informs
https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *Mathematics of Operations Research*

ISSN 0364-765X, EISSN 1526-5471

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
https://hhaji.github.io/MyWeb/

Ronald Ortner
Lehrstuhl für Informationstechnologie
Montanuniversität Leoben, Austria
rortner@unileoben.ac.at
https://ortner.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 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 [7, 19] 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 the expected regret with respect to the optimal policy is bounded as for the UCRL2 algorithm [15].

Key words: reinforcement learning, Markov decision process, regret

MSC2000 subject classification: Primary: 68Q32; secondary: 68T05, 90C40

OR/MS subject classification: Primary: Computer science: Artificial intelligence; secondary: Analysis of algorithms: Suboptimal algorithms; Probability: Markov processes

History:

1. Introduction

Work in reinforcement learning (RL) typically follows the optimization paradigm. That is, the problem setting is designed as a Markov decision process (MDP), and the goal of the learner is to acquire an optimal policy in this underlying MDP. Historically, this can be explained easily, as the MDP setting is well investigated, so that reinforcement learning was able to resort to the rich theory on MDPs developed by the operations research community. On

the other hand, when taking a closer look, many reinforcement learning problems are about learning to perform a special task for which there are in general plenty of good ways and no best way actually stands out. Typical things humans do in their everyday life are of that kind, e.g., go to work, wash the dishes, mow the lawn, shop a list of items, etc. Even for tasks where algorithms already do quite well, such as game playing or autonomous driving, one rather compares to the level humans reach and is usually happy with respective suboptimal performance. Thus playing a game like chess better than most humans is considered to be sufficient, just as driving a car as good as a human. In many of these tasks it even is difficult to make up an optimality criterion, or it is not feasible computing a respective optimal policy. Accordingly, the concept of satisficing, where one seeks a solution that is just good enough, makes a lot of sense in the context of reinforcement learning.

In this paper we want to investigate the question, whether there is an advantage in pursuing this more modest goal of satisficing when doing reinforcement learning in an MDP under the average reward criterion. Instead of aiming at optimal average reward, 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 briefly discuss related work, define the setting, and recall some preliminaries in Section 2. Then in Section 3 we present our algorithm for the case when the optimal average reward ρ^* of the underlying MDP (which we assume to be communicating) is above σ. In Section 4 this algorithm is shown to learn with only constant regret (i.e., independent of the number of steps) with respect to the desired level σ. This result hightlights the alleviation when considering satisficing instead of optimization, as classic regret bounds for reinforcement learning usually grow with the number of considered time steps (cf. e.g. [15]). Moreover, if only the optimal policy is above σ, this entails even constant regret with respect to ρ^*, which generalizes a result of [7] from the bandit to the general MDP setting. Section 5 presents an alternative algorithm as well as a bound on the approximation error for estimating the diameter –a key quanitity– of an MDP.

In Section 6, 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 $\rho^* \leq \sigma$, the same algorithm can be shown to enjoy bounds on the classic regret (i.e., with respect to ρ^*) just like the well-known UCRL2 algorithm [15].

1.1. Related Work While satisficing objectives have been considered before, most respective investigations have been made in the much simpler bandit setting [1, 24, 19]. Here [19] is closest to our approach and shows constant regret with respect to the satisfaction level σ if the latter is reachable. Otherwise, combining the suggested algorithm with UCB one can achieve logarithmic bounds on the classic pseudo-regret similar as those known for UCB [5]. A more general notion of regret which also considers the learner's degree of belief in a Bayesian setting is considered in [24], which beside various upper bounds also provides lower bounds, cf. also the respective discussion in [19].

While [19, 24] employ a notion of *pseudo*-regret with respect to σ, [1] suggests to consider the actual rewards and introduces a second-level reward of 1, if the actual reward is above a sufficiency level, which need not be constant. The given regret bounds are logarithmic in the horizon and depend on the reward distributions of the single arms. For more discussion of related work in the setting of multi-armed bandits (such as conservative and thresholding bandits) we refer to the detailed related work section of [19].

For the general MDP setting there is little work on satisficing. A satisficing variant of Q-learning is presented in [14], for which however only some experimental evaluation is given. The problem of finding a satisficing policy in robust MDPs is considered in [25], which provides an alternative to the usual worst-case analysis in this setting.

A notion of satisficing regret is introduced in [4] for an episodic and Bayesian RL setting. As one may not have access to enough information needed to determine an optimal policy, the goal is to consider an MDP approximation that allows to compute a satisficing policy. Satisficing in this context is not defined with respect to a fixed level as in our case but with respect to the currently available information. This is formalized using rate distortion theory, and the satisficing regret basically corresponds to the respective approximation error. The paper proposes a generalization of the PSRL algorithm [28] and gives bounds on the classic (Bayesian) regret that generalize respective results for bandits [26, 2, 3] and in a special case correspond to those available for PSRL [21].

More relevant to our setting is work on constrained MDPs (CMDPs). Typically, the goal in this setting is to optimize some reward function as in standard MDPs, but on the other hand satisfy some cost constraint for an additional cost function. Concerning the latter one can consider the respective regret with respect to constraint violations, which basically corresponds to our regret with respect to a satisfaction level σ.

Most work on RL in CMDPs considers the discounted [11] or the episodic setting [23, 16, 18, 10]. However, there is some relevant work considering the average reward setting as we do. Usually, some strong assumptions are made, such as that the transition probabilities are known to the learner [31], when it is easy to guarantee that constraints are met. Ergodic CMDPs are considered in [27], however the regret from constraint violations after T steps is of order $T^{2/3}$ and thus quite large.

The state of the art is [8], whose algorithm on ergodic MDPs has regret (with respect to the optimal average reward) of order $\sqrt{T}$ and a constant violation of the cost constraint. For weakly communicating MDPs, bounds of order $T^{2/3}$ are shown for regret as well as cost constraint violation. Both bounds are high probability bounds, but the considered reward and cost functions are assumed to be deterministic.

Our setting can be considered to only take into account the cost constraint in a CMDP. That is, the optimization of the rewards is trivial (e.g., assuming that all rewards are known to be 0). The respective regret bounds we give are in expectation, however unlike in [8] our costs are random, when no constant high probability bounds on the regret are possible (cf. the discussion in Section 1.2.1 below). The constant constraint violation regret of [8] corresponds to our Theorem 2 below, however we achieve it not only for ergodic but more generally for communicating MDPs.

1.2. 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 $r(s,a)$ 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 [15] 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 $T(s'|M,\pi,s)$ 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

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

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 $\rho_\pi(M, s_1) := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(r_t^{\pi,s_1})$,

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. Moreover, considering nonstationary policies does not increase the optimal average reward [22]. 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 related transition parameter.

DEFINITION 2. For any stationary policy π we set

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

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.

1.2.1. 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

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

We 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. [15].

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

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

The σ-regret generalizes similar notions such as satisficing regret from the bandit to the MDP setting, see e.g. [19]. We will also consider the classic regret with respect to ρ^* after any T steps, defined as in [15] as

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

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

In the end we will be interested in bounds on the *expected* regret. We note that when operating on MDPs with random rewards it is not possible to obtain constant regret, as the sum of T accumulated rewards in general will deviate from the mean by a term of order $\sqrt{T}$ (see e.g. Theorem 13.1 of [17], which specifies this for the bandit setting).

Beside the gaps Δ^σ_{π,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 define the gaps $\Delta^\sigma_* := \rho^* - \sigma$ between the optimal average reward and σ, and $\Delta_g := \rho^* - \max_{\pi:\rho_\pi<\rho^*} \rho_\pi$ between the optimal average reward and the average reward of the best suboptimal policy. As before, in the definition of Δ_g it is sufficient to consider policies π with state independent average reward ρ_π.

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 [15] 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. [15]**). *The number of episodes of* UCRL2 *up to step $T \geq AS$ is upper bounded by*

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

[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'. (For a definition of *irreducible class* see footnote 3 below.)

More importantly, for UCRL2 one can give bounds on the classic online regret as defined in (1). We will use the following two bounds, a high probability bound as well as a gap-dependent bound on the expected regret.

THEOREM 1 **(Jaksch et al. [15]).** *With probability at least $1-\delta$, for all T the regret of UCRL2 run with confidence parameter δ is bounded by*

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

Moreover, the expected regret of UCRL2 *run with confidence parameter $\delta = \frac{1}{3T}$ is bounded by*

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

where T_π is the smallest natural number such that for all $T \geq T_\pi$ 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 [29] 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 in [29], 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, [29] provides another algorithm (based on GOSPRL), which computes an approximation of the diameter of the underlying MDP. This algorithm, which we call GOSPRL-Diam in the following, 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 it holds that $D \leq \bar{D} \leq (1+2\varepsilon_g(1+\varepsilon_g))(1+\varepsilon_g)D$.

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 satisficing policy when given a satisfaction level σ, provided that $\rho^* > \sigma$.

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 GOSPRL-Diam is run to estimate the MDP's diameter

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 chosen $\bar{b}(s,a)$ times **do**
6: Run GOSPRL($\bar{b}, \delta_g$).
7: For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, set $\bar{b}(s,a) := \max\{0, b - N(s,a)\}$.
8: **end while**
9: **while** the diameter of the estimated MDP M_0 is infinite **do**
10: Run GOSPRL-Diam(δ_g, ε_g) 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** $\rho_{\pi_k}(M_k, s_k) \geq \sigma$ **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 $\bar{b}(s,a) := \max\{0, b - N(s,a)\}$.
20: Run GOSPRL($\bar{b}, \delta_g$).
21: **end while**
22: **end if**
23: **end for**

(lines 9–11). This is only done to guarantee that the diameter of the empirical MDP M_0 is finite, that is, M_0 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[2] and can be assumed to have a unique irreducible class[3] I_{π_k}.

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 below σ, 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 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 [17]). The suggested value iteration usually works well in practice, however does not have polynomial time guarantees, cf. [12, 6] for respective results in the discounted case.

Before we proceed to analyze SAT-RL, we note that in Appendix C we present an alternative algorithm called SAT-RL2. Instead of running GOSPRL-Diam to estimate the diameter, SAT-RL2 uses a result about the diameter of MDP approximations (Theorem 3), which might be of interest in itself. Further details can be found in Section 5 below.

4. Regret Bound for SAT-RL

In this section we provide a proof of the following bound on the σ-regret of SAT-RL. For the sake of readability, some technical details have been deferred to 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.

The second term in the bound of Theorem 2 is small and hence negligible, unless A is very large compared to the other terms. The appearance of Δ_*^σ and $\Delta^{\sigma,-}$ in the denominator of the first and the third term is similar as in the upper bounds derived in [19] for satisficing in the bandit setting.

[2] Using value iteration (VI) introduces an approximation error ε, which however can easily be controlled: The error ε of the current VI iteration is bounded by the span of the difference of two successive value vectors. In practice, it would even be sufficient to stop VI as soon as it can be guaranteed that the current policy π satisfies $\rho_\pi - \varepsilon > \sigma$.

[3] We recall that the irreducible classes of a Markov chain are the equivalence classes with respect to the relation of being communicating, where two states are communicating if one can reached from the other with positive probability after a finite number of $t \geq 0$ steps.

Indeed, writing ρ_i for the mean reward of arm i and $\Delta_i^\sigma := |\sigma - \rho_i|$ for the gap to σ, [19] provides a simple lower bound of order $\sum_{i:\rho_i<\sigma} 1/\Delta_i^\sigma$ on the σ-regret for bandits. It is straightforward to adapt the respective example (originally suggested in [7]) to the MDP setting to obtain a preliminary lower bound of order $AS/\Delta^{\sigma,-}$ for the σ-regret in MDPs, cf. Proposition 2 below. Considering the available lower bounds for reinforcement learning [15] it also seems intuitive that the diameter has to appear. A construction similar to the one used for the proof of Theorem 5 in [15] looks promising to get an additional factor of D in the lower bound of Proposition 2. On the other hand, we think that the factor of $\Delta^{\sigma,+}/\Delta^{\sigma,-}$ in the third term of Theorem 2 is an artifact of the analysis. Similarly, it seems plausible that the parameter D_W should have a smaller exponent and in general should be replaced by the smaller diameter D.

Concerning the first term in Theorem 2, it is an open question whether the dependence on Δ_*^σ is necessary. While it is plausible that the problem becomes harder when the optimal arm/policy is close to σ, a respective lower bound is also still missing in the bandit setting. The quadratic dependence as in Theorem 2 seems to be another artifact of our proof. The upper bounds of [19] only have a dependence of order $1/\Delta_*^\sigma$.

PROPOSITION 2. *For all natural numbers $S \geq 1$, $A > 1$ and each RL algorithm $\mathfrak{A}$ there is a communicating MDP with S states and A actions in which $\mathfrak{A}$ suffers expected σ-regret of order $AS/\Delta^{\sigma,-}$.*

Proof. Consider a bandit problem with $S(A-1)$ arms where each suboptimal arm has distance Δ to the satisfaction level σ and the unique optimal arm is the only sufficient arm with average reward above σ. Generalizing Theorem 5 of [7] to more than two arms, the expected regret of any bandit algorithm can be shown to be at least of order AS/Δ for this problem. It remains to construct an MDP with S states and A actions that corresponds to this bandit setting. This can be easily accomplished by distributing the $S(A-1)$ arms uniformly over S states. Then in each state there are $A-1$ arms available, which corresponds to an MDP with S states and $A-1$ actions (with the average rewards defined as in the original bandit setting).

If the learner at each step were allowed to pick a state and an action directly this obviously corresponds to the original bandit problem. In order to obtain a communicating MDP we have to define suitable transition probabilities. First, for each action a we set $p(s|s,a) = 1$ for all states s and accordingly $p(s'|s,a) = 0$ for $s \neq s'$. In order to connect the states we arrange them in a cycle $s_1, s_2, \ldots, s_S$ where we add in each state s_i a special action that gives 0 reward and deterministically

leads to state s_{i+1} ($1 \le i \le S$ with $s_{S+1} := s_1$). The arising MDP has S states and A actions, and $\Delta^{\sigma,-} = \Delta$. Due to the necessary transitions between states learning in the MDP is harder than in the original bandit problem, so that the regret is again at least of order $AS/\Delta^{\sigma,-}$. □

Before we finally give a proof, we note that Theorem 2 immediately implies the following generalization of a result of [7] from the bandit to the general MDP setting.

COROLLARY 1. *When the learner is given a reference value σ with $\rho^* > \sigma > \rho^* - \Delta_g$, then the expected regret (with respect to ρ^*) of* SAT-RL *with satisfaction level σ is bounded by a constant.*

Proof. As the chosen satisfaction level σ is able to distinguish any suboptimal policy from an optimal one, reaching the level σ is equivalent to playing an optimal policy. Accordingly, an episode of SAT-RL will only contribute to the regret (with respect to ρ^*) if it also has nonzero σ-regret. Therefore, the expected regret of SAT-RL is bounded by a constant. □

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 expected σ-regret after n episodes can be bounded by

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

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

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

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

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

Intuitively, the higher the frequency the more samples have been gathered for state-action pairs occurring in the current episode. Accordingly, a higher frequency implies higher precision of the empirical model used.

Upper Bound on Exploitation Regret By Theorem 4.8 of [9], the cover 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, following 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.

Further, $\Delta^{\sigma}_{\pi_k}$ can be upper bounded by $\Delta^{\sigma,+}$, so that

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

It remains to analyze the term

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

where we choose

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

Intuitively, θ is the number of samples needed for each state-action pair so that the policy π_k is satisficing with high probability, cf. Appendix B.1 for details. More precisely, Lemma 8 in Appendix B.1 shows that we can bound (5) as

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

Concerning (4), first note that after the initialization phase $\mathrm{freq}_k \geq S+1$ for all episodes k. Further, Lemma 4 in Appendix A.3 shows that for any natural number f with $S+1 \leq f \leq \theta$, there are at most AS episodes k for which $\mathrm{freq}_k = f$. As there are precisely $\theta - S$ such numbers f, we can bound (4) as

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) \geq \sigma \wedge \rho(\pi_k, s_k) < \sigma \wedge \mathrm{freq}_k \leq \theta\big) \leq AS(\theta - S) \leq \tilde{O}\left(\frac{AS^2 D_W^2}{(\Delta^{\sigma,-})^2}\right). \tag{8}$$

Taking together eqs. (3) – (7) we obtain

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

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, and $\frac{1}{\delta_g}$. 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 GOSPRL-Diam(δ_g, ε_g) 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

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

Upper Bound on Exploration Regret Analogously to the analysis of R_{Exploit}, we first bound

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

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

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

where we set

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

Similar to the definition of θ in (6), θ_* is the number of samples needed to identify the optimal policy in the underlying MDP as satisficing in the empirical MDP M_k, cf. Appendix B.2 for details. In particular, Lemma 11 in Appendix B.2 shows that (12) is bounded as

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \ge \theta_* + 1\big) \le \frac{2A}{\theta_*^{S-1} \log(2\theta_*)}. \tag{14}$$

Further, 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,

$$\sum_{k=1}^{n} \mathbb{P}\big(\rho_k(\pi_k, s_k) < \sigma \wedge \text{freq}_k \le \theta_*\big) \le \left\lceil \frac{\theta_*}{S} \right\rceil. \tag{15}$$

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

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

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

5. An Error Bound for Estimating the Diameter and SAT-RL2

As already mentioned, SAT-RL employs GOSPRL-Diam to estimate the diameter of the underlying MDP (in lines 9–11 of SAT-RL) in order to guarantee that the empirical MDP is communicating before proceeding. Here we discuss an alternative approach to achieve that. In fact, instead one can simply perform an

ordinary exploration episode using GOSPRL (as in lines 17–21 of SAT-RL), in case the empirical MDP is not communicating. The respective complete algorithm, which we call SAT-RL2, is shown as Algorithm 3 in Appendix C.1.

For the regret analysis of SAT-RL2 one has to consider the steps in these additional exploration episodes, until the empirical MDP becomes communicating. The following theorem provides a bound on the approximation error for the diameter estimate one obtains from the empirical MDP. This result is of interest in itself and might be useful in other contexts.

THEOREM 3. *Let $M = (\mathcal{S}, \mathcal{A}, r, p)$ be a communicating MDP with diameter D and consider another MDP $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{r}, \hat{p})$ over the same state-action space such that*[4] *for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$*

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

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

A proof of Theorem 3 is given in Appendix C.2. Appendix C.3 provides a bound on the expected regret of SAT-RL2 that is of the same order as the bound given in Theorem 2 for SAT-RL.

6. 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 those given in Theorem 1 for UCRL2. In the following, we present the algorithm SAT-UCRL, which exactly achieves that: When σ is below ρ^*, we have constant σ-regret just as for SAT-RL. If however $\rho^* \leq \sigma$, we show bounds on the expected classic regret with the same dependency on T as those given in Theorem 1. These results generalize the findings of [19] from the bandit to the MDP setting.

6.1. 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 (internal) 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. As by definition of UCRL2

[4] The following condition means that $\hat{M}$ is *environmentally an ε-approximation* of M, as defined in Appendix A.2.

Algorithm 2 SAT-UCRL for the general RL setting

1: **Input:** state space $\mathcal{S}$, action space $\mathcal{A}$, satisfaction level σ, horizon T
2: **Initialization:**
3: Set confidence level $\delta_g := \frac{1}{2}$ and accuracy level $\varepsilon_g := \frac{1}{2}$.
4: Set initial sampling number $b := S + 1$ and $b_u := \frac{AS}{8}$.
5: Define function $\bar{b} : \mathcal{S} \times \mathcal{A} \to \mathbb{N}$ to be $\bar{b}(s, a) = b$ for any (s, a).
6: **while** an action $a \in \mathcal{A}$ at some state $s \in \mathcal{S}$ has not been chosen $\bar{b}(s, a)$ times **do**
7: Run GOSPRL($\bar{b}, \delta_g$).
8: For any $(s, a) \in \mathcal{S} \times \mathcal{A}$, set $\bar{b}(s, a) := \max\{0, b - N(s, a)\}$.
9: **end while**
10: **while** the diameter of the estimated MDP M_0 is infinite **do**
11: Run GOSPRL-Diam(δ_g, ε_g) to estimate the diameter of M.
12: **end while**
13: **for** episodes $k = 1, 2, \ldots$ **do**
14: Compute an optimal policy π_k on M_k that induces a unique irreducible class I_{π_k}.
15: **if** $\rho_{\pi_k}(M_k, s_k) \geq \sigma$ **then** perform *exploitation episode*:
16: Play π_k until all states in I_{π_k} have been visited at least once.
17: **else** perform *exploration episode* using UCRL2 with confidence parameter $\delta = \frac{1}{3T}$:
18: Set $b_u := 8b_u$.
19: **while** the length of the current episode is below b_u **do**
20: Run a **sub-episode** of UCRL2.
21: **end while**
22: **end if**
23: **end for**

these sub-episodes increase with time, also the length of our exploration episodes is chosen to be increasing (cf. line 18). That way, in an exploration episode of SAT-UCRL in general several sub-episodes of UCRL2 are run, which guarantees that different policies are employed.

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. In practice, it would of course speed up convergence and hence improve the algorithm, if UCRL2 would use these samples as well. Concerning computational complexity, similar to SAT-UCRL the computationally most

elaborate 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, cf. [15]. Still, the computation can be done in polynomial time as shown in Section 38.5.2 of [17].

6.2. Regret Bounds for SAT-UCRL Now we present the promised bounds on the (σ-) regret of SAT-UCRL. We start with the bound on the standard online regret, when σ cannot be attained by any policy.

THEOREM 4. *Let $\sigma \geq \rho^*$. For any initial state s_1 and any $T > 1$, the expected regret of* SAT-UCRL *with respect to ρ^* is bounded by*

$$1 + 34DS\sqrt{AT\log(3T^2)} + \tilde{O}\left(\frac{AS^2D_W^3}{(\Delta^{\sigma,-})^2}\right),$$

as well as

$$\frac{34^2AS^2D^2\log(T)}{\Delta_g} + \sum_{a,s}\left[1 + \log_2(\max_{\pi:\pi(s)=a} T_\pi)\right]\max_{\pi:\pi(s)=a} T_\pi + \tilde{O}\left(\frac{AS^2D_W^3}{(\Delta^{\sigma,-})^2}\right),$$

where T_π is as defined in Theorem 1. 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 initialization phase can be bounded just as in (10) in the proof of Theorem 2. Similarly, the regret accumulated in exploitation episodes can be analyzed just 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

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

This term 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 [15] also works when the initial state of an episode does not coincide with the last visited state of the previous episode but is 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. Note that as in Corollary 1, this implies constant regret with respect to ρ^* in case $\sigma > \rho^* - \Delta_g$.

THEOREM 5. *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. This 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. Writing $b_u^{(m)}$ for the value of b_u in exploration episode m, by lines 4 and 18 of the algorithm, it holds that $b_u^{(m)} = 2^{3(m-1)} AS$. This also provides a lower bound on the number of steps in episode m (cf. line 19 of the algorithm). In order to obtain an upper bound, we note that by the (sub-)episode termination criterion[5] of UCRL2, a sub-episode of UCRL2 starting at some step t has length at most t. (Here we only consider steps in exploration episodes.) Therefore, the maximal length $\ell_+^{(m)}$ of exploration episode m is upper bounded by $\sum_{i=1}^{m-1} 2\ell_+^{(i)} + 2b_u^{(m)}$ and a simple proof by induction on m shows that $\ell_+^{(m)} \le 3b_u^{(m)} < 2^{3m-1} AS$.

Now we distinguish between long and short exploration episodes. That is, defining θ_M, θ'_M as in Appendix D, and θ_* as in (13), we set β to be the smallest positive integer for which $\frac{2^{3\beta-5}}{\beta} \ge \max\{\theta_M, (\theta_*+1)^2 \theta'_M\}$. 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 $2^{3m-1} AS$, we can bound the first term by

$$\sum_{m=1}^{\beta} \mathbb{E}[L_m \Delta^{\sigma}_{\pi_m}] \le \sum_{m=1}^{\beta} 2^{3m-1} AS \Delta^{\sigma,+} \le \frac{2^{3\beta+2}}{7} AS \Delta^{\sigma,+}. \tag{19}$$

[5] A (sub-)episode of UCRL2 terminates when the number of visits to some state-action pair has doubled [15].

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 $m \geq \beta$ 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 $2^{3m-1}AS$ 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

$$\begin{aligned}\sum_{m>\beta} \mathbb{P}(\overline{A_{m-1}})\, \mathbb{E}\big[L_m \Delta^\sigma_{\pi_m} \big| \overline{A_{m-1}}\big] &\leq \sum_{m>\beta} (2^{3m-1}AS\Delta^{\sigma,+})(\tfrac{1}{2})^{(\theta_*+1)2^{m-1-\beta}-1} \\ &\leq 2^{3\beta+5-\theta_*} AS\Delta^{\sigma,+}. \end{aligned} \tag{21}$$

For an upper bound on the second term of (20), we apply the same proof technique as for Lemmas 9, 10, and 11 in Appendix B. That is, we set $d(m) = 2^{m-\beta}$ and obtain, recalling that A_{m-1} holds in case of B_{m-1},

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

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

7. Conclusion Let us briefly summarize our findings, cf. also Table 1 for a quick overview. When $\rho^* > \sigma$ we have constant σ-regret (Theorems 2 and 5). Moreover, if $\rho^* > \sigma > \rho^* - \Delta_g$ then even the expected regret with respect to ρ^* is constant (Corollary 1). Otherwise, when $\sigma \leq \rho^* - \Delta_g$, our algorithms may exploit a suboptimal policy, so that the regret with respect to ρ^* will be linear. However, we note that it is in general not possible to have a consistent algorithm which at the same time has constant σ-regret and sublinear regret with respect to ρ^*. This already holds in the bandit setting, where any consistent algorithm will keep playing each suboptimal arm (see e.g. Section 16 of [17]), hence accumulating σ-regret.

When $\rho^* < \sigma$, the σ-regret necessarily will be linear, but it is possible to obtain bounds on the expected regret with respect to ρ^* as for UCRL2 (Theorem 4). For the case $\rho^* = \sigma$ we however cannot guarantee constant σ-regret. Although this seems to be possible, a different approach would

TABLE 1. Dependence of expected (σ-) regret on T.

Regime	σ-regret	regret wrt ρ^*
$\rho^* > \sigma > \rho^* - \Delta_g$	constant*	constant*
$\sigma \leq \rho^* - \Delta_g$	constant*	linear$^{(*)}$
$\rho^* < \sigma$	linear*	as for UCRL2* [15]
$\rho^* = \sigma$	linear	as for UCRL2* [15]

The table shows the dependence of the upper bounds for SAT-UCRL on the expected (σ-)regret on the horizon T. A star indicates that the bound is optimal with respect to T. When $\sigma \leq \rho^* - \Delta_g$, then no consistent algorithm can have constant σ-regret and sublinear regret with respect to ρ^*.

be necessary to achieve it. We note that in the bandit setting, [19] points out an alternative algorithm adapted from [13] that achieves constant σ-regret in the bandit setting when σ coincides with the optimal reward.

While we achieve best possible bounds with respect to the horizon T in all cases except when $\sigma = \rho^*$, we did not try to optimize the other parameters in the constant regret bounds, and they are unlikely to be optimal. Moreover, for the distance Δ_*^σ of the optimal policy to the level σ, it is unclear even in the bandit setting whether an appearance is necessary, that is, respective lower bounds are still missing.

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.

Appendix 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*

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

and

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

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 $a \in \mathcal{A}$,

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

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

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

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. [20]**). *Let M be a communicating MDP with optimal policy π^* satisfying the Bellman equations. If $\hat{M}$ is an ε-approximation of M, then for any initial state s_1,*

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

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

$$\left|\hat{r}(s,\pi(s)) - r(s,\pi(s))\right| < \varepsilon.$$

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,*

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

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 also has 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 satisfies the Bellman equations. Since $\hat{M}'$ is an ε-approximation of M', the claim follows by Lemma 2. □

A.3. A Combinatorial Lemma Let $U_1, U_2, \ldots, U_n$ 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 $U_1, U_2, \ldots, U_n$ 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 $1 \le f_1 < f_2 < \cdots < f_\ell \le n$ be distinct positive integers, such that the frequency of each of $U_{f_1}, U_{f_2}, \ldots, U_{f_\ell}$ is f. By definition, for any $1 \le j \le \ell$ 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

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

This completes the proof. □

A.4. Another Useful Lemma

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

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

Proof. Note that $\frac{\log(2z)}{z}$ is decreasing when $z \ge \frac{e}{2}$. 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. □

Appendix B: Analysis of SAT-RL This appendix collects results concering the quality of MDP approximations that are 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 non-satisficing policy is low, as soon as the frequency freq_k of the episodes k becomes large enough. Let us set

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

Intuitively, ε is the accuracy needed to guarantee that the policy π_k is above σ (cf. Proposition 3 below), while θ will be seen to be a sufficient frequency to achieve this accuracy with high probability (cf. Lemmas 6 and 7).

PROPOSITION 3. *If M_k is an (ε, I_{π_k})-approximation of M, then π_k has average reward above σ in M.*

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

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

so that

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

Accordingly, as soon as M_k is an (ε, I_{π_k})-approximation of M, we have by Lemma 3 that

$$\rho(\pi_k, s_k) \ge \rho_k(\pi_k, s_k) - \varepsilon(D_W+1) > \sigma - \Delta^{\sigma}_{\pi_k}. \tag{23}$$

Now if π_k had average reward below σ, then $\Delta^{\sigma}_{\pi_k} > 0$ and we would obtain from (23) the contradiction

$$\rho(\pi_k, s_k) > \sigma - \Delta^{\sigma}_{\pi_k} = \rho(\pi_k, s_k).$$

This shows that $\Delta^{\sigma}_{\pi_k} = 0$ and (23) implies that policy π_k has average reward above σ on M. □

In Lemmas 6 and 7 below, we show that when $\mathrm{freq}_k > \theta$, then with high probability M_k is an (ε, I_{π_k})-approximation of M. This implies that 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 $\sum_{s'\in\mathcal{S}} |p_k(s'|s,\pi_k(s)) - p(s'|s,\pi_k(s))| \ge \varepsilon$ or $|r_k(s,\pi_k(s)) - r(s,\pi_k(s))| \ge \varepsilon$.

Recall that $\bar{r}_t(s,a)$ and $\bar{p}_t(s'\big|s,a)$ 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*

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

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

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

Accordingly,

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

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

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

Proof. In [30], it is shown that for the L^1-deviation of the true distribution and the empirical distribution over m distinct events from t samples, it holds that

$$\mathbb{P}\big(\|\bar{p}_t(\cdot) - p(\cdot)\|_1 \geq \varepsilon\big) \leq (2^m - 2)\exp\big(-\tfrac{t\varepsilon^2}{2}\big). \tag{24}$$

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 (24) for any $t \geq \theta+1$,

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

Hence,

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

Now, we are ready to show that the total probability of choosing a non-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*

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

Proof. From Proposition 3 we know that when M_k is an (ε, I_{π_k})-approximation of M, then π_k has average reward above σ in M. Accordingly, if $\rho_{\pi_k}(M_k,s_k)\geq\sigma$ 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

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

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

$$\varepsilon_*=\sqrt{\frac{2S\log(2\theta_*)}{\theta_*}},\ \text{where}\ \theta_*=\left\lceil\frac{4S(D+1)^2}{(\Delta_*^\sigma)^2}\log\Big(\tfrac{8S(D+1)^2}{(\Delta_*^\sigma)^2}\Big)\right\rceil. \tag{25}$$

Proposition 4 below shows that accuracy ε_* is sufficient in order to identify an optimal policy π^* as satisficing. Further, Lemmas 9 and 10 will show that accuracy ε_* will be reached with high probability, when the frequency of the respective episode exceeds θ_*.

PROPOSITION 4. *If M_k is an ε_*-approximation of M, then $\rho_k(\pi^*, s_k) > \sigma$.*

Proof. 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

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

Accordingly, as soon as M_k is an ε_*-approximation of M, no exploration episode is played anymore (cf. line 14 of the algorithm). Subsequently, Lemmas 9 and 10 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 of 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 $\sum_{s' \in \mathcal{S}} |p_k(s'|s,a) - p(s'|s,a)| \geq \varepsilon_*$ or $|r_k(s,a) - r(s,a)| \geq \varepsilon_*$.

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

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

Proof. If $\mathrm{freq}_k \geq d\theta_* + 1$, 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

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

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

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

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 (24) for any $t \geq d\theta_* + 1$,

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

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

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

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

Proof. If $\rho_{\pi^*}(M_k, s_k) < \sigma$, then M_k cannot be an ε_*-approximation of M by Proposition 4. Hence, in this case V_k^* cannot be empty, and we have by Lemmas 9 and 10,

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

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

1: **Input:** state space $\mathcal{S}$, action space $\mathcal{A}$, satisfaction level σ
2: **Initialization:**
3: Set confidence level $\delta_g := \frac{1}{2}$, 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 chosen $\bar{b}(s,a)$ times **do**
6: Run GOSPRL($\bar{b}, \delta_g$).
7: For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, define $\bar{b}(s,a) := \max\{0, b - N(s,a)\}$.
8: **end while**
9: **for** episodes $k = 1, 2, \ldots$ **do**
10: Compute an optimal policy π_k on M_k that induces a unique irreducible class I_{π_k}.
11: **if** M_k is communicating and $\rho_{\pi_k}(M_k, s_k) \geq \sigma$ **then** perform *exploitation episode*:
12: Play π_k until all states in I_{π_k} have been visited at least once.
13: **else** perform *exploration episode*:
14: Set $b := b + S$.
15: **while** $N(s,a) < b$ for some state-action pair (s,a) **do**
16: For any (s,a), set $\bar{b}(s,a) := b - N(s,a)$.
17: Run GOSPRL($\bar{b}, \delta_g$).
18: **end while**
19: **end if**
20: **end for**

Appendix 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 GOSPRL-Diam 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 GOSPRL-Diam to estimate the diameter of the underlying MDP (i.e., lines 9–11 in SAT-RL). As already discussed, GOSPRL-Diam is employed 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 3 derived in the following section provides a bound on the approximation error for the diameter estimate one obtains from the empirical MDP. This result 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 give a proof of Theorem 3, 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 $T(s'|M, \pi^+, s)$ 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

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

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

PROPOSITION 5. *For any MDP M,*

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

As Proposition 5 demonstrates, the notions of semi-stationary policies and semi-diameter do not add anything substantial to the ordinary notions of stationary policy and diameter. However, they are practical for the proof of Theorem 3, which we restate here for the sake of readability.

THEOREM 3. *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 $\ell \geq 3$. 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 $1 \le i \le \ell$.

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 higher than $\frac{1}{\ell}$. For any sequence $s_0 s_1 \cdots s_\ell$, let us consider the probabilities

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

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,

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

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,

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

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

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

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 would follow 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

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

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 higher 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 5. □

C.3. Regret Bound for SAT-RL2 As for SAT-RL we have a constant bound on the σ-regret of SAT-RL2. The bound differs from Theorem 2 only in the constants.

THEOREM 6. *If $\rho^* > \sigma$, then the expected σ-regret of* SAT-RL2 *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 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 3 as follows. Choosing $\ell = 3$ in Theorem 3 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 4)

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

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

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

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

Hence, we can replace θ_* by $c\theta_*$ in the derivations in Section B.2, and 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

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

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. □

Appendix D: Details for the Proof of Theorem 5 We start looking at the two parameters θ_M, θ'_M. These are related to episodes which are sufficiently long to guarantee optimality or visits in all states of the irreducible class, respectively.

D.1. Parameter θ_M The regret bounds of Theorem 1 also imply that when a sub-episode of UCRL2 has sufficient length, the policy used in this sub-episode has to be optimal. Indeed, by Theorem 1 the per-step regret after T steps is $\frac{34 \cdot DS\sqrt{A\log(T/\delta)}}{\sqrt{T}}$ with high probability. Accordingly, 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 1 with probability at least $1-\frac{1}{3T}$, any sub-episode of length $\geq \theta_M$ in an exploration episode of SAT-UCRL will use an optimal policy. If $\rho^* \geq \sigma$, then the optimal policy is satisficing. Therefore, 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 $m > \beta \geq 4$ contains a reliable sub-episode.*

Proof. We have already seen in the proof of Theorem 5 that the number of steps of the m-th exploration episode is at least $2^{3m-3}AS$ and at most $2^{3m-1}AS$. By Proposition 1, one can conclude that the number of sub-episodes in episode m is at most $(3m+2)AS \leq 4mAS$ for $m > \beta \geq 4$. (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 the beginning of some sub-epsiode.)

It follows that there is a sub-episode of length at least $\frac{2^{3m-5}}{m}$. Further, if $\frac{2^{3m-5}}{m} \geq \theta_M$, 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 $\tau_\ell(C)$. We have already noted that by Theorem 4.8 of [9] it holds that $\tau_1(C) \leq D(C)(1+\log(S))$, so that $\tau_\ell(C) \leq \ell D(C)(1+\log(S))$.

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. Then by Markov's inequality, any random walk of length $2\ell D(M_\pi)(1+\log(S))$ 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

$$\theta'_M = \max_{\pi:\rho(\pi)=\rho^*} 2D(M_\pi)(1+\log(S)) + 2D_W.$$

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 $2\ell\ell' D(M_{\pi_*})(1+\log(S))$ 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 (25) of Section B.2.

LEMMA 14. *For any $m \geq \beta \geq 4$,*

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

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

Acknowledgments. This work was supported by the Austrian Science Fund (FWF): TAI 590-N. We would like to thank the anonymous reviewers for their valuable comments, which greatly helped to improve the paper.

References

[1] Abernethy JD, Amin K, Zhu R (2016) Threshold bandits, with and without censored feedback. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS 2016*, 4889–4897.

[2] Arumugam D, Roy BV (2021) Deciding what to learn: A rate-distortion approach. *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, 373–382 (PMLR).

[3] Arumugam D, Roy BV (2021) The value of information when deciding what to learn. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 9816–9827.

[4] Arumugam D, Roy BV (2022) Deciding what to model: Value-equivalent sampling for reinforcement learning. *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 9024–9044.

[5] Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47(2-3):235–256.

[6] Balaji N, Kiefer S, Novotný P, Pérez GA, Shirmohammadi M (2019) On the complexity of value iteration. *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019*, volume 132 of *LIPIcs*, 102:1–102:15 (Schloss Dagstuhl - Leibniz-Zentrum für Informatik).

[7] Bubeck S, Perchet V, Rigollet P (2013) Bounded regret in stochastic multi-armed bandits. *COLT 2013 – The 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, 122–134 (PMLR).

[8] Chen L, Jain R, Luo H (2022) Learning infinite-horizon average-reward Markov decision process with constraints. *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, 3246–3270 (PMLR).

[9] Dabbs B (2009) Markov chains and mixing times. *University of Chicago VIGRE REU* 1–20, URL https://math.uchicago.edu/~may/VIGRE/VIGRE2009/REUPapers/Dabbs.pdf.

[10] Ding D, Wei X, Yang Z, Wang Z, Jovanovic MR (2021) Provably efficient safe exploration via primal-dual policy optimization. *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021*, volume 130 of *Proceedings of Machine Learning Research*, 3304–3312 (PMLR).

[11] Ding D, Zhang K, Basar T, Jovanovic MR (2020) Natural policy gradient primal-dual method for constrained Markov decision processes. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 8378–8390.

[12] Feinberg EA, Huang J (2014) The value iteration algorithm is not strongly polynomial for discounted dynamic programming. *Oper. Res. Lett.* 42(2):130–131.

[13] Garivier A, Ménard P, Stoltz G (2019) Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.* 44(2):377–399.

[14] Goodrich MA, Quigley M (2004) Satisficing Q-learning: Efficient learning in problems with dichotomous attributes. *Proceedings of the 2004 International Conference on Machine Learning and Applications – ICMLA 2004*, 65–72 (IEEE Computer Society).

[15] Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.* 11:1563–1600.

[16] Kalagarla KC, Jain R, Nuzzo P (2021) A sample-efficient algorithm for episodic finite-horizon MDP with constraints. *35th AAAI Conference on Artificial Intelligence, AAAI 2021, 33rd Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The 11th Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*, 8030–8037 (AAAI Press).

[17] Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press).

[18] Liu T, Zhou R, Kalathil D, Kumar PR, Tian C (2021) Learning policies with zero or bounded constraint violation for constrained MDPs. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 17183–17193.

[19] Michel T, Hajiabolhassan H, Ortner R (2023) Regret bounds for satisficing in multi-armed bandit problems. *Transact. Mach. Learn. Res.* 08/2023.

[20] Ortner R, Maillard O, Ryabko D (2014) Selecting near-optimal approximate state representations in reinforcement learning. *Algorithmic Learning Theory – 25th International Conference, ALT 2014, Proceedings*, volume 8776 of *Lecture Notes in Computer Science*, 140–154 (Springer).

[21] Osband I, Roy BV (2017) Why is posterior sampling better than optimism for reinforcement learning? *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, 2701–2710 (PMLR).

[22] Puterman ML (2005) *Markov decision processes: Discrete stochastic dynamic programming* (John Wiley & Sons).

[23] Qiu S, Wei X, Yang Z, Ye J, Wang Z (2020) Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 15277–15287.

[24] Reverdy P, Srivastava V, Leonard NE (2017) Satisficing in multi-armed bandit problems. *IEEE Trans. Autom. Control.* 62(8):3788–3803.

[25] Ruan H, Zhou S, Chen Z, Ho CP (2023) Robust satisficing MDPs. *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, 29232–29258 (PMLR).

[26] Russo D, Van Roy B (2022) Satisficing in time-sensitive bandit learning. *Math. Oper. Res.* 47(4):2815–2839.

[27] Singh R, Gupta A, Shroff NB (2023) Learning in constrained Markov decision processes. *IEEE Trans. Control Netw. Syst.* 10(1):441–453.

[28] Strens MJA (2000) A Bayesian framework for reinforcement learning. *Proceedings of the 17th International Conference on Machine Learning, ICML 2000*, 943–950 (Morgan Kaufmann).

[29] Tarbouriech J, Pirotta M, Valko M, Lazaric A (2021) A provably efficient sample collection strategy for reinforcement learning. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 7611–7624.

[30] Weissman T, Ordentlich E, Seroussi G, Verdu S, Weinberger MJ (2003) Inequalities for the L1 deviation of the empirical distribution. *Information Theory Research Group HP Laboratories* URL https://www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf.

[31] Zheng L, Ratliff LJ (2020) Constrained upper confidence reinforcement learning. *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control, L4DC 2020*, volume 120 of *Proceedings of Machine Learning Research*, 620–629 (PMLR).