Online Regret Bounds for Satisficing in MDPs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider general reinforcement learning under the average reward criterion in Markov decision processes (MDPs) when the learner's goal is not to learn an optimal policy but accepts any policy whose average reward is above a certain given satisfaction level σ. We show that with this more modest objective it is possible to have algorithms that only have constant regret with respect to the level σ, provided that there is a policy above this level. This result generalizes findings of Bubeck et al. [2013] from the bandit setting to MDPs.

Further, we present a more general algorithm that achieves the best of both worlds: If the optimal policy has average reward above σ this algorithm has bounded regret with respect to σ. On the other hand, if all policies are below σ then we can show logarithmic bounds on the expected regret with respect to the optimal policy.

1 Introduction

Learning optimal policies in real-world reinforcement learning (RL) problems is usually intricate. In this paper we want to investigate the question whether there is an advantage in pursuing a more modest goal: Instead of aiming at optimal performance the learner is content with average reward above a specified satisfaction level σ. As performance criterion we consider online regret with respect to this level σ. That is, as long as the agent follows a policy π whose average reward ρ_π is above σ there is no regret, otherwise the per-step regret is $\sigma - \rho_\pi$.

In the following, we show in Section 3 that when the optimal average reward ρ^* of the underlying MDP (which we assume to be communicating) is above σ, learning with only constant regret (i.e., independent of the number of steps) is possible. This generalizes a result of Bubeck et al. [2013] from the bandit to the general MDP setting. We proceed to the general case when ρ^* may be below σ. Here we provide an algorithm that on the one hand also only suffers constant regret when $\rho^* > \sigma$. On the other hand, when $\rho^* < \sigma$ the same algorithm can be shown to have classic regret (i.e., with respect to ρ^*) that is bounded logarithmically in the number of steps just as for state-of-the-art RL algorithms in this setting.

While satisficing objectives have been considered before, most respective investigations have been made in the much simpler bandit setting [Abernethy et al., 2016, Reverdy et al., 2017, Michel et al., 2022]. For the general MDP setting, beside some related work on multi-objective RL [cf. Roijers et al., 2013, for an overview] we are only aware of some experimental work on a satisficing variant of Q-learning [Goodrich and Quigley, 2004]. In particular, as far as we know we present the first theoretical results on RL in MDPs under the satisficing criterion.

1.1 Setting and Notation

Let $M = (\mathcal{S}, \mathcal{A}, r, p)$ be an MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, mean rewards $\mu(s,a)$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$, and transition probabilities $p(s'|s,a)$ for $(s',s,a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$. The

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

random rewards are assumed to be bounded, i.e., contained in $[0,1]$. We set $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$. Beside S and A the diameter $D(M)$ as introduced in [Jaksch et al., 2010] is an important parameter of the MDP.

Definition 1. *Consider the stochastic process defined by a stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ operating on an MDP M with initial state s. Let $T(s'\big|M,\pi,s)$ be the random variable for the first time step in which state s' is reached in this process. Then the diameter of M is defined as*

$$D(M) := \max_{s,s'\in\mathcal{S}} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}(T(s'\big|M,\pi,s)).$$

In the following we assume that the diameter $D(M)$ is finite. This guarantees that a learner operating in M always is able to recover from a mistake, as any state is reachable from another state. Indeed, let us define the average reward of a stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ starting in initial state s_1 to be $\rho_\pi(M,s_1) := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r_t^{\pi,s_1}$, where r_t^{π,s_1} is the random reward obtained by the policy π at step t when starting in s_1. Then the optimal average reward ρ^* in M is independent of the initial state when $D(M)$ is finite. Further, considering nonstationary policies does not increase the optimal average reward [Puterman, 2005]. In the following, π^* denotes a respective optimal policy in M such that $\rho_{\pi^*}(M,s_1) = \rho^*$ for any initial state s_1. Further, for any policy π whose average reward is independent of the initial state s_1, we write ρ_π for $\rho_\pi(M,s_1)$.

Beside the standard diameter we also consider a similar transition parameter.

Definition 2. *For any stationary policy π we set*

$$D_\pi(M) := \max_{\substack{s\neq s'\in\mathcal{S}:\\ \mathbb{E}(T(s'|M,\pi,s)<\infty}} \mathbb{E}(T(s'\big|M,\pi,s))$$

to be the maximal finite distance between any two states under π. Then the worst-case diameter is defined as

$$D_W(M) := \max_\pi D_\pi(M).$$

In the following, we often drop the notation for the MDP and write e.g. D instead of $D(M)$ whenever M is understood from the context.

1.1.1 Regret and σ-regret

We are interested in policies whose average reward is above a given satisfaction level σ. Accordingly, for a policy π and an initial state s_1 we define the gap to σ as $\Delta^\sigma_{\pi,s_1} := \max\{0, \sigma - \rho_\pi(M,s_1)\}$. If the average reward for ρ is independent of the initial state, we drop the latter in the notation and simply write Δ^σ_π. Intuitively, Δ^σ_π is the average per-step regret an agent suffers when playing policy π. Accordingly, we define the *σ-regret* of a policy π starting in state s_1 after T steps as

$$R^{\sigma,T}_{\pi,s_1} := T\Delta^\sigma_{\pi,s_1}.$$

Note that the respective expected accumulated reward may deviate from $T\rho_\pi(M,s_1)$ at most by a term of oder $D_\pi(M)$, cf. [Jaksch et al., 2010].

More generally we are interested in the σ-regret of more general episodic algorithms $\mathcal{A}$ which stick to the same stationary policy for a certain number of steps before changing to another stationary policy. That is, if algorithm $\mathcal{A}$ plays policy π_k in episode k starting in state s_k at step T_k (with $k = 1,2,\ldots$), the respective regret after n episodes is defined as

$$R^{\sigma,T_{n+1}}_{\mathcal{A},s_1} := \sum_{k=1}^{n} (T_{k+1} - T_k)\, \Delta^\sigma_{\pi,s_k}.$$

Beside the σ-regret we will also consider the standard regret after any T steps as defined in [Jaksch et al., 2010] as

$$R^T_{\mathcal{A},s_1} := T\rho^* - \sum_{t=1}^{T} r_t^{\mathcal{A},s_1},$$

where similar as before $r_t^{\mathcal{A},s_1}$ denotes the random reward obtained by algorithm $\mathcal{A}$ at step t when starting in state s_1.

Beside the gaps $\Delta^{\sigma}_{\pi,s_1}$ we also consider the quantities $\Delta^{\sigma,-} := \min_{\pi:\Delta^{\sigma}_{\pi}>0} \Delta^{\sigma}_{\pi}$ and $\Delta^{\sigma,+} := \max_{\pi:\Delta^{\sigma}_{\pi}>0} \Delta^{\sigma}_{\pi}$ where both $\max$ and $\min$ range over policies with average reward independent of the intial state.[1] Further, the gap $\Delta_g := \rho^* - \max_{\pi:\rho_\pi<\rho^*} \rho_\pi$ between the optimal average reward and the average reward of the best suboptimal policy.

2 Preliminaries

Our proposed approach is based on the two RL algorithms UCRL2 and GOSPRL that we will employ in a blackbox manner. Accordingly, in the following we will only briefly recall some basic properties that we will need for our purposes.

2.1 UCRL2

UCRL2 [Jaksch et al., 2010] is a well-konwn RL algorithm which based on the idea of employing *optimism in the face of uncertainty*. UCRL2 proceeds in episodes in which a fixed stationary policy is executed. Based on the episode termination criterion used by UCRL2, the following bound on the number of episodes holds.

Proposition 1. [Jaksch et al., 2010] *The number of episodes of* UCRL2 *up to step $T \geqslant AS$ is upper bounded by*

$$AS \log_2 \left(\tfrac{8T}{AS}\right).$$

More importantly, for UCRL2 one can give bounds on the classic online regret (i.e., with respect to the optimal average reward ρ^*) in communicating MDPs. We will use the following bound on the *expected* regret.

Theorem 1. [Jaksch et al., 2010] *For any initial state $s_1 \in \mathcal{S}$ and any $T > 1$, the expected regret of* UCRL2 *run with confidence parameter $\delta = \frac{1}{3T}$ is bounded by*

$$\mathbb{E}[R^T_{\mathrm{UCRL2},s_1}] \leqslant \frac{34^2 AS^2 D^2 \ln(T)}{\Delta_g} + \sum_{a,s} \big[1 + \log_2(\max_{\pi:\pi(s)=a} T_\pi)\big] \max_{\pi:\pi(s)=a} T_\pi,$$

where T_π is the smallest natural number such that for all $T \geqslant T_\pi$ the expected average reward after T-steps is $\frac{\Delta_g}{2}$-close to the average reward of π.

2.2 GOSPRL

Unlike UCRL2, GOSPRL [Tarbouriech et al., 2021] is an exploration algorithm whose goal is to collect a specified number of samples in an unknown communicating MDP. That is, for a given function $\bar{b} : \mathcal{S} \times \mathcal{A} \to \mathbb{N}$ and a confidence parameter δ_g, GOSPRL($\bar{b}, \delta_g$) for any action $a \in \mathcal{A}$ and any state $s \in \mathcal{S}$ collects at least $\bar{b}(s,a)$ samples with overall success probability at least $1 - \delta_g$. As shown by Tarbouriech et al. [2021], this is accomplished after $\tilde{O}(\bar{B}D + AS^2D^{\frac{3}{2}})$ steps, where $\bar{B} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{b}(s,a)$ and the $\tilde{O}$ notation hides logarithmic dependencies on $S, A, D, \bar{B}$, and $\frac{1}{\delta_g}$. Furthermore, based on GOSPRL Tarbouriech et al. [2021] further provide an algorithm that computes an approximation of the diameter of the underlying MDP. This algorithm takes a confidence parameter δ_g and a precision parameter ε_g as input, and after $\tilde{O}(\frac{AS^2D^3}{\varepsilon_g^2})$ steps with probability at least $1 - \delta_g$ outputs an estimate $\bar{D}$ of the diameter for which $D \leqslant \bar{D} \leqslant (1 + 2\varepsilon_g(1+\varepsilon_g))(1+\varepsilon_g)D$.

3 Algorithm SAT-RL

In this section, we introduce our algorithm SAT-RL (shown as Algorithm 1) which is designed to find and keep playing a satificing policy when given a satisfaction level σ, provided that $\rho^* > \sigma$.

[1] Note that for any policy π and any initial state s_1 there is a policy π' such that the average reward of π' is independent of the initial state and $\rho_{\pi'} = \rho_\pi(M, s_1)$: Since M is assumed to be communicating, for states s not in the same irreducible class $I_\pi(s_1)$ as s_1, one can choose actions for $\pi'(s)$ that eventually lead to $I_\pi(s_1)$, so that there is only a single irreducible class under π'.

Algorithm 1 SAT-RL for satisficing in RL

1: **Input:** state space $\mathcal{S}$, action space $\mathcal{A}$, satisfaction level σ
2: **Initialization:**
3: Set confidence level $\delta_g := \frac{1}{2}$, accuracy level $\varepsilon_g := \frac{1}{2}$, and initial sampling number $b := S + 1$.
4: Define function $\bar{b} : \mathcal{S} \times \mathcal{A} \to \mathbb{N}$ to be $\bar{b}(s, a) = b$ for any (s, a).
5: **while** an action $a \in \mathcal{A}$ at some state $s \in \mathcal{S}$ has not been run $\bar{b}(s, a)$ times **do**
6: Run GOSPRL($\bar{b}, \delta_g$).
7: For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, define $\bar{b}(s, a) := b - N(s, a)$.
8: **end while**
9: **while** the diameter of the estimated MDP M_k is infinite **do**
10: Run GOSPRL-based(δ_g, ε_g) procedure to estimate the diameter of M.
11: **end while**
12: **for** episodes $k = 1, 2, \ldots$ **do**
13: Compute an optimal policy π_k on M_k that induces a unique irreducible class I_k.
14: **if** $\rho_{\pi_k}(M_k, s_k) \geqslant \sigma$ **then**
15: Play π_k until all states in I_k have been visited at least once.
16: **else**
17: Set $b := b + S$.
18: **while** for any state $s \in \mathcal{S}$ and $a \in \mathcal{A}$, $N(s, a) < b$ **do**
19: For any (s, a), set $\bar{b}(s, a) := b - N(s, a)$.
20: Run GOSPRL($\bar{b}, \delta_g$).
21: **end while**
22: **end if**
23: **end for**

SAT-RL starts by collecting some initial samples for each state-action pair using GOSPRL. That is, first at least $S + 1$ samples for each state-action pair are collected (lines 5–8). Here we use $N(s, a)$ to denote the current number of samples of a state-action pair s, a. If the diameter of the estimated MDP is infinite, in addition the procedure to estimate the MDP's diameter is run (lines 9–11). This is only done to guarantee that the diameter of the empirical MDP is finite, that is, it is communicating.

After this initialization phase, the algorithm proceeds in episodes k in which at first the optimal policy π_k in the estimated MDP M_k is computed (line 13). As the diameter of M_k is finite at this step, this optimal policy can be computed by value iteration (VI) and can be assumed to have a unique irreducible class I_k. Furthermore, by running another instance of VI we can further assume that for all states not in I_k the computed policy π_k will move to I_k as fast as possible, i.e. in expected time at most D.

If the average reward of π_k on M_k is above σ, SAT-RL performs the policy π_k in an *exploitation episode*, which ends after all states reachable under π_k have been visited (line 15). Otherwise, if the average reward of π_k on M_k is $< \sigma$, SAT-RL performs an *exploration episode*, in which GOSPRL is used to collect another S samples from each state-action pair.

Before we proceed to analyze SAT-RL, we note that in Appendix C we present an alternative algorithm called SAT-RL2 that instead of using the GOSPRL procedure to estimate diameter uses a result about the diameter of MDP approximations (Lemma 5) that might be of interest in itself.

4 Regret Bound for SAT-RL

In this section we will give a proof sketch for the following bound on the σ-regret of SAT-RL. Details can be found in Appendix B.

Theorem 2. *If $\rho_{\pi^*} > \sigma$, then the expected σ-regret of* SAT-RL *after any number of steps is bounded by*

$$\tilde{O}\left(\frac{AS^2D^{\frac{7}{2}}}{(\Delta_*^\sigma)^2} + \frac{(\Delta_*^\sigma)^{2S-2}A^2}{D^{S-\frac{5}{2}}S^{S-3}} + \frac{\Delta^{\sigma,+}AS^2D_W^3}{(\Delta^{\sigma,-})^2}\right),$$

where the $\tilde{O}$-notation hides logarithmic dependencies on $A, S, D_W, \Delta^{\sigma,-}$, and Δ_*^σ.

4.1 Proof of Theorem 2

Proof. Let us first introduce some notation. For any episode k and any state-action pair (s, a), we write $r_k(s, a)$, $p_k(\cdot|s, a)$, and $N_k(s, a)$ for the empirical average reward, the empirical transition probability distribution, and the number of times action a has been chosen in state s before the start of episode k. Also, M_k denotes the estimated MDP and s_k is the initial state at start of episode k. We set $\rho_k(\pi_k, s_k) = \rho_{\pi_k}(M_k, s_k)$.

Let L_k be a random variable for the number of steps *before* episode k ($k > 0$). Then the regret after n episodes can be bounded by

$$\mathbb{E}[L_0] + \sum_{k=1}^{n} \mathbb{E}\big[\mathbb{1}\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma\} L_k \Delta^{\sigma}_{\pi_k}\big] + \sum_{k=1}^{n} \mathbb{E}\big[\mathbb{1}\{\rho_k(\pi_k, s_k) < \sigma\} L_k\big]. \quad (1)$$

We call the three terms in this sum *initialization regret* R_{Init}, *exploitation regret* R_{Exploit}, and *exploration regret* R_{Explore}, respectively. In the following, we derive bounds for each term separately.

Note that in the initialization phase as well as in exploration episodes we perform GOSPRL, which in general does not execute a stationary policy. Accordingly, the regret of these episodes is actually not well defined. so that we simply consider a regret of 1 per step (which can be considered as 'episodes' of length 1 according to our regret definition) and simply bound the (expected) number of steps L_k of the respective episodes of the algorithm. This is already reflected in (1).

In the following we define freq_k to be the number of visits in the state-action pair (s, a) that has the fewest visits before episode k among the state action-pairs that will be regularly visited during episode k. That is, for exploration episodes k we set

$$\text{freq}_k := \min_{s,a} N_k(s, a),$$

while for exploitation episodes k in which policy π_k is played we set

$$\text{freq}_k := \min_{s \in I_k} N_k(s, \pi_k(s)).$$

Upper Bound on Exploitation Regret

By Theorem 4.8 of Dabbs [2009] the covering time (i.e., the first time at which all states have been visited) of an irreducible Markov chain with S states and diameter at most D is less than $D(1+\ln(S))$. Consequently, for a fixed policy π_k, we need in average at most D_{π_k} steps to reach the irreducible part I_{π_k} and at most $D_{\pi_k}(1 + \ln(S))$ steps to cover it following policy π_k. Further, $\Delta^{\sigma}_{\pi_k}$ can be upper bounded by $\Delta^{\sigma,+}$, so that

$$\sum_{\pi} \mathbb{E}\big[L_k \Delta^{\sigma}_{\pi_k} \big| \pi_k = \pi \wedge \rho_k(\pi, s_k) \geqslant \sigma \wedge \rho(\pi, s_k) < \sigma\big] \;\leqslant\; D_W(1 + \ln(S))\Delta^{\sigma,+}. \quad (2)$$

It remains to analyze the term

$$\begin{aligned}
&\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma\big\} \\
&\quad = \sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \text{freq}_k \leqslant \theta\big\} \quad (3) \\
&\qquad + \sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant \theta + 1\big\}, \quad (4)
\end{aligned}$$

where we choose

$$\theta = \left\lceil \frac{4S(D_W + 1)^2}{(\Delta^{\sigma,-})^2} \ln \frac{8S(D_W + 1)^2}{(\Delta^{\sigma,-})^2} \right\rceil. \quad (5)$$

Lemma 4 in Appendix A.3 shows that for any number f with $S + 1 \leqslant f \leqslant \theta$, there are at most AS episodes k for which $\text{freq}_k = f$, so that (3) can be bounded as

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \text{freq}_k \leqslant \theta\big\} \;\leqslant AS(\theta - S) \;\leqslant\; \tilde{O}\Big(\frac{AS^2 D_W^2}{(\Delta^{\sigma,-})^2}\Big). \quad (6)$$

Further, by Lemma 8 in Appendix B.1, we can bound (4) as

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta + 1\big\} \;\leqslant\; \frac{2A}{\theta^{S-1}\ln(2\theta)}. \tag{7}$$

Taking together eqs. (2)–(7) we obtain

$$R_{\mathrm{Exploit}} \leqslant \tilde{O}\bigg(\frac{\Delta^{\sigma,+} A S^2 D_W^3}{(\Delta^{\sigma,-})^2}\bigg).$$

Upper Bound on Initialization Regret

The sample complexity of GOSPRL($\bar{b}, \delta_g$) is $\tilde{O}(\bar{B}D + AS^2D^{\frac{3}{2}})$, where $\bar{B} = \sum_{s,a} \bar{b}(s,a)$ and the $\tilde{O}$-notation hides logarithmic dependencies on S, A, $\frac{1}{\delta_g}$. In our case $\bar{B} = AS(S+1)$, $\delta_g = \frac{1}{2}$, and as GOSPRL is run until each state has been visited at least $S+1$ times, the expected regret of the first part of the initialization phase (lines 5–8 of the algorithm) is at most

$$\sum_{i=1}^{\infty} (\tfrac{1}{2})^{i-1} \tilde{O}(\bar{B}D + AS^2D^{\frac{3}{2}}) = \tilde{O}(AS^2D + AS^2D^{\frac{3}{2}}) = \tilde{O}(AS^2D^{\frac{3}{2}}).$$

For estimating the diameter (lines 9–11), the sample complexity of the GOSPRL-based(δ_g, ε_g) procedure is $\tilde{O}\big(\frac{AS^2D^3}{\varepsilon_g^2}\big)$. Similar as before, since $\delta_g = \varepsilon_g = \frac{1}{2}$, the respective regret is at most

$$\sum_{i=1}^{\infty} (\tfrac{1}{2})^{i-1}\, \tilde{O}\bigg(\frac{AS^2D^3}{\varepsilon_g^2}\bigg) = \tilde{O}(AS^2D^3),$$

so that we can bound the total regret in the initialization phase as

$$R_{\mathrm{Init}} \;\leqslant\; \tilde{O}(AS^2D^{\frac{3}{2}}) + \tilde{O}(AS^2D^3) \;=\; \tilde{O}(AS^2D^3).$$

Upper Bound on Exploration Regret

Similar to the analysis of R_{Exploit} we first bound

$$\mathbb{E}\big[L_k \big| \rho_k(\pi_k, s_k) \leqslant \sigma\big] \;=\; \tilde{O}(AS^2D^{\frac{3}{2}}) \tag{8}$$

according to the sample complexity of GOSPRL (cf. also the analysis of R_{Init}) and it remains to bound

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma\big\} \;=\; \sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma) \wedge \mathrm{freq}_k \leqslant \theta_*\big\} \tag{9}$$

$$+ \sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta_* + 1\big\}, \tag{10}$$

where we set

$$\theta_* = \bigg\lceil \frac{4S(D+1)^2}{(\Delta_*^{\sigma})^2} \ln \frac{8S(D+1)^2}{(\Delta_*^{\sigma})^2} \bigg\rceil. \tag{11}$$

By definition of the algorithm, in any exploration episode each state-action pair is visited at least S times so that also freq will increase by S. Accordingly,

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \leqslant \theta_*\big\} \;\leqslant \bigg\lceil \frac{\theta_*}{S} \bigg\rceil. \tag{12}$$

Concerning (10), Lemma 11 in Appendix B.2 shows that

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta_* + 1\big\} \leqslant \frac{2A}{\theta_*^{S-1}\ln(2\theta_*)}. \tag{13}$$

Consequently, summarizing (8)–(13) we obtain

$$R_{Explore} \leqslant \tilde{O}\bigg(\frac{AS^2D^{\frac{7}{2}}}{(\Delta_*^{\sigma})^2} + \frac{(\Delta_*^{\sigma})^{2S-2}A^2}{D^{S-\frac{5}{2}}S^{S-3}}\bigg)$$

and summing up the three regret terms yields the claimed regret bound of the theorem. □

5 The General Case

We have seen that when the satisfaction level σ can be attained by the optimal policy, we can have constant σ-regret. What can we hope for when it is not known whether $\rho^* > \sigma$? Obviously, when $\rho^* < \sigma$ we it is not possible to have constant σ-regret anymore, as the latter will always be linear in T. However, a reasonable aim in this case would be to re-establish standard online regret bounds with respect to ρ^* just as the one given in Theorem 1 for UCRL2. In the following, we indeed present the algorithm SAT-UCRL, which precisely achieves that: When σ can be realized, we have constant regret just as for SAT-RL. If however $\rho^* \geqslant \sigma$, we show a bound with the same dependency on T as the one given in Theorem 1.

5.1 Algorithm SAT-UCRL

Our proposed algorithm SAT-UCRL is shown as Algorithm 2. It resembles SAT-RL, only that now in exploration episodes we do not use GOSPRL but UCRL2. As already mentioned, UCRL2 itself uses episodes in which it employs a fixed policy. In order to differentiate between episodes of SAT-UCRL and these internal episodes of UCRL2, in the following we will refer to the latter as *sub-episodes* of UCRL2. Note that in an exploration episode of SAT-UCRL several sub-episodes of UCRL2 are run.

Algorithm 2 SAT-UCRL for the general RL setting

Input: state space $\mathcal{S}$, action space $\mathcal{A}$, satisfaction level σ, horizon T
Initialization:
Set confidence level $\delta_g := \frac{1}{2}$ and accuracy level $\varepsilon_g := \frac{1}{2}$.
Set initial sampling number $b := S + 1$ and $b_u := \frac{AS}{8}$.
Define function $\bar{b} : \mathcal{S} \times \mathcal{A} \to \mathbb{N}$ to be $\bar{b}(s, a) = b$ for any (s, a).
while an action $a \in \mathcal{A}$ at some state $s \in \mathcal{S}$ has not been run $\bar{b}(s, a)$ times **do**
    Run GOSPRL($\bar{b}, \delta_g$).
    For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, set $\bar{b}(s, a) := b - N(s, a)$.
end while
while the diameter of the estimated MDP M_k is infinite **do**
    Run GOSPRL-based(δ_g, ε_g) procedure to estimate the diameter of M.
end while
for episodes $k = 1, 2, \ldots$ **do**
    Compute an optimal policy π_k on M_k that induces a unique irreducible class I_k.
    if $\rho_{\pi_k}(M_k, s_k) \geqslant \sigma$ **then**
        Play π_k until all states in I_k have been visited at least once.
    else start a new **episode** of UCRL2 with confidence parameter $\delta = \frac{1}{3T}$.
        Set $b_u := 8b_u$.
        while the number of running of new steps is less than b_u **do**
            Run a **sub-episode** of UCRL2.
            Terminate current episode.
        end while
    end if
end for

In order to facilitate the analysis, in the following we assume that the exploration episodes employing UCRL2 do not use any samples of the exploitation episodes, which in practice of course would speed up convergence and hence improve the algorithm.

5.2 Regret Bound for SAT-UCRL

Here we present the two promised bounds on the (σ-)regret. We start with the bound on the standard online regret when σ cannot be attained by any policy.

Theorem 3. *Let $\sigma \geqslant \rho^*(M)$. For any initial state s_1, and any $T > 1$, the expected regret of* SAT-UCRL *(with respect to ρ^*) is bounded by*

$$\frac{34^2 AS^2 D^2 \ln(T)}{\Delta_g} + \tilde{O}\left(\frac{AS^2 D_W^3}{(\Delta^{\sigma,-})^2}\right),$$

where the $\tilde{O}$ notation hides logarithmic dependencies on $A, S, D_W, \Delta^{\sigma,-}$, and Δ^{σ}_{*}.

Proof. We set

$$\theta = \left\lceil \frac{4S(D_W+1)^2}{(\Delta^{\sigma,-})^2} \ln \frac{8S(D_W+1)^2}{(\Delta^{\sigma,-})^2} \right\rceil \qquad \text{and} \qquad \varepsilon = \sqrt{\frac{2S\ln(2\theta)}{\theta}}$$

and first consider the case when in episode k we run policy π_k with $\rho_k(\pi_k, s_k) > \sigma$ and $\rho(\pi_k, s_k) < \sigma$. If the total frequency freq_k is less than θ, then we consider the maximum amount for the expected regret, otherwise one can show that this event occurs with low probability. By Lemma 3, one can see that if $\varepsilon(D_W+1) < \Delta^{\sigma}_{\pi}$ and M_k is an ε-approximation of M, then

$$\rho_k(\pi) \leqslant \rho(\pi) + \varepsilon(D_W+1) \leqslant \sigma - \Delta^{\sigma}_{\pi} + \varepsilon(D_W+1) < \sigma,$$

which contradicts our assumption $\rho_k(\pi) > \sigma$. Analogous to Lemma 8, one can show that

$$\sum_{k=1}^{n} \mathbb{E}\big[\mathbb{1}\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta + 1\}\big] \leqslant \frac{2A}{\theta^{S-1}\ln(2\theta)}.$$

Consequently, if $\rho_k(\pi_k, s_k) > \sigma$ then as for the exploitation regret in the proof of Theorem 2 one can derive (but now replacing $\Delta^{\sigma,+}$ by 1) a respective upper bound of

$$\tilde{O}\left(\frac{AS^2 D_W^3}{(\Delta^{\sigma,-})^2}\right).$$

Finally, in order to bound the accumulated regret of episodes in which $\rho_k(\pi_k, s_k) \leqslant \sigma$ we can simply apply Theorem 1, noting that proof of Jaksch et al. [2010] also works when the initial states of an episode do not coincide with the last visited state of the previous episode but are chosen arbitrarily. □

If $\sigma < \rho^*(M)$, we can show also for SAT-UCRL that the σ-regret is bounded by a constant.

Theorem 4. *If $\sigma < \rho^*(M)$, then the σ-regret of* SAT-UCRL *is bounded by a constant independent of T.*

Proof. As in eq. (1) in the proof of Theorem 2, we first decompose the regret into three terms, the initialization regret, the exploitation regret, and the exploration regret. The initialization regret (of lines 6–12 of the algorithm) as well as the exploitation regret (line 16) can simply be analyzed as in the proof of Theorem 2 to get an upper bound on both terms of

$$\tilde{O}\left(\frac{\Delta^{\sigma,+} AS^2 D_W^3}{(\Delta^{\sigma,-})^2}\right). \tag{14}$$

What remains to do is to analyze the exploration regret due to episodes in which UCRL2 is played (lines 17–22 of the algorithm), which happens when all policies are empirically below σ. In the following, we consider only exploration episodes in which UCRL2 is executed and renumber these episodes using the variable $m = 1, 2, \ldots$ instead of k in order to indicate that episode m is the m-th episode in which UCRL2 is played. Then by definition of the algorithm, the number of steps of the m-th of these episodes is at least $2^{3m-3}AS$ and at most $2^{4m-3}AS$.

Now we distinguish between long and short exploration episodes and set β to be the smallest positive integer for which $\frac{2^{3\beta-5}}{\beta} \geqslant \max\{\theta_M, (\theta_*+1)^2\theta'_M\}$, where θ_M, θ'_M are defined in Appendix D.1, while θ_* is as defined in (11). Then we decompose the exploitation regret with respect to β into

$$\sum_{m=1}^{\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}] + \sum_{m>\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}], \tag{15}$$

now using the changed episode numbers and hence slightly abusing notation, so that e.g. L_m denotes the episode length for the m-th exploration episode. Using the maximal episode length of $2^{4m-3}AS$ we can bound the first term by

$$\sum_{m=1}^{\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}] \leqslant \sum_{m=1}^{\beta} 2^{4m-3} AS\Delta^{\sigma,+} \leqslant \frac{2^{4\beta+1}}{15} AS\Delta^{\sigma,+}. \tag{16}$$

Thus let us consider the regret caused by exploration episodes $m > \beta$. As shown in Lemma 12 in Appendix D.1 each such exploration episode $m > \beta$ consists a *reliable* sub-episode with at least $\lceil \frac{2^{3m-5}}{m} \rceil$ steps, which by definition of β will employ an optimal policy with probability at least $1 - \frac{1}{T}$. In the following we will indeed assume that each considered exploration episode $m > \beta$ has a reliable sub-episode in which the optimal policy is played.

For $m > \beta$ we define the following events:

- A_m denotes the event that each state-action pair of the irreducible class of the optimal policy (played in the reliable sub-episode) has been visited at least $2^{m-\beta}\theta_*$ times during episode m, where θ_* is as defined in (11).
- B_m denotes the event that A_m and that rewards or transition probabilities of some state-action pair in the irreducible class of the optimal policy are not estimated with accuracy ε^* after episode m, where $\varepsilon_* := \sqrt{\frac{2S\ln(2\theta_*)}{\theta_*}}$, cf. Appendix B.2.

Note that when A_m and $\overline{B_m}$ hold, an accuracy of ε^* has been reached, which guarantees that the optimal policy on the empirical MDP will be satisficing w.h.p., cf. Appendix B.2. Accordingly, exploration episode m playing UCRL2 will only happen when for the previous exploration episode $m-1$ we have $\overline{A_m}$ or B_m. (Recall that samples that may have been collected in the meantime in exploitation episodes are not used in exploration episodes.) Accordingly, we have

$$\sum_{m>\beta} \mathbb{E}\big[L_m \Delta^\sigma_{\pi_m}\big] = \sum_{m>\beta} \mathbb{P}(\overline{A_{m-1}})\, \mathbb{E}\big[L_m \Delta^\sigma_{\pi_m} | \overline{A_{m-1}}\big] + \sum_{m>\beta} \mathbb{P}(B_{m-1})\, \mathbb{E}\big[L_m \Delta^\sigma_{\pi_m} | B_{m-1}\big]. \quad (17)$$

Concerning the first term of (17) we again use the upper bound of $2^{4m-3}AS$ on the length of exploration episode m. Further, by Lemma 13 from Appendix D, a random walk of length at least $(2^{m-\beta}(\theta_*+1))^2\theta'_M$ will visit each each state-action pair in the irreducible class no less than $2^{m-\beta}(\theta_*+1)$ times with a probability of at least $1 - (\frac{1}{2})^{(\theta_*+1)2^{m-\beta}-1}$, so that

$$\begin{aligned}\sum_{m>\beta} \mathbb{P}(\overline{A_{m-1}})\, \mathbb{E}[L_m \Delta^\sigma_{\pi_m} | \overline{A_{m-1}}] &\leqslant \sum_{m>\beta} (2^{4m-3}AS\Delta^{\sigma,+})(\tfrac{1}{2})^{(\theta_*+1)2^{m-1-\beta}-1} \\ &\leqslant 2^{4\beta+4-\theta_*}AS\Delta^{\sigma,+}.\end{aligned}$$

For an upper bound on the second term of (17), we can apply the same proof technique as for Lemmas 9 and 10 to obtain (cf. Appendix D.2 for details)

$$\sum_{m>\beta} \mathbb{P}(B_{m-1})\, \mathbb{E}\big[L_m \Delta^\sigma_{\pi_m} | B_{m-1}\big] \leqslant \sum_{m>\beta} \frac{2^{4m-2}A^2S\Delta^{\sigma,+}}{2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}, \quad (18)$$

where $d = 2^{m-\beta}$. Collecting all regret terms (14)–(18) completes the proof of the theorem. □

6 Conclusion

While it is nice to have an algorithm that gives constant σ-regret when $\rho^* > \sigma$ and for which one obtains regret bounds as for UCRL2 otherwise, there is still some work to be done. In particular, we did not try to optimize the parameters in the constant regret bound, which are hence probably not optimal. Having a respective lower bound to compare with would help to decide whether e.g. D_W could be replaced by D and which of the gap parameters have to appear in an optimal upper bound. In any case just as in the bandit setting, while it is possible to have constant σ-regret, it still seems to be unavoidable (at least in the worst case) to have a dependence on the whole state-action space.

Another improvement that seems desirable and not out of reach is to design an algorithm that does not take the state space as input, but instead only works with the part of the state space it has discovered by itself so far.

References

Jacob D. Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandits, with and without censored feedback. In *Advances in Neural Information Processing Systems 29*, pages 4889–4897, 2016.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *COLT 2013 – The 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 122–134, 2013.

Siu On Chan, Qinghua Ding, and Sing Hei Li. Learning and testing irreducible Markov chains via the k-cover time. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 458–480. PMLR, 2021.

Beau Dabbs. Markov chains and mixing times. *University of Chicago VIGRE REU*, pages 1–20, 2009. URL `https://math.uchicago.edu/~may/VIGRE/VIGRE2009/REUPapers/Dabbs.pdf`.

Michael A. Goodrich and Morgan Quigley. Satisficing Q-learning: efficient learning in problems with dichotomous attributes. In *Proceedings of the 2004 International Conference on Machine Learning and Applications - ICMLA 2004*, pages 65–72. IEEE Computer Society, 2004.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010. ISSN 1532-4435.

Thomas Michel, Hossein Hajiabolhassan, and Ronald Ortner. Regret bounds for satisficing in multi-armed bandit problems. In *15th European Workshop on Reinforcement Learning (EWRL)*, 2022.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *Algorithmic Learning Theory*, pages 140–154. Springer, 2014.

Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

Paul Reverdy, Vaibhav Srivastava, and Naomi Ehrich Leonard. Satisficing in multi-armed bandit problems. *IEEE Trans. Autom. Control.*, 62(8):3788–3803, 2017.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res.*, 48:67–113, 2013.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 7611–7624. Curran Associates, Inc., 2021.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Information Theory Research Group Hp Laboratories*, 2003. URL `https://www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf`.

A Useful Results

A.1 Concentration Inequalities

The following concentration inequalities are derived from the Hoeffding-Chernoff bound.

Lemma 1. *Let $\bar{r}_t(s,a)$ and $\bar{p}_t(s'|s,a)$ be the empirical average reward and the empirical transition probability after observing t samples. Then*

$$\mathbb{P}\big(\big|\bar{r}_t(s,a) - \mu(s,a)\big| \geqslant \varepsilon\big) \leqslant 2\exp(-2t\varepsilon^2)$$

and

$$\mathbb{P}\big(\big|\bar{p}_t(s'|s,a) - p(s'|s,a)\big| \geqslant \varepsilon\big) \leqslant 2\exp(-2t\varepsilon^2).$$

A.2 MDP Approximations

Definition 3. *An MDP $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{r}, \hat{p})$ is environmentally an ε-approximation of another MDP $M = (\mathcal{S}, \mathcal{A}, r, p)$ if they have the same state space and for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$*

$$\sum_{s' \in \mathcal{S}} |\hat{p}(s'|s,a) - p(s'|s,a)| < \varepsilon.$$

Moreover, if in addition for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$

$$|\hat{\mu}(s,a) - \mu(s,a)| < \varepsilon$$

then $\hat{M}$ is termed an ε-approximation of M.

The following result bounds the error in optimal average reward when working with an ε-approximation.

Lemma 2. *Ortner et al. [2014] Let M be a communicating MDP with optimal policy π^* whose diameter is D and initial state s_1. If $\hat{M}$ is an ε-approximation of M, then*

$$|\rho^*(M) - \rho^*(\hat{M}, s_1)| \leqslant |\rho_{\pi^*}(M, s_1) - \rho_{\pi^*}(\hat{M}, s_1)| \leqslant \varepsilon(D+1).$$

Let π^* be a policy of $\hat{M}$ and consider an irreducible class $I_{\pi^*} \subseteq \mathcal{S}$ of π^*. If the definition of ε-approximation holds just for all states of the irreducible class I_{π^*} and actions restricted to π^* then $\hat{M}$ is called an (ε, I_{π^*})-*approximation* of M. Precisely, for all $s \in I_{\pi^*}$ we have

$$\sum_{s' \in \mathcal{S}} |\hat{p}(s'|s,\pi^*(s)) - p(s'|s,\pi^*(s))| < \varepsilon$$

and

$$|\hat{\mu}(s,\pi^*(s)) - \mu(s,\pi^*(s))| < \varepsilon.$$

The following result is a consequence of Lemma 2.

Lemma 3. *Let $M = (\mathcal{S}, \mathcal{A}, r, p)$ and $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{r}, \hat{p})$ be two MDPs with the same state space and action space, where M is communicating. Suppose that for any $(s', s, a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ that if $\hat{p}(s'|s,a) > 0$ then also $p(s'|s,a) > 0$. Let π^* be an optimal policy of $\hat{M}$ containing a unique irreducible class $\hat{I}_{\pi^*}$. If $\hat{M}$ is an $(\varepsilon, \hat{I}_{\pi^*})$-approximation of M with $\varepsilon > 0$, then for any initial state s_1,*

$$\rho_{\pi^*}(M, s_1) \geqslant \rho_{\pi^*}(\hat{M}, s_1) - \varepsilon(D_W(M) + 1).$$

Proof. Since $\hat{p}(s'|s,\pi^*(s)) > 0$ implies that $p(s'|s,\pi^*(s)) > 0$, one can conclude that policy π^* in M has also a unique irreducible class I_{π^*} containing $\hat{I}_{\pi^*}$. This means that by starting from any state and following π^* in M, we reach $\hat{I}_{\pi^*}$ after a while.

Now construct two new MDPs $M' = (I_{\pi^*}, \{\pi^*\}, r', p')$ and $\hat{M}' = (I_{\pi^*}, \{\pi^*\}, \hat{r}', \hat{p}')$ as follows. In both MDPs the state space is I_{π^*} and in each state s there is a unique action $\pi^*(s)$ available. Furthermore, for any state $s \in \hat{I}_{\pi^*}$ and $s' \in I_{\pi^*}$, the transition probabilities $p'(s'|s,\pi^*(s))$ and $\hat{p}'(s'|s,\pi^*(s))$ are the same as $p(s'|s,\pi^*(s))$ and $\hat{p}(s'|s,\pi^*(s))$, respectively. Similarily the rewards $r'(s,\pi^*(s))$ and $\hat{r}'(s,\pi^*(s))$ are the same as $r(s,\pi^*(s))$ and $\hat{r}(s,\pi^*(s))$, respectively. For any other pair s, s' where $s \notin \hat{I}_{\pi^*}$, the transition probabilities $p'(s'|s,\pi^*(s))$ and $\hat{p}'(s'|s,\pi^*(s))$ are the same as $p(s'|s,\pi^*(s))$. Also for the respective rewards $r'(s,\pi^*(s))$ and $\hat{r}'(s,\pi^*(s)))$ are the same as $r(s,\pi^*(s))$ for $s \notin \hat{I}_{\pi^*}$. One can check that M' is communicating with diameter at most $D_{\pi^*}(M)$. As M' only has a single policy π^* it is optimal, and since $\hat{M}'$ is an ε-approximation of M', the claim follows by Lemma 2. □

A.3 A Combinatorial Lemma

Let $T_1, T_2, \ldots, T_n$ be a sequence of non-empty multi-subsets of a universal set T. For any $x \in T$, let $N_i(x)$ denote the number of occurrences of the element x in T_j's, where $j < i$ (prior to the

multi-set T_i). Note that since each T_j is a multi-subset, for computing the number of occurrences of the element s in T_j, we count all of its occurrences. In this regard, the total frequency of T_i is defined as

$$\mathrm{freq}(T_i) = \min_{x \in T_i} N_i(x).$$

Lemma 4. *Let $T_1, T_2, \ldots, T_n$ be a sequence of non-empty multi-subsets of a universal set T. For any non-negative integer f, there are at most $|T|$ members of this sequence such that the total frequency of each of them is f.*

Proof. Let $1 \leqslant f_1 < f_2 < \cdots < f_\ell \leqslant n$ be distinct positive integers such that the total frequency of each of $T_{f_1}, T_{f_2}, \ldots, T_{f_\ell}$ is f. By definition of total frequency, for any $1 \leqslant j \leqslant \ell$ there exists an $x_{f_j} \in T_{f_j}$ such that $\mathrm{freq}(T_{f_j}) = f = N_{f_j}(s_{f_j})$. On the other hand, for any $f_i > f_j$ we have $s_{f_i} \neq s_{f_j}$, since otherwise

$$\mathrm{freq}(T_{f_i}) = N_{f_i}(x_{f_i}) \geqslant 1 + N_{f_j}(x_{f_i}) = 1 + \mathrm{freq}(T_{f_j}) = f + 1,$$

which is a contradiction. This completes the proof. □

A.4 Another Useful Lemma

Lemma 5. *Let $c > \frac{1}{4}$ be a real number. If $f \geqslant \lceil 2c \ln(4c) \rceil$, then*

$$\frac{\ln(2f)}{f} < \frac{1}{c}.$$

Proof. Note that the $\frac{\ln(2f)}{f}$ is decreasing when $f \geqslant \frac{e}{2}$, where e is Euler's number. If we set $f = 2c\ln(4c)$, then it is straightforward to check that the inequality $\frac{\ln(4c\ln(4c))}{2c\ln(4c)} < \frac{1}{c}$ holds. □

B Analysis of SAT-RL

We start the analysis of our algorithm SAT-RL by deriving some results concering the quality of MDP approximations that later will be used to bound the error by using the empirical MDP.

B.1 Estimated Environment in Exploitation Episode

In this section, we show that in an exploitation episode, the probability of running an unsatisfying policy is low, provided that the total frequency of its irreducible class is large enough. Setting

$$\theta = \left\lceil \frac{4S(D_W+1)^2}{(\Delta^{\sigma,-})^2} \ln \frac{8S(D_W+1)^2}{(\Delta^{\sigma,-})^2} \right\rceil \qquad \text{and} \qquad \varepsilon = \sqrt{\frac{2S\ln(2\theta)}{\theta}},$$

it holds that $\varepsilon(D_{\pi_k}+1) \leqslant \varepsilon(D_W+1) < \Delta^{\sigma,-} \leqslant \Delta^{\sigma}_{\pi_k}$ by Lemma 5, noting that it is sufficient to show

$$\frac{\ln(2\theta)}{\theta} < \frac{(\Delta^{\sigma,-})^2}{2S(D_W+1)^2}.$$

Consequently, if M_k is an (ε, I_{π_k})-approximation of M, then by Lemma 3

$$\rho(\pi_k) \geqslant \rho_k(\pi_k) - \varepsilon(D_{\pi_k}+1) > \sigma - \Delta^{\sigma}_{\pi_k} \geqslant \rho(\pi_k),$$

which is a contradiction. In the sequel, we show that with high probability, M_k is an (ε, I_{π_k})-approximation of M provided that $\mathrm{freq}_k \geqslant \theta + 1$, and consequently, π_k is satisfying with high probability. Recall that in any exploitation episode k, I_k is the irreducible class π_k induces on M_k, i.e., $I_k = \{(s, \pi_k(s)) | s \in I_{\pi_k}\}$. Let V_k the set of all state-action pairs $(s, \pi_k(s))$ such that $s \in I_{\pi_k}$ and either rewards or transition probabilities are not estimated well enough at the start of exploitation episode k. That is, for $s \in I_{\pi_k}$ we have $\sum_{s' \in \mathcal{S}} |p_k(s'|s, \pi_k(s)) - p(s'|s, \pi_k(s))| \geqslant \varepsilon$ or $|r_k(s, \pi_k(s)) - \mu(s, \pi_k(s))| \geqslant \varepsilon$.

Recall that $\bar{r}_t(s,a)$ and $\bar{p}_t(s'|s,a)$ stand for the empirical average reward and the empirical transition probability after observing exactly t samples.

Lemma 6. *For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, we have*

$$\sum_{t \geqslant \theta+1} \mathbb{P}\big(|\bar{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon\big) \leqslant \frac{1}{S\theta^{S-1}\ln(2\theta)}.$$

Proof. For any state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, and positive integer $t \geqslant \theta + 1$, by Lemma 1,

$$\begin{aligned}
\mathbb{P}\big(|\bar{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon\big) &= \mathbb{P}\Big(|\bar{r}_t(s,\pi(s)) - \mu(s,\pi(s))| \geqslant \sqrt{\tfrac{2S\ln(2\theta)}{\theta}}\Big) \\
&\leqslant 2\exp^{-2t\left(\sqrt{\frac{2S\ln(2\theta)}{\theta}}\right)^2} \\
&\leqslant 2\exp^{\frac{-4St\ln(2\theta)}{\theta}}.
\end{aligned}$$

Accordingly,

$$\begin{aligned}
\sum_{t \geqslant \theta+1} \mathbb{P}\big(|\bar{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon\big) &\leqslant \sum_{\theta+1}^{\infty} 2\exp^{\frac{-4St\ln(2\theta)}{\theta}} \\
&\leqslant \int_{\theta}^{\infty} 2\exp^{\frac{-4St\ln(2\theta)}{\theta}}\, dt \\
&\leqslant \frac{1}{S2^{4S+1}\theta^{4S-1}\ln(2\theta)} \\
&\leqslant \frac{1}{S\theta^{S-1}\ln(2\theta)}.
\end{aligned}$$

□

Lemma 7. *For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, , we have*

$$\sum_{t \geqslant \theta+1} \sum_{s'} \mathbb{P}\big(|\bar{p}_t(s'|s,a) - p(s'|s,a)| \geqslant \varepsilon\big) \leqslant \frac{1}{S\theta^{S-1}\ln(2\theta)}.$$

Proof. Weissman et al. [2003] show that for the L^1-deviation of the true distribution and the empirical distribution over m distinct events from n samples it holds that

$$\mathbb{P}\big(||\bar{p}_n(\cdot) - p(\cdot)||_1 \geqslant \varepsilon\big) \leqslant (2^m - 2)\exp^{-\frac{n\varepsilon^2}{2}}. \tag{19}$$

For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, the number of $s' \in \mathcal{S}$ for which $p(s'|s,a) > 0$ is at most S. Accordingly, by (19) for any $t \geqslant \theta + 1$,

$$\begin{aligned}
\mathbb{P}\Big(||\bar{p}_t(\cdot|s,a) - p(\cdot|s,a)||_1 \geqslant \sqrt{\tfrac{2S\ln(2\theta)}{\theta}}\Big) &\leqslant 2^S \exp^{-\frac{t\left(\sqrt{\frac{2S\ln(2\theta)}{\theta}}\right)^2}{2}} \\
&\leqslant 2^S \exp^{-\frac{St\ln(2\theta)}{\theta}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sum_{t \geqslant \theta+1} \sum_{s'} \mathbb{P}\big(|\bar{p}_t(s'|s,a) - p(s'|s,a)| \geqslant \varepsilon\big) &\leqslant \sum_{t \geqslant \theta+1} 2^S \exp^{-\frac{St\ln(2\theta)}{\theta}} \\
&\leqslant \int_{\theta}^{\infty} 2^S \exp^{-\frac{St\ln(2\theta)}{\theta}}\, dt \\
&\leqslant \frac{1}{S\theta^{S-1}\ln(2\theta)}
\end{aligned}$$

□

Now, we show that the total probability of choosing an unsatisfying policy in exploitation episodes is bounded by a constant provided that total frequency of each episode is high enough.

Lemma 8. *For any positive integer n, we have*

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta + 1\big\} \leqslant \frac{2A}{\theta^{S-1}\ln(2\theta)}.$$

Proof. If $\rho_{\pi_k}(M_k, s_k) \geqslant \sigma$ and $\rho(\pi_k, s_k) < \sigma$, then there exists a state $s \in I_{\pi_k}$ for which $(s, \pi_k(s)) \in V_k$ since otherwise by Lemma 3 we obtain the contradiction

$$\begin{aligned}\rho_k(\pi_k, s_k) = \rho_{\pi_k}(M_k, s_k) &\leqslant \rho_{\pi_k}(M, s_k) + \varepsilon(D_{\pi_k} + 1)\\ &\leqslant \rho_{\pi_k}(M, s_k) + \varepsilon(D_W + 1)\\ &< \rho_{\pi_k}(M, s_k) + \Delta^{\sigma,-}\\ &\leqslant \rho_{\pi_k}(M, s_k) + \Delta^{\sigma}_{\pi_k} = \sigma\end{aligned}$$

Hence, by Lemmas 6 and 7 we have

$$\begin{aligned}&\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta + 1\big\}\\ &\quad\leqslant \sum_{k=1}^{n} \mathbb{P}\big\{\exists (s,a) \in V_k : \pi_k(s) = a \wedge N_k(s,a) \geqslant \theta + 1\big\}\\ &\quad\leqslant \sum_{k=1}^{n}\sum_{s,a} \mathbb{P}\big\{|r_k(s,a) - \mu(s,a)| \geqslant \varepsilon \wedge s \in I_{\pi_k} \wedge \pi_k(s) = a \wedge N_k(s,a) \geqslant \theta + 1\big\}\\ &\qquad + \sum_{k=1}^{n}\sum_{s,a} \mathbb{P}\left\{\big|\big|p_k(\cdot|s,a) - p(\cdot|s,a)\big|\big|_1 \geqslant \varepsilon \wedge s \in I_{\pi_k} \wedge \pi_k(s) = a \wedge N_k(s,a) \geqslant \theta + 1\right\}\\ &\quad\leqslant \sum_{s,a}\sum_{t \geqslant \theta+1} \mathbb{P}\left\{|\bar{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon\right\} + \sum_{s,a}\sum_{t \geqslant \theta+1} \mathbb{P}\left\{\big|\big|\bar{p}_t(\cdot|s,a) - p(\cdot|s,a)\big|\big|_1 \geqslant \varepsilon\right\}\\ &\quad\leqslant \frac{2A}{\theta^{S-1}\ln(2\theta)}.\end{aligned}$$

□

B.2 Estimated Environment in Exploration Episode

Now, we show that after a certain number of exploration episodes the probability of having another exploration episode is low. We set

$$\theta_* = \left\lceil \frac{4S(D+1)^2}{(\Delta^{\sigma}_*)^2} \ln \frac{8S(D+1)^2}{(\Delta^{\sigma}_*)^2} \right\rceil \qquad \text{and} \qquad \varepsilon_* = \sqrt{\frac{2S\ln(2\theta_*)}{\theta_*}}. \tag{20}$$

By Lemma 5, one can see that $\varepsilon_*(D+1) < \Delta^{\sigma}_*$, and consequently, if M_k is an ε_*-approximation of M, then

$$\rho_k(\pi^*) \geqslant \rho(\pi^*) - \varepsilon_*(D+1) \geqslant \sigma + \Delta^{\sigma}_* - \varepsilon_*(D+1) > \sigma,$$

so that no exploration episode would be played, whihc only happens if $\rho_k(\pi^*) < \sigma$. In the sequel, we show that with high probability, M_k is an ε_*-approximation of M provided that $\mathrm{freq}_k \geqslant \theta_* + 1$, and consequently, we run exploration episodes with low probability. Note that in exploration episodes, I_k is $\mathcal{S} \times \mathcal{A}$. Let $\bar{V}_k$ be the set of all state-action pairs (s,a) for which rewards or transition probabilities are not estimated well enough at the start of exploration episode k. That is, for (s,a) we have $\sum_{s' \in \mathcal{S}} |p_k(s'|s,a) - p(s'|s,a)| \geqslant \varepsilon_*$ or $|r_k(s,a) - \mu(s,a)| \geqslant \varepsilon_*$.

Lemma 9. *For any state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, and positive integer $d \geqslant 1$, we have*

$$\sum_k \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant d\theta_* + 1 \wedge |\bar{r}_k(s,a) - \mu(s,a)| \geqslant \varepsilon_*\big) \leqslant \frac{1}{S2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}.$$

Proof. If $\text{freq}_k \geqslant d\theta_* + 1$, then any state-action pair has been visited at least $d\theta_* + 1$ times prior to episode k. For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, we have

$$\begin{aligned}
\sum_k \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant d\theta_* + 1 \wedge |\bar{r}_k(s,a) - \mu(s,a)| \geqslant \varepsilon_*\big) \\
&\leqslant \sum_{t \geqslant d\theta_*+1} \mathbb{P}\big(|\bar{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon_*\big) \\
&\leqslant \sum_{t \geqslant d\theta_*+1} 2\exp^{-2t\left(\sqrt{\frac{2S\ln(2\theta_*)}{\theta_*}}\right)^2} \\
&\leqslant \sum_{t \geqslant d\theta_*+1} 2\exp^{\frac{-4St\ln(2\theta_*)}{\theta_*}} \\
&\leqslant \int_{d\theta_*}^{\infty} 2\exp^{\frac{-4St\ln(2\theta_*)}{\theta_*}}\, dt \\
&\leqslant \frac{1}{S2^{4dS+1}\theta_*^{4dS-1}\ln(2\theta_*)} \\
&< \frac{1}{S2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}
\end{aligned}$$

□

Lemma 10. *For any state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, and positive integer $d \geqslant 1$, we have*

$$\sum_k \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant d\theta_* + 1 \wedge ||\bar{p}_k(\cdot|s,a) - p(\cdot|s,a)||_1 \geqslant \varepsilon_*\big) \leqslant \frac{1}{S2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}.$$

Proof. For any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, the number of $s' \in \mathcal{S}$ for which $p(s'|s,a) > 0$ is at most S. Accordingly, by (19) for any $t \geqslant d\theta_* + 1$

$$\begin{aligned}
\sum_k \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant d\theta_* + 1 \wedge ||\bar{p}_k(\cdot|s,a) - p(\cdot|s,a)||_1 \geqslant \varepsilon_*\big) \\
&\leqslant \sum_{t \geqslant d\theta_*+1} \mathbb{P}\big(||\bar{p}_t(\cdot|s,a) - p(\cdot|s,a)||_1 \geqslant \varepsilon_*\big) \\
&\leqslant \sum_{t \geqslant d\theta_*+1} 2^S \exp^{-\frac{t\left(\sqrt{\frac{2S\ln(2\theta_*)}{\theta_*}}\right)^2}{2}} \\
&\leqslant \sum_{t \geqslant d\theta_*+1} 2^S \exp^{-\frac{St\ln(2\theta_*)}{\theta_*}} \\
&\leqslant \int_{d\theta_*}^{\infty} 2^S \exp^{-\frac{St\ln(2\theta_*)}{\theta_*}}\, dt \\
&\leqslant \frac{1}{S2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}
\end{aligned}$$

□

Now, we show that the total probability of running exploration episodes is bounded by a constant.

Lemma 11. *If $\rho_{\pi^*} > \sigma$, then*

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant d\theta_* + 1\big\} \leqslant \frac{2A}{2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}.$$

Proof. If $\rho_{\pi^*}(M_k, s_k) < \sigma$, then there exists an $(s,a) \in \bar{V}_k$ since otherwise by Lemma 2 we would obtain the contradiction

$$\begin{aligned}
\rho_{\pi^*}(M_k, s_k) &\geqslant \rho_{\pi^*}(M, s_k) - \varepsilon_*(D+1) \\
&> \rho_{\pi^*}(M, s_k) - \Delta_*^{\sigma} = \sigma.
\end{aligned}$$

Hence, by Lemmas 9 and 10,

$$
\begin{aligned}
\sum_{k=1}^{n}\mathbb{P}\big\{\rho_k(\pi_k,s_k)<\sigma\wedge \mathrm{freq}_k\geqslant d\theta_*+1\big\} &\leqslant \sum_{k=1}^{n}\mathbb{P}\big\{\rho_k(\pi^*,s_k)<\sigma\wedge \mathrm{freq}_k\geqslant d\theta_*+1\big\}\\
&\leqslant \sum_{k=1}^{n}\mathbb{P}\left\{\exists (s,a)\in \bar{V}_k : N_k(s,a)\geqslant d\theta_*+1\right\}\\
&\leqslant \sum_{k=1}^{n}\sum_{s,a}\mathbb{P}\left\{|r_k(s,a)-\mu(s,a)|\geqslant \varepsilon_*\wedge N_k(s,a)\geqslant d\theta_*+1\right\} \qquad (21)\\
&\quad+\sum_{k=1}^{n}\sum_{s,a}\mathbb{P}\left\{\big|\big|p_k(\cdot|s,a)-p(\cdot|s,a)\big|\big|_1\geqslant \varepsilon_*\wedge N_k(s,a)\geqslant d\theta_*+1\right\}\\
&\leqslant \sum_{s,a}\sum_{t\geqslant d\theta_*+1}\mathbb{P}\left\{|\bar{r}_t(s,a)-\mu(s,a)|\geqslant\varepsilon_*\right\}+\sum_{s,a}\sum_{t\geqslant d\theta_*+1}\mathbb{P}\left\{\big|\big|\bar{p}_t(\cdot|s,a)-p(\cdot|s,a)\big|\big|_1\geqslant\varepsilon_*\right\}\\
&\leqslant \frac{2A}{S2^{dS-S}\theta_*^{dS-1}\ln(2\theta_*)}
\end{aligned}
$$

□

C SAT-RL2

In this part of the appendix we present an alternative algorithm that does not need to use GOSPRL to estimate the diameter of the underlying MDP in order to guarantee that the empirical MDP is communicating.

C.1 SAT-RL2: Estimation of the Diameter Without GOSPRL

The algorithm SAT-RL2 is shown as Algorithm 3 skips the part of SAT-RL which uses GOSPRL to estimate the diameter of the underlying MDP. As already dsicussed, this is done to guarantee that the empirical MDP is communicating. Now SAT-RL2 instead of using the GOSPRL procedure to estimate the diameter performs an ordinary exploration episode using GOSPRL.

Algorithm 3 SAT-RL2: Satisficing without using GOSPRL for diameter estimation

Input: state space $\mathcal{S}$, action space $\mathcal{A}$, satisfaction level σ
Initialization:
Set confidence level $\delta_g := \frac{1}{2}$,and initial sampling number $b := S+1$.
Define function $\bar{b} : \mathcal{S}\times\mathcal{A}\to\mathbb{N}$ to be $\bar{b}(s,a)=b$ for any (s,a).
while an action $a\in\mathcal{A}$ at some state $s\in\mathcal{S}$ has not been run $\bar{b}(s,a)$ times **do**
    Run GOSPRL($\bar{b},\delta_g$).
    For any $(s,a)\in\mathcal{S}\times\mathcal{A}$, define $\bar{b}(s,a) := b - N(s,a)$.
end while
for episodes $k=1,2,\ldots$ **do**
    Compute an optimal policy π_k on M_k.
    if M_k is communicating and $\rho_{\pi_k}(M_k,s_k)\geqslant\sigma$ **then**
        π_k can be chosen to induce a unique irreducible class I_k.
        Play π_k until all states in I_k have been visited at least once.
    else
        Set $b := b+S$.
        while for any state $s\in\mathcal{S}$ and $a\in\mathcal{A}$, $N(s,a)<b$ **do**
            For any (s,a), set $\bar{b}(s,a) := b - N(s,a)$.
            Run GOSPRL($\bar{b},\delta_g$).
        end while
    end if
end for

C.2 Approximation Error for the Empirical Diameter

In this section we derive a bound on the approximation error when estimating the diameter by considering the diameter on an empirically estimated MDP.

We start with some definitions. Let Π be a multi-set consisting of S stationary policies such that for each state $s \in \mathcal{S}$ there exists a unique policy $\pi_s \in \Pi$. Consider an agent following a policy Π for a while and then changing it by switching to policy $\pi_s \in \Pi$ when being in some state s. By iterating this procedure, we obtain a non-stationary policy. We call such a policy *semi-stationary* and denote the set of semi-stationary policies of M by $\Pi^{Sem}(M)$. Naturally, we can generalize the definition of diameter of an MDP as follows.

Definition 4. *Consider the stochastic process defined by a semi-stationary policy $\pi^+ \in \Pi^{Sem}(M)$ operating on an MDP M with initial state s. Let $T(s'\big|M, \pi^+, s)$ be the random variable for the first time step in which state s' is reached in this process. Then the semi-diameter of M is defined as*

$$D^{Sem}(M) = \max_{s \neq s' \in \mathcal{S}} \min_{\pi^+ \in \Pi^{Sem}(M)} \mathbb{E}\big[T(s'\big|M, \pi^+, s)\big].$$

Obviously, $D^{Sem}(M) \leqslant D(M)$ in any MDP M. Not surprisingly, the two notions coincide in general.

Proposition 2. *For any MDP M,*

$$D^{Sem}(M) = D(M).$$

Theorem 5. *Let $M = (\mathcal{S}, \mathcal{A}, r, p)$ be a communicating MDP with diameter D and $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{r}, \hat{p})$ be environmentally an ε-approximation of M over the same state-action space. Then for any positive integer $l \geqslant 3$ and any $\varepsilon < \frac{l-2}{l(lD-1)}$ the diameter of $\hat{M}$ is at most $l^2 D - l$.*

Proof. Since the diameter of M is D, for any two states s and s' there exists a policy $\pi_{s,s'}$ such that by following $\pi_{s,s'}$ starting in s, we reach s' in at most D steps (on average). Hereafter, by $\pi_{s,s'}$-exploration in M, we mean the process of following policy $\pi_{s,s'}$ starting in s. A $\pi_{s,s'}$-exploration of l steps in M generates a sequence $s_0 s_1 \cdots s_l$ such that $s_0 = s$ and $p\big(s_i\big|s_{i-1}, \pi_{s,s'}(s_{i-1})\big) > 0$ for any $1 \leqslant i \leqslant l$. Now, for any $s, s' \in \mathcal{S}$, we show that with probability more than $\frac{1}{l}$, a $\pi_{s,s'}$-exploration of $lD - 1$ steps in $\hat{M}$ contains s'. For any sequence $s_0 s_1 \cdots s_l$ we consider the probabilities

$$\mathbb{P}_M^{\pi_{s,s'}}\big(s_0 s_1 \cdots s_l\big) = \prod_{i=1}^{l} p\big(s_i\big|s_{i-1}, \pi_{s,s'}(s_{i-1})\big),$$

$$\mathbb{P}_{\hat{M}}^{\pi_{s,s'}}\big(s_0 s_1 \cdots s_l\big) = \prod_{i=1}^{l} \hat{p}\big(s_i\big|s_{i-1}, \pi_{s,s'}(s_{i-1})\big).$$

Further, let

$$\mathbb{P}_{\min}^{\pi_{s,s'}}\big(s_0 s_1 \cdots s_l\big) = \prod_{i=1}^{l} \min\{p\big(s_i\big|s_{i-1}, \pi_{s,s'}(s_{i-1})\big), \hat{p}\big(s_i\big|s_{i-1}, \pi_{s,s'}(s_{i-1})\big)\},$$

which is the minimal probability of generating the sequence $s_0 s_1 \cdots s_l$ if we follow policy $\pi_{s,s'}$ for l steps in M (resp. $\hat{M}$) starting in s_0.

For any finite sample space Ω, any function $f : \Omega \to \mathbb{R}$, and an event $W \subseteq \Omega$, we set $f(W) := \sum_{w \in W} f(w)$. Consider the sample spaces consisting of all possible $\pi_{s,s'}$-explorations of l steps in M and $\hat{M}$, respectively:

$$W_M(s, l) = \{s_0 s_1 \cdots s_l \big| s_0 = s, s_i \in \mathcal{S}, \mathbb{P}_M^{\pi_{s,s'}}(s_0 s_1 \cdots s_l) > 0\},$$

$$W_{\hat{M}}(s, l) = \{s_0 s_1 \cdots s_l \big| s_0 = s, s_i \in \mathcal{S}, \mathbb{P}_{\hat{M}}^{\pi_{s,s'}}(s_0 s_1 \cdots s_l) > 0\}.$$

Further, let $W_M(s, s', l) \subseteq W_M(s, l)$ consist of all $\pi_{s,s'}$-explorations of length l in M that contain s'. That is,

$$W_M(s, s', l) = \big\{s_0 s_1 \cdots s_l \big| s_0 = s, s_j = s' \text{ for some } 0 \leqslant j \leqslant l, s_i \in \mathcal{S}, \mathbb{P}_M^{\pi_{s,s'}}(s_0 s_1 \cdots s_l) > 0\big\}.$$

Denoting by $\overline{W_M(s,s',l)} \subseteq W_M(s,l)$ the complement of $W_M(s,s',l)$ we obviously have

$$W_M(s,l) = W_M(s,s',l) \cup \overline{W_M(s,s',l)}.$$

Note that for any positive integer l, the probability of choosing a sequence in $W_M(s,s',lD-1)$ is at least $\frac{l-1}{l}$, i.e., $\mathbb{P}_M^{\pi_{s,s'}}(W_M(s,s',lD-1)) \geqslant \frac{l-1}{l}$: Otherwise s' would be reached from s in more than D steps on average, contradicting our assumption. Since $\hat{M}$ is environmentally an ε-approximation of M, we have

$$\mathbb{P}_{\min}^{\pi_{s,s'}}\big(W_M(s,lD-1) \cap W_{\hat{M}}(s,lD-1)\big) \geqslant (1-\varepsilon)^{lD-1} \geqslant 1-(lD-1)\varepsilon > \frac{2}{l}.$$

Now, we claim that

$$\mathbb{P}_{\min}^{\pi_{s,s'}}\big(W_M(s,s',lD-1) \cap W_{\hat{M}}(s,lD-1)\big) > \frac{1}{l}, \tag{22}$$

since otherwise $\mathbb{P}_{\min}^{\pi_{s,s'}}\big(\overline{W_M(s,s',lD-1)} \cap W_{\hat{M}}(s,lD-1)\big) > \frac{1}{l}$ and consequently

$$\begin{aligned}\mathbb{P}_M^{\pi_{s,s'}}\big(\overline{W_M(s,s',lD-1)}\big) &\geqslant \mathbb{P}_M^{\pi_{s,s'}}\big(\overline{W_M(s,s',lD-1)} \cap W_{\hat{M}}(s,lD-1)\big)\\ &\geqslant \mathbb{P}_{\min}^{\pi_{s,s'}}\big(\overline{W_M(s,s',lD-1)} \cap W_{\hat{M}}(s,lD-1)\big).\\ &> \frac{1}{l}.\end{aligned}$$

This leads us to a contradiction since

$$\begin{aligned}1 &= \mathbb{P}_M^{\pi_{s,s'}}\big(W_M(s,lD-1)\big)\\ &= \mathbb{P}_M^{\pi_{s,s'}}\big(W_M(s,s',lD-1)\big) + \mathbb{P}_M^{\pi_{s,s'}}\big(\overline{W_M(s,s',lD-1)}\big)\\ &> \frac{l-1}{l} + \frac{1}{l} = 1.\end{aligned}$$

Now, by (22), one can conclude that

$$\mathbb{P}_{\hat{M}}^{\pi_{s,s'}}\big(W_M(s,s',lD-1) \cap W_{\hat{M}}(s,lD-1)\big) > \frac{1}{l}.$$

This means that if we run a $\pi_{s,s'}$-exploration of $lD-1$ steps in environment $\hat{M}$, then this exploration contains s' with probability more than $\frac{1}{l}$. Now, consider the following policy to estimate the diameter of $\hat{M}$. For any two states s and s', start from s and follow the policy $\pi_{s,s'}$ for $lD-1$ steps in $\hat{M}$. If after at most $lD-1$ steps, the state s' has not been reached then for the current state s'' follow the policy $\pi_{s'',s'}$ for next $lD-1$ steps. Iterate this procedure until s' is reached. In view of the expectation of the geometric distribution and Lemma 2, the expected number of steps to reach s' then is at most l^2D-l. □

C.3 Regret Bound for SAT-RL2

As for SAT-RL we have a constant bound on the σ-regret of SAT-RL2.

Theorem 6. *If $\rho_{\pi^*} > \sigma$, then the expected σ–regret of* SAT-RL2 *satisfies for all $n \geqslant 1$*

$$R_n^S = \tilde{O}\bigg(\frac{AS^2D^{\frac{7}{2}}}{(\Delta_*^\sigma)^2} + \frac{(\Delta_*^\sigma)^{2S-2}A^2}{D^{S-\frac{5}{2}}S^{S-3}} + \frac{\Delta^{\sigma,+}AS^2D_W^3}{(\Delta^{\sigma,-})^2}\bigg),$$

where logarithmic dependencies on $A, S, D_W, \Delta^{\sigma,-}$, and Δ_*^σ are not shown.

Proof. The theorem is derived analogously to Theorem 2. What has to be additionally shown is a bound on the number of steps in exploration episodes which guarantee that the empirical MDP is communicating with high probability. This bound is obtained using Theorem 5: Setting $l = 3$ in Theorem 5, if the estimated environment M_k is an ε-approximation of M with $\varepsilon < \frac{1}{9D}$, then M_k has finite diameter, i.e. is communicating. Set $\varepsilon = \frac{1}{9}\sqrt{\frac{2S\ln(2\theta_*)}{\theta_*}}$. By Lemma 5, we have $9\varepsilon D < \varepsilon_*(D+1) < \Delta_*^\sigma \leqslant 1$. On the other hand,

$$\varepsilon = \frac{1}{9}\sqrt{\frac{2S\ln(2\theta_*)}{\theta_*}} \leqslant \sqrt{\frac{2S\ln(162\theta_*)}{81\theta_*}}.$$

By setting $d = 1$ and replacing θ_* by $81\theta_*$ in Lemma 11 one can prove that

$$\sum_{k=1}^{n} \mathbb{P}\big\{\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant 81\theta_* + 1\big\} \leqslant \frac{2A}{(81\theta_*)^{S-1}\ln(162\theta_*)}.$$

This means that if we run the exploration part of algorithm for $\lceil \frac{81\theta_*}{S} \rceil$ times, then the estimated environment is an ε-approximation of M with probability at least $1 - \frac{2A}{(81\theta_*)^{S-1}\ln(162\theta_*)}$. This will only cause an additional factor of 81 in the the regret term of the exploration part, so that the claimed bound holds. □

D Details for the Proof of Theorem 4

D.1 Sufficient Episode Lengths

D.1.1 Parameter θ_M

Let θ_M be the smallest positive integer such that for any $t \geqslant \theta_M$, we have

$$\frac{t\Delta_g}{2} > 34^2 \frac{AS^2D^2\ln(t)}{\Delta_g} + \sum_{a,s}\big[1 + \log_2(\max_{\pi:\pi(s)=a} T_\pi)\big] \max_{\pi:\pi(s)=a} T_\pi.$$

If $\rho^*(M) \geqslant \sigma$, then executing SAT-UCRL for T steps guarantees that the likelihood of encountering an unsatisfying policy π for $t \geqslant \theta_M$ consecutive steps within a segment of UCRL2 is no more than $\delta = \frac{1}{3T}$.

Lemma 12. *With probability at least $1 - \frac{1}{T}$, any exploration episode $m > \beta$ contains a reliable sub-episode.*

Proof. By definition of the algorithm, the number of steps of the m-th exploration episode is at least $2^{3m-3}AS$ and at most $2^{4m-3}AS$. So, by Proposition 1, one can conclude that the number of sub-episodes in episode m is at most $4mAS$. Thus, there is a sub-episode with a minimum of $\frac{2^{3m-5}}{m}$ running steps, and accordingly, if $\frac{2^{3m-5}}{m} \geqslant \theta_M$, then the policy of this particular episode is deemed optimal with a probability of at least $1 - \frac{1}{T}$. □

D.1.2 Parameter θ'_M

The expected duration of a random walk to visit each state no less than k times, in a Markov chain M with S states, is known as the k-cover time and denoted by $\tau_k(M)$. In Chan et al. [2021], it was proved that the k-cover time of an irreducible Markov chain M is at most $(e^2 + e\ln(S))(\tau_1(M) + \frac{k}{\zeta_M})$, where e is Euler's number and ζ_M is the minimum stationary probability of states.

Lemma 13. *Any random walk of length $\ell\ell'\theta'_M$ will visit each state-action pair in the irreducible class no less than ℓ times with a probability of at least $1 - (\frac{1}{2})^{\ell'-1}$.*

Proof. For any policy π with a unique irreducible class I_π, consider the irreducible Markov chain M_π restricted to states of I_π in which transition from one state to another is according to π. Set $t_\pi = (e^2 + e\ln(S))(\tau_1(M_\pi) + \frac{1}{\zeta_{M_\pi}})$. Note that ℓt_π is an upper bound for $\tau_\ell(M_\pi)$. One can conclude that any random walk of length $2\ell t_\pi$ starting in the irreducible class visit each state at least ℓ times with probability at least $\frac{1}{2}$. On the other hand, one can reach to any irreducible class with D_W steps (in average). Accordingly, set

$$\theta'_M = \max_{\pi:\rho(\pi)=\rho_*} \big(2e^2 + 2e\ln(S)\big)\Big(\tau_1(M_\pi) + \tfrac{1}{\zeta_{M_\pi}}\Big) + 2D_W.$$

All together, given a policy (with unique irreducible class) and starting from any state, it can be inferred that a random walk of length $\ell\ell'\theta'_M$ will inevitably reach the irreducible class with a probability of $1 - (\frac{1}{2})^{\ell\ell'}$ during the first $2\ell\ell' D_W$ steps. Next, it will visit each state-action pair in that class no less than ℓ times with a probability of at least $1 - (\frac{1}{2})^{\ell'}$. Accordingly, any random walk of length $\ell\ell'\theta'_M$ will visit each state-action pair in the irreducible class no less than ℓ times with a probability of at least $1 - (\frac{1}{2})^{\ell'-1}$. □

D.1.3 Bounding $\mathbb{P}(\overline{A_m})$

In the following we use the definition of θ_* as in (20) of Section B.2.

Lemma 14. *For any $m \geqslant \beta \geqslant 4$,*

$$\mathbb{P}(\overline{A_m}) \leqslant \left(\tfrac{1}{2}\right)^{(\theta_*+1)2^{m-\beta}-1}.$$

Proof. By Lemma 13, $\mathbb{P}(\overline{A_\beta}) \leqslant (\frac{1}{2})^{\theta_*}$. For any $k > \beta \geqslant 4$, $\lceil \frac{2^{3k-5}}{k} \rceil \geqslant 4 \lceil \frac{2^{3(k-1)-5}}{k-1} \rceil$. Also, $\frac{2^{3\beta-5}}{\beta} \geqslant \max\{\theta_M, (\theta_*+1)^2 \theta'_M\}$. So for any $m > \beta \geqslant 4$, $\lceil \frac{2^{3m-5}}{m} \rceil \geqslant (2^{m-\beta}\theta_*)^2 \theta'_M$. Therefore, for any $m \geqslant \beta \geqslant 4$, it holds that $\mathbb{P}(\overline{A_m}) \leqslant (\frac{1}{2})^{(\theta_*+1)2^{m-\beta}-1}$. □

D.2 Derivation of (18)

Let ε_* be as in (20) of Section B.2. Then using the same proof technique as for Lemmas 9 and 10 we have

$$\begin{aligned}
&\sum_{m>\beta} \mathbb{P}(B_{m-1})\, \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m} | B_{m-1}] \\
&\quad \leqslant \sum_{m>\beta} 2^{4m-3} AS\Delta^{\sigma,+} \Big(\sum_{s,a} \mathbb{P}\left\{ |r_m(s,a) - \mu(s,a)| \geqslant \varepsilon_* \wedge N_m(s,a) \geqslant 2^{m-\beta}(\theta_*+1) \right\} \Big) \\
&\quad + \sum_{m>\beta} 2^{4m-3} AS\Delta^{\sigma,+} \Big(\sum_{s,a} \mathbb{P}\left\{ \left|\left| p_m(\cdot|s,a) - p(\cdot|s,a) \right|\right|_1 \geqslant \varepsilon_* \wedge N_m(s,a) \geqslant 2^{m-\beta}(\theta_*+1) \right\} \Big) \\
&\quad \leqslant \sum_{m>\beta} 2^{4m-3} AS\Delta^{\sigma,+} \Big(\sum_{s,a} \sum_{t \geqslant 2^{m-\beta}(\theta_*+1)} \mathbb{P}\left\{ |\overline{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon_* \right\} \Big) \\
&\quad + \sum_{m>\beta} 2^{4m-3} AS\Delta^{\sigma,+} \Big(\sum_{s,a} \sum_{t \geqslant 2^{m-\beta}(\theta_*+1)} \mathbb{P}\left\{ \left|\left| \overline{p}_t(\cdot|s,a) - p(\cdot|s,a) \right|\right|_1 \geqslant \varepsilon_* \right\} \Big) \\
&\quad \leqslant \sum_{m>\beta} \frac{2^{4m-2} A^2 S \Delta^{\sigma,+}}{2^{dS-S} \theta_*^{dS-1} \ln(2\theta_*)},
\end{aligned}$$

where $d = 2^{m-\beta}$.