi \cup \pi_k$
    if $N(s_h^k, a_h^k) \geq N$, $(s_h^k, a_h^k) \in \mathcal{X}$, and $(s_h^k, a_h^k) \notin \mathcal{Y}$ **then**
      $\mathcal{Y} \leftarrow \mathcal{Y} \cup (s_h^k, a_h^k)$
      $r_h^{j+1}(s,a) \leftarrow 1$ for $(s,a) \in \mathcal{X}\backslash\mathcal{Y}$ and 0 otherwise
      $j \leftarrow j + 1$
      Restart Euler
  return $\mathcal{Y}, \Pi$

Learn2Explore is the backbone of our sample collection procedure and is called both in Line 4 of Moca-SE as well as in CollectSamples. We provide a full analysis of Learn2Explore in Appendix D.

A.3.2. Moca-SE Overview

Given this formal description of Learn2Explore, we are ready to formally describe the Moca-SE (single-epoch Moca) procedure. Assume that we run Moca-SE with tolerance ϵ and confidence δ. We begin by calling Learn2Explore on Line 4, which allows us to form an estimate of $W_h(s)$, the maximum reachability of (s,h). This in turn allows us to determine which states are efficiently reachable. We let $\mathcal{Z}_h$ denote the set of all such efficiently reachable states at stage h: $W_h(s) \geq \frac{\epsilon}{2H^2S}, \forall s \in \mathcal{Z}_h$. All other states have little

Algorithm 4 Monte Carlo Action Elimination - Single Epoch (MOCA-SE(ϵ, δ, FinalRound))

input: tolerance ϵ, confidence δ, final round flag FinalRound
initialize $\epsilon_{\exp} \leftarrow \frac{\epsilon}{2H^2S}$, $\mathcal{Z}_h \leftarrow \emptyset$, $\iota_{\exp} = \lceil \log \frac{1}{\epsilon_{\exp}} \rceil$
for each (s,h) **do** // loop over all s,h to learn maximum reachability
  $\{(\mathcal{X}_j^{sh}, \Pi_j^{sh}, N_j^{sh})\}_{j=1}^{\iota_{\exp}} \leftarrow$ LEARN2EXPLORE($\{(s,a)\}, h, \frac{\delta}{SH}, \frac{1}{2}, \epsilon_{\exp}$) for arbitrary $a \in \mathcal{A}$
  if $\mathcal{X}_j^{sh} = \{(s,a)\}$ for $j \in [\iota_{\exp}]$ **then** $\widehat{W}_h(s) \leftarrow \frac{N_j^{sh}}{2|\Pi_j^{sh}|} = \frac{1}{16 \cdot 2^j}$, $\mathcal{Z}_h \leftarrow \mathcal{Z}_h \cup \{s\}$
set $\iota_\epsilon \leftarrow \lceil \log \frac{64}{H^2 S \epsilon} \rceil$, $\iota_\delta \leftarrow \log \frac{SAH\iota_\epsilon(\ell_\epsilon+1)}{\delta}$, $\ell_\epsilon \leftarrow \lceil \log \frac{H}{\epsilon} \rceil$, $\widehat{\pi}_h(s) \leftarrow$ arbitrary action, $\mathcal{A}_h^0(s) \leftarrow \mathcal{A}$.
for $h = H, H-1, \ldots, 1$ **do** // loop over horizon
  for $i = 1, 2, \ldots, \iota_\epsilon$ **do** // loop over estimated maximum reachability
    $\mathcal{Z}_{hi} \leftarrow \{s \in \mathcal{Z}_h \ : \ \widehat{W}_h(s) \in [2^{-i}, 2^{-i+1}]\}$
    for $\ell = 1, \ldots, \ell_\epsilon$ **do** // loop over tolerance ϵ_ℓ
      $\epsilon_\ell \leftarrow H2^{-\ell}$, $\mathcal{Z}_{hi}^\ell \leftarrow \{(s,a) \ : \ s \in \mathcal{Z}_{hi}, a \in \mathcal{A}_h^{\ell-1}(s), |\mathcal{A}_h^{\ell-1}(s)| > 1\}$
      $n_{ij}^\ell \leftarrow \frac{2^{18}H^2\iota_\delta}{2^{2i}\epsilon_\ell^2}$, $\gamma_{ij}^\ell \leftarrow \frac{2^i\epsilon_\ell}{2^8}$ for $j = 1, \ldots, \iota_\epsilon$
      $\mathfrak{D}_{hi}^\ell, \{\mathcal{X}_{hij}^\ell\}_{j=1}^{\iota_\epsilon} \leftarrow$ COLLECTSAMPLES($\mathcal{Z}_{hi}^\ell, \{n_{ij}^\ell\}_{j=1}^{\iota_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H\iota_\epsilon\ell_\epsilon}, \frac{\epsilon_{\exp}}{32}$)
      $\{\mathcal{A}_h^\ell(s)\}_{s \in \mathcal{Z}_{hi}} \leftarrow$ ELIMINATEACTIONS($\mathcal{Z}_{hi}^\ell, \{\mathcal{X}_{hij}^\ell\}_{j=1}^{\iota_\epsilon}, \mathfrak{D}_{hi}^\ell, \{\mathcal{A}_h^{\ell-1}(s)\}_{s \in \mathcal{Z}_{hi}}, h, \{\gamma_{ij}^\ell\}_{j=1}^{\iota_\epsilon}$)
  if FinalRound is true **then** // ensure $\widehat{\pi}$ ϵ-optimal
    $\mathcal{Z}_h^{\ell_\epsilon+1} \leftarrow \{(s,a) \ : \ s \in \mathcal{Z}_h, a \in \mathcal{A}_h^{\ell_\epsilon}(s), |\mathcal{A}_h^{\ell_\epsilon}(s)| > 1\}$
    $n_j^{\ell_\epsilon+1} \leftarrow \frac{64H^4\iota_\delta\iota_\epsilon^2 2^{2(-j+1)}}{\epsilon^2}$, $\gamma_j^{\ell_\epsilon+1} \leftarrow \frac{\epsilon}{4H\iota_\epsilon 2^{-j+1}}$ for $j = 1, \ldots, \iota_\epsilon$
    $\mathfrak{D}_h^{\ell_\epsilon+1}, \{\mathcal{X}_{hj}^{\ell_\epsilon+1}\}_{j=1}^{\iota_\epsilon} \leftarrow$ COLLECTSAMPLES($\mathcal{Z}_h^{\ell_\epsilon+1}, \{n_j^{\ell_\epsilon+1}\}_{j=1}^{\iota_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H}, \frac{\epsilon_{\exp}}{32}$)
    $\{\mathcal{A}_h^{\ell_\epsilon+1}(s)\}_{s \in \mathcal{Z}_h^{\ell_\epsilon+1}} \leftarrow$ ELIMINATEACTIONS($\mathcal{Z}_h^{\ell_\epsilon+1}, \{\mathcal{X}_{hj}^{\ell_\epsilon+1}\}_{j=1}^{\iota_\epsilon}, \mathfrak{D}_h^{\ell_\epsilon+1}, \{\mathcal{A}_h^{\ell_\epsilon}(s)\}_{s \in \mathcal{Z}_h^{\ell_\epsilon+1}}, h, \{\gamma_j^{\ell_\epsilon+1}\}_{j=1}^{\iota_\epsilon}$)
  else
    $\mathcal{A}_h^{\ell_\epsilon+1}(s) \leftarrow \mathcal{A}_h^{\ell_\epsilon}(s)$ for all $s \in \mathcal{Z}_h$
  Set $\widehat{\pi}_h(s)$ to any action in $\mathcal{A}_h^{\ell_\epsilon+1}(s)$ for all $s \in \mathcal{Z}_h$
return $\widehat{\pi}$, $\max_{s,h} |\mathcal{A}_h^{\ell_\epsilon+1}(s)|$

effect on the performance of any policy and can henceforth be ignored. The following claim shows that our estimate of $W_h(s)$ is in fact accurate for $s \in \mathcal{Z}_h$.

Claim A.1 (Informal) *If running* MOCA-SE*, with high probability* $\widehat{W}_h(s) \le W_h(s) \le 32\widehat{W}_h(s)$ *for all* $s \in \mathcal{Z}_h$.

We then proceed to our main loop over h in Line 7. For a fixed h, we loop over i and form the partition $\mathcal{Z}_{hi}$ which contains all $s \in \mathcal{Z}_h$ with $\widehat{W}_h(s) \sim 2^{-i}$. We then proceed to our action elimination procedure and loop over $\ell \in \mathbb{N}$, where we eliminate actions at tolerance $\epsilon_\ell = H2^{-\ell}$. For each such ℓ, we define $\mathcal{Z}_{hi}^\ell \subseteq \mathcal{S} \times \mathcal{A}$ as the set of (s,a) for $s \in \mathcal{Z}_{hi}$, and a we have not yet determined are $\epsilon_{\ell-1}/W_h(s)$-suboptimal. We next run COLLECTSAMPLES on $\mathcal{Z}_{hi}^\ell$ and seek to collect $n_{ij}^\ell = \mathcal{O}(H^2/(2^{2i}\epsilon_\ell^2)) = \mathcal{O}(H^2 W_h(s)^2/\epsilon_\ell^2)$ from each $(s,a) \in \mathcal{Z}_{hi}^\ell$.

Note that every $(s,a) \in \mathcal{Z}_{hi}^\ell$ has similar maximum reachability $W_h(s) \sim 2^{-i}$, determined by index i. Nevertheless, as outlined in Section 6.1, to obtain the proper scaling in S,

we may still need to group states in a way that allows nearby states to be explored effectively. Calling `Learn2Explore` in `CollectSamples` does just this, efficiently traversing the MDP to guarantee enough samples are collected from all states in tandem. After running `CollectSamples`, we run `EliminateActions` to eliminate suboptimal actions, yielding a set of candidate $\epsilon_\ell/W_h(s)$-suboptimal actions for each (s,h), denoted $\mathcal{A}_h^\ell(s)$.

The `FinalRound` flag. Single-episode Moca is called multiple times by our main algorithm (Algorithm 5), each with geometrically decreasing tolerance ϵ. For all but the smallest such ϵ, `Moca-SE` is run with `FinalRound` = `false`, and terminates after the previously described loop over h, i, ℓ terminates. The last call to `Moca-SE` constitutes the "final round", where we set `FinalRound` = `true`; this calls `CollectSamples` and `EliminateActions` one more time for each h.

While the loop with the `FinalRound` = `false` is able to eliminate suboptimal actions, it does not shrink the action set enough to guarantee that the returned policy is ϵ-optimal. In particular, while each (s,h) pair upon entering this final-round loop is sub-optimal by at most $\epsilon_h(s) = \mathcal{O}(\epsilon/W_h(s))$, Proposition 7 suggests that we actually need $\epsilon_h(s) \le \mathcal{O}(\epsilon/H \cdot \sup_\pi \sum_{s' \in \mathcal{X}} w_h^\pi(s'))$. To remedy this, `FinalRound` = `true` invokes a final step to ensure the latter bound holds. Critically, while in the previous step we only sampled (s,a) in proportion with $W_h(s)^2$, the individual maximum reachability of that state, in this step we sample each (s,a) in proportion with the *reachability of the partition containing* (s,a). *This subtlety is indispensable for attaining our instance-dependent sample complexity.*

In other words, after forming our set $\mathcal{Z}_h^{\ell_\epsilon+1}$ of active states and actions corresponding to the minimal error-resolution index $\ell = \ell_\epsilon$ (from the previous argument, this will only contain states we have not determined the optimal action for and actions that satisfy $\Delta_h(s,a) \le \frac{3\epsilon}{2W_h(s)}$) and partitioning it into $\{\mathcal{X}_{hj}^{\ell_\epsilon+1}\}_j$ by calling `Learn2Explore`, we seek to collect $\mathcal{O}(H^4 2^{-2j}/\epsilon^2)$ from every $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$. By Theorem 8, $\mathcal{X}_{hj}^{\ell_\epsilon+1}$ satisfies $\sup_\pi \sum_{(s,a)\in\mathcal{X}_{hj}^{\ell_\epsilon+1}} w_h^\pi(s,a) \le 2^{-j+1}$, so sampling (s,a) $\mathcal{O}(H^4 2^{-2j}/\epsilon^2)$ times means we sample it in proportion to its group reachability squared.

A.3.3. Moca Overview

Algorithm 5 MOnte **C**arlo **A**ction Elimination (Moca)

input: tolerance ϵ_{tol}, confidence δ_{tol}
$\mathcal{A}_h^0(s) \leftarrow \mathcal{A}$ for all s, h
for $m = 1, \dots, \lceil \log(H/\epsilon_{\mathrm{tol}}) \rceil - 1$ **do**
  $\epsilon_{\mathrm{tol}(m)} \leftarrow H2^{-m}, \delta_{\mathrm{tol}(m)} \leftarrow \frac{\delta_{\mathrm{tol}}}{36m^2}$
  $\widehat{\pi}^m, \texttt{MaxOpt} \leftarrow \texttt{Moca-SE}(\epsilon_{\mathrm{tol}(m)}, \delta_{\mathrm{tol}(m)}, \texttt{false})$
  if `MaxOpt` = 1 **then**
    return $\widehat{\pi}^m$
$\widehat{\pi}, \texttt{MaxOpt} \leftarrow \texttt{Moca-SE}(\epsilon_{\mathrm{tol}}, \frac{\delta_{\mathrm{tol}}}{36\lceil \log(H/\epsilon_{\mathrm{tol}}) \rceil^2}, \texttt{true})$
return $\widehat{\pi}$

We turn now to our main algorithm, Moca. Moca takes as input a tolerance ϵ_{tol} and confidence δ_{tol}. Were our goal simply to find an ϵ_{tol}-optimal policy, from the above argument

we could call MOCA-SE with tolerance ϵ_{tol} and FinalRound = true. However, if ϵ_{tol} is small enough that MOCA-SE identifies the optimal action in *every* state, this may result in overexploring—since once we have identified the optimal action in every state we can terminate and output the optimal policy. To remedy this, we instead call MOCA-SE with exponentially decreasing tolerance and FinalRound = false. If it returns a set of actions for every s, h with $|\mathcal{A}_h(s)| = 1$, we can guarantee we have identified the optimal policy, and simply terminate without overexploring. Note also in this stage, since FinalRound = false, we do not pay for the $\widetilde{\mathcal{O}}(\frac{H^4}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}|)$ term. If this condition is never met, we simply call MOCA-SE a final time at the end with FinalRound = true to ensure the policy we return is ϵ_{tol}-optimal.

A.3.4. Helper Function Descriptions

Algorithm 6 MOCA Helper Functions

function CollectSamples(active set $\mathcal{X}$, allocation $\{n_j\}_{j=1}^{\lceil \log 1/\epsilon_{\mathtt{cs}} \rceil}$, step h, policy $\widehat{\pi}$, tolerance $\delta_{\mathtt{cs}}$, precision $\epsilon_{\mathtt{cs}}$)
  $\{(\mathcal{X}_j, \Pi_j, N_j)\}_{j=1}^{\lceil \log 1/\epsilon_{\mathtt{cs}} \rceil} \leftarrow$ Learn2Explore$(\mathcal{X}, h, \delta_{\mathtt{cs}}, \frac{\delta_{\mathtt{cs}}}{\lceil \log 1/\epsilon_{\mathtt{cs}} \rceil \max_j n_j}, \epsilon_{\mathtt{cs}})$, $\mathfrak{D} \leftarrow \emptyset$
  for $j = 1, \dots, \lceil \log 1/\epsilon_{\mathtt{cs}} \rceil$ **do**
    for $\pi \in \Pi_j$ **do**
      Run π for $T = \lceil 2n_j/N_j \rceil$ times up to level h, then play $\widehat{\pi}$
      Collect reward rollouts $\mathfrak{D} \leftarrow \mathfrak{D} \cup \{s_h^t, a_h^t, \mathring{Q}_h^{\widehat{\pi},t}(s_h^t, a_h^t) := \sum_{h'=h}^H R_{h'}^t\}_{t=1}^T$
  return $\mathfrak{D}$, $\{\mathcal{X}_j\}_{j=1}^{\lceil \log 1/\epsilon_{\mathtt{cs}} \rceil}$

function EliminateActions(active set $\mathcal{X}$, partition $\{\mathcal{X}_j\}_{j=1}^k$, dataset $\mathfrak{D}$, active actions $\{\mathcal{A}_h(s)\}_{s\in\mathcal{Z}}$, level h, thresholds $\{\gamma_j\}_{j=1}^k$)
  for $(s,a) \in \mathcal{X}$ **do**
    $N_h(s,a) \leftarrow \sum_{(s_h^t, a_h^t, \mathring{Q}_h^{\widehat{\pi},t}(s_h^t,a_h^t)) \in \mathfrak{D}} \mathbb{I}\{(s_h^t, a_h^t) = (s,a)\}$
    $\widehat{Q}_h^{\widehat{\pi}}(s,a) \leftarrow \frac{1}{N_h(s,a)} \sum_{(s_h^t, a_h^t, \mathring{Q}_h^{\widehat{\pi},t}(s_h^t,a_h^t)) \in \mathfrak{D}} \mathbb{I}\{(s_h^t, a_h^t) = (s,a)\} \cdot \mathring{Q}_h^{\widehat{\pi},t}(s_h^t, a_h^t)$
  for $j = 1, \dots, k$ **do**
    for s s.t. $\exists a$ with $(s,a) \in \mathcal{X}_j$ **do**
      $j(s) \leftarrow \arg\max_{j'} j'$ s.t. $\exists a', (s,a') \in \mathcal{X}_{j'}$
      $\mathcal{A}_h(s) \leftarrow \{a \in \mathcal{A}_h(s) \ : \ \max_{a' \in \mathcal{A}_h(s)} \widehat{Q}_h^{\widehat{\pi}}(s,a') - \widehat{Q}_h^{\widehat{\pi}}(s,a) \le \gamma_{j(s)}\}$
  return $\{\mathcal{A}_h(s)\}_{s\in\mathcal{Z}}$

Description of CollectSamples. CollectSamples takes as input a set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, an allocation $\{n_j\}_j$, a timestep h, and a policy $\widehat{\pi}$. In short, CollectSamples first calls Learn2Explore on $\mathcal{X}$ to obtain a partition $\{\mathcal{X}_j\}_j$, and then reruns the policies returned by Learn2Explore enough times to ensure that every $(s,a) \in \mathcal{X}_j$ is reached at least n_j times at timestep h. After reaching (s,a,h), $\widehat{\pi}$ is played, to obtain a Monte Carlo estimate $\mathring{Q}_h^{\widehat{\pi},t}(s,a)$ of $Q_h^{\widehat{\pi}}(s,a)$. CollectSamples then returns the data collected and the partition returned by Learn2Explore.

Description of `EliminateActions`. `EliminateActions` takes as input a set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, a partition of this set $\{\mathcal{X}_j\}_j$, a dataset $\mathfrak{D}$ generated by `CollectSamples`, a set of active actions $\{\mathcal{A}_h(s)\}_s$, a timestep h, and a threshold $\{\gamma_j\}_j$. For each $(s,a) \in \mathcal{X}$, it forms an estimate of $Q_h^{\widehat{\pi}}(s,a)$ from the rollouts in $\mathfrak{D}$. Given these estimates, for s such that there exists a with $(s,a) \in \mathcal{X}_j$, it removes actions from $\mathcal{A}_h(s)$ that are more than $\gamma_{j(s)}$-suboptimal.

Appendix B. MDP Technical Results

Proof [Proof of Proposition 7] This is a direct consequence of Lemma B.1 since we can apply this lemma to get that, for arbitrary π',

$$V_0^\star - V_0^{\widehat{\pi}} = \sum_s P_0(s)(V_1^\star(s) - V_1^{\widehat{\pi}}(s)) = \sum_s w_1^{\pi'}(s)(V_1^\star(s) - V_1^{\widehat{\pi}}(s)) \le \sum_{h=1}^H \sup_\pi \sum_s w_h^\pi(s)\epsilon_h(s)$$

where we note that $w_1^{\pi'}(s) = \mathbb{P}_{\pi'}[s_1 = s] = P_0(s)$. ■

Lemma B.1 *Assume that for each h and s, $\widehat{\pi}$ plays an action which satisfies*

$$\max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \le \epsilon_h(s). \tag{B.1}$$

Then for any h and π',

$$\sum_s w_h^{\pi'}(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s)) \le \sum_{h'=h}^H \sup_\pi \sum_s w_{h'}^\pi(s)\epsilon_{h'}(s).$$

Proof We proceed by backwards induction. The base case, $h = H$, is trivial. Assume that at level h, for any π,

$$\sum_s w_h^\pi(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s)) \le \sum_{h'=h}^H \sup_{\pi'} \sum_{s'} w_{h'}^{\pi'}(s')\epsilon_{h'}(s')$$

and that at level $h-1$, for each s (B.1) holds. By definition,

$$\begin{aligned} V_{h-1}^\star(s) - V_{h-1}^{\widehat{\pi}}(s) &= Q_{h-1}^\star(s, \pi_{h-1}^\star(s)) - Q_{h-1}^{\widehat{\pi}}(s, \widehat{\pi}_{h-1}(s)) \\ &= Q_{h-1}^\star(s, \pi_{h-1}^\star(s)) - Q_{h-1}^{\widehat{\pi}}(s, \pi_{h-1}^\star(s)) + Q_{h-1}^{\widehat{\pi}}(s, \pi^\star(s)) - \max_a Q_{h-1}^{\widehat{\pi}}(s,a) \\ &\qquad + \max_a Q_{h-1}^{\widehat{\pi}}(s,a) - Q_{h-1}^{\widehat{\pi}}(s, \widehat{\pi}_{h-1}(s)). \end{aligned}$$

Clearly, $Q_{h-1}^{\widehat{\pi}}(s, \pi^\star(s)) - \max_a Q_{h-1}^{\widehat{\pi}}(s,a) \le 0$ and by assumption $\max_a Q_{h-1}^{\widehat{\pi}}(s,a) - Q_{h-1}^{\widehat{\pi}}(s, \widehat{\pi}_{h-1}(s)) \le \epsilon_{h-1}(s)$. Furthermore,

$$Q_{h-1}^\star(s, \pi_{h-1}^\star(s)) - Q_{h-1}^{\widehat{\pi}}(s, \pi_{h-1}^\star(s)) = \sum_{s'} P_{h-1}(s'|s, \pi_{h-1}^\star(s))(V_h^\star(s') - V_h^{\widehat{\pi}}(s')).$$

Then, for any π,

$$\begin{aligned}\sum_s w_{h-1}^{\pi}(s)(V_{h-1}^{\star}(s) - V_{h-1}^{\widehat{\pi}}(s)) &\le \sum_s w_{h-1}^{\pi}(s)\epsilon_{h-1}(s) \\ &\quad + \sum_s \sum_{s'} w_{h-1}^{\pi}(s) P_{h-1}(s'|s, \pi_{h-1}^{\star}(s))(V_h^{\star}(s') - V_h^{\widehat{\pi}}(s')) \\ &= \sum_s w_{h-1}^{\pi}(s)\epsilon_{h-1}(s) + \sum_s w_h^{\pi'}(s)(V_h^{\star}(s) - V_h^{\widehat{\pi}}(s)) \\ &\le \sum_{h'=h-1}^{H} \sup_{\pi'} \sum_{s'} w_{h'}^{\pi'}(s')\epsilon_{h'}(s')\end{aligned}$$

where the last inequality follows by the inductive hypothesis and we have used that

$$\sum_s w_{h-1}^{\pi}(s) P_{h-1}(s'|s, \pi_{h-1}^{\star}(s)) = w_h^{\pi'}(s').$$

where $\pi'_{h'}(s) = \pi_{h'}(s)$ for all $h' \le h-2$ and $\pi'_{h'}(s) = \pi^{\star}_{h'}(s)$ for $h' \ge h-1$. The conclusion then follows. ■

Lemma B.2 *Assume that*

$$\sup_{\pi} \sum_{s'} w_h^{\pi}(s')(V_h^{\star}(s') - V_h^{\widehat{\pi}}(s')) \le \epsilon \quad \textit{and} \quad \sup_{\pi} \sum_{s'} w_{h+1}^{\pi}(s')(V_{h+1}^{\star}(s') - V_{h+1}^{\widehat{\pi}}(s')) \le \epsilon.$$

Then, for any s,

$$|\Delta_h(s,a) - \Delta_h^{\widehat{\pi}}(s,a)| \le \epsilon / W_h(s).$$

Proof By definition,

$$\begin{aligned}|\Delta_h(s,a) - \Delta_h^{\widehat{\pi}}(s,a)| &= |V_h^{\star}(s) - Q_h^{\star}(s,a) - (\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a))| \\ &\le \max\{|V_h^{\star}(s) - \max_{a'} Q_h^{\widehat{\pi}}(s,a')|, |Q_h^{\widehat{\pi}}(s,a) - Q_h^{\star}(s,a)|\}.\end{aligned}$$

where the last inequality follows since

$$V_h^{\star}(s) - Q_h^{\star}(s,a) - (\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a)) \le V_h^{\star}(s) - \max_{a'} Q_h^{\widehat{\pi}}(s,a')$$

and

$$-(V_h^{\star}(s) - Q_h^{\star}(s,a) - (\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a))) \le Q_h^{\star}(s,a) - Q_h^{\widehat{\pi}}(s,a).$$

Now,

$$\begin{aligned}V_h^{\star}(s) - \max_{a'} Q_h^{\widehat{\pi}}(s,a') &= V_h^{\star}(s) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) + Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) - \max_{a'} Q_h^{\widehat{\pi}}(s,a') \\ &\le V_h^{\star}(s) - V_h^{\widehat{\pi}}(s)\end{aligned}$$

where the inequality follows since, by definition, $V_h^{\widehat{\pi}}(s) = Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s))$ and $Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) - \max_{a'} Q_h^{\widehat{\pi}}(s, a') \le 0$. By assumption,

$$\sup_{\pi} \sum_{s'} w_h^{\pi}(s')(V_h^{\star}(s') - V_h^{\widehat{\pi}}(s')) \le \epsilon$$

and furthermore, for any s,

$$\sup_{\pi} \sum_{s'} w_h^{\pi}(s')(V_h^{\star}(s') - V_h^{\widehat{\pi}}(s')) \ge W_h(s)(V_h^{\star}(s) - V_h^{\widehat{\pi}}(s))$$

so it follows that $|V_h^{\star}(s) - V_h^{\widehat{\pi}}(s)| \le \epsilon / W_h(s)$. By definition,

$$Q_h^{\star}(s,a) - Q_h^{\widehat{\pi}}(s,a) = \sum_{s'} P_h(s'|s,a)(V_{h+1}^{\star}(s') - V_{h+1}^{\widehat{\pi}}(s'))$$

so

$$\begin{aligned} W_h(s)(Q_h^{\star}(s,a) - Q_h^{\widehat{\pi}}(s,a)) &= \sum_{s'} P_h(s'|s,a) W_h(s)(V_{h+1}^{\star}(s') - V_{h+1}^{\widehat{\pi}}(s')) \\ &\le \sup_{\pi} \sum_{s'} w_{h+1}^{\pi}(s')(V_{h+1}^{\star}(s') - V_{h+1}^{\widehat{\pi}}(s')) \end{aligned}$$

where the inequality follows since $V_{h+1}^{\star}(s') \ge V_{h+1}^{\widehat{\pi}}(s')$, and since

$$P_h(s'|s,a)W_h(s) = \mathbb{P}[s_{h+1} = s' | s_h = s, a_h = a]\mathbb{P}_{\pi}[s_h = s] = \mathbb{P}_{\pi'}[s_{h+1} = s', s_h = s] \le \mathbb{P}_{\pi'}[s_{h+1}]$$

where π denotes the policy achieving $\mathbb{P}_{\pi}[s_h = s] = W_h(s)$ and π' plays π up to h and then $\pi'_h(s) = a$. Thus, if $\sup_{\pi} \sum_{s'} w_{h+1}^{\pi}(s')(V_{h+1}^{\star}(s') - V_{h+1}^{\widehat{\pi}}(s')) \le \epsilon$, rearranging the inequalities gives the result.

■

We are aware of several works which obtain the following result for non-episodic MDPs (Zimin and Neu, 2013; Puterman, 2014), but present the result for episodic MDPs for completeness.

Proposition 12 *Fix some MDP $\mathcal{M}$. Then:*

1. *The set of valid state-action visitation distributions on $\mathcal{M}$ is convex.*
2. *For any valid state-action visitation distribution on $\mathcal{M}$, there exists some policy which realizes it.*

Proof The set of valid state-action visitation distributions, $\mathcal{W}$, is defined as

$$\mathcal{W} := \Big\{ w \in [0,1]^{SAH} \ : \ \exists \pi \in \Pi \text{ s.t. } w_h(s,a) = \pi_h(a|s) \cdot \sum_{s',a'} P_{h-1}(s|s',a') w_{h-1}(s',a'), \forall h \ge 1,$$
$$w_0(s,a) = \pi_0(a|s) P_0(s), \quad \sum_{s,a} w_h(s,a) = 1, \forall h \ge 0 \Big\}$$

where here $\Pi = \triangle(\mathcal{A})^{SH}$.

Fix some state-action visitation distributions $w, w' \in \mathcal{W}$, and let π and π' denote their correponding policies as above. Furthermore, denote $w_h(s) = \sum_a w_h(s,a)$ (and similarly for w'). Our goal is to show that for any $t \in [0,1]$, $\widetilde{w} = (1-t)w + tw' \in \mathcal{W}$. First, we show that there exists some policy $\widetilde{\pi}$ such that

$$(1-t)w_0(s,a) + tw'_0(s,a) = \widetilde{\pi}_0(a|s)P_0(s).$$

Note that we can take $\widetilde{\pi}_0(a|s) = (1-t)\pi_0(a|s) + t\pi'_0(a|s)$, since

$$((1-t)\pi_0(a|s) + t\pi'_0(a|s))P_0(s) = (1-t)w_0(s,a) + tw'_0(s,a).$$

By construction, for any $h \geq 1$,

$$\widetilde{w}_h(s) = \sum_a \widetilde{w}_h(s,a) = (1-t)\sum_a w_h(s,a) + t\sum_a w'_h(s,a) = (1-t)w_h(s) + tw'_h(s).$$

Furthermore, since w is a valid state-action distribution,

$$w_h(s) = \sum_{s',a'} P_{h-1}(s|s',a')w_{h-1}(s',a')$$

and similarly for w'. Let $\widetilde{\pi}_h(a|s) = \widetilde{w}_h(s,a)/\widetilde{w}_h(s)$ (where we define $0/0 = 0$), and note that this is a valid distribution since by definition $\sum_a \widetilde{w}_h(s,a) = \widetilde{w}_h(s)$. Then,

$$\begin{aligned}
\widetilde{w}_h(s,a) &= \widetilde{\pi}_h(a|s)\widetilde{w}_h(s) \\
&= \widetilde{\pi}_h(a|s)((1-t)w_h(s) + tw'_h(s)) \\
&= \widetilde{\pi}_h(a|s)\sum_{s',a'} P_{h-1}(s|s',a')((1-t)w_{h-1}(s',a') + tw'_{h-1}(s',a')) \\
&= \widetilde{\pi}_h(a|s)\sum_{s',a'} P_{h-1}(s|s',a')\widetilde{w}_{h-1}(s',a')
\end{aligned}$$

where the last equality follows by the definition of $\widetilde{w}_{h-1}$. The other constraints are trivial to verity, so $\widetilde{w} \in \mathcal{W}$. This proves the first result.

For the second result, take some $w \in \mathcal{W}$, and let $\pi_h(a|s) = w_h(s,a)/w_h(s)$. By definition this is a valid distribution. Furthermore, it trivially holds that $w_0^\pi(s,a) = w_0(s,a)$. Assume that $w_{h-1}^\pi(s,a) = w_{h-1}(s,a)$ for all (s,a). By definition and the inductive hypothesis,

$$\begin{aligned}
w_h^\pi(s,a) &= \pi_h(a|s)\sum_{s',a'} P_{h-1}(s|s',a')w_{h-1}^\pi(s',a') \\
&= \pi_h(a|s)\sum_{s',a'} P_{h-1}(s|s',a')w_{h-1}(s,a) \\
&= \pi_h(a|s)w_h(s) \\
&= w_h(s,a),
\end{aligned}$$

which proves the second result. ■

Appendix C. Proof of Theorem 2

In this section we give a formal proof of Theorem 2.

Notation. Throughout the proof, we let ϵ_{tol} denote the tolerance and δ_{tol} the confidence given as an input to MOCA, and $\epsilon = \epsilon_{\text{tol}(m)}$ and $\delta = \delta_{\text{tol}(m)}$ the tolerance and confidence given as an input to `MOCA-SE` at epoch m of MOCA, respectively. For convenience, we will also define $\epsilon_0 = H$. For a single call of `MOCA-SE`, we will use the following notation:

- For a given h, i, and ℓ, consider the call to `COLLECTSAMPLES` on Line 13, and let $\{\mathcal{X}^{\ell}_{hij}\}_{j=1}^{\iota_\epsilon}$ denote the partition returned by calling `LEARN2EXPLORE` on Line 2 of `COLLECTSAMPLES`. Similarly, let $\{\Pi^{\ell}_{hij}\}_{j=1}^{\iota_\epsilon}$ and $\{N^{\ell}_{hij}\}_{j=1}^{\iota_\epsilon}$ denote the policies and minimum number of samples returned by `LEARN2EXPLORE`, respectively.
- For a given h, consider the call to `COLLECTSAMPLES` on Line 18, and let $\{\mathcal{X}^{\ell_\epsilon}_{hj}\}_{j=1}^{\iota_\epsilon}$ denote the partition returned by calling `LEARN2EXPLORE` on Line 2 of `COLLECTSAMPLES`. As before, let $\{\Pi^{\ell_\epsilon+1}_{hj}\}_{j=1}^{\iota_\epsilon}$ and $\{N^{\ell_\epsilon+1}_{hj}\}_{j=1}^{\iota_\epsilon}$ denote the policies and minimum number of samples.

Good Events. We next define the good events, which we will assume hold throughout the remainder of the proof.

First, let $\mathcal{E}_{\text{exp}}$ be the event on which, for all calls to `MOCA-SE` simultaneously:

- For every $h = 1, \dots, H$, $i = 1, \dots, \iota_\epsilon$, $\ell = 1, \dots, \ell_\epsilon$, we collect at least n^{ℓ}_{i1} samples from each $(s,a) \in \mathcal{Z}^{\ell}_{hi}$. Furthermore, $\cup_{j=1}^{\iota_\epsilon} \mathcal{X}^{\ell}_{hij} = \mathcal{Z}^{\ell}_{hi}$ and $\mathcal{X}^{\ell}_{hij}$ satisfy
$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}^{\ell}_{hij}} w^{\pi}_h(s,a) \le 2^{-j+1}.$$
- For every $h = 1, \dots, H$, if `MOCA-SE` is run with `FinalRound` $=$ `true`, then we collect at least $n^{\ell_\epsilon+1}_{j}$ samples from each $(s,a) \in \mathcal{X}^{\ell_\epsilon+1}_{hj}$. Furthermore, $\cup_{j=1}^{\iota_\epsilon} \mathcal{X}^{\ell_\epsilon+1}_{hj} = \mathcal{Z}^{\ell_\epsilon+1}_{h}$ and $\mathcal{X}^{\ell_\epsilon+1}_{hj}$ satisfies
$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}^{\ell_\epsilon+1}_{hj}} w^{\pi}_h(s,a) \le 2^{-j+1}.$$
- $\widehat{W}_h(s) \le W_h(s) \le 32\widehat{W}_h(s)$ for all $s \in \mathcal{Z}_h$.
- Following Line 7 of `MOCA-SE`, $\mathcal{Z}_h$ satisfies, for all h,
$$\sup_{\pi} \max_{s\in\mathcal{Z}_h^c} w^{\pi}_h(s) \le \frac{\epsilon}{2H^2S}.$$

Next, let $\mathcal{E}_{\text{est}}$ be the event on which, for all calls to `MOCA-SE`,

$$|\widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a) - Q^{\widehat{\pi}}_h(s,a)| \le \sqrt{\frac{H^2\iota_\delta}{N^{hi\ell}_h(s,a)}}, \quad \forall (s,a) \in \mathcal{Z}^{\ell}_{hi}, \forall h \in [H], i \in [\iota_\epsilon], \ell \in [\ell_\epsilon]$$

$$|\widehat{Q}^{\widehat{\pi}}_{h,\ell_\epsilon+1}(s,a) - Q^{\widehat{\pi}}_h(s,a)| \le \sqrt{\frac{H^2\iota_\delta}{N_h^{h(\ell_\epsilon+1)}(s,a)}}, \quad \forall (s,a) \in \mathcal{Z}_h^{\ell_\epsilon}, \forall h \in [H]$$

where $\widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a)$ is the estimate of $Q^{\widehat{\pi}}_h(s,a)$ formed on Line 12 of `EliminateActions`, $N_h^{hi\ell}(s,a)$ is the number of samples collected from (s,a,h) at iteration (h,i,ℓ), and $\widehat{Q}^{\widehat{\pi}}_{h,\ell_\epsilon+1}(s,a)$ and $N_h^{h(\ell_\epsilon+1)}(s,a)$ are the analogous quantities for the sampling done if `FinalRound` = `true`.

We can think of $\mathcal{E}_{\text{exp}}$ as the event on which we *explore* successfully—we reach every state the desired number of times—and $\mathcal{E}_{\text{est}}$ the event on which we *estimate* correctly—our Monte Carlo estimates of $Q^{\widehat{\pi}}_h(s,a)$ concentrate. The following lemma shows that these events hold with high probability.

Lemma C.1 *If we run* MOCA, $\mathbb{P}[\mathcal{E}_{\text{exp}} \cap \mathcal{E}_{\text{est}}] \ge 1 - \delta_{\text{tol}}$.

Proof [Proof Sketch] That $\mathcal{E}_{\text{est}}$ holds is simply a consequence of Hoeffding's inequality since $Q^{\widehat{\pi}}_h(s,a)$ will be in $[0,H]$ almost surely. That $\mathcal{E}_{\text{exp}}$ holds is a direct consequence of the correctness of our exploration procedure, as described in Appendix D. We give the full proof of this result in Appendix C.4. ■

C.1. Correctness of `Moca-SE`.

We next establish that the policy returned by `Moca-SE` run with tolerance ϵ and `FinalRound` = `true` is ϵ-optimal. To this end, we first show that any action in the active set, $\mathcal{A}_h^\ell(s)$, will satisfy a certain suboptimality bound.

Lemma C.2 (Formal Statement of Lemma 6.1 and Lemma 6.3) *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, if `Moca-SE` is run with tolerance ϵ, for any $h \in [H]$ and $\ell \in [\ell_\epsilon + 1]$, if $|\mathcal{A}_h^\ell(s)| = 1$, then for $a \in \mathcal{A}_h^\ell(s)$,*

$$\max_{a'} Q^{\widehat{\pi}}_h(s,a') - Q^{\widehat{\pi}}_h(s,a) = 0.$$

Furthermore, if $|\mathcal{A}_h^\ell(s)| > 1$, $\ell \le \ell_\epsilon$, and $s \in \mathcal{Z}_{hi}$ for some i, then any $a \in \mathcal{A}_h^\ell(s)$ satisfies

$$\Delta_h(s,a) \le \frac{3\epsilon_\ell}{2W_h(s)}.$$

Finally, if $|\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1$ and $s \in \mathcal{Z}_h$, then any $a \in \mathcal{A}_h^{\ell_\epsilon+1}(s)$ satisfies

$$\Delta^{\widehat{\pi}}_h(s,a) \le \frac{\epsilon}{2H\iota_\epsilon \cdot 2^{-j(s)+1}}$$

where $j(s) = \arg\max_j j$ s.t. $\exists a', (s,a') \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$.

Proof We first claim that the optimal action with respect to $\widehat{\pi}$ must always be active.

Claim C.3 *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, for any h, s, and $\ell \in [\ell_\epsilon + 1]$, we will have that $\widehat{a}^\star_h(s) \in \mathcal{A}_h^\ell(s)$ where $\widehat{a}^\star_h(s) = \arg\max_a Q^{\widehat{\pi}}_h(s,a)$.*

We prove this claim in Appendix C.4. By construction, we will always have that $|\mathcal{A}_h^\ell(s)| \geq 1$. If $|\mathcal{A}_h^\ell(s)| = 1$, from Claim C.3 it follows that $\mathcal{A}_h^\ell(s) = \{\widehat{a}_h^\star(s)\}$, and thus $\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) = 0$.

Assume then that $|\mathcal{A}_h^\ell(s)| > 1$, $\ell \leq \ell_\epsilon$, and $s \in \mathcal{Z}_{hi}$. The result is trivial when $\ell = 0$, since in this case $\epsilon_\ell = H$, and we will always have $\Delta_h(s,a) \leq H, W_h(s) \leq 1$. On the event $\mathcal{E}_{\mathrm{exp}}$, for all $i \in [\iota_\epsilon]$ we will collect at least $n_{i1}^\ell = 2^{18} \cdot 2^{-2i} H^2 \iota_\delta / \epsilon_\ell^2$ samples from (s,a) for each $a \in \mathcal{A}_h^\ell(s)$, and on $\mathcal{E}_{\mathrm{est}}$ we will then have that

$$|\widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,a)| \leq \sqrt{\frac{H^2 \iota_\delta}{n_{i1}^\ell}} = 2^i \epsilon_\ell / 2^9.$$

Thus, for any $a \in \mathcal{A}_h^\ell(s)$, we have

$$\begin{aligned}\max_{a' \in \mathcal{A}_h^\ell(s)} \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a') - \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) &\geq \max_{a' \in \mathcal{A}_h^\ell(s)} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) - 2 \cdot 2^i \epsilon_\ell / 2^9 \\ &= \max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) - 2 \cdot 2^i \epsilon_\ell / 2^9\end{aligned}$$

where the equality follows since $\widehat{a}_h^\star(s) \in \mathcal{A}_h^\ell(s)$. It follows that if

$$\Delta_h^{\widehat{\pi}}(s,a) = \max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) \geq 4 \cdot 2^i \epsilon_\ell / 2^9$$

then

$$\max_{a' \in \mathcal{A}_h^\ell(s)} \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a') - \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) \geq 2 \cdot 2^i \epsilon_\ell / 2^9.$$

so the exit condition on Line 16 for `EliminateActions` is met for our choice of $\gamma_{ij}^\ell = 2^i \epsilon_\ell / 2^8$ (note that in this case, since γ_{ij}^ℓ is the same for all ℓ, Line 15 has no effect), and therefore $a \notin \mathcal{A}_h^{\ell+1}(s)$. Thus, any $a \in \mathcal{A}_h^{\ell+1}(s)$ must satisfy

$$\Delta_h^{\widehat{\pi}}(s,a) \leq 2^i \epsilon_\ell / 2^7.$$

By construction, we will have that $\widehat{W}_h(s) \in [2^{-i}, 2^{-i+1}]$ and on $\mathcal{E}_{\mathrm{exp}}$, $\widehat{W}_h(s) \leq W_h(s) \leq 32 \widehat{W}_h(s)$. Thus, we can upper bound

$$\Delta_h^{\widehat{\pi}}(s,a) \leq 2^i \epsilon_\ell / 2^7 \leq \frac{2\epsilon_\ell}{\widehat{W}_h(s) 2^7} \leq \frac{32 \cdot 2\epsilon_\ell}{W_h(s) 2^7} = \frac{\epsilon_\ell}{2 W_h(s)}.$$

Finally, the following claim, proved in Appendix C.4, allows us to relate $\Delta_h^{\widehat{\pi}}(s,a)$ to $\Delta_h(s,a)$:

Claim C.4 *On the event $\mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_{\mathrm{exp}}$, for any (s,a,h), we will have $|\Delta_h^{\widehat{\pi}}(s,a) - \Delta_h(s,a)| \leq \epsilon / W_h(s)$.*

Applying Claim C.4, we can lower bound $\Delta_h^{\widehat{\pi}}(s,a) \geq \Delta_h(s,a) - \epsilon / W_h(s) \geq \Delta_h(s,a) - \epsilon_\ell / W_h(s)$. Rearranging this gives the second conclusion.

The argument for the third conclusion is similar to the preceding argument. However, we now have the extra subtlety that for $a \neq a'$ with $a, a' \in \mathcal{A}_h^{\ell_\epsilon}(s)$, we may collect a different number of samples from (s,a) and (s,a') since it's possible that $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ and $(s,a') \in \mathcal{X}_{hj'}^{\ell_\epsilon+1}$ for $j \neq j'$. Denote

$$j(s) = \arg\max_j j \quad \text{s.t.} \quad \exists a, (s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}.$$

Note that, on $\mathcal{E}_{\text{exp}}$, we are guaranteed that there exists some $j \in [\iota_\epsilon]$ such that $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ so $j(s)$ is always well-defined. We can repeat the above argument, but now we can only guarantee that

$$|\widehat{Q}_{h,\ell_\epsilon+1}^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,a)| \leq \sqrt{\frac{H^2\iota_\delta}{n_{j(s)}^{\ell_\epsilon+1}}} = \frac{\epsilon}{8H\iota_\epsilon 2^{-j(s)+1}}.$$

since we can only guarantee we collect $n_{j(s)}^{\ell_\epsilon+1}$ samples from each $(s,a), a \in \mathcal{A}_h^{\ell_\epsilon}(s)$. It again follows that if

$$\Delta_h^{\widehat{\pi}}(s,a) \geq 4 \cdot \frac{\epsilon}{8H\iota_\epsilon 2^{-j(s)+1}}$$

then

$$\max_{a' \in \mathcal{A}_h^{\ell_\epsilon}(s)} \widehat{Q}_{h,\ell_\epsilon+1}^{\widehat{\pi}}(s,a') - \widehat{Q}_{h,\ell_\epsilon+1}^{\widehat{\pi}}(s,a) \geq 2 \cdot \frac{\epsilon}{8H\iota_\epsilon 2^{-j(s)+1}}.$$

As this is precisely the elimination criteria used in `EliminateActions`, it follows that a will be eliminated. Thus, all $a \in \mathcal{A}_h^{\ell_\epsilon+1}(s)$ must satisfy

$$\Delta_h^{\widehat{\pi}}(s,a) \leq 4 \cdot \frac{\epsilon}{8H\iota_\epsilon 2^{-j(s)+1}}$$

which gives the third conclusion.

■

Lemma C.2 and the definition of $\mathcal{E}_{\text{exp}}$ then let us prove that Moca returns an ϵ-optimal policy.

Lemma C.5 (Formal Statement of Lemma 6.5) *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, if* `Moca-SE` *is run with tolerance ϵ and* `FinalRound = true`*, then the policy $\widehat{\pi}$ returned by* `Moca-SE` *is ϵ-suboptimal.*

Proof Proposition 7 gives that, if $\widehat{\pi}$ satisfies $\max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \leq \epsilon_h(s)$ for all h and s, then $\widehat{\pi}$ is at most

$$\sum_{h=1}^{H} \sup_\pi \sum_s w_h^\pi(s)\epsilon_h(s) \tag{C.1}$$

suboptimal. When running Algorithm 4, for a particular h every state s can be classified in one of three ways:

- $s \notin \mathcal{Z}_h$: In this case, on $\mathcal{E}_{\text{exp}}$ we will have $\sup_\pi w_h^\pi(s) \le \epsilon/(2H^2S)$ and $\epsilon_h(s) \le H$.
- $s \in \mathcal{Z}_h$ and $|\mathcal{A}_h^{\ell_\epsilon+1}(s)| = 1$: In this case, by Lemma C.2, since $\widehat{\pi}$ only takes actions that are in $\mathcal{A}_h^{\ell_\epsilon+1}(s)$, we will have $\epsilon_h(s) = \max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) = 0$.
- $s \in \mathcal{Z}_h$, $|\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1$: Then we can apply Lemma C.2 to get

$$\epsilon_h(s) = \max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \le \frac{\epsilon}{2H\iota_\epsilon \cdot 2^{-j(s)+1}}$$

Let $\widetilde{\mathcal{X}}_j = \{s \ : \ j(s) = j\}$ and note that $\{s \in \mathcal{Z}_h \ : \ |\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1\} \subseteq \cup_{j=1}^{\iota_\epsilon} \widetilde{\mathcal{X}}_j$ since, on $\mathcal{E}_{\text{exp}}$, for every s satisfying $s \in \mathcal{Z}_h, |\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1$, we will have $(s,a) \in \mathcal{Z}_h^{\ell_\epsilon+1}$ for some a, so we must have that $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ for some $j \in [\iota_\epsilon]$. Furthermore, by definition of $j(s)$, if $s \in \widetilde{\mathcal{X}}_j$, then $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ for some a. Then, plugging all of this into Equation (C.1), on $\mathcal{E}_{\text{exp}}$,

$$\begin{aligned}
\sum_{h=1}^H \sup_\pi \sum_s w_h^\pi(s)\epsilon_h(s) &\le \sum_{h=1}^H \sup_\pi \sum_{j=1}^{\iota_\epsilon} \sum_{s \in \widetilde{\mathcal{X}}_j} w_h^\pi(s)\epsilon_h(s) + H\sum_{h=1}^H \sup_\pi \sum_{s \in \mathcal{Z}_h^c} w_h^\pi(s) \\
&\le \frac{\epsilon}{2H\iota_\epsilon} \sum_{h=1}^H \sup_\pi \sum_{j=1}^{\iota_\epsilon} \sum_{s \in \widetilde{\mathcal{X}}_j} w_h^\pi(s) 2^{j(s)-1} + H\sum_{h=1}^H \sup_\pi \sum_{s \in \mathcal{Z}_h^c} w_h^\pi(s) \\
&\overset{(a)}{\le} \frac{\epsilon}{2H\iota_\epsilon} \sum_{h=1}^H \sum_{j=1}^{\iota_\epsilon} 2^{j-1} \sup_\pi \sum_{(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}} w_h^\pi(s,a) + H\sum_{h=1}^H \sup_\pi \sum_{s \in \mathcal{Z}_h^c} w_h^\pi(s) \\
&\le \frac{\epsilon}{2H\iota_\epsilon} \sum_{h=1}^H \sum_{j=1}^{\iota_\epsilon} 2^{j-1}2^{-j+1} + H\sum_{h=1}^H \sum_{s \in \mathcal{Z}_h^c} \frac{\epsilon}{2H^2S} \\
&\le \epsilon
\end{aligned}$$

where (a) holds since for $s \in \widetilde{\mathcal{X}}_j$, $j(s) = j$, and since we can always choose π so that $\pi_h(s) = a$ so $w_h^\pi(s,a) = w_h^\pi(s)$. It follows that $\widehat{\pi}$ is at most ϵ-suboptimal. ■

C.2. Sample Complexity

We turn now to establishing a bound on the sample complexity of Moca. We first bound the complexity of a *single* call to `CollectSamples`.

Lemma C.6 `CollectSamples`$(\mathcal{Z}_{hi}^\ell, \{n_{ij}^\ell\}_{j=1}^{\iota_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H\iota_\epsilon\ell_\epsilon}, \frac{\epsilon_{\text{exp}}}{32})$ *terminates in at most*

$$\frac{cH^2\iota_\delta\iota_\epsilon}{\epsilon_\ell^2} \sum_{j=1}^{\iota_\epsilon} 2^j \sum_{(s,a) \in \mathcal{X}_{hij}^\ell} W_h(s)^2 + \frac{\text{poly}(S,A,H,\log 1/\delta, \log 1/\epsilon)}{\epsilon}$$

episodes and `CollectSamples`$(\mathcal{Z}_h^{\ell_\epsilon+1}, \{n_j^{\ell_\epsilon+1}\}_{j=1}^{\iota_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H}, \frac{\epsilon_{\text{exp}}}{32})$ *terminates in at most*

$$\frac{cH^4\iota_\delta\iota_\epsilon^2}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\text{poly}(S,A,H,\log 1/\delta, \log 1/\epsilon)}{\epsilon}$$

episodes.

Proof Recall that $\epsilon_{\text{exp}} = \frac{\epsilon}{2H^2S}$. The complexity of $\texttt{CollectSamples}(\mathcal{Z}^\ell_{hi}, n^\ell_i, h, \widehat{\pi}, \frac{\delta}{H\iota_\epsilon\ell_\epsilon}, \frac{\epsilon}{64H^2S})$ can be bounded by the sum of the complexity of calling $\texttt{Learn2Explore}$ to learn a set of exploration policies, and the complexity of playing these policies to collect samples. By Theorem 13, we can bound the complexity of calling $\texttt{Learn2Explore}$ by

$$C_K(\tfrac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, \iota_\epsilon)\frac{256H^2S}{\epsilon}$$

where $\delta_{\text{samp}} = \frac{\delta}{H\iota_\epsilon\ell_\epsilon} \cdot \frac{1}{\iota_\epsilon \max_j n^\ell_{ij}} \le \frac{\delta\epsilon^2_\ell}{2^{17}H^3\iota_\delta\iota^2_\epsilon\ell_\epsilon}$. As shown in Appendix D, $C_K(\frac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, \iota_\epsilon)$ is $\text{poly}(S, A, H, \log 1/\epsilon, \log 1/\delta)$, so this entire term is $\frac{\text{poly}(S,A,H,\log 1/\epsilon,\log 1/\delta)}{\epsilon}$.

Since rerunning the policies in Π^ℓ_{hij} yields at least $N^\ell_{hij}/2$ samples from each (s,a) in X^ℓ_{hij}, if we desire n^ℓ_{ij} samples from each (s,a), the complexity of running the policies returned by $\texttt{Learn2Explore}$ in order to collect the desired samples is clearly given by

$$\sum_{j=1}^{\iota_\epsilon} |\Pi^\ell_{hij}|\lceil 2n^\ell_{ij}/N^\ell_{hij}\rceil.$$

By the construction of Π^ℓ_{hij} and definition of N^ℓ_{hij} given in $\texttt{Learn2Explore}$, we have that

$$|\Pi^\ell_{hij}| = 2^j C_K(\tfrac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, j), \quad N^\ell_{hij} = \frac{|\Pi^\ell_{hij}|}{4M^\ell_{hij}2^j}.$$

where $M^\ell_{hij} = \sum_{j'=j}^{\iota_\epsilon+1} |\mathcal{X}^\ell_{hij'}|$ and $\mathcal{X}^\ell_{hi(\iota_\epsilon+1)} = \mathcal{Z}^\ell_{hi} \setminus \cup_{j=1}^{\iota_\epsilon} \mathcal{X}^\ell_{hij}$. As we are on $\mathcal{E}_{\text{exp}}$, $\mathcal{Z}^\ell_{hi} = \cup_{j=1}^{\iota_\epsilon}\mathcal{X}^\ell_{hij}$, so $|\mathcal{X}^\ell_{hi(\iota_\epsilon+1)}| = 0$. It follows that the complexity can be upper bounded as

$$\begin{aligned}
\sum_{j=1}^{\iota_\epsilon} |\Pi^\ell_{hij}|\lceil 2n^\ell_{ij}/N^\ell_{hij}\rceil &\le 8\sum_{j=1}^{\iota_\epsilon} 2^j M^\ell_{hij} n^\ell_{ij} + \sum_{j=1}^{\iota_\epsilon} 2^j C_K(\tfrac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, j) \\
&\le 8\sum_{j=1}^{\iota_\epsilon} 2^j M^\ell_{hij} n^\ell_{ij} + 2^{\iota_\epsilon+1} C_K(\tfrac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, \iota_\epsilon) \\
&= 8\frac{2^{17}H^2\iota_\delta}{2^{2i}\epsilon^2_\ell}\sum_{j=1}^{\iota_\epsilon} 2^j M^\ell_{hij} + 2^{\iota_\epsilon+1} C_K(\tfrac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, \iota_\epsilon)
\end{aligned}$$

The term $2^{\iota_\epsilon+1}C_K(\frac{\delta}{H\iota_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, \iota_\epsilon)$ is $\frac{\text{poly}(S,A,H,\log 1/\epsilon,\log 1/\delta)}{\epsilon}$ by definition of ι_ϵ and C_K. Furthermore,

$$\sum_{j=1}^{\iota_\epsilon} 2^j M^\ell_{hij} = \sum_{j=1}^{\iota_\epsilon} 2^j \sum_{j'=j}^{\iota_\epsilon} |\mathcal{X}^\ell_{hij'}| \le \iota_\epsilon \sum_{j=1}^{\iota_\epsilon} 2^j |\mathcal{X}^\ell_{hij}|.$$

We can therefore bound

$$\frac{2^{17}H^2\iota_\delta}{2^{2i}\epsilon^2_\ell}\sum_{j=1}^{\iota_\epsilon} 2^j M^\ell_{hij} \le \frac{cH^2\iota_\delta\iota_\epsilon}{\epsilon^2_\ell}\sum_{j=1}^{\iota_\epsilon} 2^{j-2i}|\mathcal{X}^\ell_{hij}|.$$

Finally, using that on $\mathcal{E}_{\mathrm{exp}}$ $W_h(s) \geq \widehat{W}_h(s)$, and that all $(s,a) \in \mathcal{X}^{\ell}_{hij}$ have a value of $\widehat{W}_h(s)$ within a factor of 2 of every other, we can upper bound $2^{-i} \leq 4W_h(s)$ for any $(s,a) \in \mathcal{X}^{\ell}_{hij}$. This completes the proof of the first claim.

The second claim follows similarly. By the same argument as above, we can upper bound the sample complexity of calling $\texttt{CollectSamples}(\mathcal{Z}_h^{\ell_\epsilon+1}, \{n_j^{\ell_\epsilon+1}\}_{j=1}^{\iota_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H}, \frac{\epsilon_{\mathrm{exp}}}{32})$ as

$$\begin{aligned}
&\sum_{j=1}^{\iota_\epsilon} |\Pi_{hj}^{\ell_\epsilon+1}| \lceil 2n_j^{\ell_\epsilon+1}/N_{hj}^{\ell_\epsilon+1} \rceil + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon} \\
&\qquad \leq 8\sum_{j=1}^{\iota_\epsilon} 2^j M_{hj}^{\ell_\epsilon+1} n_j^{\ell_\epsilon+1} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon} \\
&\qquad \overset{(a)}{=} \tfrac{cH^4\iota_\delta\iota_\epsilon^2}{\epsilon^2} \sum_{j=1}^{\iota_\epsilon} 2^{-j} M_{hj}^{\ell_\epsilon+1} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon} \\
&\qquad \overset{(b)}{\leq} \tfrac{cH^4\iota_\delta\iota_\epsilon^2}{\epsilon^2} |\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon}
\end{aligned}$$

where (a) follows by our setting of $n_j^{\ell_\epsilon+1}$ and (b) follows since $M_{hj}^{\ell_\epsilon+1} \leq |\mathcal{Z}_h^{\ell_\epsilon+1}|$. The second conclusion follows. ■

Using this, we show our main sample complexity lemma.

Lemma C.7 (Formal Statement of Lemma 6.2) *On the event $\mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_{\mathrm{exp}}$, for a given h and i, the loop over ℓ on Line 10 of* `MOCA-SE` *will take at most*

$$cH^2\iota_\delta\iota_\epsilon^2\ell_\epsilon \inf_\pi \max_{s\in\mathcal{Z}_{hi}} \max_a \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}\right\}$$

episodes. Furthermore, the total complexity of calling `MOCA-SE` *with* `FinalRound` $=$ `false` *is bounded by:*

$$H^2 c\iota_\delta\iota_\epsilon^3\ell_\epsilon \cdot \sum_{h=1}^H \inf_\pi \max_{s,a} \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}\right\} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon,\log 1/\delta)}{\epsilon}$$

for a universal constant c.

Proof With `FinalRound` $=$ `false`, the complexity of `MOCA-SE` is given by the complexity incurred calling `LEARN2EXPLORE` on Line 4 and calling `COLLECTSAMPLES` on Line 13. By Theorem 13 and since we call `LEARN2EXPLORE` at most SH times, we can bound the complexity of calling `LEARN2EXPLORE` by

$$\frac{\mathrm{poly}(S,A,H,\log 1/\epsilon,\log 1/\delta)}{\epsilon}.$$

Next, we turn to upper bounding the sample complexity of `LEARN2EXPLORE`. We can lower bound

$$|\mathcal{X}_{hij}^{\ell}| \sup_\pi \min_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^\pi(s,a) \leq \sup_\pi \sum_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^\pi(s,a).$$

so, on $\mathcal{E}_{\mathrm{exp}}$, $2^j \le 2(|\mathcal{X}^\ell_{hij}| \sup_\pi \min_{(s,a)\in\mathcal{X}^\ell_{hij}} w^\pi_h(s,a))^{-1}$. Plugging this into the bound given in Lemma C.6, we can bound the leading term in the sample complexity of a single call to `CollectSamples` as

$$\begin{aligned}\frac{cH^2\iota_\delta\iota_\epsilon}{\epsilon_\ell^2}\sum_{j=1}^{\iota_\epsilon} 2^j \sum_{(s,a)\in\mathcal{X}^\ell_{hij}} W_h(s)^2 &\le \frac{cH^2\iota_\delta\iota_\epsilon}{\epsilon_\ell^2}\sum_{j=1}^{\iota_\epsilon}\frac{1}{|\mathcal{X}^\ell_{hij}|\sup_\pi \min_{(s,a)\in\mathcal{X}^\ell_{hij}} w^\pi_h(s,a)}\sum_{(s,a)\in\mathcal{X}^\ell_{hij}} W_h(s)^2 \\ &\overset{(a)}{\le} \frac{cH^2\iota_\delta\iota_\epsilon}{\epsilon_\ell^2}\sum_{j=1}^{\iota_\epsilon}\inf_\pi \max_{(s,a)\in\mathcal{X}^\ell_{hij}}\frac{W_h(s)^2}{w^\pi_h(s,a)} \\ &\le \frac{cH^2\iota_\delta\iota_\epsilon^2}{\epsilon_\ell^2}\inf_\pi \max_{j\in\{1,\dots,\iota_\epsilon\}}\max_{(s,a)\in\mathcal{X}^\ell_{hij}}\frac{W_h(s)^2}{w^\pi_h(s,a)}\end{aligned}$$

where (a) holds since all $s \in \mathcal{X}^\ell_{hij}$ have values of $\widehat{W}_h(s)$ within a constant factor of each other, and since on $\mathcal{E}_{\mathrm{exp}}$ $\widehat{W}_h(s) \le W_h(s) \le 32\widehat{W}_h(s)$, which together imply that

$$\max_{s\in\mathcal{X}^\ell_{hij}} W_h(s) \le c \min_{s\in\mathcal{X}^\ell_{hij}} W_h(s).$$

If $(s,a) \in \mathcal{X}^\ell_{hij}$, then we must have that $(s,a) \in \mathcal{Z}^\ell_{hi}$ since $\mathcal{X}^\ell_{hij} \subseteq \mathcal{Z}^\ell_{hi}$, and, by the definition of $\mathcal{Z}^\ell_{hi}$, $a \in \mathcal{A}^{\ell-1}_h(s)$ and $|\mathcal{A}^{\ell-1}_h(s)| > 1$. Lemma C.2 gives that any $a \in \mathcal{A}^{\ell-1}_h(s)$ satisfies $\Delta_h(s,a) \le 3\epsilon_{\ell-1}/(2W_h(s))$. Since $|\mathcal{A}^{\ell-1}_h(s)| > 1$, it follows there exists a, a', $a \ne a'$, such that

$$\Delta_h(s,a) \le 3\epsilon_{\ell-1}/(2W_h(s)) \quad \text{and} \quad \Delta_h(s,a') \le 3\epsilon_{\ell-1}/(2W_h(s)).$$

Thus, if $(s,a) \in \mathcal{X}^\ell_{hij}$, $\frac{1}{4\epsilon_\ell^2} = \frac{1}{\epsilon_{\ell-1}^2} \le \frac{9}{4W_h(s)^2\Delta_h(s,a)^2}$ and $\frac{1}{4\epsilon_\ell^2} = \frac{1}{\epsilon_{\ell-1}^2} \le \frac{9}{4W_h(s)^2\Delta_h(s,a')^2}$, which implies $\frac{1}{4\epsilon_\ell^2} \le \frac{9}{4W_h(s)^2\max\{\Delta_h(s,a)^2,\Delta_h(s,a')^2\}}$. Note that $\max\{\Delta_h(s,a)^2, \Delta_h(s,a')^2\} \ge \widetilde{\Delta}_h(s,a)^2$ since if $\Delta_h(s,a) = 0$, we will have $\max\{\Delta_h(s,a)^2, \Delta_h(s,a')^2\} = \Delta_h(s,a')^2$, so either a is the unique optimal action at (s,h), in which case $\Delta_h(s,a') \ge \Delta_{\min}(s,h) = \widetilde{\Delta}_h(s,a)$, or there are multiple optimal actions, in which case $\Delta_h(s,a') \ge 0 = \widetilde{\Delta}_h(s,a)$. Thus,

$$\begin{aligned}&\frac{cH^2\iota_\delta\iota_\epsilon^2}{\epsilon_\ell^2}\inf_\pi \max_{j\in\{1,\dots,\iota_\epsilon\}}\max_{(s,a)\in\mathcal{X}^\ell_{hij}}\frac{W_h(s)^2}{w^\pi_h(s,a)} \\ &\qquad\le cH^2\iota_\delta\iota_\epsilon^2\inf_\pi \max_{j\in\{1,\dots,\iota_\epsilon\}}\max_{(s,a)\in\mathcal{X}^\ell_{hij}}\min\left\{\frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon_\ell^2}\right\} \\ &\qquad\le cH^2\iota_\delta\iota_\epsilon^2\inf_\pi \max_{(s,a)\in\mathcal{Z}^\ell_{hi}}\min\left\{\frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon_\ell^2}\right\}.\end{aligned}$$

Summing over ℓ and using that for all $(s,a) \in \mathcal{Z}^\ell_{hi}$, $s \in \mathcal{Z}_{hi}$, proves the first conclusion. Summing over i, and h gives

$$\sum_{h=1}^H\sum_{i=1}^{\iota_\epsilon} cH^2\iota_\delta\iota_\epsilon^2\ell_\epsilon \inf_\pi \max_{s\in\mathcal{Z}_{hi},a}\min\left\{\frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon^2}\right\}$$

$$\leq cH^2\iota_\delta\iota_\epsilon^3\ell_\epsilon\sum_{h=1}^{H}\inf_{\pi}\max_{s,a}\min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}\right\}.$$

This proves the result. ■

Finally, we bound the complexity of calling `MOCA-SE` with `FinalRound` = `true`.

Lemma C.8 (Formal Statement of Lemma 6.4) *On the event $\mathcal{E}_{\mathrm{est}}\cap\mathcal{E}_{\mathrm{exp}}$, if* `MOCA-SE` *is called with* ***`FinalRound` = `true`****, the procedure within the if statement on Line 15 will terminate after collecting at most*

$$\frac{cH^4\iota_\delta\iota_\epsilon^2}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\mathrm{poly}(S,A,H,\log 1/\delta,\log 1/\epsilon)}{\epsilon}$$

episodes. Furthermore, the total complexity of calling `MOCA-SE` *with* ***`FinalRound` = `true`*** *is bounded by:*

$$H^2c\iota_\delta\iota_\epsilon^3\ell_\epsilon\cdot\mathcal{C}(\mathcal{M},\epsilon) + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon,\log 1/\delta)}{\epsilon}$$

for a universal constant c.

Proof The only additional samples taken when running `MOCA-SE` with `FinalRound` = `true` as compared to running it with `FinalRound` = `false` is incurred by calling `COLLECTSAMPLES` on Line 18 of `MOCA-SE`. Thus, the total complexity can be bounded by adding the complexity bound from Lemma C.7 to this additional cost.

In particular, by Lemma C.6, this additional call of `COLLECTSAMPLES` will require at most

$$\frac{cH^4\iota_\delta\iota_\epsilon^2}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\mathrm{poly}(S,A,H,\log 1/\delta,\log 1/\epsilon)}{\epsilon}$$

episodes to terminate, from which the first conclusion follows. We can repeat the argument from the proof of Lemma C.7 to get that $\mathcal{Z}_h^{\ell_\epsilon+1}\subseteq\mathcal{W}_h^{\ell_\epsilon+1}$, where we define $\mathcal{W}_h^{\ell_\epsilon} := \{(s,a) : s\in\mathcal{Z}_h, \exists a'\neq a, \max\{\Delta_h(s,a),\Delta_h(s,a')\}\leq 3\epsilon_{\ell_\epsilon-1}/(2W_h(s))\}$. However, note that $\epsilon_{\ell_\epsilon-1}\leq 2\epsilon$, and the condition $\exists a'\neq a, \max\{\Delta_h(s,a),\Delta_h(s,a')\}\leq 3\epsilon_{\ell_\epsilon-1}/(2W_h(s))$ implies $\widetilde{\Delta}_h(s,a)\leq 3\epsilon_{\ell_\epsilon-1}/(2W_h(s))$. It follows that

$$\mathcal{W}_h^{\ell_\epsilon+1}\subseteq\left\{(s,a) \;:\; \widetilde{\Delta}_h(s,a)\leq 3\epsilon/W_h(s)\right\} =: \mathrm{OPT}(\epsilon,h)$$

Summing over h gives the result.

■

C.3. Proof of Theorem 2

We are finally ready to complete the proof of Theorem 2.

Proof [Proof of Theorem 2] Note that $\mathbb{P}[\mathcal{E}_{\mathrm{est}}\cap\mathcal{E}_{\mathrm{exp}}]\geq 1-\delta$ by Lemma C.1. We will assume for the remainder of the proof that this event holds.

Case 1: $\epsilon_{\mathrm{tol}} \geq \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$**.** In this case, that the policy returned is ϵ_{tol}-optimal is guaranteed by Lemma C.5 since the final call to `MOCA-SE` is run with `FinalRound` = `true`. To bound the sample complexity, we can then simply combine Lemma C.7 and Lemma C.8, which gives that the total sample complexity is bounded as (using that $\epsilon_{\mathrm{tol}(m)} \geq \epsilon_{\mathrm{tol}}$ and that $\delta_{\mathrm{tol}(m)} \geq \delta_{\mathrm{tol}}/(36\lceil \log H/\epsilon_{\mathrm{tol}}\rceil^2) =: \delta'$):

$$\sum_{m=1}^{\lceil \log H/\epsilon_{\mathrm{tol}}\rceil - 1} H^2 c\iota_{\delta_{\mathrm{tol}(m)}} \iota^3_{\epsilon_{\mathrm{tol}(m)}} \ell_{\epsilon_{\mathrm{tol}(m)}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon^2_{\mathrm{tol}(m)}} \right\}$$
$$+ H^2 c\iota_{\delta_{\mathrm{tol}(m)}} \iota^3_{\epsilon_{\mathrm{tol}}} \ell_{\epsilon_{\mathrm{tol}}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon^2_{\mathrm{tol}}} \right\}$$
$$+ \frac{cH^4 \iota_{\delta_{\mathrm{tol}}} \iota^2_{\epsilon_{\mathrm{tol}}} |\mathrm{OPT}(\epsilon_{\mathrm{tol}})|}{\epsilon^2_{\mathrm{tol}}} + \frac{\lceil \log H/\epsilon_{\mathrm{tol}}\rceil \cdot \mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}}, \log 1/\delta_{\mathrm{tol}})}{\epsilon_{\mathrm{tol}}}.$$

This can be upper bounded as

$$\lceil \log H/\epsilon_{\mathrm{tol}}\rceil \cdot H^2 c\iota_{\delta'} \iota^3_{\epsilon_{\mathrm{tol}}} \ell_{\epsilon_{\mathrm{tol}}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon^2_{\mathrm{tol}}} \right\}$$
$$+ \frac{cH^4 \iota_{\delta'} \iota^2_{\epsilon_{\mathrm{tol}}} |\mathrm{OPT}(\epsilon_{\mathrm{tol}})|}{\epsilon^2_{\mathrm{tol}}} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}}, \log 1/\delta_{\mathrm{tol}})}{\epsilon_{\mathrm{tol}}}.$$

This and the definition of $\mathcal{C}(\mathcal{M},\epsilon)$ gives the first conclusion of Theorem 2.

Case 2: $\epsilon_{\mathrm{tol}} < \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$**.** As we showed in the proof of Lemma C.8, we will have that $\mathcal{Z}_h^{\ell_\epsilon+1} \subseteq \mathrm{OPT}(\epsilon, h)$. Therefore, if for all (s,a), $\widetilde{\Delta}_h(s,a) > 3\epsilon/W_h(s)$, we will have that $|\mathcal{Z}_h^{\ell_\epsilon+1}| = 0$, which implies that for every $s \in \mathcal{Z}_h$, $|\mathcal{A}_h^{\ell_\epsilon}(s)| = 1$. Furthermore, on $\mathcal{E}_{\mathrm{exp}}$, we will have that $\mathcal{Z}_h = \mathcal{S} \times \mathcal{A}$ if $\frac{\epsilon}{2H^2 S} < \min_s W_h(s)$. If each of these conditions hold for all h, then the returned sets $\mathcal{A}_h^{\ell_\epsilon+1}(s)$ will satisfy $|\mathcal{A}_h^{\ell_\epsilon+1}(s)|$ for all s and h.

It follows then that if $\epsilon_{\mathrm{tol}} < \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$, either $\epsilon_{\mathrm{tol}(m)} < \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$ for some m, in which case the above condition will be met, and the termination criteria on Line 6 of MOCA will be satisfied, or

$$\epsilon_{\mathrm{tol}(m)} \geq \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\},$$

and MOCA will reach the final call of `MOCA-SE` with `FinalRound` = `true`. In the former case, letting $\bar{m}$ denote the value of m at which MOCA terminates, the total sample complexity will be bounded as, using the same argument as in Case 1,

$$\sum_{m=1}^{\bar{m}} H^2 c\iota_{\delta_{\mathrm{tol}(\bar{m})}} \iota^3_{\epsilon_{\mathrm{tol}(\bar{m})}} \ell_{\epsilon_{\mathrm{tol}(\bar{m})}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w^\pi_h(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w^\pi_h(s,a)\epsilon^2_{\mathrm{tol}(m)}} \right\}$$
$$+ \frac{\bar{m} \cdot \mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}(\bar{m})}, \log 1/\delta_{\mathrm{tol}(\bar{m})})}{\epsilon_{\mathrm{tol}(\bar{m})}}$$

$$\leq \bar{m}H^2 c\iota_{\delta_{\mathrm{tol}(\bar{m})}}\iota^3_{\epsilon_{\mathrm{tol}(\bar{m})}}\ell_{\epsilon_{\mathrm{tol}(\bar{m})}} \cdot \sum_{h=1}^{H} \inf_{\pi}\max_{s,a}\min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2_{\mathrm{tol}(\bar{m})}}\right\} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}(\bar{m})},\log 1/\delta_{\mathrm{tol}(\bar{m})})}{\epsilon_{\mathrm{tol}(\bar{m})}}$$

and note that $\epsilon_{\mathrm{tol}(\bar{m}-1)} \geq \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$, since we did not terminate at round $\bar{m}-1$, implying that $\epsilon_{\mathrm{tol}(\bar{m})} \geq 2\min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$. Note also that $\delta_{\mathrm{tol}(\bar{m})} = \frac{\delta}{36\log^2 \epsilon_{\mathrm{tol}(\bar{m})}}$, so we can also bound

$$\log 1/\delta_{\mathrm{tol}(\bar{m})} \leq \mathcal{O}(\log 1/\delta_{\mathrm{tol}} + \log\log(2\min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\})).$$

Together these give the bound stated in Theorem 2.

In the latter case, when we do not terminate early at Line 6, the same sample complexity bound applies but with $\epsilon_{\mathrm{tol}(\bar{m})}$ replaced by ϵ_{tol}, since if $|\mathcal{Z}_h^{\ell_\epsilon+1}| = 0$, the final call to CollectSamples in Line 18 of Moca-SE will not collect any samples. As before, in this case we can lower bound

$$\epsilon_{\mathrm{tol}} \geq 2\min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$$

from which the bound follows.

It remains to show that $\widehat{\pi} = \pi^\star$. This follows inductively from Lemma C.2 since if $|\mathcal{A}_H^\ell(s)| = 1$, this implies that for $a \in \mathcal{A}_H^\ell(s)$, $a = \pi_H^\star(s)$. Then if we assume that $\widehat{\pi}_{h'}(s) = \pi^\star_{h'}(s)$ for all s and $h' > h$, if $|\mathcal{A}_h^\ell(s)| = 1$ this implies that for $a \in \mathcal{A}_h^\ell(s)$, $a = \pi_h^\star(s)$ since, by Lemma C.2, in this case

$$\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) = 0$$

but $Q_h^{\widehat{\pi}}(s,a'') = Q_h^\star(s,a'')$. Thus, it follows that $\widehat{\pi} = \pi^\star$, which completes the proof.

■

C.4. Proofs of Additional Lemmas and Claims

Proof [Proof of Lemma C.1] $\mathcal{E}_{\mathrm{est}}$ **holds.** That $\mathcal{E}_{\mathrm{est}}$ holds with probability $1-\delta_{\mathrm{tol}}/2$ follows directly from Hoeffding's inequality and a union bound, since $\mathring{Q}_h^{\widehat{\pi},t}(s_h^t,a_h^t) \leq H$ almost surely. In particular, note that for any given call to Moca-SE, we will form at most $SAH\iota_\epsilon(\ell_\epsilon+1)$ estimates of $Q_h^{\widehat{\pi}}(s,a)$. By Hoeffding's inequality and our choice of ι_δ, that each of these estimates concentrates as given on $\mathcal{E}_{\mathrm{est}}$ then holds with probability

$$1 - SAH\iota_\epsilon(\ell_\epsilon+1)\cdot\frac{\delta}{SAH\iota_\epsilon(\ell_\epsilon+1)} = 1-\delta.$$

With our choice of $\delta_{\mathrm{tol}(m)} = \frac{\delta_{\mathrm{tol}}}{36m^2}$, union bounding over this holding for each call to Moca-SE, we then have that $\mathcal{E}_{\mathrm{est}}$ holds with probability at least

$$1 - \sum_{m=1}^{\lceil \log H/\epsilon\rceil} \frac{\delta_{\mathrm{tol}}}{36m^2} \geq 1 - \frac{\delta_{\mathrm{tol}}}{2},$$

which is the desired result.

$\mathcal{E}_{\exp}$ holds. We show that the desired events hold for a single call of MOCA-SE, then union bound over all calls to MOCA-SE to get the final result. Let $\mathcal{E}_{\exp}^m$ denote the event on which all conditions of $\mathcal{E}_{\exp}$ hold for the mth call to MOCA-SE.

Assume that we run MOCA-SE with tolerance $\epsilon_{\mathrm{tol}(m)}$ and confidence $\delta_{\mathrm{tol}(m)}$. Let $\mathcal{E}_{\mathrm{L2E}}^{sh}$ denote the success event of calling LEARN2EXPLORE on Line 4, $\mathcal{E}_{\mathrm{L2E}}^{hi\ell}$ denote the success event of calling LEARN2EXPLORE in the call to COLLECTSAMPLES at iteration (h, i, ℓ) on Line 13, and $\mathcal{E}_{\mathrm{L2E}}^{h}$ the success event of calling LEARN2EXPLORE in the call to COLLECTSAMPLES on Line 18. By Theorem 13 and the confidence with which we call LEARN2EXPLORE, we have that $\mathbb{P}[\mathcal{E}_{\mathrm{L2E}}^{sh}] \geq 1 - \delta_{\mathrm{tol}(m)}/SH$, $\mathbb{P}[\mathcal{E}_{\mathrm{L2E}}^{hi\ell}] \geq 1 - \delta_{\mathrm{tol}(m)}/(H\iota_\epsilon\ell_\epsilon)$, and $\mathbb{P}[\mathcal{E}_{\mathrm{L2E}}^{h}] \geq 1 - \delta_{\mathrm{tol}(m)}/H$. Union bounding over these events, and using that there are at most $H\iota_\epsilon\ell_\epsilon$ indices (h, i, ℓ), we get that the event

$$(\cap_{s,h}\mathcal{E}_{\mathrm{L2E}}^{sh}) \cap (\cap_{h=1}^{H} \cap_{i=1}^{\iota_\epsilon} \cap_{\ell=1}^{\ell_\epsilon} \mathcal{E}_{\mathrm{L2E}}^{hi\ell}) \cap (\cap_{h=1}^{H} \mathcal{E}_{\mathrm{L2E}}^{h})$$

holds with probability at least $1 - 3\delta_{\mathrm{tol}(m)}$.

That

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^{\pi}(s,a) \leq 2^{-j+1}$$

for $j \in [\iota_\epsilon]$, is a direct consequence of $\mathcal{E}_{\mathrm{L2E}}^{hi\ell}$ holding, and similarly that

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}_{hj}^{\ell_\epsilon+1}} w_h^{\pi}(s,a) \leq 2^{-j+1}$$

holds for $j \in [\iota_\epsilon]$, is a direct consequence of $\mathcal{E}_{\mathrm{L2E}}^{h}$. In addition, that

$$\sup_{\pi} \max_{s\in\mathcal{Z}_h^c} w_h^{\pi}(s) \leq \frac{\epsilon}{2H^2S}$$

holds for all h is immediate on $\cap_{s,h}\mathcal{E}_{\mathrm{L2E}}^{sh}$.

On the event $\mathcal{E}_{\mathrm{L2E}}^{hi\ell}$, if we run the policies returned by LEARN2EXPLORE for some $j \in \{1, \ldots, \iota_\epsilon\}$, Π_{hij}^{ℓ}, Theorem 13 and our choice of δ_{samp} gives that we will collect at least $\frac{1}{2}N_{hij}^{\ell}$ samples from each $(s,a) \in \mathcal{X}_{hij}^{\ell}$ with probability at least $1 - \delta_{\mathrm{tol}(m)}/(H\iota_\epsilon^2\ell_\epsilon n_{i1}^{\ell})$. As COLLECTSAMPLES runs each policy $\lceil 2n_{i1}^{\ell}/N_{hij}^{\ell} \rceil$ times, it follows that we will collect at least $\lceil 2n_{i1}^{\ell}/N_{hij}^{\ell} \rceil \cdot \frac{1}{2}N_{hij}^{\ell} \geq n_{i1}^{\ell}$ samples from each $(s,a) \in \mathcal{X}_{hij}^{\ell}$ with probability at least $1 - \delta_{\mathrm{tol}(m)}/(H\iota_\epsilon^2\ell_\epsilon n_{i1}^{\ell}) \cdot \lceil 2n_{1i}^{\ell}/N_{hij}^{\ell} \rceil \geq 1 - 3\delta_{\mathrm{tol}(m)}/(H\iota_\epsilon^2\ell_\epsilon)$. Union bounding over this for each h, i, ℓ and $j \in [\iota_\epsilon]$ gives that with probability at least $1 - 3\delta_{\mathrm{tol}(m)}$, we collect at least n_{i1}^{ℓ} samples from each $(s,a) \in \mathcal{X}_{hij}^{\ell}$. The same argument gives that with probability at least $1 - 3\delta_{\mathrm{tol}(m)}$ we collect at least $n_j^{\ell_\epsilon+1}$ samples from each $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$, $j = 1, \ldots, \iota_\epsilon, h \in [H]$.

Relating $\widehat{W}_h(s)$ to $W_h(s)$. It remains to show that $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$ for all $s \in \mathcal{Z}_h$, $\cup_{j=1}^{\iota_\epsilon}\mathcal{X}_{hij}^{\ell} = \mathcal{Z}_{hi}^{\ell}$, and $\cup_{j=1}^{\iota_\epsilon}\mathcal{X}_{hj}^{\ell_\epsilon+1} = \mathcal{Z}_h^{\ell_\epsilon+1}$.

We first show $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$. Consider running LEARN2EXPLORE with $\mathcal{X} = \{(s,a)\}$ for arbitrary a and assume that $\mathcal{X}_j^{sh}$ is the returned partition containing (s,a). By Theorem 13, on $\mathcal{E}_{\mathrm{L2E}}^{sh}$ we will have that

$$W_h(s) \leq 2^{-j+1}$$

and, furthermore, that with probability at least $1/2$, if we rerun all policies in Π_j^{sh} returned by LEARN2EXPLORE, we will obtain at least $N_j^{sh}/2 = |\Pi_j^{sh}|/(8|\mathcal{X}|2^j) = |\Pi_j^{sh}|/(8 \cdot 2^j)$ samples from (s, a, h).

Let X be a random variable which is the count of total samples collected in (s, a, h) when running $\pi_k \in \Pi_j^{sh}$. Then Markov's inequality and the above property of Π_j^{sh} gives

$$\frac{1}{2} \le \mathbb{P}[X \ge N_j^{sh}/2] \le \frac{2\mathbb{E}[X]}{N_j^{sh}} = \frac{2}{N_j^{sh}} \sum_{\pi \in \Pi_j^{sh}} w_h^\pi(s,a) \le \frac{2|\Pi_j^{sh}|}{N_j^{sh}} W_h(s) = 8 \cdot 2^j W_h(s).$$

Rearranging this and recalling that we set $\widehat{W}_h(s) = \frac{1}{16 \cdot 2^j}$, we have that $\widehat{W}_h(s) \le W_h(s)$. However, we also have

$$W_h(s) \le 2^{-j+1} = 32\widehat{W}_h(s).$$

This proves that $\widehat{W}_h(s) \le W_h(s) \le 32\widehat{W}_h(s)$.

Now note that any $s \in \mathcal{Z}_h$ has $\widehat{W}_h(s) \ge \frac{\epsilon_{\mathrm{tol}}(m)}{32H^2S}$, which, combined with the above, implies that $W_h(s) \ge \frac{\epsilon_{\mathrm{tol}}(m)}{32H^2S}$. Fix (h, i, ℓ), and note that the call to LEARN2EXPLORE in the call to COLLECTSAMPLES for index (h, i, ℓ) uses input tolerance $\frac{\epsilon_{\mathrm{tol}}(m)}{64H^2S}$. Theorem 13 then gives that, on $\mathcal{E}_{\mathtt{L2E}}^{hi\ell}$, we will have

$$\sup_\pi \sum_{(s,a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hij}^\ell)} w_h^\pi(s,a) \le \frac{\epsilon_{\mathrm{tol}}(m)}{64H^2S}.$$

However, as $W_h(s') \le \sup_\pi \sum_{(s,a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hij}^\ell)} w_h^\pi(s,a)$ for any $(s', a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hij}^\ell)$, we will have that any $(s, a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hij}^\ell)$ has $W_h(s) \le \frac{\epsilon_{\mathrm{tol}}(m)}{64H^2S}$. This is a contradiction since we know $W_h(s) \ge \frac{\epsilon_{\mathrm{tol}}(m)}{32H^2S}$ for any $(s, a) \in \mathcal{Z}_{hi}^\ell$. Thus, we must have that $\mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hij}^\ell) = \emptyset$ so $\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hij}^\ell = \mathcal{Z}_{hi}^\ell$. The same argument shows that $\cup_{j=1}^{\iota_\epsilon} \mathcal{X}_{hj}^{\ell_\epsilon+1} = \mathcal{Z}_h^{\ell_\epsilon+1}$.

Completing the proof. We have therefore shown that $\mathbb{P}[\mathcal{E}_{\mathrm{exp}}^m] \ge 1 - 9\delta_{\mathrm{tol}(m)}$. Union bounding over all m, by our choice of $\delta_{\mathrm{tol}(m)} = \frac{\delta_{\mathrm{tol}}}{36m^2}$, we have that

$$\mathbb{P}[\mathcal{E}_{\mathrm{exp}}] = \mathbb{P}[\cap_{m=1}^{\lceil \log H/\epsilon \rceil} \mathcal{E}_{\mathrm{exp}}^m] \ge 1 - \sum_{m=1}^{\lceil \log H/\epsilon \rceil} 9\frac{\delta_{\mathrm{tol}}}{36m^2} \ge 1 - \delta_{\mathrm{tol}}/2.$$

Union bounding over $\mathcal{E}_{\mathrm{exp}}$ and $\mathcal{E}_{\mathrm{est}}$ then gives the result. ■

Proof [Proof of Claim C.3] We proceed by induction. Consider some $s \in \mathcal{Z}_{hi}$. The base case is trivial as $\mathcal{A}_h^0(s) = \mathcal{A}$. Fix some $\ell \le \ell_\epsilon$ and assume that $\widehat{a}_h^\star(s) \in \mathcal{A}_h^{\ell-1}(s)$ and $|\mathcal{A}_h^{\ell-1}(s)| > 1$. Then, on $\mathcal{E}_{\mathrm{exp}}$, we can guarantee that we will collect at least $\frac{2^{18}H^2\iota_\delta}{2^{2i}\epsilon_\ell^2}$ samples from (s, a) for each $a \in \mathcal{A}_h^{\ell-1}$. On the event $\mathcal{E}_{\mathrm{est}}$, it then follows that for each $a \in \mathcal{A}_h^{\ell-1}(s)$,

$$|\widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,a)| \le 2^i\epsilon_\ell/2^9.$$

Thus, since by assumption $\widehat{a}^\star_h(s) \in \mathcal{A}^{\ell-1}_h(s)$,

$$\begin{aligned}\max_{a\in\mathcal{A}^{\ell-1}_h(s)} \widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a) - \widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,\widehat{a}^\star_h(s)) &\le \max_{a\in\mathcal{A}^{\ell-1}_h(s)} Q^{\widehat{\pi}}_h(s,a) - Q^{\widehat{\pi}}_h(s,\widehat{a}^\star_h(s)) + 2\cdot 2^i\epsilon_\ell/2^9 \\ &\le 2\cdot 2^i\epsilon_\ell/2^9 \\ &= \gamma^\ell_{ij}\end{aligned}$$

for any j, so the exit condition on Line 16 of EliminateActions is not met for $\widehat{a}^\star_h(s)$, and thus $\widehat{a}^\star_h(s) \in \mathcal{A}^\ell_h(s)$. The result follows analogously if $\ell = \ell_\epsilon + 1$, in which case we simply use the different values of n and γ.

Now if $(s,a) \notin \mathcal{Z}^\ell_{hi}$ for all a, that means we will never remove arms from $\mathcal{A}^\ell_h(s)$ again. However, by the above inductive argument, if ℓ' is the last round such that $(s,a) \in \mathcal{Z}^{\ell'}_{hi}$ for some a, we will have that $\widehat{a}^\star_h(s) \in \mathcal{A}^{\ell'}_h(s)$, so it follows that $s \in \mathcal{A}^\ell_h(s)$.

Finally, if $s \notin \mathcal{Z}_h$, then we will never remove an arm from $\mathcal{A}^0_h(s)$, and since $\mathcal{A}^0_h(s) = \mathcal{A}$, the conclusion follows trivially. ■

Proof [Proof of Claim C.4] In Lemma C.5, we showed that the local suboptimality bounds of $\widehat{\pi}$, $\epsilon_h(s)$, satisfy

$$\sum_{h=1}^H \sup_\pi \sum_s w^\pi_h(s)\epsilon_h(s) \le \epsilon.$$

By Lemma B.1, it follows that for any π' and any h,

$$\sum_s w^{\pi'}_h(s)(V^\star_h(s) - V^{\widehat{\pi}}_h(s)) \le \sum_{h'=h}^H \sup_\pi \sum_s w^\pi_{h'}(s)\epsilon_{h'}(s) \le \epsilon.$$

The result then follows from Lemma B.2. ■

Appendix D. Learning to Explore

Define the following value:

$$\begin{aligned}K_i(\delta,\delta_{\mathrm{samp}}) = \Bigg\lceil 2^i \max\Bigg\{ &288c^2_{\mathrm{eu}}S^2A^2H(i+3)\log(576c_{\mathrm{eu}}SAH(i+3)), 288c^2_{\mathrm{eu}}S^2A^2H\log\frac{2SAH}{\delta}, \\ &2048S^2A^2\log\frac{4SAH}{\delta_{\mathrm{samp}}}, 256c_{\mathrm{eu}}S^3A^2H^4(i+9)^3\log^3\left(512c_{\mathrm{eu}}SAH(i+9)\right), \\ &128c_{\mathrm{eu}}S^3A^2H^4\log^3\frac{2SAH}{\delta} + 8H\log\frac{4}{\delta}\Bigg\}\Bigg\rceil \\ =: 2^iC_K(\delta,\delta_{\mathrm{samp}},i)\end{aligned} \tag{D.1}$$

and note that $C_K(\delta,\delta_{\mathrm{samp}},i) = \mathrm{poly}(S,A,H,\log 1/\delta,\log 1/\delta_{\mathrm{samp}},i)$.

Remark D.1 *The exploration procedure of* `FindExplorableSets` *is potentially quite wasteful as we restart* Euler *every time the desired number of samples for a given state is collected. This could likely be improved on by instead running a regret-minimization algorithm that is able to handle time-varying rewards, such as the algorithm presented in Zhang et al. (2020a). As the focus of this work is not in optimizing the lower-order terms, we chose to instead simply use* Euler.

Theorem 13 (Formal Statement of Theorem 8) *Consider running* `Learn2Explore` *with tolerance* $\epsilon_{L2E} \leftarrow \epsilon$ *and confidence* δ *and obtaining a partition* $\mathcal{X}_i \subseteq \mathcal{S} \times \mathcal{A}$ *and policies* Π_i, $i \in \{1, 2, \dots, \lceil \log(1/\epsilon) \rceil\}$. *Let* $\mathcal{E}_{\mathtt{L2E}}$ *be the event on which, for all* i *simultaneously:*

1. *Sets* $\mathcal{X}_i$ *satisfy:*
$$\sup_\pi \sum_{(s,a) \in \mathcal{X}_i} w_h^\pi(s,a) \leq 2^{-(i-1)}$$
2. *For any* i, *if all policies in* Π_i *are each rerun once, we will collect* $\frac{1}{2}N_i$ *samples from each* $(s,a) \in \mathcal{X}_i$ *with probability* $1 - \delta_{\mathrm{samp}}$, *where we recall* $N_i = K_i(\delta/\lceil \log(1/\epsilon) \rceil, \delta_{\mathrm{samp}})/(4 \cdot 2^i |\mathcal{X} \setminus \cup_{i'=1}^{i-1} \mathcal{X}_{i'}|)$.
3. *The remaining states,* $\mathcal{X} \setminus (\cup_{i=1}^{\lceil \log(1/\epsilon) \rceil} \mathcal{X}_i)$ *satisfy,*
$$\sup_\pi \sum_{(s,a) \in (\mathcal{X} \setminus (\cup_{i=1}^{\lceil \log(1/\epsilon) \rceil} \mathcal{X}_i))} w_h^\pi(s,a) \leq \epsilon.$$

Then $\mathbb{P}[\mathcal{E}_{\mathtt{L2E}}] \geq 1 - \delta$. *Furthermore, Algorithm 3 takes at most*
$$C_K\left(\frac{\delta}{\lceil \log 1/\epsilon \rceil}, \delta_{\mathrm{samp}}, \lceil \log 1/\epsilon \rceil\right) \frac{4}{\epsilon}$$
episodes to terminate.

Proof This directly follows by induction and Lemma D.1. For $i = 1$, it will clearly be the case that
$$\sup_\pi \sum_{(s,a) \in \mathcal{X}} w_h^\pi(s,a) \leq 2^{-(i-1)} = 1$$
since $\sum_{s,a} w_h^\pi(s,a) = 1$ for any π and h. Now consider an epoch i and assume that
$$\sup_\pi \sum_{(s,a) \in \mathcal{X}} w_h^\pi(s,a) \leq 2^{-(i-1)}.$$
By Lemma D.1, running `FindExplorableSets` will produce a set $\mathcal{X}_i$ and policies Π_i such that
$$\sup_\pi \sum_{(s,a) \in \mathcal{X}_i} w_h^\pi(s,a) \leq 2^{-(i-1)}, \quad \sup_\pi \sum_{(s,a) \in \mathcal{X} \setminus \mathcal{X}_i} w_h^\pi(s,a) \leq 2^{-i}$$

and rerunning every policy in Π_i at once will allow us to collect at least $\frac{1}{2}N_i$ samples from each $(s,a) \in \mathcal{X}_i$. As $\mathcal{X} \leftarrow \mathcal{X} \backslash \mathcal{X}_i$, the hypothesis will then be met at the next epoch, $i+1$. Union bounding over epochs completes the first part of the proof. That

$$\sup_{\pi} \sum_{(s,a) \in (\mathcal{X} \backslash (\cup_{i=1}^{\lceil \log(1/\epsilon) \rceil} \mathcal{X}_i))} w_h^\pi(s,a) \leq \epsilon$$

follows on this same event by Lemma D.1 and since we run until $i = \lceil \log(1/\epsilon) \rceil$ which implies $2^{-\lceil \log(1/\epsilon) \rceil} \leq \epsilon$. Union bounding over each i gives the result.

The sample complexity bound follows by bounding

$$\begin{aligned}\sum_{i=1}^{\lceil \log(1/\epsilon) \rceil} K_i(\delta/\lceil \log 1/\epsilon \rceil, \delta_{\text{samp}}) &\leq C_K \left(\frac{\delta}{\lceil \log 1/\epsilon \rceil}, \delta_{\text{samp}}, \lceil \log 1/\epsilon \rceil \right) \sum_{i=1}^{\lceil \log(1/\epsilon) \rceil} 2^i \\ &\leq C_K \left(\frac{\delta}{\lceil \log 1/\epsilon \rceil}, \delta_{\text{samp}}, \lceil \log 1/\epsilon \rceil \right) \frac{4}{\epsilon}.\end{aligned}$$

■

Lemma D.1 *Assume that $\mathcal{X}$ satisfies*

$$\sup_{\pi} \sum_{(s,a) \in \mathcal{X}} w_h^\pi(s,a) \leq 2^{-(i-1)}.$$

Then, if `FindExplorableSets`$(\mathcal{X}, h, \delta, K_i, N_i)$ *returns partition $\mathcal{X}_i$ and policies Π_i, with probability $1-\delta$ the returned partition $\mathcal{X}_i$ will satisfy*

$$\sup_{\pi} \sum_{(s,a) \in \mathcal{X}_i} w_h^\pi(s,a) \leq 2^{-(i-1)}, \quad \sup_{\pi} \sum_{(s,a) \in \mathcal{X} \backslash \mathcal{X}_i} w_h^\pi(s,a) \leq 2^{-i}.$$

Furthermore, if all policies in Π_i are each rerun once, we will collect $\frac{1}{2}N_i$ samples from each $(s,a,h) \in \mathcal{X}_i$ with probability $1 - \delta_{\text{samp}}$.

Proof The structure of this proof takes inspiration from the proof presented in Zhang et al. (2020a). The first conclusion is trivial since $\mathcal{X}_i \subseteq \mathcal{X}$ and by our assumption on $\mathcal{X}$.

We will simply denote $K_i := K_i(\delta, \delta_{\text{samp}})$ throughout the proof. In addition, we will let K_{ij} denote the total number of epochs taken for fixed j, and will let m_i denote the total number of times j is incremented. Therefore,

$$K_i = \sum_{j=1}^{m_i} K_{ij}.$$

Let $V_0^{\star,ij}$ denote the optimal value function on the reward function r_h^j at stage j of epoch i. By our assumption on $\mathcal{X}$ and the definition of our reward function we can bound

$$V_0^{\star,ij} \leq \sup_{\pi} \mathbb{E}_\pi[\mathbb{I}\{(s_h, a_h) \in \mathcal{X}\}] = \sup_{\pi} \sum_{(s,a) \in \mathcal{X}} w_h^\pi(s,a) \leq 2^{-(i-1)}. \tag{D.2}$$

As `FindExplorableSets` runs Euler, by Lemma D.4 we will have, with probability at least $1-\delta$, for any fixed K and j,

$$\Big(\sum_{k=1}^{K} V_0^{\star,ij} - \sum_{k=1}^{K} V_0^{k,ij}\Big)|\mathcal{F}_{j-1} \le c_{\mathrm{eu}}\sqrt{SAHV_0^{\star,i1}K\log\frac{SAHK}{\delta}} + c_{\mathrm{eu}}S^2AH^4\log^3\frac{SAHK}{\delta} \tag{D.3}$$

where $\mathcal{F}_{j-1}$ denotes the filtration of up to iteration j, and we have used that $V_0^{\star,ij} \le V_0^{\star,i1}$ for all j since the reward function can only decrease as j increases. `FindExplorableSets` terminates and restarts Euler if the condition on Line 14 is met, but this is a *random* stopping condition. As such, to guarantee that (D.3) holds for any possible value of this stopping time, we union bound over all values. Since `FindExplorableSets` runs for at most K_i epochs, it suffices to union bound over K_i stopping times. We then have that

$$\Big(\sum_{k=1}^{K} V_0^{\star,ij} - \sum_{k=1}^{K} V_0^{k,ij}\Big)|\mathcal{F}_{j-1} \le 2c_{\mathrm{eu}}\sqrt{SAHV_0^{\star,i1}K\log\frac{2SAHK_i}{\delta}} + 8c_{\mathrm{eu}}S^2AH^4\log^3\frac{2SAHK_i}{\delta}$$

with probability at least $1-\frac{\delta}{2SA}$ for all $K\in[1,K_i]$ simultaneously. Since $m_i \le SA$, union bounding over all j we then have that, with probability at least $1-\delta/2$,

$$\begin{aligned}\sum_{j=1}^{m_i}\Big(\sum_{k=1}^{K_{ij}} V_0^{\star,ij} - \sum_{k=1}^{K_{ij}} V_0^{k,ij}\Big) &\le \sum_{j=1}^{m_i} 2c_{\mathrm{eu}}\sqrt{SAHV_0^{\star,i1}K_{ij}\log\frac{2SAHK_i}{\delta}} + 8c_{\mathrm{eu}}S^3A^2H^4\log^3\frac{2SAHK_i}{\delta}\\ &\le 2c_{\mathrm{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}} + 8c_{\mathrm{eu}}S^3A^2H^4\log^3\frac{2SAHK_i}{\delta}\end{aligned}$$

where the final inequality follows from Jensen's inequality. Using the same calculation as in the proof of Lemma D.4, we can bound

$$\mathbb{E}_{\pi_k}[(\sum_{h=1}^{H} R_h^j(s_h,a_h) - V_0^{k,ij})^2] \le 4V_0^{k,ij}$$

By (D.2), $4V_0^{k,ij} \le 4/2^{i-1}$, so we can apply Lemma D.5 with $\sigma_V^2 = 4/2^{i-1}$, to get that, with probability at least $1-\delta/2$,

$$\left|\sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}}\sum_{h=1}^{H} R_h^j(s_h^{j,k},a_h^{j,k}) - \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} V_0^{k,ij}\right| \le \sqrt{32K_i2^{-i}\log\frac{4}{\delta}} + 2H\log\frac{4}{\delta}.$$

Putting this together and union bounding over these events, we have that with probability at least $1-\delta$,

$$\sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}}\sum_{h=1}^{H} R_h^j(s_h^{j,k},a_h^{j,k}) \ge \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} V_0^{\star,ij} - \sqrt{64K_i2^{-i}\log\frac{4}{\delta}} - 2c_{\mathrm{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}}$$

where we denote

$$C_{\mathcal{R}} := 8c_{\mathrm{eu}}S^3A^2H^4\log^3\frac{2SAHK_i}{\delta} + 2H\log\frac{4}{\delta}.$$

Assume that $V_0^{\star,im_i} > 2^{-i}$. Using that the reward decreases monotonically so $V_0^{\star,im_i} \le V_0^{\star,ij}$ for any $j \le m_i$, we can lower bound the above as

$$\begin{aligned}
&\ge 2^{-i}K_i - \sqrt{64K_i2^{-i}\log\frac{4}{\delta}} - 2c_{\mathrm{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}} \\
&\ge 2^{-i}K_i - 3c_{\mathrm{eu}}\sqrt{S^2A^2H2^{-i}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}}
\end{aligned}$$

where the second inequality follows by (D.2) and since $\sqrt{64K_i2^{-i}\log\frac{4}{\delta}}$ will then be dominated by the regret term, $c_{\mathrm{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}}$. Lemma D.2 gives

$$K_i \ge 2^i\max\left\{4C_{\mathcal{R}}, 144c_{\mathrm{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}\right\}$$

which implies

$$\frac{1}{4}2^{-i}K_i - C_{\mathcal{R}} \ge 0$$

and

$$\begin{aligned}
&\frac{1}{4}2^{-i}K_i - 3c_{\mathrm{eu}}\sqrt{S^2A^2H2^{-i}K_i\log\frac{2SAHK_i}{\delta}} \\
&\quad\ge \frac{2^i\cdot 144c_{\mathrm{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}}{4\cdot 2^i} - 3c_{\mathrm{eu}}\sqrt{S^2A^2H2^{-i}\log\frac{2SAHK_i}{\delta}\cdot 2^i144c_{\mathrm{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}} \\
&\quad= 0.
\end{aligned}$$

Thus, we can lower bound the above as

$$2^{-i}K_i - 3c_{\mathrm{eu}}\sqrt{S^2A^2H2^{-i}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}} \ge \frac{1}{2}2^{-i}K_i.$$

Note that we can collect a total reward of at most $|\mathcal{X}|N_i$. However, by our choice of $N_i = K_i/(4|\mathcal{X}|\cdot 2^i)$, we have that

$$|\mathcal{X}|N_i = \frac{1}{4\cdot 2^i}K_i < \frac{1}{2\cdot 2^i}K_i.$$

This is a contradiction. Thus, we must have that $V_0^{\star,im_i} \le 1/2^i$. The second conclusion follows from this by definition of $V_0^{\star,im_i}$.

For the third conclusion, we can apply Lemma D.3. By construction, we will only add some (s,a,h) to $\mathcal{X}_i$ if we visit N_i times. It follows by Lemma D.3 that, with probability $1-\delta_{\mathrm{samp}}/(SAH)$, if we rerun all policies, we will collect at least

$$N_i - \sqrt{8K_i\max_k w_h^{\pi_k}(s,a)\log\frac{4SAH}{\delta_{\mathrm{samp}}}} - \frac{4}{3}\log\frac{4SAH}{\delta_{\mathrm{samp}}}$$

samples from (s,a,h). Note that $\max_k w_h^{\pi_k}(s,a) \le 2^{-i}$ by our assumption on $\mathcal{X}$. Given our choice of N_i, we can then guarantee that we will collect at least

$$\frac{K_i}{4|\mathcal{X}|2^i} - \sqrt{\frac{8K_iH}{2^i}\log\frac{4SAH}{\delta_{\text{samp}}}} - \frac{4}{3}\log\frac{4SAH}{\delta_{\text{samp}}}$$

samples. Since $K_i \ge 2048S^2A^2\log\frac{4SAH}{\delta_{\text{samp}}}$, and $|\mathcal{X}| \le SA$, we will have that

$$\frac{K_i}{4|\mathcal{X}|2^i} - \sqrt{\frac{8K_iH}{2^i}\log\frac{4SAH}{\delta_{\text{samp}}}} - \frac{4}{3}\log\frac{4SAH}{\delta_{\text{samp}}} \ge \frac{K_i}{8|\mathcal{X}|2^i} = \frac{1}{2}N_i$$

The third conclusion follows by union bounding over every $(s,a,h) \in \mathcal{X}_i$. ■

Remark D.2 (Improving lower order term to $\log 1/\delta \cdot \log\log 1/\delta$) *In Section 4 we noted that relying on* StrongEuler *instead of* Euler *in the exploration phase would allow us to reduce the lower order term from $\log^3 1/\delta$ to $\log 1/\delta \cdot \log\log 1/\delta$. We briefly sketch out that argument here.*

As shown in Simchowitz and Jamieson (2019), the lower order term in StrongEuler *scales as $H^4SA(S\vee H)\log\frac{SAHK}{\delta}\cdot\min\{\log\frac{SAHK}{\delta}, \log\frac{SAH}{\Delta_{\min}}\}$. This already achieves the correct scaling in $\log 1/\delta$ but unfortunately relies on an instance-dependent quantity, $\Delta_{\min}$, which is unknown (indeed, note that since we are running this on the MDP with reward function set to induce exploration, $\Delta_{\min}$ here is different than the minimum gap on the original reward function). As such, since* Learn2Explore *relies on knowing the regret bound of the algorithm it is running, this bound cannot be applied directly.*

Fundamentally, the lower order term arises from summing over the lower order term in the Bernstein-style bonuses which scale as $\mathcal{O}(\frac{\log 1/\delta}{N_h(s,a)})$, where $N_h(s,a)$ is the visitation count of (s,a,h). Intuitively, by summing this bonus over all s,a,h and episodes K, we can obtain a term scaling as $\mathrm{poly}(S,A,H)\log(1/\delta)\log K$. Indeed, we see that the original proof of StrongEuler *in Simchowitz and Jamieson (2019) relies on an integration lemma which does just this (Lemma B.9). However, by modifying the proof of this lemma slightly, we obtain a scaling in the lower-order term of $\log^2 K + \log K\cdot\log 1/\delta$. We then apply the observation from Lemma D.2 that $x \ge C^i(i+3j)^j\log^j(C(i+3j))$ implies $x \ge C^i\log^j x$ to get that we need only*

$$K \gtrsim C\log(1/\delta)\log(C\log(1/\delta)), \quad K \gtrsim C\log^2(C)$$

to ensure that $K \gtrsim C(\log^2 K + \log K\cdot\log 1/\delta)$. It follows that using the lower order term of StrongEuler *in the definition of $C_{\mathcal{R}}$ in Lemma D.1, we can guarantee that $K_i \ge 2^iC_{\mathcal{R}}$ while only requiring that $K_i \gtrsim \log(1/\delta)\log(\log(1/\delta))$. This allows us to reduce the $\log 1/\delta$ dependence in the definition of C_K, which allows us to then reduce the dependence on $\log 1/\delta$ in the lower-order term of Theorem 2.*

D.1. Technical Lemmas

Lemma D.2 *We will have that*

$$K_i(\delta, \delta_{\text{samp}}) \geq 2^i \max\left\{32c_{\text{eu}}S^3A^2H^4\log^3\frac{2SAHK_i(\delta,\delta_{\text{samp}})}{\delta} + 8H\log\frac{4}{\delta}, \right.$$
$$\left. 144c_{\text{eu}}^2S^2A^2H\log\frac{2SAHK_i(\delta)}{\delta}\right\}.$$

Proof Note that for any $i, j > 0$ and $C > 0$, if $x \geq C^i(i+3j)^j\log^j(C(i+3j))$, then $x \geq C^i\log^j x$ since

$$\begin{aligned} C^i\log^j x = C^i\log^j[C^i(i+3j)^j\log^j(C(i+3j))] &\leq C^i\log^j[C^{i+j}(i+3j)^{2j}] \\ &\leq C^i(i+3j)^j\log[C(i+3j)] \\ &= x \end{aligned}$$

and, furthermore, $\frac{d}{dy}y|_{y=C^{i+j}(\max\{i+j,2j\})^{2j}} = 1$, while

$$\frac{d}{dy}C^i\log^j y|_{y=C^{i+j}(\max\{i+j,2j\})^{2j}} = \frac{C^i\log^{j-1}y}{y}|_{y=C^{i+j}(\max\{i+j,2j\})^{2j}} \leq 1$$

and since the derivative of poly log functions decreases monotonically.

It follows that

$$K_i(\delta,\delta_{\text{samp}}) \geq 2^i \cdot 256c_{\text{eu}}S^3A^2H^4\log^3 K_i(\delta,\delta_{\text{samp}})$$

as long as

$$K_i(\delta,\delta_{\text{samp}}) \geq 2^i \cdot 256c_{\text{eu}}S^3A^2H^4(i+9)^3\log^3(512c_{\text{eu}}SAH(i+9))$$

So

$$\begin{aligned} K_i(\delta,\delta_{\text{samp}}) &\geq 2\cdot 2^i\max\{128c_{\text{eu}}S^3A^2H^4\log^3 K_i(\delta,\delta_{\text{samp}}), 128c_{\text{eu}}S^3A^2H^4\log^3\frac{2SAH}{\delta} + 8H\log\frac{4}{\delta}\} \\ &\geq 2^i(32c_{\text{eu}}S^3A^2H^4\log^3\frac{2SAHK_i(\delta,\delta_{\text{samp}})}{\delta} + 8H\log\frac{4}{\delta}) \end{aligned}$$

if

$$K_i(\delta,\delta_{\text{samp}}) \geq \max\{2^i\cdot 256c_{\text{eu}}S^3A^2H^4(i+9)^3\log^3(512c_{\text{eu}}SAH(i+9)), 128c_{\text{eu}}S^3A^2H^4\log^3\frac{2SAH}{\delta} + 8H\log\frac{4}{\delta}$$

Similarly,

$$K_i(\delta,\delta_{\text{samp}}) \geq 2^i\cdot 144c_{\text{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}$$

if

$$K_i(\delta,\delta_{\text{samp}}) \geq \max\{2^i\cdot 288c_{\text{eu}}^2S^2A^2H(i+3)\log(576c_{\text{eu}}SAH(i+3)), 288c_{\text{eu}}^2S^2A^2H\log\frac{2SAH}{\delta}\}.$$

The result then follows recalling the definition of $K_i(\delta,\delta_{\text{samp}})$ given in (D.1). ■

Lemma D.3 *Consider a set of policies $\{\pi_k\}_{k=1}^K$. Assume that running each of these policies once, we collect at least N samples from some (s,a,h). Then, if we rerun each of these policies once, we will collect, with probability $1-\delta$, at least*

$$N - \sqrt{8K \max_k w_h^{\pi_k}(s,a) \log 4/\delta} - 4/3 \log 4/\delta$$

samples from (s,a,h).

Proof Note that when running π_k, the expected number of visits to (s,a,h) is $w_h^{\pi_k}(s,a)$. By Bernstein's inequality, and using that $\mathbb{I}\{(s_h^k, a_h^k) = (s,a)\} \sim \text{Bernoulli}(w_h^{\pi_k}(s,a))$, we then have that, with probability at least $1-\delta$,

$$\left| \sum_{k=1}^K w_h^{\pi_k}(s,a) - \sum_{k=1}^K \mathbb{I}\{(s_h^k, a_h^k) = (s,a)\} \right| \le \sqrt{2K \max_k w_h^{\pi_k}(s,a) \log 2/\delta} + 2/3 \log 2/\delta$$

As our first draw from the policies yielded a value of at least N, we can apply Proposition 14, which gives that, with probability at least $1-2\delta$,

$$\sum_{k=1}^K \mathbb{I}\{(s_h^k, a_h^k) = (s,a)\} \ge N - 2\sqrt{2K \max_k w_h^{\pi_k}(s,a) \log 2/\delta} - 4/3 \log 2/\delta$$

The result follows. ■

Lemma D.4 (Lemma 3.4 of Jin et al. (2020)) *If r_h^k is non-zero for at most one h per episode, the regret of* Euler *(Zanette and Brunskill, 2019) will be bounded, with probability at least $1-\delta$, as*

$$\sum_{k=1}^K V_0^\star - \sum_{k=1}^K V_0^{\pi_k} \le c_{\text{eu}} \sqrt{SAHV_0^\star K \log \frac{SAHK}{\delta}} + c_{\text{eu}} S^2 AH^4 \log^3 \frac{SAHK}{\delta}$$

for some absolute constant c_{eu}.

Proof The proof of this is identical to the proof of Lemma 3.4 in Jin et al. (2020) but we include it for completeness. We therefore repeat their analysis, using an alternative upper bound for equation (156) in Zanette and Brunskill (2019):

$$\begin{aligned}
\frac{1}{KH} \sum_{k=1}^K \mathbb{E}_{\pi_k} \left[(\sum_{h=1}^H r_h^k - V_0^{\pi_k})^2 \right] &\le \frac{2}{KH} \sum_{k=1}^K \mathbb{E}_{\pi_k} \left[(\sum_{h=1}^H r_h^k)^2 + (V_0^{\pi_k})^2 \right] \\
&\overset{(a)}{\le} \frac{2}{KH} \sum_{k=1}^K \mathbb{E}_{\pi_k} \left[\sum_{h=1}^H (r_h^k)^2 + V_0^{\pi_k} \right] \\
&\overset{(b)}{\le} \frac{2}{KH} \sum_{k=1}^K \mathbb{E}_{\pi_k} \left[\sum_{h=1}^H r_h^k + V_0^{\pi_k} \right]
\end{aligned}$$

$$= \frac{4}{KH} \sum_{k=1}^{K} V_0^{\pi_k}$$
$$\leq 4V_0^\star / H$$

where (a) follows since r_h^k is nonzero for at most one h and (b) follows since $r_h^k \leq 1$. Thus, we can replace $\mathcal{G}^2$ in Theorem 1 of Zanette and Brunskill (2019) with $4V_0^\star$. As Zanette and Brunskill (2019) assume a stationary MDP while ours is non-stationary, we must replace S in their bound with SH. This gives the result.

■

Lemma D.5 *Consider some set of policies $\{\pi_k\}_{k=1}^K$ where π_k is $\mathcal{F}_{k-1}$ measurable. Let $\sum_{h=1}^H R_h^k$ denote the (random) reward obtained running π_k on the MDP $\mathcal{M}_k$, and let V_0^k denote the value function of running π_k on $\mathcal{M}_k$. Assume that*

$$\mathbb{E}_{\pi_k}[(\sum_{h=1}^{H} R_h^k - V_0^k)^2 | \mathcal{F}_{k-1}] \leq \sigma_V^2$$

for all k and constant σ_V^2 which is $\mathcal{F}_0$-measurable. Then, with probability at least $1-\delta$,

$$\left| \sum_{k=1}^{K} \sum_{h=1}^{H} R_h^k - \sum_{k=1}^{K} V_0^k \right| \leq \sqrt{8K\sigma_V^2 \log \frac{2}{\delta}} + 2H \log \frac{2}{\delta}.$$

Proof By definition, $V_0^k = \mathbb{E}[\sum_{h=1}^H R_h^k | \mathcal{F}_{k-1}]$ and $|\sum_{h=1}^H R_h^k - V_0^k| \leq H$ almost surely. The result then follows directly from Freedman's Inequality (Freedman, 1975).

■

Proposition 14 *Consider some distribution $\mathbf{P}$ and assume that $\mathbb{P}_{x \sim \mathbf{P}}[x \in [\mu - c, \mu + c]] \geq 1 - \delta$. Then $\mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[x \geq x' - 2c] \geq 1 - 2\delta$.*

Proof

$$\begin{aligned}
\mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[x \geq x' - 2c] &= \mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[x' - \mu + \mu - x \leq 2c] \\
&\geq \mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[|x' - \mu| + |\mu - x| \leq 2c] \\
&\geq \mathbb{P}_{x \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[|\mu - x| \leq c] \mathbb{P}_{x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[|x' - \mu| \leq c] \\
&\geq (1 - \delta)^2 \\
&\geq 1 - 2\delta.
\end{aligned}$$

■

Appendix E. Proof that Low-Regret is Suboptimal for PAC

E.1. Proof of Proposition 1

Instance Class E.1 *Given gap parameters $\Delta_1, \Delta_2 > 0$ and transition probability $p \in (0, 1/2)$, consider an MDP with $H = S = A = 2$ which always starts in state s_0 and has rewards and transitions defined as (where we drop the horizon subscript for simplicity):*

$$P(s_1|s_0, a_1) = 1 - p, \quad P(s_2|s_0, a_1) = p, \quad P(s_1|s_0, a_2) = 0, \quad P(s_2|s_0, a_2) = 1$$
$$R(s_0, a_1) \sim \textit{Bernoulli}(1), \quad R(s_0, a_2) \sim \textit{Bernoulli}(0)$$
$$R(s_i, a_1) \sim \textit{Bernoulli}(0.5 + \Delta_i), \quad R(s_i, a_2) \sim \textit{Bernoulli}(0.5), i \in \{1, 2\}$$

At $h = 2$, we can then think of each state as simply a two-armed bandit with gap Δ_i. We assume that $p < 1/2$, so that $1-p$ can be thought of as a constant. This instance is illustrated in Figure 1.

Proposition 15 (Formal Statement of Proposition 1) *Given any MDP in Instance Class E.1, any learner executing Protocol 5.1 which computes an optimal policy with probability at least $1 - \delta$ must collect at least*

$$K \geq \Omega\left(\frac{\log 1/\delta}{\Delta_1^2} + \frac{\log 1/\delta}{p\Delta_2^2}\right)$$

episodes, as long as $\frac{\log 1/\delta}{\Delta_2^2} \geq c \max\{C_2, C_1^{\frac{1}{1-\alpha}} p^{\frac{-\alpha}{1-\alpha}}\}$, for a universal constant c. However, on this instance,

$$\mathcal{C}^\star(\mathcal{M}) \leq \mathcal{O}\left(\frac{1}{\Delta_1^2} + \frac{1}{\Delta_2^2}\right)$$

and so, with probability $1 - \delta$, Moca *will terminate in at most $K \leq \widetilde{\mathcal{O}}(\mathcal{C}^\star(\mathcal{M}) \cdot \log 1/\delta)$ episodes and return the optimal policy.*

Proof [Proof of Proposition 15] To get the complexity bound of Moca, we apply Theorem 2 and Proposition 9. The stated complexity follows since $W_2(s_1) = 1-p \geq 1/2$ and $W_2(s_2) = 1$, from which the stated complexity follows directly.

Complexity of Low-Regret Algorithms. The expected regret of any algorithm is given by

$$N_1(s_0, a_2) + \Delta_1 N_2(s_1, a_2) + \Delta_2 N_2(s_2, a_2)$$

where we let $N_h(s_i, a_j)$ denote the expected number of times action a_j is taken in state s_i at timestep h. Our assumption on the regret implies that $N_1(s_1, a_2) \leq C_1 K^\alpha + C_2$.

From standard lower bounds on bandits (Theorem 4 of Kaufmann et al. (2016)), and using that for small Δ KL(Bernoulli(0.5)||Bernoulli(0.5 + Δ)) = $\Theta(\Delta^2)$, to solve the bandit in s_1 with probability at least $1 - \delta$, we must have that $N_2(s_1) \geq c\frac{\log 1/\delta}{\Delta_1^2}$, and similarly, to solve the bandit in s_2, we must have that $N_2(s_2) \geq c\frac{\log 1/\delta}{\Delta_2^2}$, for an absolute constant c.

Note that $N_2(s_1) = (1-p)N_1(s_0,a_1)$ and $N_2(s_2) = pN_1(s_0,a_1) + N_1(s_0,a_2)$, and that the total number of episodes run is $N_1(s_0,a_1) + N_1(s_0,a_2)$. This implies that we must have

$$N_1(s_0,a_1) \geq \frac{c\log 1/\delta}{(1-p)\Delta_1^2}, \quad pN_1(s_0,a_1) + N_1(s_0,a_2) \geq \frac{c\log 1/\delta}{\Delta_2^2}.$$

However, since $N_1(s_0,a_2) \leq C_1K^\alpha + C_2$, $N_1(s_0,a_1)$ must at least satisfy

$$pN_1(s_0,a_1) + C_1K^\alpha + C_2 \geq \frac{\log 1/\delta}{\Delta_2^2} \implies N_1(s_0,a_1) \geq \frac{1}{p}\left(\frac{c\log 1/\delta}{\Delta_2^2} - C_1K^\alpha - C_2\right).$$

Thus, we need

$$K = N_1(s_0,a_1) + N_1(s_0,a_2) \geq N_1(s_0,a_1) \geq \max\left\{\frac{c\log 1/\delta}{(1-p)\Delta_1^2}, \frac{1}{p}\left(\frac{c\log 1/\delta}{\Delta_2^2} - C_1K^\alpha - C_2\right)\right\}.$$

Assume that $\frac{c\log 1/\delta}{\Delta_2^2} \geq 2C_2$, then

$$K \geq \frac{1}{p}\left(\frac{c\log 1/\delta}{\Delta_2^2} - C_1K^\alpha - C_2\right)$$

implies

$$K \geq \frac{1}{p}\left(\frac{c\log 1/\delta}{2\Delta_2^2} - C_1K^\alpha\right) \implies 2\max\{pK, C_1K^\alpha\} \geq \frac{c\log 1/\delta}{2\Delta_2^2}.$$

The second expression is equivalent to

$$K \geq \min\left\{\frac{c\log 1/\delta}{4p\Delta_2^2}, (\frac{c\log 1/\delta}{4C_1\Delta_2^2})^{1/\alpha}\right\}$$

and we will have that the minimizer of this is $\frac{c\log 1/\delta}{4p\Delta_2^2}$ as long as $\frac{\log 1/\delta}{\Delta_2^2} \geq c' C_1^{\frac{1}{1-\alpha}} p^{\frac{-\alpha}{1-\alpha}}$. The result follows. ∎

E.2. Proof for Instance Class 5.1

Instance Class E.2 (Formal Definition of Instance Class 5.1) *Given a number of states $S \in \mathbb{N}$, consider MDP with horizon $H = 2$, S states, and $S+1$ actions. We assume we always start in state s_0 and define our transition kernel and reward function as follows:*

$$P(s_i|s_0,a^\star) = \frac{2^{-i}}{1-2^{-S}}, \quad P(s_i|s_0,a_i) = 1, i \in [S]$$

$$R(s_0,a^\star) \sim \textit{Bernoulli}(1), \quad R(s_0,a_i) \sim \textit{Bernoulli}(0), i \in [S]$$
$$\forall i: \quad R(s_i,a^\star) \sim \textit{Bernoulli}(0.9), \quad R(s_i,a_j) \sim \textit{Bernoulli}(0.1), j \in [S].$$

Note that $a^\star$ is the optimal action in every state.

Proposition 16 (Formal Statement of Proposition 6) *For the MDP in Instance Class E.2 with S states, and any*

$$\epsilon \in [2^{-S}, c\min\{C_1^{-1/\alpha}(S\log 1/\delta)^{\frac{1-\alpha}{\alpha}}, C_2^{-1}S\log 1/\delta\}]$$

where c is an absolute constant, to find an ϵ-optimal policy with probability $1-\delta$ any learner executing Protocol 5.1 with a low-regret algorithm satisfying Definition 5.1 must collect at least

$$\Omega\left(\frac{S\log 1/\delta}{\epsilon}\right)$$

episodes. In contrast, on this example $\mathcal{C}^\star(\mathcal{M}) = \mathcal{O}(S^2)$ and $\epsilon^\star = 1/3$, so, for $\epsilon \in [2^{-S}, 1/3]$, with probability $1-\delta$, Moca *will terminate and output $\pi^\star$ in $\widetilde{\mathcal{O}}(C_{\mathrm{LOT}}(1/3))$ episodes.*

Randomized to deterministic policies. Assume we are given some randomized policy π which for every (s,h) choose action a with probability $\pi_h(a|s)$. Then we define the deterministic policy $\widetilde{\pi}$ given this randomized policy as

$$\widetilde{\pi}_h(s) = \arg\max_a \pi_h(a|s).$$

We will use this mapping in our lower bound.

Proof [Proof of Proposition 16] The complexity for Algorithm 5 follows directly from Theorem 2 and Proposition 9 and since in this example we will have $W_2(s) = 1$ for each s and so $\epsilon^\star = 1/3$. Furthermore, $\mathcal{C}^\star(\mathcal{M}) = \mathcal{O}(S^2)$. The stated complexity follows.

Complexity of Low-Regret Algorithms. Let $\Delta_{\mathrm{KL}} := \mathrm{KL}(\mathrm{Bernoulli}(0.1)||\mathrm{Bernoulli}(0.9)) \approx 1.76$ denote the KL divergence between the reward distributions of the optimal and suboptimal actions at any state for $h=2$, and $\Delta := 0.9 - 0.1$ the suboptimality gap.

Assume that a policy π takes action $a^\star$ in s_0. Then, the total suboptimality of the policy is given by

$$\sum_{i=1}^{S} \frac{2^{-i}}{1-2^{-S}}\epsilon_2(s_i,\pi)$$

where $\epsilon_2(s_i,\pi)$ denotes the suboptimality of policy π in $s_i, i \in [S]$. In particular, for any i_ϵ, to guarantee our policy is ϵ-good we need

$$\frac{2^{-i_\epsilon}}{1-2^{-S}}\epsilon_2(s_{i_\epsilon},\pi) \le \epsilon.$$

By the structure of the reward in any state s_{i_ϵ}, the total suboptimality in this state will be

$$\epsilon_2(s_{i_\epsilon},\pi) = (1 - \sum_{j=1}^{S}\pi_2(a_j|s_{i_\epsilon}))\Delta$$

It follows that if $\epsilon_2(s_{i_\epsilon}, \pi) < \Delta/4$, then we will have that $\widetilde{\pi}_2(s_{i_\epsilon}) = a^\star$, where $\widetilde{\pi}$ is the deterministic policy derived from π. Choose $i_\epsilon = \lfloor -\log_2(2\epsilon(1-2^{-S})/\Delta) - 1 \rfloor$. Then it follows that,

$$\frac{2^{-i_\epsilon}}{1-2^{-S}}(1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon}))\Delta \geq 4\epsilon(1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon}))$$

and thus, for the policy to be ϵ-optimal, we must have that $(1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon})) \leq 1/4$. This implies that $\widetilde{\pi}_2(s_{i_\epsilon}) = a^\star$, so we have therefore derived a deterministic policy from our stochastic one that is optimal in $(s_{i_\epsilon}, 2)$. By Theorem 4 of Kaufmann et al. (2016), to identify the optimal action in state s_{i_ϵ} with probability $1-\delta$ we must have that

$$N_2(s_{i_\epsilon}) \geq \frac{(S+1)}{\Delta_{\text{KL}}} \log \frac{1}{2.4\delta}$$

where $N_2(s_{i_\epsilon})$ is the expected number of samples collected in s_{i_ϵ} at $h=2$. As we have deterministically derived $\widetilde{\pi}$ from π, and since $\widetilde{\pi}$ will play the optimal action in s_{i_ϵ} for any ϵ-optimal π, it follows that this lower bound on $N_2(s_{i_\epsilon})$ applies here.

If our low-regret algorithm has regret bounded as $C_1 K^\alpha + C_2$, then we must have that

$$\sum_{i=1}^{S} N_1(s_1, a_i) \leq C_1 K^\alpha + C_2$$

since every time action $a_i \neq a^\star$ is taken we will incur a loss of 1. This implies that

$$N_2(s_{i_\epsilon}) \leq C_1 K^\alpha + C_2 + \frac{2^{-i_\epsilon}}{1-2^{-S}} K$$

since if action $a^\star$ is taken in state s_1, we will only reach state s_{i_ϵ} with probability $\frac{2^{-i_\epsilon}}{1-2^{-S}}$. Combining these, to ensure that the optimal action is learned in s_{i_ϵ}, we will need that

$$\frac{(S+1)}{\Delta_{\text{KL}}} \log \frac{1}{2.4\delta} \leq C_1 K^\alpha + C_2 + \frac{2^{-i_\epsilon}}{1-2^{-S}} K \leq C_1 K^\alpha + C_2 + \frac{4\epsilon}{\Delta} K$$

where the second inequality follows by our choice of i_ϵ. It follows that we need

$$\begin{aligned} K \geq \frac{\Delta}{4\epsilon}\left(\frac{(S+1)}{\Delta_{\text{KL}}} \log \frac{1}{2.4\delta} - C_1 K^\alpha - C_2\right) &\geq \frac{(S+1)\log 1/2.4\delta}{12\epsilon} - C_1 K^\alpha - C_2 \\ &\geq \frac{(S+1)\log 1/2.4\delta}{24\epsilon} - C_1 K^\alpha \end{aligned}$$

where the final inequality holds as long as $\frac{(S+1)\log 1/2.4\delta}{12\epsilon} \geq 2C_2$. This implies

$$2\max\{K, C_1 K^\alpha\} \geq \frac{(S+1)\log 1/2.4\delta}{24\epsilon}$$

which is equivalent to

$$K \geq \min\left\{\frac{(S+1)\log 1/2.4\delta}{48\epsilon}, \left(\frac{(S+1)\log 1/2.4\delta}{48 C_1 \epsilon}\right)^{1/\alpha}\right\}.$$

For

$$\epsilon \le \mathcal{O}\left(C_1^{-1/\alpha}(S\log 1/\delta)^{\frac{1-\alpha}{\alpha}}\right)$$

we will have that the minimizer is the first term, and

$$K \ge \Omega\left(\frac{S\log 1/\delta}{\epsilon}\right).$$

■

Appendix F. Lower Bounds on Best Policy Identification

Lemma F.1 *Consider MDPs $\mathcal{M}$ and $\mathcal{M}'$ with the same state space $\mathcal{S}$, actions space $\mathcal{A}$, horizon H, and initial state distribution P_0. Fix some $(s,h) \in \mathcal{S}\times[H]$, and for any $a\in\mathcal{A}$ let $\nu_h(s,a)$ denote the law of the joint distribution of (s',R) where $s'\sim P_{\mathcal{M}}(\cdot|s,a)$ and $R\sim R_{\mathcal{M}}(s,a)$. Define the law $\nu'_h(s,a)$ analogously with respect to $\mathcal{M}'$. For any almost-sure stopping time τ with respect to $(\mathcal{F}_k)$,*

$$\sum_{s,a,h}\mathbb{E}_{\mathcal{M}}[N_h^\tau(s,a)]\mathrm{KL}(\nu_h(s,a),\nu'_h(s,a)) \ge \sup_{\mathcal{E}\in\mathcal{F}_\tau} d(\mathbb{P}_{\mathcal{M}}(\mathcal{E}),\mathbb{P}_{\mathcal{M}'}(\mathcal{E}))$$

where $d(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$ and $N_h^\tau(s,a)$ denotes the number of visits to (s,a,h) in the τ episodes.

Proof This is the MDP analogue of Lemma 1 of Kaufmann et al. (2016) and its proof follows identically. ■

Definition F.1 *We say an algorithm is δ-correct if, for any MDP $\mathcal{M}\in\mathfrak{M}$, we have that $\mathcal{M}$ terminates at some (possibly random) episode K_δ and outputs $\pi^\star$, with probability at least $1-\delta$.*

F.1. Proof of Proposition 3

MDP Construction. Fix some $\bar{h}\in[H]$, gaps $\{\mathsf{gap}(s,a)\}_{s\in[S],a\in[A-1]}\subseteq(0,1/2)^{SA}$, and arbitrary transition kernels $\{P_h\}_{h=1}^{\bar{h}-1}$. For each s, fix a single a and set $\mathsf{gap}(s,a)=0$. Let $\mathcal{M}$ denote the MDP with transitions $\{P_h\}_{h=1}^{\bar{h}-1}$, and for $h\ge\bar{h}$ define

$$P_h(s|s,a) = 1, \quad \forall a\in\mathcal{A}.$$

Then let the rewards be defined as follows. For all $h>\bar{h}$ and all s, choose any a' and set $R_h(s,a')=1$, and $R_h(s,a)=0$ for all $a\neq a'$. For $h=\bar{h}$, set

$$R_h(s,a)\sim\mathrm{Bernoulli}(3/4-\mathsf{gap}(s,a)).$$

For $h < \bar{h}$, let

$$\pi_h^\star(s) = \arg\max_a \sum_{s'} P_h(s'|s,a) V_{h+1}^\star(s')$$

where $V_{h+1}^\star(s')$ is the optimal value function at step $h+1$ (note that the MDP is now fully specified for $h' > h$ so this is well-defined). Then set $R_h(s, \pi_h^\star(s)) = 1$ and $R_h(s,a) = 0$ for $a \neq \pi_h^\star(s)$ (if $\pi_h^\star(s)$ is not unique, simply choose some $\pi^\star$ out of all $\pi_h^\star(s)$ arbitrarily, set $R_h(s, \pi^\star) = 1$, and all other $R_h(s,a) = 0$).

Note that we could have just as easily encoded the gaps in the transition function and set the rewards to be, for example, deterministic at level $\bar{h}$.

Lemma F.2 *The MDP constructed above has gaps which satisfy*

$$\Delta_{\bar{h}}(s,a) = \mathsf{gap}(s,a), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, a \neq \pi_h^\star(s)$$
$$\Delta_h(s,a) \geq 1, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, h \neq \bar{h}$$

Furthermore, for each s and $h > \bar{h}$, we have $W_h(s) = W_{\bar{h}}(s)$.

Proof We begin with level $\bar{h}$. Since the action take at $(s, \bar{h})$ does not effect the outgoing transition, we have that, for $a \neq \pi_h^\star(s)$,

$$\Delta_{\bar{h}}(s,a) = \max_{a'} Q_{\bar{h}}^\star(s,a') - Q_{\bar{h}}^\star(s,a) = 3/4 - (3/4 - \mathsf{gap}(s,a)) = \mathsf{gap}(s,a).$$

For $h > \bar{h}$, we again have that the outgoing transition is not effected by the action taken, so it follows that the gap depends exclusively on the reward function at this state. Since the reward is set to 1 for a single action and 0 otherwise, it follows that the gaps are all 1.

For $h \leq \bar{h}$, we will have that

$$\begin{aligned}
\Delta_h(s,a) &= \max_{a'} Q_h^\star(s,a') - Q_h^\star(s,a) \\
&= 1 + \max_{a'} \sum_{s'} P_h(s'|s,a') V_{h+1}^\star(s') - \sum_{s'} P_h(s'|s,a) V_{h+1}^\star(s') \\
&\geq 1.
\end{aligned}$$

Finally, that $W_h(s) = W_{\bar{h}}(s)$ for all s and $h > \bar{h}$ follows since for all steps after $\bar{h}$, state s transitions to state s with probability 1. ■

Lemma F.3 *On this example,*

$$\mathcal{C}^\star(\mathcal{M}) \leq \inf_\pi \max_{s,a} \frac{1}{w_{\bar{h}}^\pi(s,a)\Delta_h(s,a)^2} + \max_{s,h} \frac{SAH}{W_h(s)}.$$

Proof By definition,

$$\mathcal{C}^\star(\mathcal{M}) = \sum_{h=1}^H \inf_\pi \max_{s,a} \frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}.$$

By Lemma F.2, we can bound

$$\sum_{h\neq\bar{h}} \inf_{\pi} \max_{s,a} \frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2} \leq \sum_{h\neq\bar{h}} \inf_{\pi} \max_{s,a} \frac{1}{w_h^\pi(s,a)}.$$

Consider the policy π' which is the mixture over the policies π^{sh} where $w_h^{\pi^{sh}}(s) = W_h(s)$. Then,

$$\sum_{h\neq\bar{h}} \inf_{\pi} \max_{s,a} \frac{1}{w_h^\pi(s,a)} \leq \sum_{h\neq\bar{h}} \max_{s,a} \frac{1}{w_h^{\pi'}(s,a)} \leq \sum_{h\neq\bar{h}} \max_{s} \frac{SA}{W_h(s)} \leq \max_s \frac{SAH}{W_h(s)}.$$

■

Lemma F.4 *On the MDP constructed above, any δ-correct algorithm will have*

$$\begin{aligned}\mathbb{E}_{\mathcal{M}}[K_\delta] &\geq \inf_{\pi} \max_{s,a} \frac{1}{6w_{\bar{h}}^\pi(s,a)\Delta_{\bar{h}}(s,a)^2} \cdot \log\frac{1}{2.4\delta} \\ &\gtrsim \mathcal{C}^\star(\mathcal{M}) \cdot \log\frac{1}{2.4\delta} - \max_{s,h} \frac{SAH}{W_h(s)}.\end{aligned}$$

Proof We will apply Lemma F.1 on our MDP, $\mathcal{M}$, and MDP $\mathcal{M}'$ which is identical to $\mathcal{M}$ except that, for some (s,a), $a \neq \pi^\star_{\bar{h}}(s)$, we set $R_{\bar{h}}(s,a) \sim \text{Bernoulli}(3/4+\alpha)$ for small α. Note that in this case we have that the optimal policy on $\mathcal{M}$ and $\mathcal{M}'$ differ at $(s,\bar{h})$. Since $\mathcal{M}$ and $\mathcal{M}'$ are identical at all points but this one, we have

$$\begin{aligned}&\sum_{s,a,h} \mathbb{E}_{\mathcal{M}}[N_h^\tau(s,a)]\mathrm{KL}(\nu_h(s,a),\nu'_h(s,a)) \\ &\quad= \mathbb{E}_{\mathcal{M}}[N_{\bar{h}}^\tau(s,a)]\mathrm{KL}\big(\text{Bernoulli}(3/4-\mathsf{gap}(s,a)),\text{Bernoulli}(3/4+\alpha)\big).\end{aligned}$$

Let $\pi^\star(\mathcal{M})$ denote the optimal policy on $\mathcal{M}$, and $\widehat{\pi}$ denote the policy returned by our algorithm. Let $\mathcal{E} = \{\widehat{\pi} = \pi^\star(\mathcal{M})\}$. Since we assume our algorithm is δ-correct, and since the optimal policies on $\mathcal{M}$ and $\mathcal{M}'$ differ, we have $\mathbb{P}_{\mathcal{M}}(\mathcal{E}) \geq 1-\delta$ and $\mathbb{P}_{\mathcal{M}'}(\mathcal{E}) \leq \delta$. By Kaufmann et al. (2016), we can then lower bound

$$d(\mathbb{P}_{\mathcal{M}}(\mathcal{E}), \mathbb{P}_{\mathcal{M}'}(\mathcal{E})) \geq \log\frac{1}{2.4\delta}.$$

Thus, by Lemma F.1, we have shown that, for any (s,a), $a \neq \pi^\star_{\bar{h}}(s)$,

$$\mathbb{E}_{\mathcal{M}}[N_{\bar{h}}^\tau(s,a)] \geq \frac{1}{\mathrm{KL}\big(\text{Bernoulli}(3/4-\mathsf{gap}(s,a)),\text{Bernoulli}(3/4+\alpha)\big)} \cdot \log\frac{1}{2.4\delta}.$$

For small α, we can bound (see e.g. Lemma 2.7 of Tsybakov (2009))

$$\mathrm{KL}\big(\text{Bernoulli}(3/4-\mathsf{gap}(s,a)),\text{Bernoulli}(3/4+\alpha)\big) \leq 6(\mathsf{gap}(s,a)-\alpha)^2.$$

Taking $\alpha \to 0$, we have

$$\mathbb{E}_{\mathcal{M}}[N_h^\tau(s,a)] \geq \frac{1}{6\mathsf{gap}(s,a)^2} \cdot \log \frac{1}{2.4\delta}.$$

We can write $\mathbb{E}_{\mathcal{M}}[N_h^\tau(s,a)] = \mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau w_h^{\pi_k}(s,a)]$ where π_k denotes the policy our algorithm played at episode k. Note that all state-visitation distributions lie in a convex set in $[0,1]^{SA}$ and that for any valid state-visitation distribution, there exists some policy that realizes it, by Proposition 12. By Caratheodory's Theorem, it follows that there exists some set of policies Π with $|\Pi| \leq SA+1$ such that, for any π and all s,a, $w_h^\pi(s,a) = \sum_{\pi' \in \Pi} \lambda_{\pi'} w_h^{\pi'}(s,a)$, for some $\lambda \in \triangle_\Pi$. Letting λ^k denote this distribution satisfying the above inequality for π_k, it follows that

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau w_h^{\pi_k}(s,a)] &= \mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau \sum_{\pi \in \Pi} \lambda_\pi^k w_h^\pi(s,a)] \\
&= \sum_{\pi \in \Pi} \mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau \lambda_\pi^k] w_h^\pi(s,a) \\
&= \mathbb{E}_{\mathcal{M}}[\tau] \sum_{\pi \in \Pi} \frac{\mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau \lambda_\pi^k]}{\mathbb{E}_{\mathcal{M}}[\tau]} w_h^\pi(s,a).
\end{aligned}$$

Note that $\sum_{\pi \in \Pi} \mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau \lambda_\pi^k] = \mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau \sum_{\pi \in \Pi} \lambda_\pi^k] = \mathbb{E}_{\mathcal{M}}[\tau]$ so it follows that $(\frac{\mathbb{E}_{\mathcal{M}}[\sum_{k=1}^\tau \lambda_\pi^k]}{\mathbb{E}_{\mathcal{M}}[\tau]})_{\pi \in \Pi} \in \triangle_\Pi$. Thus, a δ-correct algorithm must satisfy, for all s,a and some $\lambda \in \triangle_\Pi$,

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq \frac{1}{6\mathsf{gap}(s,a)^2 \cdot \sum_{\pi \in \Pi} \lambda_\pi w_h^\pi(s,a)} \cdot \log \frac{1}{2.4\delta}.$$

Since the set of state visitation distributions is convex, and since for any state-visitation distribution we can find some policy realizing that distribution, for any $\lambda \in \triangle_\Pi$, it follows that there exists some π' such that, for all s,a, $\sum_{\pi \in \Pi} \lambda_\pi w_h^\pi(s,a) = w_h^{\pi'}(s,a)$. So, we need, for all s,a

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq \frac{1}{6\mathsf{gap}(s,a)^2 \cdot w_h^{\pi'}(s,a)} \cdot \log \frac{1}{2.4\delta}.$$

It follows that *every* δ-correct algorithm must satisfy

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq \inf_\pi \max_{s,a} \frac{1}{6\mathsf{gap}(s,a)^2 \cdot w_h^\pi(s,a)} \cdot \log \frac{1}{2.4\delta},$$

from which the first inequality follows. The second follows from Lemma F.3. ■