Online Regret Bounds for Satisficing in MDPs

Hossein Hajiabolhassan
Institute of Human Genetics
Diagnostic and Research Center for Molecular Biomedicine
Medical University of Graz
Austria
hossein.hajiabolhassan@medunigraz.at

Ronald Ortner
Lehrstuhl für Informationstechnologie
Montanuniversität Leoben
Austria
rortner@unileoben.ac.at

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 give 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., 2023]. For the general MDP setting, beside some related work on multi-objective RL [cf. Roijers

16th European Workshop on Reinforcement Learning (EWRL 2023).

et al., 2013, for an overview] and experimental work on a satisficing variant of Q-learning [Goodrich and Quigley, 2004] we are only aware of [Arumugam and Roy, 2022], which proposes an algorithm that uses rate distortion theory for satisficing in episodic RL. For this algorithm also regret bounds are derived, which however are not directly comparable to our results, as the setting considered in [Arumugam and Roy, 2022] is Bayesian and the bounds accordingly are for the Bayesian regret.

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 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, that is, the underlying MDP M is communicating. 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 connected 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 with respect to σ 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 order $D_\pi(M)$, cf. [Jaksch et al., 2010].

More generally we are interested in the σ-regret of 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

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

Beside the σ-regret we will also consider the classic regret with respect to ρ^* after any T steps as defined in [Jaksch et al., 2010] as

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

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 initial state.[1]

Further, we set $\Delta^{\sigma}_{*} := \rho^* - \sigma$ and $\Delta_g := \rho^* - \max_{\pi:\rho_\pi<\rho^*} \rho_\pi$ to be the gaps between the optimal average reward and σ resp. 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 briefly recall some basic properties that we will need for our purposes.

2.1 UCRL2

UCRL2 [Jaksch et al., 2010] is a well-known RL algorithm which is 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 as defined in (1). 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 $\delta = \frac{1}{3T}$ in a communicating MDP is bounded by*

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

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

[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'.

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 $\rho^* > \sigma$.

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_{π_k}.
14: **if** $\rho_{\pi_k}(M_k, s_k) \geqslant \sigma$ **then** perform *exploitation episode*:
15: Play π_k until all states in I_{π_k} have been visited at least once.
16: **else** perform *exploration episode*:
17: Set $b := b + S$.
18: **while** $N(s, a) < b$ for some state-action pair (s, a) **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 and can be assumed to have a unique irreducible class I_{π_k}. Furthermore, by running another instance of value iteration we can further assume that for all states not in I_{π_k} the computed policy π_k will move to I_{π_k} as fast as possible, i.e., in expected time at most D.

If the average reward of π_k on M_k is at least σ, SAT-RL plays 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 (lines 17–21).

Concerning computational complexity, the computationally most elaborate step of SAT-RL is the (repeated) calculation of the optimal policy of the empirical MDP (line 13). Similarly, GOSPRL has to repeatedly solve a stochastic shortest path problem, which is a special instance of finding an optimal policy in an MDP. This problem can be solved in polynomial time e.g. by LP algorithms (cf. Section 38.3.1 of Lattimore and Szepesvári [2020]). The suggested value iteration usually works well in practice, however does not have polynomial time guarantees [Feinberg and Huang, 2014, Balaji et al., 2019].

Before we proceed to analyze SAT-RL, we note that in Appendix C we present an alternative algorithm called SAT-RL2 that instead of running the GOSPRL procedure to estimate the 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 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 $\rho^* > \sigma$, 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

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. Similarly, M_k denotes the estimated MDP and s_k is the initial state at start of episode k. Further, we set $\rho_k(\pi_k, s_k) := \rho_{\pi_k}(M_k, s_k)$ and $\rho(\pi_k, s_k) := \rho_{\pi_k}(M, s_k)$.

Let L_k be a random variable for the number of steps in episode k (for $k > 0$) and in the initialization phase (for $k = 0$), respectively. Then the regret after n episodes can be bounded by

$$\mathbb{E}[L_0] + \sum_{k=1}^{n} \mathbb{E}\big[\mathbb{1}\{\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma\} L_k \Delta_{\pi_k}^\sigma\big] + \sum_{k=1}^{n} \mathbb{E}\big[\mathbb{1}\{\rho_k(\pi_k, s_k) < \sigma\} L_k\big]. \quad (2)$$

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. In order to repair this we simply consider a regret of 1 per step in these episodes k and accordingly simply bound the (expected) number of steps L_k. This is already reflected in (2).

In the following we define the *frequency* freq_k of episode 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

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

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 $D(1 + \log(S))$. 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 $D_{\pi_k}(1+\log(S))$ steps to cover it following policy π_k. Further, $\Delta^{\sigma}_{\pi_k}$ can be upper bounded by $\Delta^{\sigma,+}$, so that

$$\sum_{\pi} \mathbb{E}\big[L_k \Delta^{\sigma}_{\pi_k} \big| \pi_k = \pi \wedge \rho_k(\pi, s_k) \geqslant \sigma \wedge \rho(\pi, s_k) < \sigma\big] \;\leqslant\; D_W(2+\log(S))\,\Delta^{\sigma,+}. \tag{3}$$

It remains to analyze the term

$$\begin{aligned}
&\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma\big) \\
&\quad = \sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \leqslant \theta\big) \qquad (4)\\
&\qquad\qquad + \sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta+1\big), \qquad (5)
\end{aligned}$$

where we choose

$$\theta = \left\lceil \frac{4S(D_W+1)^2}{(\Delta^{\sigma,-})^2}\, \log\!\left(\tfrac{8S(D_W+1)^2}{(\Delta^{\sigma,-})^2}\right) \right\rceil. \tag{6}$$

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 $\mathrm{freq}_k = f$, so that (4) can be bounded as

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \leqslant \theta\big) \;\leqslant AS(\theta - S) \;\leqslant\; \tilde{O}\!\left(\frac{AS^2 D_W^2}{(\Delta^{\sigma,-})^2}\right). \tag{7}$$

Further, by Lemma 8 in Appendix B.1, we can bound (5) as

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) \geqslant \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant \theta+1\big) \;\leqslant\; \frac{2A}{\theta^{S-1}\log(2\theta)}. \tag{8}$$

Taking together eqs. (3) – (8) we obtain

$$R_{\mathrm{Exploit}} \leqslant \tilde{O}\!\left(\frac{\Delta^{\sigma,+} AS^2 D_W^3}{(\Delta^{\sigma,-})^2}\right). \tag{9}$$

Upper Bound on Initialization Regret

The sample complexity of GOSPRL($\bar{b}, \delta_g$) is $\tilde{O}(\bar{B}D + AS^2 D^{\frac{3}{2}})$, 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 $\bar{B} = AS(S+1)$ and $\delta_g = \frac{1}{2}$. 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

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

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

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

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

$$R_{\mathrm{Init}} \;\leqslant\; \tilde{O}(AS^2 D^{\frac{3}{2}}) + \tilde{O}(AS^2 D^3) \;=\; \tilde{O}(AS^2 D^3). \tag{10}$$

Upper Bound on Exploration Regret

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

$$\mathbb{E}\big[L_k \big| \rho_k(\pi_k, s_k) \leqslant \sigma\big] \;=\; \tilde{O}(AS^2 D^{\frac{3}{2}}) \tag{11}$$

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

$$\begin{aligned}\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma\big) \;&=\; \sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma) \wedge \text{freq}_k \leqslant \theta_*\big) \\ &\quad + \sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant \theta_* + 1\big),\end{aligned} \tag{12}$$

where we set

$$\theta_* = \left\lceil \frac{4S(D+1)^2}{(\Delta_*^\sigma)^2} \log\Big(\tfrac{8S(D+1)^2}{(\Delta_*^\sigma)^2}\Big) \right\rceil. \tag{13}$$

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,

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \leqslant \theta_*\big) \;\leqslant \left\lceil \frac{\theta_*}{S} \right\rceil. \tag{14}$$

Concerning (12), Lemma 11 in Appendix B.2 shows that

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \geqslant \theta_* + 1\big) \leqslant \frac{2A}{\theta_*^{S-1} \log(2\theta_*)}. \tag{15}$$

Consequently, summarizing (11) – (15) we obtain

$$R_{Explore} \leqslant \tilde{O}\left(\frac{AS^2 D^{\frac{7}{2}}}{(\Delta_*^\sigma)^2} + \frac{(\Delta_*^\sigma)^{2S-2} A^2}{D^{S-\frac{5}{2}} S^{S-3}} \right) \tag{16}$$

and summing up the three regret terms (9), (10), and (16) yields the claimed regret bound of the theorem. □

5 The General Case

We have seen that when the satisfaction level σ is 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 $\rho^* < \sigma$ 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 present the algorithm SAT-UCRL, which precisely achieves that: When σ is below ρ^*, we have constant regret just as for SAT-RL. If however $\rho^* \leqslant \sigma$, we show a bound with the same dependency on T as the one given in Theorem 1. These results generalize the findings of Michel et al. [2023] from the bandit to the MDP setting.

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 and these exploration episodes have increasing length (cf. line 18). As already mentioned, UCRL2 itself uses episodes in which it follows 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_{π_k}.
    if $\rho_{\pi_k}(M_k, s_k) \geqslant \sigma$ **then** perform *exploitation episode*:
        Play π_k until all states in I_{π_k} have been visited at least once.
    else perform *exploration episode* using UCRL2 with confidence parameter $\delta = \frac{1}{3T}$:
        Set $b_u := 8b_u$.
        while the length of the current episode is below b_u **do**
            Run a **sub-episode** of UCRL2.
        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. Concerning computational complexity, similar to SAT-UCRL the computationally most eloborate step of SAT-UCRL is the calculation of the optimal policy in an MDP. Due to the use of UCRL2 this also concerns MDPs with continuous action space. Still, the computation can be done in polynomial time as shown in Section 38.5.2 of Lattimore and Szepesvári [2020].

5.2 Regret Bound for SAT-UCRL

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

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

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

Proof. Similar to eq. (2) in the proof of Theorem 2, we first decompose the regret into three terms, the regret accumulated in the initialization phase and in exploitation episodes as well as the regret of exploration episodes. The regret in the intialization phase can be bounded just as in (10) in the proof of Theorem 2. Similarly, the regret accumulated in exploitation episodes can be analyzed as the respective exploitation regret in the proof of Theorem 2 with the only difference that we now consider a per-step regret of 1 instead of $\Delta^{\sigma,+}$. Accordingly, we obtain an upper bound on the regret in exploitation episodes of

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

a term that also subsumes the already mentioned regret in the initialization phase.

Finally, in order to bound the accumulated regret of exploration episodes we can simply apply Theorem 1, noting that the 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. Summing up the three regret terms gives the claimed bound. □

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

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

Proof. Once more we decompose the regret into three terms, the initialization regret, the exploitation regret, and the exploration regret. As the algorithm is the same as SAT-RL in the initialization phase and exploitation episodes, the first two regret terms can simply be analyzed as in the proof of Theorem 2, which yields an upper bound on both terms of

$$\tilde{O}\left(\frac{\Delta^{\sigma,+} AS^2 D_W^3}{(\Delta^{\sigma,-})^2}\right). \tag{17}$$

What remains to do is to analyze the exploration regret due to episodes in which UCRL2 is played, which happens when all policies are empirically below σ. In the following, we consider only these exploration episodes and renumber them using the variable $m = 1, 2, \ldots$ instead of k in order to indicate that episode m is the m-th exploration episode. 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, while θ_* is as defined in (13). Then we decompose the exploitation regret with respect to β into

$$\sum_{m=1}^{\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}] + \sum_{m>\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}], \tag{18}$$

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

$$\sum_{m=1}^{\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}] \leqslant \sum_{m=1}^{\beta} 2^{4m-3} AS \Delta^{\sigma,+} \leqslant \frac{2^{4\beta+1}}{15} AS \Delta^{\sigma,+}. \tag{19}$$

Thus let us consider the regret caused by exploration episodes $m > \beta$. As shown by Lemma 12 in Appendix D each such exploration episode $m > \beta$ contains a *reliable* sub-episode of length at least $\lceil \frac{2^{3m-5}}{m} \rceil$ which employs an optimal policy with probability at least $1 - \frac{1}{3T}$. 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 $m \geqslant \beta$ 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 (13).
- B_m denotes the event that A_m holds 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 $\varepsilon_* := \sqrt{\frac{2S\log(2\theta_*)}{\theta_*}}$, 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 of the empirical MDP will be satisficing with high probability, cf. Appendix B.2. Accordingly, exploration episode m playing UCRL2 will only occur when for the previous exploration

episode $m-1$ we have $\overline{A_{m-1}}$ or B_{m-1}. (Recall that samples that may have been collected in the meantime in exploitation episodes are not used in exploration episodes.) Therefore we have

$$\sum_{m>\beta} \mathbb{E}\big[L_m \Delta^{\sigma}_{\pi_m}\big] = \sum_{m>\beta} \mathbb{P}(\overline{A_{m-1}})\, \mathbb{E}\big[L_m \Delta^{\sigma}_{\pi_m} \big| \overline{A_{m-1}}\big] + \sum_{m>\beta} \mathbb{P}(B_{m-1})\, \mathbb{E}\big[L_m \Delta^{\sigma}_{\pi_m} \big| B_{m-1}\big]. \tag{20}$$

Concerning the first term of (20) we again use the upper bound of $2^{4m-3}AS$ on the length of exploration episode m. Further, by Lemma 14 in Appendix D, the probability of $\overline{A_m}$ is bounded by $(\frac{1}{2})^{(\theta_*+1)2^{m-\beta}-1}$ so that

$$\begin{aligned}\sum_{m>\beta} \mathbb{P}(\overline{A_{m-1}})\, \mathbb{E}\big[L_m \Delta^{\sigma}_{\pi_m} \big| \overline{A_{m-1}}\big] &\leqslant \sum_{m>\beta} (2^{4m-3}AS\Delta^{\sigma,+})(\tfrac{1}{2})^{(\theta_*+1)2^{m-1-\beta}-1} \\ &\leqslant 2^{4\beta+4-\theta_*} AS\Delta^{\sigma,+}. \end{aligned} \tag{21}$$

For an upper bound on the second term of (20), we can apply the same proof technique as for Lemmas 9 and 10 and setting $d(m) = 2^{m-\beta}$ to obtain

$$\begin{aligned}
&\sum_{m>\beta} \mathbb{P}(B_{m-1})\, \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m} | B_{m-1}] \\
&\quad\leqslant \sum_{m>\beta} 2^{4m-3}AS\Delta^{\sigma,+} \sum_{s,a} \mathbb{P}\Big(|r_m(s,a) - \mu(s,a)| \geqslant \varepsilon_* \wedge N_m(s,a) \geqslant 2^{m-\beta}(\theta_*+1)\Big) \\
&\qquad + \sum_{m>\beta} 2^{4m-3}AS\Delta^{\sigma,+} \sum_{s,a} \mathbb{P}\Big(\big\|p_m(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \geqslant \varepsilon_* \wedge N_m(s,a) \geqslant 2^{m-\beta}(\theta_*+1)\Big) \\
&\quad\leqslant \sum_{m>\beta} 2^{4m-3}AS\Delta^{\sigma,+} \sum_{s,a} \sum_{t\geqslant 2^{m-\beta}(\theta_*+1)} \mathbb{P}\big(|\overline{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon_*\big) \\
&\qquad + \sum_{m>\beta} 2^{4m-3}AS\Delta^{\sigma,+} \sum_{s,a} \sum_{t\geqslant 2^{m-\beta}(\theta_*+1)} \mathbb{P}\Big(\big\|\overline{p}_t(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \geqslant \varepsilon_*\Big) \\
&\quad\leqslant \sum_{m>\beta} \frac{2^{4m-2}A^2S\Delta^{\sigma,+}}{2^{d(m)S-S}\,\theta_*^{d(m)S-1}\log(2\theta_*)} < \text{const}\cdot 2^{\beta}A^2S\Delta^{\sigma,+}.
\end{aligned} \tag{22}$$

Collecting all regret terms (17) – (22) and noting that none of them depends on the horizon T completes the proof of the theorem. □

6 Conclusion

While it is satisfactory 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.

Acknowledgments This work was supported by the Austrian Science Fund (FWF): TAI 590-N.

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.

Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. In *Advances in Neural Information Processing Systems 36*, 2022.

Nikhil Balaji, Stefan Kiefer, Petr Novotný, Guillermo A. Pérez, and Mahsa Shirmohammadi. On the complexity of value iteration. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019*, volume 132 of *LIPIcs*, pages 102:1–102:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

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 *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`.

Eugene A. Feinberg and Jefferson Huang. The value iteration algorithm is not strongly polynomial for discounted dynamic programming. *Oper. Res. Lett.*, 42(2):130–131, 2014.

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.

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

Thomas Michel, Hossein Hajiabolhassan, and Ronald Ortner. Regret bounds for satisficing in multi-armed bandit problems. *Transact. Mach. Learn. Res.*, 08 2023.

Ronald Ortner, Odalric-Ambrym Maillard, and Daniil Ryabko. Selecting near-optimal approximate state representations in reinforcement learning. In *Algorithmic Learning Theory - 25th International Conference, ALT 2014, Proceedings*, volume 8776 of *Lecture Notes in Computer Science*, 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, 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 probabilities 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

This section collects results about the error in average reward when working with MDP approximations that have slightly different rewards and transition probabilities.

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 and action space and for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$*

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

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 called 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^*. If $\hat{M}$ is an ε-approximation of M, then for any initial state s_1,*

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

Consider two MDPs M, $\hat{M}$ on the same state and action space and let π be an arbitrary policy that induces an irreducible class $I_\pi \subseteq \mathcal{S}$ on $\hat{M}$. Assume that the definition of ε-approximation holds just for the states of I_π and the actions of π, that is, for all $s \in I_\pi$ we have

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

and

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

Then we call $\hat{M}$ an (ε, I_π)*-approximation* of M. 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, and assume that M is communicating. Suppose that for any $(s', s, a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ if $\hat{p}(s'|s,a) > 0$ then also $p(s'|s,a) > 0$. Let π^* be an optimal policy of $\hat{M}$ inducing a unique irreducible class $\hat{I}_{\pi^*}$. If $\hat{M}$ is an $(\varepsilon, \hat{I}_{\pi^*})$-approximation of M, 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 $M' = (I_{\pi^*}, \{a^*\}, r', p')$ and $\hat{M}' = (I_{\pi^*}, \{a^*\}, \hat{r}', \hat{p}')$ as follows. The state space of both MDPs is I_{π^*} and in each state s there is a unique action $a^* := \pi^*(s)$ available. For any states $s \in \hat{I}_{\pi^*}$ and $s' \in I_{\pi^*}$, the transition probabilities $p'(s'|s,a^*)$ and $\hat{p}'(s'|s,a^*)$ are the same as $p(s'|s,\pi^*(s))$ and $\hat{p}(s'|s,\pi^*(s))$, respectively. Similarly, the rewards $r'(s,a^*)$ and $\hat{r}'(s,a^*)$ 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 $p'(s'|s,a^*)$ and $\hat{p}'(s'|s,a^*)$ are the same as $p(s'|s,\pi^*(s))$. Also the respective rewards $r'(s,a^*)$ and $\hat{r}'(s,a^*)$ are the same as $r(s,\pi^*(s))$ for $s \notin \hat{I}_{\pi^*}$. It is easy to check that M' is communicating with diameter at most $D_{\pi^*}(M)$ and that the average reward of π^* in M' and $\hat{M}'$ coincides with $\rho_{\pi^*}(M, s_1)$ and $\rho_{\pi^*}(M', s_1)$, respectively. As M' only has a single policy π^* this policy is optimal, and since $\hat{M}'$ is an ε-approximation of M', the claim follows by Lemma 2. $\square$

A.3 A Combinatorial Lemma

Let $U_1, U_2, \dots, U_n$ be a sequence of non-empty multi-subsets of a universal set U. For any $x \in U$, let $N_i(x)$ denote the number of occurrences of the element x in sets U_j with $j < i$. Here all occurrences of x in the multi-subsets U_j are counted. We define the *frequency* of a set U_i to be

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

Lemma 4. *Let $U_1, U_2, \dots, U_n$ be a sequence of non-empty multi-subsets of a universal set U. For any non-negative integer f, there are at most $|U|$ members of this sequence that have frequency f.*

Proof. Let $1 \leqslant f_1 < f_2 < \cdots < f_\ell \leqslant n$ be distinct positive integers such that the frequency of each of $U_{f_1}, U_{f_2}, \dots, U_{f_\ell}$ is f. By definition, for any $1 \leqslant j \leqslant \ell$ there exists an $x_{f_j} \in U_{f_j}$ such that $\mathrm{freq}(U_{f_j}) = f = N_{f_j}(x_{f_j})$. Note however that for any two distinct sets U_{f_i}, U_{f_j} with $f_i > f_j$ we have $x_{f_i} \neq x_{f_j}$, since otherwise we would obtain the contradiction

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

This completes the proof. □

A.4 Another Useful Lemma

Lemma 5. *Let c, z be real numbers such that $z \geqslant \lceil 2c\log(4c) \rceil \geqslant \frac{\mathrm{e}}{2}$, where e is Euler's number. Then*

$$\frac{\log(2z)}{z} < \frac{1}{c}\,.$$

Proof. Note that $\frac{\log(2z)}{z}$ is decreasing when $z \geqslant \frac{\mathrm{e}}{2}$. For $z = 2c\log(4c)$ it is straightforward to check that the inequality $\frac{\log(4c\log(4c))}{2c\log(4c)} < \frac{1}{c}$ holds, which completes the proof. □

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 when using the empirical MDP instead of the true one.

B.1 The Empirical MDP in Exploitation Episodes

In this section, we show that in an exploitation episode, the probability of running a not satisficing policy is low, as soon as the frequency freq_k of the episodes k becomes large enough. Let us set

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

Intuitively, ε is the accuracy needed to guarantee that the policy π_k is above σ, while θ will be seen to be the frequency needed to guarantee this accuracy with high probability.

Proposition 2. *If M_k is an (ε, I_{π_k})-approximation of M, then π_k has average reward above σ on M.*

Proof. Setting $c = 2S(D_W+1)^2/(\Delta^{\sigma,-})^2$ we have $\theta = \lceil 2c\log(4c) \rceil$ and by Lemma 5

$$\varepsilon < \sqrt{\frac{2S}{c}} = \frac{\Delta^{\sigma,-}}{(D_W+1)},$$

so that

$$\varepsilon(D_W+1) < \Delta^{\sigma,-} \leqslant \Delta^{\sigma}_{\pi_k}.$$

Accordingly, as soon as M_k is an (ε, I_{π_k})-approximation of M, the policy π_k has average reward above σ on M, as otherwise by Lemma 3 we would get the contradiction

$$\rho(\pi_k, s_k) \geqslant \rho_k(\pi_k, s_k) - \varepsilon(D_W+1) > \sigma - \Delta^{\sigma}_{\pi_k} = \rho(\pi_k, s_k).$$ □

In the sequel, we show that when $\text{freq}_k > \theta$ then with high probability M_k is an (ε, I_{π_k})-approximation of M and hence π_k is satisficing. Let V_k be 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*

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

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

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

Accordingly,

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

□

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

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

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 t samples it holds that

$$\mathbb{P}\big(\|\bar{p}_t(\cdot) - p(\cdot)\|_1 \geqslant \varepsilon\big) \leqslant (2^m - 2)\exp\big(-\tfrac{t\varepsilon^2}{2}\big). \tag{23}$$

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 (23) for any $t \geqslant \theta + 1$,

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

Hence,

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

□

Now, we show that the total probability of choosing a not satisficing policy in exploitation episodes is bounded by a constant, provided that the frequency of each episode is sufficiently large.

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

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

Proof. From Proposition 2 we know that when M_k is an (ε, I_{π_k})-approximation of M, then π_k has average reward above σ on M. Accordingly, if $\rho_{\pi_k}(M_k, s_k) \geqslant \sigma$ and $\rho(\pi_k, s_k) < \sigma$, then there has to be a state $s \in I_{\pi_k}$ for which $(s, \pi_k(s)) \in V_k$. Hence, by Lemmas 6 and 7 we have

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

□

B.2 The Empirical MDP in Exploration Episodes

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

$$\varepsilon_* = \sqrt{\frac{2S\log(2\theta_*)}{\theta_*}}, \text{ where } \theta_* = \left\lceil \frac{4S(D+1)^2}{(\Delta_*^\sigma)^2} \log\Big(\tfrac{8S(D+1)^2}{(\Delta_*^\sigma)^2}\Big) \right\rceil. \tag{24}$$

As we will see below, ε_* is the accuracy needed in order to identify an optimal policy π^* as satisficing. Further, accuracy ε_* will be reached with high probability when the frequency of the respective episode exceeds θ_*.

Proposition 3. *If M_k is an ε_*-approximation of M, then $\rho_k(\pi^*, s_k) > \sigma$.*

Proof. Indeed, setting $c = 2S(D+1)^2/(\Delta^{\sigma,*})^2$ in Lemma 5, one can see that $\varepsilon_*(D+1) < \Delta_*^\sigma$. Consequently, if M_k is an ε_*-approximation of M, then by Lemma 2

$$\rho_k(\pi^*, s_k) \geqslant \rho^* - \varepsilon_*(D+1) = \sigma + \Delta_*^\sigma - \varepsilon_*(D+1) > \sigma.$$

□

Accordingly, as soon as M_k is an ε_*-approximation of M no exploration episode is played anymore (cf. line 14 of the algorithm). In the following, we show that with high probability M_k is indeed an ε_*-approximation of M when $\mathrm{freq}_k > \theta_*$. The following arguments are similar, yet a bit more general than those given in Section B.1 and will later also be needed in the analysis for the general algorithm.

Let 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) in V_k^* 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 $\mathrm{freq}_k \geqslant d\theta_* + 1$, 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

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

□

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

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

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 (23) for any $t \geqslant d\theta_* + 1$

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

□

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

Lemma 11. *If $\rho^* > \sigma$, then*

$$\sum_k \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant d\theta_* + 1\big) \leqslant \frac{2A}{2^{dS-S}\theta_*^{dS-1}\log(2\theta_*)}.$$

Proof. If $\rho_{\pi^*}(M_k, s_k) < \sigma$, then M_k cannot be an ε_*-approximation of M by Proposition 3. Hence, in this case V_k^* cannot be empty and we have by Lemmas 9 and 10,

$$\begin{aligned}
\sum_k \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant d\theta_* + 1\big) &\leqslant \sum_k \mathbb{P}\big(\rho_k(\pi^*, s_k) < \sigma \wedge \mathrm{freq}_k \geqslant d\theta_* + 1\big) \\
&\leqslant \sum_k \mathbb{P}\left(\exists (s,a) \in V_k^* : N_k(s,a) \geqslant d\theta_* + 1\right) \\
&\leqslant \sum_k \sum_{s,a} \mathbb{P}\left(|r_k(s,a) - \mu(s,a)| \geqslant \varepsilon_* \wedge N_k(s,a) \geqslant d\theta_* + 1\right) \\
&\quad + \sum_k \sum_{s,a} \mathbb{P}\left(\big|\big|p_k(\cdot|s,a) - p(\cdot|s,a)\big|\big|_1 \geqslant \varepsilon_* \wedge N_k(s,a) \geqslant d\theta_* + 1\right) \\
&\leqslant \sum_{s,a} \sum_{t \geqslant d\theta_* + 1} \mathbb{P}\left(|\bar{r}_t(s,a) - \mu(s,a)| \geqslant \varepsilon_*\right) + \sum_{s,a} \sum_{t \geqslant d\theta_* + 1} \mathbb{P}\left(\big|\big|\bar{p}_t(\cdot|s,a) - p(\cdot|s,a)\big|\big|_1 \geqslant \varepsilon_*\right) \\
&\leqslant \frac{2A}{S 2^{dS-S} \theta_*^{dS-1} \log(2\theta_*)}
\end{aligned} \tag{25}$$

□

C An Error Bound for Estimating the Diameter and SAT-RL2

In this part of the appendix we present an alternative algorithm that does not resort to the GOSPRL-procedure to estimate the diameter of the underlying MDP in order to guarantee that the empirical MDP is communicating. Instead, we use a result that provides an error bound on how much the diameter in the empirical MDP can deviate from its counterpart in the true MDP.

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

The algorithm SAT-RL2, shown as Algorithm 3, skips the part of SAT-RL which uses the GOSPRL-procedure to estimate the diameter of the underlying MDP (i.e., lines 9–11 in SAT-RL). As already discussed, this is done to guarantee that the empirical MDP is communicating before proceeding. Instead SAT-RL2 just performs an ordinary exploration episode using GOSPRL (lines 15–19) in case the empirical MDP is not communicating (cf. line 11). Theorem 5 derived in the following section will provide a bound on the approximation error for the diameter estimate one obtains from

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** perform *exploitation episode*:
    π_k can be chosen to induce a unique irreducible class I_{π_k}.
    Play π_k until all states in I_{π_k} have been visited at least once.
  else perform *exploration episode*:
    Set $b := b + S$.
    while $N(s,a) < b$ for some state-action pair (s,a) **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

the empirical MDP. This result is of interest in itself and further allows us to bound the number of exploration episodes one has to perform until the empirical MDP becomes communicating.

C.2 Approximation Error for the Empirical Diameter

In this section we derive a bound on the approximation error when estimating the diameter of an MDP M by its counterpart in an ε-approximation of M.

We start with some auxiliary definitions. Let Π be a multi-set consisting of S stationary policies on an MDP M such that for each state $s \in \mathcal{S}$ there exists a unique policy $\pi_s \in \Pi$. Consider an agent starting in some state s following policy $\pi_s \in \Pi$ for a while and then changing 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 $\Pi^{Sem}(M)$. Accordingly, we introduce the semi-diameter, which generalizes the notion of diameter 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 4. *For any MDP M,*

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

As Proposition 4 demonstrates, the notions of semi-stationary policies and semi-diameter do not add anything substantial to the ordinary notions of startionary policy and diameter. However, they are practical for the proof of the following main result of this section.

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, where $\varepsilon < \frac{\ell-2}{\ell(\ell D-1)}$ for some positive integer $\ell \geqslant 3$. Then the diameter of $\hat{M}$ is at most $\ell^2 D - \ell$.*

Proof. Since the diameter of M is D, for any two states s and s' there exists a policy $\pi_{s,s'}$ such that when following $\pi_{s,s'}$ starting in s, we reach s' in at most D steps on average. Hereafter, we call the process of starting in s and following policy $\pi_{s,s'}$ a $\pi_{s,s'}$-exploration. Performing such a $\pi_{s,s'}$-exploration for ℓ steps in M generates a sequence $s_0 s_1 \cdots s_\ell$ with $s_0 = s$ and $p\big(s_i\big|s_{i-1}, \pi_{s,s'}(s_{i-1})\big) > 0$ for $1 \leqslant i \leqslant \ell$.

Now we are going to show that for any $s, s' \in \mathcal{S}$ a $\pi_{s,s'}$-exploration of $\ell D - 1$ steps in $\hat{M}$ visits s' with probability more than $\frac{1}{\ell}$. For any sequence $s_0 s_1 \cdots s_\ell$ let us consider the probabilities

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

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

Further, we set

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

to be the minimal probability of generating the sequence $s_0 s_1 \cdots s_\ell$ 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:

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

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

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

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

Denoting by $\overline{W_M(s,s',\ell)} \subseteq W_M(s,\ell)$ the complement of $W_M(s,s',\ell)$ we have by Markov's inequality for any positive integer ℓ,

$$\mathbb{P}_M^{\pi_{s,s'}}\big(\overline{W_M(s,s',\ell D-1)}\big) \leqslant \frac{1}{\ell}. \tag{26}$$

Since $\hat{M}$ is environmentally an ε-approximation of M, we further have by our assumption on ε

$$\mathbb{P}_{\min}^{\pi_{s,s'}}\big(W_M(s,\ell D-1) \cap W_{\hat{M}}(s,\ell D-1)\big) \geqslant (1-\varepsilon)^{\ell D-1} \geqslant 1-(\ell D-1)\varepsilon > \frac{2}{\ell}. \tag{27}$$

We claim that

$$\mathbb{P}_{\min}^{\pi_{s,s'}}\big(W_M(s,s',\ell D-1) \cap W_{\hat{M}}(s,\ell D-1)\big) > \frac{1}{\ell}. \tag{28}$$

Indeed, otherwise it follows from (27) that $\mathbb{P}_{\min}^{\pi_{s,s'}}\big(\overline{W_M(s,s',\ell D-1)} \cap W_{\hat{M}}(s,\ell D-1)\big) > \frac{1}{\ell}$ and consequently

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

which contradicts (26).

From (28) we can conclude that

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

showing that if we run a $\pi_{s,s'}$-exploration of $\ell D-1$ steps in $\hat{M}$, then s' will be visited with probability more than $\frac{1}{\ell}$. Now let us 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 $\ell D-1$ steps in $\hat{M}$. If after at most $\ell D-1$ steps, the state s' has not been reached then for the current state s'' follow the policy $\pi_{s'',s'}$ for another $\ell D-1$ steps. Iterate this procedure until s' is reached. In view of the expectation of the geometric distribution, the expected number of necessary iterations is ℓ and hence the expected number of steps until s' is visited is at most $\ell(\ell D-1)$. This holds for any pair of states s, s', so that the semi-diameter of $\hat{M}$ is bounded by $\ell^2 D - \ell$ and the theorem follows by Proposition 4. □

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 $\rho^* > \sigma$, then the expected σ-regret of* SAT-RL2 *after any number of steps is bounded by*

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

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. The main difference is that one additionally has to take into account how many exploration episodes have to be performed until the empirical MDP is communicating with high probability. The respective number of steps in these episodes can be bounded using Theorem 5 as follows. Choosing $\ell = 3$ in Theorem 5 shows that if the empirical MDP M_k is an ε'-approximation of M with $\varepsilon' < \frac{1}{9D}$, then M_k has finite diameter, i.e., is communicating. Accordingly, it is sufficient if we set $\varepsilon' = \frac{\varepsilon_*}{9}$ because this implies (cf. the proof of Proposition 2)

$$9\varepsilon' D < \varepsilon_*(D+1) < \Delta_*^\sigma \leqslant 1,$$

whence $\varepsilon' < \frac{1}{9D}$.

On the other hand, for a suitable constant $c > 0$,

$$\varepsilon' = \frac{1}{9}\sqrt{\frac{2S\log(2\theta_*)}{\theta_*}} \geqslant \sqrt{\frac{2S\log(c\theta_*)}{c\theta_*}},$$

so that if we replace θ_* by $c\theta_*$ in the derivations in Section B.2, we can show that when the frequency is at least $c\theta^*$ then accuracy ε' is achieved with high probability. In particular, an equivalent of Lemma 11 for $d = 1$ holds stating that

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

Accordingly, as in each run of an exploration episode the frequency increases by S, after $\lceil \frac{c\theta_*}{S} \rceil$ exploration episodes the empirical MDP is an ε'-approximation of M with probability at least $1 - \frac{2A}{(c\theta_*)^{S-1}\log(2\theta_*)}$. This will only cause an additional factor of c in the regret term of the exploration part, so that the claimed bound holds. □

D Details for the Proof of Theorem 4

We start looking at some of the parameters we are using which are related to episodes that are sufficiently long in order to guarantee optimality or visits in all states of the irreducible class.

D.1 Parameter θ_M

Beside the bound on the expected regret of UCRL2 in Theorem 1, Jaksch et al. [2010] also show the following high probability bound.

Theorem 7. *The regret of UCRL2 run with confidence parameter δ is bounded by*

$$34 \cdot DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}$$

for all T with probability at least $1 - \delta$.

This bound also implies that when a sub-episode of UCRL2 has sufficient length, the policy used in this sub-episode has to be optimal. Indeed, by Theorem 7 the per-step regret after T steps is $\frac{34 \cdot DS\sqrt{A\log(T/\delta)}}{\sqrt{T}}$ with high probability, so that when T is sufficiently large the per-step regret is below Δ_g. Thus, let θ_M be the smallest positive integer T such that

$$\frac{34 \cdot DS\sqrt{A\log(T/\delta)}}{\sqrt{T}} < \Delta_g.$$

Then by Theorem 7 with probability at least $1 - \frac{1}{3T}$ any sub-episode of length $\geqslant \theta_M$ in an exploration episode of SAT-UCRL will play an optimal policy. If $\rho^* \geqslant \sigma$, then the optimal policy is satisficing and the same argument shows that any policy played by SAT-UCRL for at least θ_M steps in a UCRL2-subepisode must be satisficing with high probability. This is used to show the following lemma.

Lemma 12. *With probability at least $1 - \frac{1}{3T}$, 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 $2^{4m-3}AS$. By Proposition 1, one can conclude that the number of sub-episodes in episode m is at most $4mAS$. (We note that while Proposition 1 assumes that T is the total number of steps, the claim also holds for any T consecutive steps starting at some sub-epsiode.)

It follows that there is a sub-episode of length at least $\frac{2^{3m-5}}{m}$, and accordingly, if $\frac{2^{3m-5}}{m} \geqslant \theta_M$, then the policy played in this sub-episode is optimal with an overall error probability of at most $\frac{1}{3T}$. □

D.2 Parameter θ'_M

Given a Markov chain C with S states, the expected number of steps it takes to visit each state at least ℓ times is known as the ℓ-cover time of C, denoted by $\tau_\ell(C)$. Chan et al. [2021] have shown that the ℓ-cover time of an irreducible Markov chain C is at most $(\mathrm{e}^2 + \mathrm{e}\log(S))\big(\tau_1(C) + \frac{\ell}{\zeta_C}\big)$, where e is Euler's number and ζ_C is the minimum stationary probability of a single state.

In our MDP setting, for any optimal policy π with a unique irreducible class I_π we consider the induced irreducible Markov chain M_π restricted to states in I_π. In accordance with the result of Chan et al. [2021] we set $\tau_\pi = (\mathrm{e}^2 + \mathrm{e}\log(S))\big(\tau_1(M_\pi) + \frac{1}{\zeta_{M_\pi}}\big)$ and note that $\ell\tau_\pi$ is an upper bound for $\tau_\ell(M_\pi)$. By Markov's inequality, any random walk of length $2\ell\tau_\pi$ starting in the irreducible class I_π will visit each state at least ℓ times with probability at least $\frac{1}{2}$. On the other hand, the irreducible class can be reached in at most D_W steps on average. In our case we are interested in the number of steps needed to visit all states in the irreducible class of an optimal policy and set

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

We summarize our observations in the following lemma.

Lemma 13. *Let π^* be an optimal policy that induces a unique irreducible class I_{π^*} on an MDP. Then following π^* for $\ell\ell'\theta'_M$ steps will visit each state in I_{π^*} at least ℓ times with probability at least $1 - (\frac{1}{2})^{\ell'-1}$.*

Proof. Following π^* will reach the irreducible class with a probability of $1 - (\frac{1}{2})^{\ell\ell'}$ within the first $2\ell\ell' D_W$ steps. After reaching I_{π^*}, in the remaining $\geqslant 2\ell\tau_{\pi^*}$ steps each state in I_{π^*} will be visited at least ℓ times with a probability of at least $1 - (\frac{1}{2})^{\ell'}$. □

D.3 Bounding $\mathbb{P}(\overline{A_m})$

In the following we use the definition of θ_* in (24) of Section B.2.

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

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

Proof. Since $\left\lceil \frac{2^{3m-5}}{m} \right\rceil \geqslant 4\left\lceil \frac{2^{3(m-1)-5}}{m-1} \right\rceil$ for $m > \beta \geqslant 4$, we have by definition of β that $\left\lceil \frac{2^{3m-5}}{m} \right\rceil \geqslant 4^{m-\beta}\left\lceil \frac{2^{3\beta-5}}{\beta} \right\rceil \geqslant (2^{m-\beta}(\theta_*+1))^2\,\theta'_M$. Choosing $\ell = \ell' = 2^{m-\beta}(\theta_*+1)$ in Lemma 13 then proves the claim. □