Journal of Machine Learning Research 11 (2010) 1563-1600 Submitted 7/09; Revised 11/09; Published 4/10

Near-optimal Regret Bounds for Reinforcement Learning[*]

Thomas Jaksch TJAKSCH@UNILEOBEN.AC.AT
Ronald Ortner RORTNER@UNILEOBEN.AC.AT
Peter Auer AUER@UNILEOBEN.AC.AT
Chair for Information Technology
University of Leoben
Franz-Josef-Strasse 18
8700 Leoben, Austria

Editor: Sham Kakade

Abstract

For undiscounted reinforcement learning in Markov decision processes (MDPs) we consider the *total regret* of a learning algorithm with respect to an optimal policy. In order to describe the transition structure of an MDP we propose a new parameter: An MDP has *diameter D* if for any pair of states s,s' there is a policy which moves from s to s' in at most D steps (on average). We present a reinforcement learning algorithm with total regret $\tilde{O}(DS\sqrt{AT})$ after T steps for any unknown MDP with S states, A actions per state, and diameter D. A corresponding lower bound of $\Omega(\sqrt{DSAT})$ on the total regret of any learning algorithm is given as well.

These results are complemented by a sample complexity bound on the number of suboptimal steps taken by our algorithm. This bound can be used to achieve a (gap-dependent) regret bound that is logarithmic in T.

Finally, we also consider a setting where the MDP is allowed to change a fixed number of ℓ times. We present a modification of our algorithm that is able to deal with this setting and show a regret bound of $\tilde{O}(\ell^{1/3}T^{2/3}DS\sqrt{A})$.

Keywords: undiscounted reinforcement learning, Markov decision process, regret, online learning, sample complexity

1. Introduction

In a Markov decision process (MDP) M with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$, a learner in some state $s \in \mathcal{S}$ needs to choose an action $a \in \mathcal{A}$. When executing action a in state s, the learner receives a random reward r drawn independently from some distribution on $[0,1]$ with mean $\bar{r}(s,a)$. Further, according to the transition probabilities $p(s'|s,a)$, a random transition to a state $s' \in \mathcal{S}$ occurs.

Reinforcement learning of MDPs is a standard model for learning with delayed feedback. In contrast to important other work on reinforcement learning—where the performance of the *learned* policy is considered (see, e.g., Sutton and Barto 1998, Kearns and Singh 1999, and also the discussion and references given in the introduction of Kearns and Singh 2002)—we are interested in the performance of the learning algorithm *during learning*. For that, we compare the rewards collected by the algorithm during learning with the rewards of an optimal policy.

[*] An extended abstract of this paper appeared in *Advances in Neural Information Processing Systems* 21 (2009), pp. 89–96.

An algorithm $\mathfrak{A}$ starting in an initial state s of an MDP M chooses at each time step t (possibly randomly) an action a_t. As the MDP is assumed to be unknown except the sets $\mathcal{S}$ and $\mathcal{A}$, usually an algorithm will map the history up to step t to an action a_t or, more generally, to a probability distribution over $\mathcal{A}$. Thus, an MDP M and an algorithm $\mathfrak{A}$ operating on M with initial state s constitute a stochastic process described by the states s_t visited at time step t, the actions a_t chosen by $\mathfrak{A}$ at step t, and the rewards r_t obtained ($t \in \mathbb{N}$). In this paper we will consider *undiscounted* rewards. Thus, the *accumulated reward* of an algorithm $\mathfrak{A}$ after T steps in an MDP M with initial state s, defined as

$$R(M,\mathfrak{A},s,T) := \sum_{t=1}^{T} r_t,$$

is a random variable with respect to the mentioned stochastic process. The value $\frac{1}{T}\,\mathbb{E}\left[R(M,\mathfrak{A},s,T)\right]$ then is the expected average reward of the process up to step T. The limit

$$\rho(M,\mathfrak{A},s) := \lim_{T\to\infty} \tfrac{1}{T}\,\mathbb{E}\left[R(M,\mathfrak{A},s,T)\right]$$

is called the *average reward* and can be maximized by an appropriate stationary *policy* $\pi : \mathcal{S} \to \mathcal{A}$ which determines an optimal action for each state (see Puterman, 1994). Thus, in what follows we will usually consider policies to be stationary.

The difficulty of learning an optimal policy in an MDP does not only depend on the MDP's size (given by the number of states and actions), but also on its transition structure. In order to measure this transition structure we propose a new parameter, the *diameter* D of an MDP. The diameter D is the time it takes to move from any state s to any other state s', using an appropriate policy for each pair of states s, s':

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*

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

In Appendix A we show that the diameter is at least $\log_{|\mathcal{A}|}|\mathcal{S}| - 3$. On the other hand, depending on the existence of states that are hard to reach under any policy, the diameter may be arbitrarily large. (For a comparison of the diameter to other mixing time parameters see below.)

In any case, a finite diameter seems necessary for interesting bounds on the *regret* of any algorithm with respect to an optimal policy. When a learner explores suboptimal actions, this may take him into a "bad part" of the MDP from which it may take up to D steps to reach again a "good part" of the MDP. Thus, compared to the simpler multi-armed bandit problem where each arm a is typically explored $\frac{\log T}{g}$ times (depending on the gap g between the optimal reward and the reward for arm a)—see, for example, the regret bounds of Auer et al. (2002a) for the UCB algorithms and the lower bound of Mannor and Tsitsiklis (2004)—the best one would expect for the general MDP setting is a regret bound of $\Theta\big(D|\mathcal{S}||\mathcal{A}|\log T\big)$. The alternative gap-independent regret bounds of $\tilde{O}(\sqrt{|\mathcal{B}|T})$ and $\Omega(\sqrt{|\mathcal{B}|T})$ for multi-armed bandits with $|\mathcal{B}|$ arms (Auer et al., 2002b) correspondingly translate into a regret bound of $\Theta(\sqrt{D|\mathcal{S}||\mathcal{A}|T})$ for MDPs with diameter D.

For MDPs with finite diameter (which usually are called *communicating*, see, e.g., Puterman 1994) the optimal average reward ρ^* does not depend on the initial state (cf. Puterman 1994, Section 8.3.3), and we set

$$\rho^*(M) := \rho^*(M,s) := \max_{\pi} \rho(M,\pi,s).$$

The optimal average reward is the natural benchmark[1] for a learning algorithm $\mathfrak{A}$, and we define the *total regret* of $\mathfrak{A}$ after T steps as

$$\Delta(M,\mathfrak{A},s,T) := T\rho^*(M) - R(M,\mathfrak{A},s,T).$$

In the following, we present our reinforcement learning algorithm UCRL2 (a variant of the UCRL algorithm of Auer and Ortner, 2007) which uses upper confidence bounds to choose an optimistic policy. We show that the total regret of UCRL2 after T steps is $\tilde{O}(D|\mathcal{S}|\sqrt{|\mathcal{A}|T})$. A corresponding lower bound of $\Omega(\sqrt{D|\mathcal{S}||\mathcal{A}|T})$ on the total regret of any learning algorithm is given as well. These results establish the diameter as an important parameter of an MDP. Unlike other parameters that have been proposed for various PAC and regret bounds, such as the *mixing time* (Kearns and Singh, 2002; Brafman and Tennenholtz, 2002) or the *hitting time* of an optimal policy (Tewari and Bartlett, 2008) (cf. the discussion below) the diameter only depends on the MDP's transition structure.

1.1 Relation to Previous Work

We first compare our results to the PAC bounds for the well-known algorithms E^3 of Kearns and Singh (2002), and R-Max of Brafman and Tennenholtz (2002) (see also Kakade, 2003). These algorithms achieve ε-optimal average reward with probability $1-\delta$ after time polynomial in $\frac{1}{\delta}$, $\frac{1}{\varepsilon}$, $|\mathcal{S}|$, $|\mathcal{A}|$, and the mixing time $T_\varepsilon^{\text{mix}}$ (see below). As the polynomial dependence on ε is of order $\frac{1}{\varepsilon^3}$, the PAC bounds translate into $T^{2/3}$ regret bounds at the best. Moreover, both algorithms need the ε-*return mixing time* $T_\varepsilon^{\text{mix}}$ of an optimal policy π^* as input parameter.[2] This parameter $T_\varepsilon^{\text{mix}}$ is the number of steps until the average reward of π^* over these $T_\varepsilon^{\text{mix}}$ steps is ε-close to the optimal average reward ρ^*. It is easy to construct MDPs of diameter D with $T_\varepsilon^{\text{mix}} \approx \frac{D}{\varepsilon}$. This additional dependence on ε further increases the exponent in the above mentioned regret bounds for E^3 and R-max. Also, the exponents of the parameters $|\mathcal{S}|$ and $|\mathcal{A}|$ in the PAC bounds of Kearns and Singh (2002) and Brafman and Tennenholtz (2002) are substantially larger than in our bound. However, there are algorithms with better dependence on these parameters. Thus, in the sample complexity bounds for the Delayed Q-Learning algorithm of Strehl et al. (2006) the dependence on states and actions is of order $|\mathcal{S}||\mathcal{A}|$, however at the cost of a worse dependence of order $\frac{1}{\varepsilon^4}$ on ε.

The MBIE algorithm of Strehl and Littman (2005, 2008)—similarly to our approach—applies confidence bounds to compute an optimistic policy. However, Strehl and Littman consider only a discounted reward setting. Their definition of regret measures the difference between the rewards[3] of an optimal policy and the rewards of the learning algorithm *along the trajectory taken by the learning algorithm*. In contrast, we are interested in the regret of the learning algorithm in respect to the rewards of the optimal policy *along the trajectory of the optimal policy*.[4] Generally, in discounted reinforcement learning only a finite number of steps is relevant, depending on the discount

[1] It can be shown that $\max_{\mathfrak{A}} \mathbb{E}[R(M,\mathfrak{A},s,T)] = T\rho^*(M) + O(D(M))$ and $\max_{\mathfrak{A}} R(M,\mathfrak{A},s,T) = T\rho^*(M) + \tilde{O}(\sqrt{T})$ with high probability.

[2] The knowledge of this parameter can be eliminated by guessing $T_\varepsilon^{\text{mix}}$ to be $1,2,\ldots$, so that sooner or later the correct $T_\varepsilon^{\text{mix}}$ will be reached (cf. Kearns and Singh 2002; Brafman and Tennenholtz 2002). However, since there is no condition on when to stop increasing $T_\varepsilon^{\text{mix}}$, the assumed mixing time eventually becomes arbitrarily large, so that the PAC bounds become exponential in the true $T_\varepsilon^{\text{mix}}$ (cf. Brafman and Tennenholtz, 2002).

[3] Actually, the state values.

[4] Indeed, one can construct MDPs for which these two notions of regret differ significantly. E.g., set the discount factor $\gamma = 0$. Then any policy which maximizes immediate rewards achieves 0 regret in the notion of Strehl and Littman. But such a policy may not move to states where the optimal reward is obtained.

factor. This makes discounted reinforcement learning similar to the setting with trials of constant length from a fixed initial state as considered by Fiechter (1994). For this case logarithmic online regret bounds in the number of trials have already been given by Auer and Ortner (2005). Also, the notion of regret is less natural than in undiscounted reinforcement learning: when summing up the regret in the individual visited states to obtain the total regret in the discounted setting, somehow contrary to the principal idea of discounting, the regret at each time step counts the same.

Tewari and Bartlett (2008) propose a generalization of the *index policies* of Burnetas and Katehakis (1997). These index policies choose actions optimistically by using confidence bounds only for the estimates in the current state. The regret bounds for the *index policies* of Burnetas and Katehakis (1997) and the OLP algorithm of Tewari and Bartlett (2008) are *asymptotically* logarithmic in T. However, unlike our bounds, these bounds depend on the gap between the "quality" of the best and the second best action, and these asymptotic bounds also hide an additive term which is exponential in the number of states. Actually, it is possible to prove a corresponding gap-dependent logarithmic bound for our UCRL2 algorithm as well (cf. Theorem 4 below). This bound holds uniformly over time and under weaker assumptions: While Tewari and Bartlett (2008) and Burnetas and Katehakis (1997) consider only *ergodic* MDPs in which *any* policy will reach every state after a sufficient number of steps, we make only the more natural assumption of a finite diameter.

Recently, Bartlett and Tewari (2009) have introduced the REGAL algorithm (inspired by our UCRL2 algorithm) and show—based on the methods we introduce in this paper—regret bounds where the diameter is replaced with a smaller transition parameter D_1 (that is basically an upper bound on the span of the *bias* of an optimal policy). Moreover, this bound also allows the MDP to have some *transient* states that are not reachable under any policy. However, the bound holds only when the learner knows an upper bound on this parameter D_1. In case the learner has no such upper bound, a doubling trick can be applied, but then the bound's dependence on $|\mathcal{S}|$ deteriorates from $|\mathcal{S}|$ to $|\mathcal{S}|^{3/2}$. Bartlett and Tewari (2009) also modify our lower bound example to obtain a lower bound of $\Omega(D_1\sqrt{|\mathcal{S}||\mathcal{A}|T})$ with respect to their new transition parameter D_1. Still, in the given example $D_1 = \sqrt{D}$, so that in this case their lower bound matches our lower bound.

2. Results

We summarize the results achieved for our algorithm UCRL2 (which will be described in the next section), and also state a corresponding lower bound. We assume an unknown MDP M to be learned, with $S := |\mathcal{S}|$ states, $A := |\mathcal{A}|$ actions, and finite diameter $D := D(M)$. Only $\mathcal{S}$ and $\mathcal{A}$ are known to the learner, and UCRL2 is run with confidence parameter δ.

Theorem 2 *With probability of at least* $1-\delta$ *it holds that for any initial state* $s \in \mathcal{S}$ *and any* $T > 1$*, the regret of* UCRL2 *is bounded by*

$$\Delta(M, \text{UCRL2}, s, T) \ \leq\ 34 \cdot DS\sqrt{AT \log\left(\tfrac{T}{\delta}\right)}.$$

It is straightforward to obtain from Theorem 2 the following sample complexity bound.

Corollary 3 *With probability of at least* $1-\delta$ *the average per-step regret of* UCRL2 *is at most* ε *for any*

$$T \ \geq\ 4 \cdot 34^2 \cdot \frac{D^2S^2A}{\varepsilon^2} \log\left(\frac{34DSA}{\delta\varepsilon}\right)$$

steps.

It is also possible to give a sample complexity bound on the number of suboptimal steps UCRL2 takes, which allows to derive the following gap-dependent logarithmic bound on the expected regret.

Theorem 4 *For any initial state $s \in \mathcal{S}$, any $T \geq 1$ and any $\varepsilon > 0$, with probability of at least $1-3\delta$ the regret of* UCRL2 *is*

$$\Delta(M, \text{UCRL2}, s, T) \;\leq\; 34^2 \cdot \frac{D^2 S^2 A \log\left(\frac{T}{\delta}\right)}{\varepsilon} + \varepsilon T.$$

Moreover setting

$$g := \rho^*(M) - \max_{s\in\mathcal{S}} \max_{\pi:\mathcal{S}\to\mathcal{A}} \left\{\rho(M,\pi,s) : \rho(M,\pi,s) < \rho^*(M)\right\}$$

to be the gap in average reward between best and second best policy in M, the expected regret of UCRL2 *(with parameter $\delta := \frac{1}{3T}$) for any initial state $s \in \mathcal{S}$ is*

$$\mathbb{E}\left[\Delta(M, \text{UCRL2}, s, T)\right] \;<\; 34^2 \cdot \frac{D^2 S^2 A \log(T)}{g} + 1 + \sum_{s,a} \left\lceil 1 + \log_2\big(\max_{\pi:\pi(s)=a} T_\pi\big)\right\rceil \max_{\pi:\pi(s)=a} T_\pi,$$

where T_π is the smallest natural number such that for all $T \geq T_\pi$ the expected average reward after T steps is $\frac{g}{2}$-close to the average reward of π. Using the doubling trick to set the parameter δ, one obtains a corresponding bound (with larger constant) without knowledge of the horizon T.

These new bounds are improvements over the bounds that have been achieved by Auer and Ortner (2007) for the original UCRL algorithm in various respects: the exponents of the relevant parameters have been decreased considerably, the parameter D we use here is substantially smaller than the corresponding mixing time of Auer and Ortner (2007), and finally, the ergodicity assumption is replaced by the much weaker and more natural assumption that the MDP has finite diameter.

The following is an accompanying lower bound on the expected regret.

Theorem 5 *For any algorithm $\mathfrak{A}$, any natural numbers $S, A \geq 10$, $D \geq 20 \log_A S$, and $T \geq DSA$, there is an MDP M with S states, A actions, and diameter D,*[5] *such that for any initial state $s \in \mathcal{S}$ the expected regret of $\mathfrak{A}$ after T steps is*

$$\mathbb{E}\left[\Delta(M, \mathfrak{A}, s, T)\right] \geq 0.015 \cdot \sqrt{DSAT}\;.$$

Finally, we consider a modification of UCRL2 that is also able to deal with changing MDPs.

Theorem 6 *Assume that the MDP (i.e., its transition probabilities and reward distributions) is allowed to change $(\ell - 1)$ times up to step T, such that the diameter is always at most D. Restarting* UCRL2 *with parameter $\frac{\delta}{\ell^2}$ at steps $\left\lceil \frac{i^3}{\ell^2} \right\rceil$ for $i = 1, 2, 3 \ldots$, the regret (now measured as the sum of missed rewards compared to the ℓ optimal policies in the periods during which the MDP remains constant) is upper bounded by*

$$65 \cdot \ell^{1/3} T^{2/3} D S \sqrt{A \log\left(\frac{T}{\delta}\right)}$$

with probability of at least $1-\delta$.

[5] As already mentioned, the diameter of any MDP with S states and A actions is at least $\log_A S - 3$.

For the simpler multi-armed bandit problem, similar settings have already been considered by Auer et al. (2002b), and more recently by Garivier and Moulines (2008), and Yu and Mannor (2009). The achieved regret bounds are $O(\sqrt{\ell T \log T})$ in the first two mentioned papers, while Yu and Mannor (2009) derive regret bounds of $O(\ell \log T)$ for a setting with side observations on past rewards in which the number of changes ℓ need not be known in advance.

MDPs with a different model of changing rewards have already been considered by Even-Dar et al. (2005) and Even-Dar et al. (2009), respectively. There, the transition probabilities are assumed to be fixed and known to the learner, but the rewards are allowed to change at every step (however, independently of the history). In this setting, an upper bound of $O(\sqrt{T})$ on the regret against an optimal stationary policy (with the reward changes known in advance) is best possible and has been derived by Even-Dar et al. (2005). This setting recently has been further investigated by Yu et al. (2009), who also show that for achieving sublinear regret it is essential that the changing rewards are chosen obliviously, as an opponent who chooses the rewards depending on the learner's history may inflict linear loss on the learner. It should be noted that although the definition of regret in the nonstochastic setting looks the same as in the stochastic setting, there is an important difference to notice. While in the stochastic setting the average reward of an MDP is always maximized by a stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, in the nonstochastic setting obviously a dynamic policy adapted to the reward sequence would in general earn more than a stationary policy. However, obviously no algorithm will be able to compete with the best dynamic policy for all possible reward sequences, so that—similar to the nonstochastic bandit problem, compare to Auer et al. (2002b)—one usually competes only with a finite set of experts, in the case of MDPs the set of stationary policies $\pi : \mathcal{S} \to \mathcal{A}$. For different notions of regret in the nonstochastic MDP setting see Yu et al. (2009).

Note that all our results scale linearly with the rewards. That is, if the rewards are not bounded in $[0,1]$ but taken from some interval $[r_{\min}, r_{\max}]$, the rewards can simply be normalized, so that the given regret bounds hold with additional factor $(r_{\max} - r_{\min})$.

3. The UCRL2 Algorithm

Our algorithm is a variant of the UCRL algorithm of Auer and Ortner (2007). As its predecessor, UCRL2 implements the paradigm of "optimism in the face of uncertainty". That is, it defines a set $\mathcal{M}$ of statistically *plausible* MDPs given the observations so far, and chooses an optimistic MDP $\tilde{M}$ (with respect to the achievable average reward) among these plausible MDPs. Then it executes a policy $\tilde{\pi}$ which is (nearly) optimal for the optimistic MDP $\tilde{M}$. More precisely, UCRL2 (see Figure 1) proceeds in episodes and computes a new policy $\tilde{\pi}_k$ only at the beginning of each episode k. The lengths of the episodes are not fixed a priori, but depend on the observations made. In Steps 2–3, UCRL2 computes estimates $\hat{r}_k(s,a)$ and $\hat{p}_k(s'|s,a)$ for the mean rewards and the transition probabilities from the observations made before episode k. In Step 4, a set $\mathcal{M}_k$ of plausible MDPs is defined in terms of confidence regions around the estimated mean rewards $\hat{r}_k(s,a)$ and transition probabilities $\hat{p}_k(s'|s,a)$. This guarantees that with high probability the true MDP M is in $\mathcal{M}_k$. In Step 5, *extended value iteration* (see below) is used to choose a near-optimal policy $\tilde{\pi}_k$ on an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$. This policy $\tilde{\pi}_k$ is executed throughout episode k (Step 6). Episode k ends when a state s is visited in which the action $a = \tilde{\pi}_k(s)$ induced by the current policy has been chosen *in* episode k equally often as *before* episode k. Thus, the total number of occurrences of any state-action pair is at most doubled during an episode. The counts $v_k(s,a)$ keep track of these occurrences in episode k.[6]

3.1 Extended Value Iteration: Finding Optimistic Model and Optimal Policy

In Step 5 of the UCRL2 algorithm we need to find a near-optimal policy $\tilde{\pi}_k$ for an optimistic MDP $\tilde{M}_k$. While value iteration typically calculates an optimal policy for a fixed MDP, we also need to select an optimistic MDP $\tilde{M}_k$ that gives almost maximal optimal average reward among all plausible MDPs.

3.1.1 PROBLEM FORMULATION

We can formulate this as a general problem as follows. Let $\mathcal{M}$ be the set of all MDPs with (common) state space $\mathcal{S}$, (common) action space $\mathcal{A}$, transition probabilities $\tilde{p}\,(\cdot|s,a)$, and mean rewards $\tilde{r}\,(s,a)$ such that

$$\begin{aligned} \|\tilde{p}\,(\cdot|s,a) - \hat{p}(\cdot|s,a)\|_1 &\leq d(s,a), & (1)\\ |\tilde{r}\,(s,a) - \hat{r}(s,a)| &\leq d'(s,a) & (2) \end{aligned}$$

for given probability distributions $\hat{p}(\cdot|s,a)$, values $\hat{r}(s,a)$ in $[0,1]$, $d(s,a) > 0$, and $d'(s,a) \geq 0$. Further, we assume that $\mathcal{M}$ contains at least one communicating MDP, that is, an MDP with finite diameter.

In Step 5 of UCRL2, the $d(s,a)$ and $d'(s,a)$ are obviously the confidence intervals as given by (4) and (3), while the communicating MDP assumed to be in $\mathcal{M}_k$ is the true MDP M. The task is to find an MDP $\tilde{M} \in \mathcal{M}$ and a policy $\tilde{\pi} : \mathcal{S} \to \mathcal{A}$ which maximize $\rho(\tilde{M}, \tilde{\pi}, s)$ for all states s.[7] This task is similar to *optimistic optimality* in *bounded parameter MDPs* as considered by Tewari and Bartlett (2007). A minor difference is that in our case the transition probabilities are bounded not individually but by the 1-norm. More importantly, while Tewari and Bartlett (2007) give a converging algorithm for computing the optimal value function, they do not bound the error when terminating their algorithm after finitely many steps. In the following, we will extend standard undiscounted value iteration (Puterman, 1994) to solve the set task.

First, note that we may combine all MDPs in $\mathcal{M}$ to get a single MDP with extended action set $\mathcal{A}'$. That is, we consider an MDP $\tilde{M}^+$ with continuous action space $\mathcal{A}'$, where for each action $a \in \mathcal{A}$, each admissible transition probability distribution $\tilde{p}(\cdot|s,a)$ according to (1) and each admissible mean reward $\tilde{r}(s,a)$ according to (2) there is an action in $\mathcal{A}'$ with transition probabilities $\tilde{p}(\cdot|s,a)$ and mean reward $\tilde{r}(s,a)$.[8] Then for each policy $\tilde{\pi}^+$ on $\tilde{M}^+$ there is an MDP $\tilde{M} \in \mathcal{M}$ and a policy $\tilde{\pi} : \mathcal{S} \to \mathcal{A}$ on $\tilde{M}$ such that the policies $\tilde{\pi}^+$ and $\tilde{\pi}$ induce the same transition probabilities and mean rewards on the respective MDP. (The other transition probabilities in $\tilde{M}$ can be set to $\hat{p}\,(\cdot|s,a)$.) On the other hand, for any given MDP $\tilde{M} \in \mathcal{M}$ and any policy $\tilde{\pi} : \mathcal{S} \to \mathcal{A}$ there is a policy $\tilde{\pi}^+$ on $\tilde{M}^+$ so that again the same transition probabilities and rewards are induced by $\tilde{\pi}$ on $\tilde{M}$ and $\tilde{\pi}^+$ on $\tilde{M}^+$. Thus, finding an MDP $\tilde{M} \in \mathcal{M}$ and a policy $\tilde{\pi}$ on $\tilde{M}$ such that $\rho(\tilde{M}, \tilde{\pi}, s) = \max_{M' \in \mathcal{M}, \pi, s'} \rho(M', \pi, s')$ for all initial states s, corresponds to finding an average reward optimal policy on $\tilde{M}^+$.

[6] Since the policy $\tilde{\pi}_k$ is fixed for episode k, $v_k(s,a) \neq 0$ only for $a = \tilde{\pi}_k(s)$. Nevertheless, we find it convenient to use a notation which explicitly includes the action a in $v_k(s,a)$.

[7] Note that, as we assume that $\mathcal{M}$ contains a communicating MDP, if an average reward of ρ is achievable in one state, it is achievable in all states.

[8] Note that in $\tilde{M}^+$ the set of available actions now depends on the state.

```
Input: A confidence parameter $\delta \in (0,1)$, $\mathcal{S}$ and $\mathcal{A}$.

Initialization: Set $t := 1$, and observe the initial state $s_1$.

For episodes $k = 1, 2, \ldots$ do
    Initialize episode $k$:
        1. Set the start time of episode $k$, $t_k := t$.
        2. For all $(s,a)$ in $\mathcal{S} \times \mathcal{A}$ initialize the state-action counts for episode $k$, $v_k(s,a) := 0$.
           Further, set the state-action counts prior to episode $k$,
               $N_k(s,a) := \#\{\tau < t_k : s_\tau = s, a_\tau = a\}$.
        3. For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ set the observed accumulated rewards and the transition
           counts prior to episode $k$,
               $R_k(s,a) := \sum_{\tau=1}^{t_k-1} r_\tau \mathbb{1}_{s_\tau = s, a_\tau = a}$,
               $P_k(s,a,s') := \#\left\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\right\}$.
           Compute estimates $\hat{r}_k(s,a) := \frac{R_k(s,a)}{\max\{1, N_k(s,a)\}}$, $\hat{p}_k(s'|s,a) := \frac{P_k(s,a,s')}{\max\{1, N_k(s,a)\}}$.
    Compute policy $\tilde{\pi}_k$:
        4. Let $\mathcal{M}_k$ be the set of all MDPs with states and actions as in $M$, and with transition
           probabilities $\tilde{p}(\cdot|s,a)$ close to $\hat{p}_k(\cdot|s,a)$, and rewards $\tilde{r}(s,a) \in [0,1]$ close
           to $\hat{r}_k(s,a)$, that is,
               $\left|\tilde{r}(s,a) - \hat{r}_k(s,a)\right| \leq \sqrt{\frac{7\log(2SAt_k/\delta)}{2\max\{1, N_k(s,a)\}}}$ and        (3)
               $\left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \leq \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1, N_k(s,a)\}}}$.        (4)
        5. Use extended value iteration (see Section 3.1) to find a policy $\tilde{\pi}_k$ and an optimistic
           MDP $\tilde{M}_k \in \mathcal{M}_k$ such that
               $\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} \rho(M', \pi, s') - \frac{1}{\sqrt{t_k}}$.
    Execute policy $\tilde{\pi}_k$:
        6. While $v_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_k(s_t, \tilde{\pi}_k(s_t))\}$ do
            (a) Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward $r_t$, and
                observe next state $s_{t+1}$.
            (b) Update $v_k(s_t, a_t) := v_k(s_t, a_t) + 1$.
            (c) Set $t := t + 1$.
```

Figure 1: The UCRL2 algorithm.

```
Input: Estimates $\hat{p}(\cdot|s,a)$ and distance $d(s,a)$ for a state-action pair $(s,a)$, and
    the states in $\mathcal{S}$ sorted descendingly according to their $u_i$ value.
    That is, let $\mathcal{S} := \{s'_1, s'_2, \ldots, s'_n\}$ with $u_i(s'_1) \geq u_i(s'_2) \geq \ldots \geq u_i(s'_n)$.

  1. Set
         $p(s'_1) := \min\{1, \hat{p}(s'_1|s,a) + \frac{d(s,a)}{2}\}$, and
         $p(s'_j) := \hat{p}(s'_j|s,a)$ for all states $s'_j$ with $j > 1$.

  2. Set $\ell := n$.

  3. While $\sum_{s'_j \in \mathcal{S}} p(s'_j) > 1$ do
       (a) Reset $p(s'_\ell) := \max\{0, 1 - \sum_{s'_j \neq s'_\ell} p(s'_j)\}$.
       (b) Set $\ell := \ell - 1$.
```

Figure 2: Computing the inner maximum in the extended value iteration (5).

3.1.2 EXTENDED VALUE ITERATION

We denote the state values of the i-th iteration by $u_i(s)$. Then we get for undiscounted value iteration (Puterman, 1994) on $\tilde{M}^+$ for all $s \in \mathcal{S}$:

$$\begin{aligned} u_0(s) &= 0, \\ u_{i+1}(s) &= \max_{a \in \mathcal{A}} \left\{ \tilde{r}\,(s,a) + \max_{p(\cdot) \in \mathcal{P}(s,a)} \left\{ \sum_{s' \in \mathcal{S}} p(s') \cdot u_i(s') \right\} \right\}, \end{aligned} \tag{5}$$

where $\tilde{r}\,(s,a) := \hat{r}(s,a) + d'(s,a)$ are the maximal possible rewards according to condition (2), and $\mathcal{P}(s,a)$ is the set of transition probabilities $\tilde{p}\,\left(\cdot|s,a\right)$ satisfying condition (1).

While (5) is a step of value iteration with an infinite action space, $\max_p \boldsymbol{p} \cdot \boldsymbol{u}_i$ is actually a linear optimization problem over the convex polytope $\mathcal{P}(s,a)$. This implies that (5) can be evaluated considering only the finite number of vertices of this polytope.

Indeed, for a given state-action pair the inner maximum of (5) can be computed in $O(S)$ computation steps by an algorithm introduced by Strehl and Littman (2008). For the sake of completeness we display the algorithm in Figure 2. The idea is to put as much transition probability as possible to the state with maximal value $u_i(s)$ at the expense of transition probabilities to states with small values $u_i(s)$. That is, one starts with the estimates $\hat{p}(s'_j|s,a)$ for $p(s'_j)$ except for the state s'_1 with maximal $u_i(s)$, for which we set $p(s'_1) := \hat{p}(s'_1|s,a) + \frac{1}{2}d(s,a)$. In order to make $\boldsymbol{p}$ correspond to a probability distribution again, the transition probabilities from s to states with small $u_i(s)$ are reduced in total by $\frac{1}{2}d(s,a)$, so that $\|\boldsymbol{p} - \hat{p}\,(\cdot|s,a)\,\|_1 = d(s,a)$. This is done iteratively. Updating $\sum_{s'_j \in \mathcal{S}}\ p(s'_j)$ with every change of $\boldsymbol{p}$ for the computation of $\sum_{s'_j \neq s'_\ell} p(s'_j)$, this iterative procedure takes $O(S)$ steps. Thus, sorting the states according to their value $u_i(s)$ at each iteration i once, $\boldsymbol{u}_{i+1}$ can be computed from $\boldsymbol{u}_i$ in at most $O\big(S^2A\big)$ steps.

3.1.3 CONVERGENCE OF EXTENDED VALUE ITERATION

We have seen that value iteration on the MDP $\tilde{M}^+$ with continuous action is equivalent to value iteration on an MDP with finite action set. Thus, in order to guarantee convergence, it is sufficient to assure that extended value iteration never chooses a policy with periodic transition matrix. (Intuitively, it is clear that optimal policies with periodic transition matrix do not matter as long as it is guaranteed that such a policy is not chosen by value iteration, compare to Sections 8.5, 9.4, and 9.5.3. of Puterman 1994. For a proof see Appendix B.) Indeed, extended value iteration always chooses a policy with aperiodic transition matrix: In each iteration there is a single fixed state s'_1 which is regarded as the "best" target state. For each state s, in the inner maximum an action with positive transition probability to s'_1 will be chosen. In particular, the policy chosen by extended value iteration will have positive transition probability from s'_1 to s'_1. Hence, this policy is aperiodic and has state independent average reward. Thus we obtain the following result.

Theorem 7 *Let $\mathcal{M}$ be the set of all MDPs with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probabilities $\tilde{p}\,(\cdot|s,a)$, and mean rewards $\tilde{r}\,(s,a)$ that satisfy* (1) *and* (2) *for given probability distributions $\hat{p}(\cdot|s,a)$, values $\hat{r}(s,a)$ in $[0,1]$, $d(s,a) > 0$, and $d'(s,a) \geq 0$. If $\mathcal{M}$ contains at least one communicating MDP, extended value iteration converges. Further, stopping extended value iteration when*

$$\max_{s\in S}\left\{u_{i+1}(s) - u_i(s)\right\} - \min_{s\in S}\left\{u_{i+1}(s) - u_i(s)\right\} < \varepsilon,$$

the greedy policy with respect to $\boldsymbol{u}_i$ is ε-optimal.

Remark 8 *When value iteration converges, a suitable transformation of $\boldsymbol{u}_i$ converges to the bias vector of an optimal policy. Recall that for a policy π the bias $\lambda(s)$ in state s is basically the expected advantage in total reward (for $T \to \infty$) of starting in state s over starting in the stationary distribution (the long term probability of being in a state) of π. For a fixed policy π, the Poisson equation*

$$\boldsymbol{\lambda} = \boldsymbol{r} - \rho\mathbf{1} + \boldsymbol{P}\boldsymbol{\lambda}$$

relates the bias vector $\boldsymbol{\lambda}$ to the average reward ρ, the mean reward vector $\boldsymbol{r}$, and the transition matrix $\boldsymbol{P}$. Now when value iteration converges, the vector $\boldsymbol{u}_i - \min_s u_i(s)\mathbf{1}$ converges to $\boldsymbol{\lambda} - \min_s \lambda(s)\mathbf{1}$. As we will see in inequality (11) *below, the so-called* span $\max_s u_i(s) - \min_s u_i(s)$ *of the vector $\boldsymbol{u}_i$ is upper bounded by the diameter D, so that this also holds for the span of the bias vector $\boldsymbol{\lambda}$ of the optimal policy found by extended value iteration, that is,* $\max_s \lambda(s) - \min_s \lambda(s) \leq D$*. Indeed, one can show that this holds for any optimal policy (cf. also Section 4 of Bartlett and Tewari, 2009).*

Remark 9 *We would like to note that the algorithm of Figure 2 can easily be adapted to the alternative setting of Tewari and Bartlett (2007), where each single transition probability $p(s'|s,a)$ is bounded as $0 \leq b^-(s',s,a) \leq p(s'|s,a) \leq b^+(s',s,a) \leq 1$. However, concerning convergence one needs to make some assumptions to exclude the possibility of choosing optimal policies with periodic transition matrices. For example, one may assume (apart from other assumptions already made by Tewari and Bartlett 2007) that for all s', s, a there is an admissible probability distribution $p(\cdot|s,a)$ with $p(s'|s,a) > 0$. Note that for Theorem 7 to hold, it is similarly essential that $d(s,a) > 0$. Alternatively, one may apply an aperiodicity transformation as described in Section 8.5.4 of Puterman (1994).*

Now returning to Step 5 of UCRL2, we stop value iteration when

$$\max_{s\in S}\big\{u_{i+1}(s)-u_i(s)\big\}-\min_{s\in S}\big\{u_{i+1}(s)-u_i(s)\big\}<\frac{1}{\sqrt{t_k}}, \tag{6}$$

which guarantees by Theorem 7 that the greedy policy with respect to $\boldsymbol{u}_i$ is $\frac{1}{\sqrt{t_k}}$-optimal.

4. Analysis of UCRL2 (Proofs of Theorem 2 and Corollary 3)

We start with a rough outline of the proof of Theorem 2. First, in Section 4.1, we deal with the random fluctuation of the rewards. Further, the regret is expressed as the sum of the regret accumulated in the individual episodes. That is, we set the *regret in episode k* to be

$$\Delta_k := \sum_{s,a} v_k(s,a)\big(\rho^*-\overline{r}(s,a)\big),$$

where $v_k(s,a)$ now denotes the final counts of state-action pair (s,a) in episode k. Then it is shown that the total regret can be bounded by

$$\sum_k \Delta_k + \sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)}$$

with high probability.

In Section 4.2, we consider the regret that is caused by failing confidence regions. We show that this term can be upper bounded by $\sqrt{T}$ with high probability. After this intermezzo, the regret of episodes for which the true MDP $M\in\mathcal{M}_k$ is examined in Section 4.3. Analyzing the extended value iteration scheme in Section 4.3.1 and using vector notation, we show that

$$\Delta_k \;\le\; \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k-\boldsymbol{I}\big)\boldsymbol{w}_k+2\sum_{s,a}v_k(s,a)\sqrt{\tfrac{7\log(2SAt_k/\delta)}{2\max\{1,N_k(s,a)\}}}+2\sum_{s,a}\frac{v_k(s,a)}{\sqrt{t_k}},$$

where $\tilde{\boldsymbol{P}}_k$ is the assumed transition matrix (in $\tilde{M}_k$) of the applied policy in episode k, $\boldsymbol{v}_k$ are the visit counts at the end of that episode, and $\boldsymbol{w}_k$ is a vector with $\|\boldsymbol{w}_k\|_\infty\le\frac{D(M)}{2}$. The last two terms in the above expression stem from the reward confidence intervals (3) and the approximation error of value iteration. These are bounded in Section 4.3.3 when summing over all episodes. The first term on the right hand side is analyzed further in Section 4.3.2 and split into

$$\begin{aligned}\boldsymbol{v}_k(\tilde{\boldsymbol{P}}_k-\boldsymbol{I})\boldsymbol{w}_k &= \boldsymbol{v}_k(\tilde{\boldsymbol{P}}_k-\boldsymbol{P}_k)\boldsymbol{w}_k+\boldsymbol{v}_k(\boldsymbol{P}_k-\boldsymbol{I})\boldsymbol{w}_k\\ &\le \big\|\boldsymbol{v}_k(\tilde{\boldsymbol{P}}_k-\boldsymbol{P}_k)\big\|_1\|\boldsymbol{w}_k\|_\infty+\boldsymbol{v}_k(\boldsymbol{P}_k-\boldsymbol{I})\boldsymbol{w}_k,\end{aligned}$$

where $\boldsymbol{P}_k$ is the true transition matrix (in M) of the policy applied in episode k. Substituting for $\tilde{\boldsymbol{P}}_k-\boldsymbol{P}_k$ the lengths of the confidence intervals as given in (4), the remaining term that needs analysis is $\boldsymbol{v}_k(\boldsymbol{P}_k-\boldsymbol{I})\boldsymbol{w}_k$. For the sum of this term over all episodes we obtain in Section 4.3.2 a high probability bound of

$$\sum_k \boldsymbol{v}_k(\boldsymbol{P}_k-\boldsymbol{I})\boldsymbol{w}_k \;\le\; D\sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)}+Dm,$$

where m is the number of episodes—a term shown to be logarithmic in T in Appendix C.2. Section 4.3.3 concludes the analysis of episodes with $M\in\mathcal{M}_k$ by summing the individual regret terms over all episodes k with $M\in\mathcal{M}_k$. In the final Section 4.4 we finish the proof by combining the results of Sections 4.1–4.3.

4.1 Splitting into Episodes

Recall that r_t is the (random) reward UCRL2 receives at step t when starting in some initial state s_1. For given state-action counts $N(s,a)$ after T steps, the r_t are independent random variables, so that by Hoeffding's inequality

$$\mathbb{P}\left\{ \sum_{t=1}^{T} r_t \leq \sum_{s,a} N(s,a)\bar{r}(s,a) - \sqrt{\tfrac{5}{8}T\log\left(\tfrac{8T}{\delta}\right)} \,\middle|\, \big(N(s,a)\big)_{s,a} \right\} \leq \left(\frac{\delta}{8T}\right)^{5/4} < \frac{\delta}{12T^{5/4}}. \qquad (7)$$

Thus we get for the regret of UCRL2 (now omitting explicit references to M and UCRL2)

$$\Delta(s_1,T) \;=\; T\rho^* - \sum_{t=1}^{T} r_t \;<\; T\rho^* - \sum_{s,a} N(s,a)\bar{r}(s,a) + \sqrt{\tfrac{5}{8}T\log\left(\tfrac{8T}{\delta}\right)}$$

with probability at least $1-\frac{\delta}{12T^{5/4}}$. Denoting the number of episodes started up to step T by m, we have $\sum_{k=1}^{m} v_k(s,a) = N(s,a)$ and $\sum_{s,a} N(s,a) = T$. Therefore, writing $\Delta_k := \sum_{s,a} v_k(s,a)\big(\rho^* - \bar{r}(s,a)\big)$, it follows that

$$\Delta(s_1,T) \;\leq\; \sum_{k=1}^{m} \Delta_k + \sqrt{\tfrac{5}{8}T\log\left(\tfrac{8T}{\delta}\right)} \qquad (8)$$

with probability at least $1-\frac{\delta}{12T^{5/4}}$.

4.2 Dealing with Failing Confidence Regions

Let us now consider the regret of episodes in which the set of plausible MDPs $\mathcal{M}_k$ does not contain the true MDP M, $\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k}$. By the stopping criterion for episode k we have (except for episodes where $v_k(s,a)=1$ and $N_k(s,a)=0$, when $\sum_{s,a} v_k(s,a) = 1 \leq t_k$ holds trivially)

$$\sum_{s,a} v_k(s,a) \;\leq\; \sum_{s,a} N_k(s,a) = t_k - 1.$$

Hence, denoting $\mathcal{M}(t)$ to be the set of plausible MDPs as given by (3) and (4) using the estimates available at step t, we have due to $\rho^* \leq 1$ that

$$\begin{aligned}
\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} \;&\leq\; \sum_{k=1}^{m}\sum_{s,a} v_k(s,a)\mathbb{1}_{M\notin\mathcal{M}_k} \;\leq\; \sum_{k=1}^{m} t_k \mathbb{1}_{M\notin\mathcal{M}_k} \;=\; \sum_{t=1}^{T} t \sum_{k=1}^{m} \mathbb{1}_{t_k=t, M\notin\mathcal{M}_k} \\
&\leq\; \sum_{t=1}^{T} t\mathbb{1}_{M\notin\mathcal{M}(t)} \;\leq\; \sum_{t=1}^{\lfloor T^{1/4}\rfloor} t\mathbb{1}_{M\notin\mathcal{M}(t)} + \sum_{t=\lfloor T^{1/4}\rfloor+1}^{T} t\mathbb{1}_{M\notin\mathcal{M}(t)} \\
&\leq\; \sqrt{T} + \sum_{t=\lfloor T^{1/4}\rfloor+1}^{T} t\mathbb{1}_{M\notin\mathcal{M}(t)}.
\end{aligned}$$

Now, $\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \leq \frac{\delta}{15t^6}$ (see Appendix C.1), and since

$$\sum_{t=\lfloor T^{1/4}\rfloor+1}^{T} \frac{1}{15t^6} \;\leq\; \frac{1}{15T^{6/4}} + \int_{T^{1/4}}^{\infty} \frac{1}{15t^6}\,dt \;=\; \frac{1}{15T^{6/4}} + \frac{1}{75T^{5/4}} \;\leq\; \frac{6}{75T^{5/4}} \;<\; \frac{1}{12T^{5/4}},$$

we have $\mathbb{P}\{\exists t : T^{1/4} < t \leq T : M \notin \mathcal{M}(t)\} \leq \frac{\delta}{12T^{5/4}}$. It follows that with probability at least $1-\frac{\delta}{12T^{5/4}}$,

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} \;\leq\; \sqrt{T}. \qquad (9)$$

4.3 Episodes with $M \in \mathcal{M}_k$

Now we assume that $M \in \mathcal{M}_k$ and start by considering the regret in a single episode k. The optimistic average reward $\tilde{\rho}_k$ of the optimistically chosen policy $\tilde{\pi}_k$ is essentially larger than the true optimal average reward ρ^*, and thus it is sufficient to calculate by how much the optimistic average reward $\tilde{\rho}_k$ overestimates the actual rewards of policy $\tilde{\pi}_k$. By the assumption $M \in \mathcal{M}_k$, the choice of $\tilde{\pi}_k$ and $\tilde{M}_k$ in Step 5 of UCRL2, and Theorem 7 we get that $\tilde{\rho}_k \geq \rho^* - \frac{1}{\sqrt{t_k}}$. Thus for the regret Δ_k accumulated in episode k we obtain

$$\Delta_k \;\leq\; \sum_{s,a} v_k(s,a)\big(\rho^* - \bar{r}(s,a)\big) \;\leq\; \sum_{s,a} v_k(s,a)\big(\tilde{\rho}_k - \bar{r}(s,a)\big) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}\,. \tag{10}$$

4.3.1 EXTENDED VALUE ITERATION REVISITED

To proceed, we reconsider the extended value iteration of Section 3.1. As an important observation for our analysis, we find that for any iteration i the range of the state values is bounded by the diameter of the MDP M, that is,

$$\max_s u_i(s) - \min_s u_i(s) \leq D. \tag{11}$$

To see this, observe that $u_i(s)$ is the total expected i-step reward of an optimal non-stationary i-step policy starting in state s on the MDP $\tilde{M}^+$ with extended action set (as considered for extended value iteration). The diameter of this extended MDP is at most D as it contains by assumption the actions of the true MDP M. Now, if there were states s', s'' with $u_i(s'') - u_i(s') > D$, then an improved value for $u_i(s')$ could be achieved by the following nonstationary policy: First follow a policy which moves from s' to s'' most quickly, which takes at most D steps on average. Then follow the optimal i-step policy for s''. Since only D of the i rewards of the policy for s'' are missed, this policy gives $u_i(s') \geq u_i(s'') - D$, contradicting our assumption and thus proving (11).

It is a direct consequence of Theorem 8.5.6. of Puterman (1994), that when the convergence criterion (6) holds at iteration i, then

$$|u_{i+1}(s) - u_i(s) - \tilde{\rho}_k| \leq \frac{1}{\sqrt{t_k}} \tag{12}$$

for all $s \in \mathcal{S}$, where $\tilde{\rho}_k$ is the average reward of the policy $\tilde{\pi}_k$ chosen in this iteration on the optimistic MDP $\tilde{M}_k$.[9] Expanding $u_{i+1}(s)$ according to (5), we get

$$u_{i+1}(s) \;=\; \tilde{r}_k(s, \tilde{\pi}_k(s)) + \sum_{s'} \tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) \cdot u_i(s')$$

and hence by (12)

$$\left| \Big(\tilde{\rho}_k - \tilde{r}_k(s, \tilde{\pi}_k(s)) \Big) - \left(\sum_{s'} \tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) \cdot u_i(s') - u_i(s) \right) \right| \leq \frac{1}{\sqrt{t_k}}. \tag{13}$$

Setting $\boldsymbol{r}_k := \big(\tilde{r}_k\big(s, \tilde{\pi}_k(s)\big)\big)_s$ to be the (column) vector of rewards for policy $\tilde{\pi}_k$, $\tilde{\boldsymbol{P}}_k := \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right)\right)_{s,s'}$ the transition matrix of $\tilde{\pi}_k$ on $\tilde{M}_k$, and $\boldsymbol{v}_k := \big(v_k\big(s, \tilde{\pi}_k(s)\big)\big)_s$ the (row)

[9] This is quite intuitive. We expect to receive average reward $\tilde{\rho}_k$ per step, such that the difference of the state values after $i+1$ and i steps should be about $\tilde{\rho}_k$.

vector of visit counts for each state and the corresponding action chosen by $\tilde{\pi}_k$, we can use (13)—recalling that $v_k(s,a) = 0$ for $a \neq \tilde{\pi}_k(s)$—to rewrite (10) as

$$
\begin{aligned}
\Delta_k &\leq \sum_{s,a} v_k(s,a)\big(\tilde{\rho}_k - \overline{r}(s,a)\big) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}} \\
&= \sum_{s,a} v_k(s,a)\big(\tilde{\rho}_k - \tilde{r}_k(s,a)\big) + \sum_{s,a} v_k(s,a)\big(\tilde{r}_k(s,a) - \overline{r}(s,a)\big) + \sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}} \\
&\leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{u}_i + \sum_{s,a} v_k(s,a)\big(\tilde{r}_k(s,a) - \overline{r}(s,a)\big) + 2\sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}.
\end{aligned}
$$

Since the rows of $\tilde{\boldsymbol{P}}_k$ sum to 1, we can replace $\boldsymbol{u}_i$ by $\boldsymbol{w}_k$ where we set

$$
w_k(s) := u_i(s) - \frac{\min_s u_i(s) + \max_s u_i(s)}{2},
$$

such that it follows from (11) that $\|\boldsymbol{w}_k\| \leq \frac{D}{2}$. Further, since we assume $M \in \mathcal{M}_k$, $\tilde{r}_k(s,a) - \overline{r}(s,a) \leq |\tilde{r}_k(s,a) - \hat{r}_k(s,a)| + |\overline{r}(s,a) - \hat{r}_k(s,a)|$ is bounded according to (3), so that

$$
\Delta_k \leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + 2\sum_{s,a} v_k(s,a)\sqrt{\tfrac{7\log(2SAt_k/\delta)}{2\max\{1,N_k(s,a)\}}} + 2\sum_{s,a} \frac{v_k(s,a)}{\sqrt{t_k}}. \tag{14}
$$

Noting that $\max\{1, N_k(s,a)\} \leq t_k \leq T$ we get from (14) that

$$
\Delta_k \leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + \left(\sqrt{14\log\left(\tfrac{2SAT}{\delta}\right)} + 2\right)\sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\,. \tag{15}
$$

4.3.2 THE TRUE TRANSITION MATRIX

Now we want to replace the transition matrix $\tilde{\boldsymbol{P}}_k$ of the policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{M}_k$ by the transition matrix $\boldsymbol{P}_k := \big(p(s'|s,\tilde{\pi}_k(s))\big)_{s,s'}$ of $\tilde{\pi}_k$ in the true MDP M. Thus, we write

$$
\begin{aligned}
\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k &= \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k + \boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k \\
&= \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k + \boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k. \qquad (16)
\end{aligned}
$$

The first term. Since by assumption $\tilde{M}_k$ and M are in the set of plausible MDPs $\mathcal{M}_k$, the first term in (16) can be bounded using condition (4). Thus, also using that $\|\boldsymbol{w}_k\|_\infty \leq \frac{D}{2}$ we obtain

$$
\begin{aligned}
\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k &= \sum_s \sum_{s'} v_k\big(s,\tilde{\pi}_k(s)\big)\cdot\Big(\tilde{p}_k\left(s'|s,\tilde{\pi}_k(s)\right) - p\left(s'|s,\tilde{\pi}_k(s)\right)\Big)\cdot w_k(s') \\
&\leq \sum_s v_k\big(s,\tilde{\pi}_k(s)\big)\cdot\big\|\tilde{p}_k\left(\,\cdot\,|s,\tilde{\pi}_k(s)\right) - p(\cdot|s,\tilde{\pi}_k(s))\big\|_1 \cdot \|\boldsymbol{w}_k\|_\infty \\
&\leq \sum_s v_k\big(s,\tilde{\pi}_k(s)\big)\cdot 2\sqrt{\frac{14S\log\left(2AT/\delta\right)}{\max\{1,N_k\left(s,\tilde{\pi}_k(s)\right)\}}}\cdot\frac{D}{2} \\
&\leq D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\,. \qquad (17)
\end{aligned}
$$

This term will turn out to be the dominating contribution in our regret bound.

The second term. The intuition about the second term in (16) is that the counts of the state visits $\boldsymbol{v}_k$ are relatively close to the stationary distribution $\boldsymbol{\mu}_k$ of the transition matrix $\boldsymbol{P}_k$, for which $\boldsymbol{\mu}_k\boldsymbol{P}_k = \boldsymbol{\mu}_k$, such that $\boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)$ should be small. For the proof we define a suitable martingale and make use of the Azuma-Hoeffding inequality.

Lemma 10 (Azuma-Hoeffding inequality, Hoeffding 1963) *Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \leq c$ for all i. Then for all $\varepsilon > 0$ and $n \in \mathbb{N}$,*

$$\mathbb{P}\big\{ \textstyle\sum_{i=1}^n X_i \geq \varepsilon \big\} \;\leq\; \exp\big(-\tfrac{\varepsilon^2}{2nc^2}\big).$$

Denote the unit vectors with i-th coordinate 1 and all other coordinates 0 by $\boldsymbol{e}_i$. Let $s_1, a_1, s_2, \ldots, a_T$, s_{T+1} be the sequence of states and actions, and let $k(t)$ be the episode which contains step t. Consider the sequence $X_t := \big(p(\cdot|s_t,a_t) - \boldsymbol{e}_{s_{t+1}}\big)\,\boldsymbol{w}_{k(t)}\mathbb{1}_{M\in\mathcal{M}_{k(t)}}$ for $t = 1, \ldots, T$. Then for any episode k with $M \in \mathcal{M}_k$, we have due to $\|\boldsymbol{w}_k\|_\infty \leq \frac{D}{2}$ that

$$\begin{aligned}
\boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k &= \sum_{t=t_k}^{t_{k+1}-1} \big(p(\cdot|s_t,a_t) - \boldsymbol{e}_{s_t}\big)\boldsymbol{w}_k \\
&= \left(\sum_{t=t_k}^{t_{k+1}-1} p(\cdot|s_t,a_t) - \sum_{t=t_k}^{t_{k+1}-1} \boldsymbol{e}_{s_{t+1}} + \boldsymbol{e}_{s_{t_{k+1}}} - \boldsymbol{e}_{s_{t_k}}\right)\boldsymbol{w}_k \\
&= \sum_{t=t_k}^{t_{k+1}-1} X_t + w_k(s_{t_{k+1}}) - w_k(s_{t_k}) \\
&\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + D.
\end{aligned}$$

Also due to $\|\boldsymbol{w}_k\|_\infty \leq \frac{D}{2}$, we have $|X_t| \leq (\|p(\cdot|s_t,a_t)\|_1 + \|\boldsymbol{e}_{s_{t+1}}\|_1)\frac{D}{2} \leq D$. Further, $\mathbb{E}\big[X_t \big| s_1, a_1, \ldots, s_t, a_t\big] = 0$, so that X_t is a sequence of martingale differences, and application of Lemma 10 gives

$$\mathbb{P}\left\{\sum_{t=1}^T X_t \geq D\sqrt{2T\cdot\tfrac{5}{4}\log\big(\tfrac{8T}{\delta}\big)}\right\} \leq \left(\frac{\delta}{8T}\right)^{5/4} < \frac{\delta}{12T^{5/4}}.$$

Since for the number of episodes we have $m \leq SA\log_2\big(\frac{8T}{SA}\big)$ as shown in Appendix C.2, summing over all episodes yields

$$\begin{aligned}
\sum_{k=1}^m \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k\mathbb{1}_{M\in\mathcal{M}_k} &\leq \sum_{t=1}^T X_t + mD \\
&\leq D\sqrt{\tfrac{5}{2}T\log\big(\tfrac{8T}{\delta}\big)} + DSA\log_2\big(\tfrac{8T}{SA}\big) \qquad (18)
\end{aligned}$$

with probability at least $1 - \frac{\delta}{12T^{5/4}}$.

4.3.3 SUMMING OVER EPISODES WITH $M \in \mathcal{M}_k$

To conclude Section 4.3, we sum (15) over all episodes with $M \in \mathcal{M}_k$, using (16), (17), and (18), which yields that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned}
\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\in\mathcal{M}_k} &\le \sum_{k=1}^{m} \boldsymbol{v}_k \big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big) \boldsymbol{w}_k \mathbb{1}_{M\in\mathcal{M}_k} + \sum_{k=1}^{m} \boldsymbol{v}_k \big(\boldsymbol{P}_k - \boldsymbol{I}\big) \boldsymbol{w}_k \mathbb{1}_{M\in\mathcal{M}_k} \\
&\quad + \sum_{k=1}^{m} \left(\sqrt{14 \log\left(\tfrac{2SAT}{\delta}\right)} + 2 \right) \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \\
&\le D\sqrt{14 S \log\left(\tfrac{2AT}{\delta}\right)} \cdot \sum_{k=1}^{m} \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \\
&\quad + D\sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + DSA \log_2\left(\tfrac{8T}{SA}\right) \\
&\quad + \left(\sqrt{14 \log\left(\tfrac{2SAT}{\delta}\right)} + 2 \right) \sum_{k=1}^{m} \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \,. \qquad (19)
\end{aligned}$$

Recall that $N(s,a) := \sum_k v_k(s,a)$ such that $\sum_{s,a} N(s,a) = T$ and $N_k(s,a) = \sum_{i<k} v_i(s,a)$. By the criterion for episode termination in Step 6 of the algorithm, we have that $v_k(s,a) \le N_k(s,a)$. Using that for $Z_k = \max\left\{1, \sum_{i=1}^{k} z_i\right\}$ and $0 \le z_k \le Z_{k-1}$ it holds that (see Appendix C.3)

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \le \left(\sqrt{2}+1\right) \sqrt{Z_n} \,,$$

we get

$$\sum_{s,a} \sum_{k} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \le \left(\sqrt{2}+1\right) \sum_{s,a} \sqrt{N(s,a)}.$$

By Jensen's inequality we thus have

$$\sum_{s,a} \sum_{k} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \le \left(\sqrt{2}+1\right) \sqrt{SAT}, \qquad (20)$$

and we get from (19) after some minor simplifications that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned}
\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\in\mathcal{M}_k} &\le D\sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + DSA \log_2\left(\tfrac{8T}{SA}\right) \\
&\quad + \left(2D\sqrt{14 S \log\left(\tfrac{2AT}{\delta}\right)} + 2 \right) \left(\sqrt{2}+1\right) \sqrt{SAT} \,. \qquad (21)
\end{aligned}$$

4.4 Completing the Proof of Theorem 2

Finally, evaluating (8) by summing Δ_k over all episodes, we get by (9) and (21)

$$\begin{aligned}
\Delta(s_1, T) &\le \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} + \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\in\mathcal{M}_k} + \sqrt{\tfrac{5}{8} T \log\left(\tfrac{8T}{\delta}\right)} \\
&\le \sqrt{\tfrac{5}{8} T \log\left(\tfrac{8T}{\delta}\right)} + \sqrt{T} + D\sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + DSA \log_2\left(\tfrac{8T}{SA}\right) \\
&\quad + \left(2D\sqrt{14 S \log\left(\tfrac{2AT}{\delta}\right)} + 2 \right) \left(\sqrt{2}+1\right) \sqrt{SAT} \qquad (22)
\end{aligned}$$

with probability at least $1-\frac{\delta}{12T^{5/4}}-\frac{\delta}{12T^{5/4}}-\frac{\delta}{12T^{5/4}}$. Further simplifications (given in Appendix C.4) yield that for any $T>1$ with probability at least $1-\frac{\delta}{4T^{5/4}}$

$$\Delta(s_1,T) \;\leq\; 34DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}. \tag{23}$$

Since $\sum_{T=2}^{\infty}\frac{\delta}{4T^{5/4}}<\delta$ the statement of Theorem 2 follows by a union bound over all possible values of T. ■

4.5 Proof of Corollary 3

In order to obtain the PAC bound of Corollary 3 we simply have to find a sufficiently large T_0 such that for all $T\geq T_0$ the per-step regret is smaller than ε. By Theorem 2 this means that for all $T\geq T_0$ we shall have

$$\frac{34DS\sqrt{AT\log\left(\frac{T}{\delta}\right)}}{T}<\varepsilon, \text{ or equivalently } \quad T>\frac{34^2D^2S^2A\log\left(\frac{T}{\delta}\right)}{\varepsilon^2}. \tag{24}$$

Setting $T_0:=2\alpha\log\left(\frac{\alpha}{\delta}\right)$ for $\alpha:=\frac{34^2D^2S^2A}{\varepsilon^2}$ we have due to $x>2\log x$ (for $x>0$)

$$T_0 \;=\; \alpha\log\left(\frac{\alpha}{\delta}\cdot\frac{\alpha}{\delta}\right) > \alpha\log\left(2\frac{\alpha}{\delta}\log\left(\frac{\alpha}{\delta}\right)\right) = \alpha\log\left(\tfrac{T_0}{\delta}\right),$$

so that (24) as well as the corollary follow. ■

5. The Logarithmic Bound (Proof of Theorem 4)

To show the logarithmic upper bound on the expected regret, we start with a bound on the number of steps in suboptimal episodes (in the spirit of *sample complexity bounds* as given by Kakade, 2003). We say that an episode k is ε-*bad* if its average regret is more than ε, where the average regret of an episode of length ℓ_k is $\frac{\Delta_k}{\ell_k}$ with[10] $\Delta_k=\sum_{t=t_k}^{t_{k+1}-1}(\rho^*-r_t)$. The following result gives an upper bound on the number of steps taken in ε-bad episodes.

Theorem 11 *Let $L_\varepsilon(T)$ be the number of steps taken by* UCRL2 *in ε-bad episodes up to step T. Then for any initial state $s\in\mathcal{S}$, any $T>1$ and any $\varepsilon>0$, with probability of at least* $1-3\delta$

$$L_\varepsilon(T) \;\leq\; 34^2\frac{D^2S^2A\log\left(\frac{T}{\delta}\right)}{\varepsilon^2}.$$

Proof The proof is an adaptation of the proof of Theorem 2 which gives an upper bound of $O\Big(DS\sqrt{L_\varepsilon A\log(AT/\delta)}\Big)$ on the regret $\Delta'_\varepsilon(s,T)$ in ε-bad episodes in terms of L_ε. The theorem then follows due to $\varepsilon L_\varepsilon\leq\Delta'_\varepsilon(s,T)$.

Fix some $T>1$, and let K_ε and J_ε be two random sets that contain the indices of the ε-bad episodes up to step T and the corresponding time steps taken in these episodes, respectively. Then by an application of Hoeffding's inequality similar to (7) in Section 4.1 and a union bound over all possible values of L_ε, one obtains that with probability at least $1-\delta$,

$$\sum_{k\in K_\varepsilon}\sum_{t=t_k}^{t_{k+1}-1} r_t \;\geq\; \sum_{k\in K_\varepsilon}\sum_{s,a}v_k(s,a)\bar{r}(s,a)-\sqrt{2L_\varepsilon\log\left(\tfrac{T}{\delta}\right)}\;.$$

[10] In the following we use the same notation as in the proof of Theorem 2.

Further, by summing up all error probabilities $\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \leq \frac{\delta}{15t^6}$ for $t = 1, 2, \ldots$ one has

$$\mathbb{P}\left\{ \sum_{k \in K_\varepsilon} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} > 0 \right\} \leq \delta\,.$$

It follows that with probability at least $1 - 2\delta$

$$\Delta'_\varepsilon(s,T) \;\leq\; \sqrt{2L_\varepsilon \log\left(\tfrac{T}{\delta}\right)} + \sum_{k \in K_\varepsilon} \Delta_k \mathbb{1}_{M \in \mathcal{M}_k}. \tag{25}$$

In order to bound the regret of a single episode with $M \in \mathcal{M}_k$ we follow the lines of the proof of Theorem 2 in Section 4.3. Combining (15), (16), and (17) we have that

$$\Delta_k \;\leq\; \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k + \left(2D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)} + 2\right)\sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}\,. \tag{26}$$

In Appendix D we prove an analogon of (20), that is,

$$\sum_{k \in K_\varepsilon} \sum_{s,a} \frac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \;\leq\; \left(\sqrt{2}+1\right)\sqrt{L_\varepsilon SA}\,. \tag{27}$$

Then from (25), (26), and (27) it follows that with probability at least $1 - 2\delta$

$$\begin{aligned}\Delta'_\varepsilon(s,T) \quad \leq \quad & \sqrt{2L_\varepsilon \log\left(\tfrac{T}{\delta}\right)} + \left(2D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)} + 2\right)\cdot\left(\sqrt{2}+1\right)\cdot\sqrt{L_\varepsilon SA} \\ & + \sum_{k \in K_\varepsilon} \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \mathbb{1}_{M \in \mathcal{M}_k}\,. \end{aligned} \tag{28}$$

For the regret term of $\sum_{k \in K_\varepsilon} \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \mathbb{1}_{M \in \mathcal{M}_k}$ we use an argument similar to the one applied to obtain (18) in Section 4.3.2. Here we have to consider a slightly modified martingale difference sequence

$$X_t = \left(p\left(\cdot|s_t, a_t\right) - \boldsymbol{e}_{s_{t+1}}\right)\boldsymbol{w}_{k(t)} \mathbb{1}_{M \in \mathcal{M}_{k(t)}} \mathbb{1}_{t \in J_\varepsilon}$$

for $t = 1, \ldots, T$ to get (using the bound on the number of episodes given in Appendix C.2)

$$\begin{aligned}\sum_{k \in K_\varepsilon} \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \mathbb{1}_{M \in \mathcal{M}_k} \quad & \leq \quad \sum_{t \in J_\varepsilon} X_t + DSA\log_2\left(\tfrac{8T}{SA}\right) \\ & \leq \quad \sum_{t=1}^{T(L_\varepsilon)} X_t + DSA\log_2\left(\tfrac{8T}{SA}\right), \end{aligned} \tag{29}$$

where we set $T(L) := \min\left\{t : \#\{\tau \leq t, \tau \in J_\varepsilon\} = L\right\}$. The application of the Azuma-Hoeffding inequality in Section 4.3.2 is replaced with the following consequence of Bernstein's inequality for martingales.

Lemma 12 (Freedman 1975) *Let $X_1, X_2, \ldots$ be a martingale difference sequence. Then*

$$\mathbb{P}\left\{\sum_{i=1}^n X_i \geq \kappa,\ \sum_{i=1}^n X_i^2 \leq \gamma\right\} \;\leq\; \exp\left(-\frac{\kappa^2}{2\gamma + \frac{2\kappa}{3}}\right).$$

Application of Lemma 12 with $\kappa = 2D\sqrt{L\log(T/\delta)}$ and $\gamma = D^2L$ yields that for $L \ \geq \ \frac{\log(T/\delta)}{D^2}$ it holds that

$$\mathbb{P}\left\{\sum_{t=1}^{T(L)} X_t > 2D\sqrt{L\log\left(\tfrac{T}{\delta}\right)}\ \right\} < \frac{\delta}{T}. \tag{30}$$

On the other hand, if $L < \frac{\log(T/\delta)}{D^2}$, we have

$$\sum_{t=1}^{T(L)} X_t \ \leq \ DL \ = \ D\sqrt{L}\sqrt{L} \ < \ \sqrt{L}\sqrt{\log\left(\tfrac{T}{\delta}\right)} \ < \ 2D\sqrt{L\log\left(\tfrac{T}{\delta}\right)}. \tag{31}$$

Hence, (30) and (31) give by a union bound over all possible values of L_ε that with probability at least $1-\delta$

$$\sum_{t=1}^{T(L_\varepsilon)} X_t \ \leq \ 2D\sqrt{L_\varepsilon\log\left(\tfrac{T}{\delta}\right)}.$$

Together with (29) this yields that with probability at least $1-\delta$

$$\sum_{k\in K_\varepsilon} \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \mathbb{1}_{M\in\mathcal{M}_k} \ \leq \ 2D\sqrt{L_\varepsilon\log\left(\tfrac{T}{\delta}\right)} + DSA\log_2\left(\tfrac{8T}{SA}\right)\ .$$

Thus by (28) we obtain that with probability at least $1-3\delta$

$$\begin{aligned}\Delta'_\varepsilon(s,T) \quad \leq \quad & \sqrt{2L_\varepsilon\log\left(\tfrac{T}{\delta}\right)} + \left(2D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)} + 2\right)\cdot\left(\sqrt{2}+1\right)\cdot\sqrt{L_\varepsilon SA} \\ & + 2D\sqrt{L_\varepsilon\log\left(\tfrac{T}{\delta}\right)} + DSA\log_2\left(\tfrac{8T}{SA}\right)\ .\end{aligned}$$

This can be simplified to

$$\Delta'_\varepsilon(s,T) \leq 34DS\sqrt{L_\varepsilon A\log\left(\tfrac{T}{\delta}\right)} \tag{32}$$

by similar arguments as given in Appendix C.4. Since $\varepsilon L_\varepsilon \ \leq \ \Delta'_\varepsilon(s,T)$, we get

$$L_\varepsilon \leq 34^2\cdot\frac{D^2S^2A\log\left(\frac{T}{\delta}\right)}{\varepsilon^2}, \tag{33}$$

which proves the theorem. ■

Now we apply Theorem 11 to obtain the claimed logarithmic upper bound on the expected regret.

Proof of Theorem 4 Upper bounding L_ε in (32) by (33), we obtain for the regret $\Delta'_\varepsilon(s,T)$ accumulated in ε-bad episodes that

$$\Delta'_\varepsilon(s,T) \leq 34^2\cdot\frac{D^2S^2A\log\left(\frac{T}{\delta}\right)}{\varepsilon}$$

with probability at least $1-3\delta$. Noting that the regret accumulated outside of ε-bad episodes is at most εT implies the first statement of the theorem.

For the bound on the expected regret, first note that the expected regret of each episode in which an optimal policy is executed is at most D, whereas due to Theorem 11 the expected regret in $\frac{g}{2}$-bad episodes is upper bounded by $34^2 \cdot \frac{2 \cdot D^2 S^2 A \log(T)}{g} + 1$, as $\delta = \frac{1}{3T}$. What remains to do is to consider episodes k with expected average regret smaller than $\frac{g}{2}$ in which however a non-optimal policy $\tilde{\pi}_k$ was chosen.

First, note that for each policy π there is a T_π such that for all $T \geq T_\pi$ the expected average reward after T steps is $\frac{g}{2}$-close to the average reward of π. Thus, when a policy π is played in an episode of length $\geq T_\pi$ either the episode is $\frac{g}{2}$-bad (in expectation) or the policy π is optimal. Now we fix a state-action pair (s,a) and consider the episodes k in which the number of visits $v_k(s,a)$ in (s,a) is doubled. The corresponding episode lengths $\ell_k(s,a)$ are not necessarily increasing, but the $v_k(s,a)$ are monotonically increasing, and obviously $\ell_k(s,a) \geq v_k(s,a)$. Since the $v_k(s,a)$ are at least doubled, it takes at most $\lceil 1 + \log_2(\max_{\pi:\pi(s)=a} T_\pi) \rceil$ episodes until $\ell_k(s,a) \geq v_k(s,a) \geq \max_{\pi:\pi(s)=a} T_\pi$, when any policy π with $\pi(s) = a$ applied in episode k that is not $\frac{g}{2}$-bad (in expectation) will be optimal. Consequently, as only episodes of length smaller than $\max_{\pi:\pi(s)=a} T_\pi$ have to be considered, the regret of episodes k where $v_k(s,a) < \max_{\pi:\pi(s)=a} T_\pi$ is upper bounded by $\lceil 1 + \log_2(\max_{\pi:\pi(s)=a} T_\pi) \rceil \max_{\pi:\pi(s)=a} T_\pi$. Summing over all state-action pairs, we obtain an additional additive regret term of

$$\sum_{s,a} \left\lceil 1 + \log_2 \big(\max_{\pi:\pi(s)=a} T_\pi \big) \right\rceil \max_{\pi:\pi(s)=a} T_\pi,$$

which concludes the proof of the theorem. ■

6. The Lower Bound (Proof of Theorem 5)

We first consider the two-state MDP depicted in Figure 3. That is, there are two states, the initial state s_0 and another state s_1, and $A' = \left\lfloor \frac{A-1}{2} \right\rfloor$ actions. For each action a, let the deterministic rewards be $r(s_0, a) = 0$ and $r(s_1, a) = 1$. For all but a single "good" action a^* let $p(s_1|s_0, a) = \delta := \frac{4}{D}$, whereas $p(s_1|s_0, a^*) = \delta + \varepsilon$ for some $0 < \varepsilon < \delta$ specified later in the proof. Further, let $p(s_0|s_1, a) = \delta$ for all a. The diameter of this MDP is $D' = \frac{1}{\delta} = \frac{D}{4}$. For the rest of the proof we assume that[11] $\delta \leq \frac{1}{3}$.

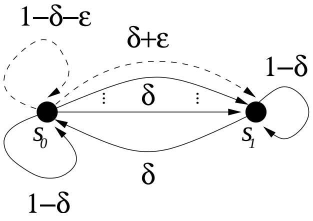

Figure 3: The MDP for the lower bound. The single action a^* with higher transition probability from state s_0 to state s_1 is shown as dashed line.

Consider $k := \left\lfloor \frac{S}{2} \right\rfloor$ copies of this MDP where only one of the copies has such a "good" action a^*. To complete the construction, we connect the k copies into a single MDP with diameter less than D,

[11] Otherwise we have $D < 12$, so that due to the made assumptions $A > 2S$. In this case we employ a different construction: Using $S-1$ actions, we connect all states to get an MDP with diameter 1. With the remaining $A-S+1$ actions we set up a bandit problem in each state as in the proof of the lower bound of Auer et al. (2002b) where only one state has a better action. This yields $\Omega(\sqrt{SAT})$ regret, which is sufficient, since D is bounded in this case.

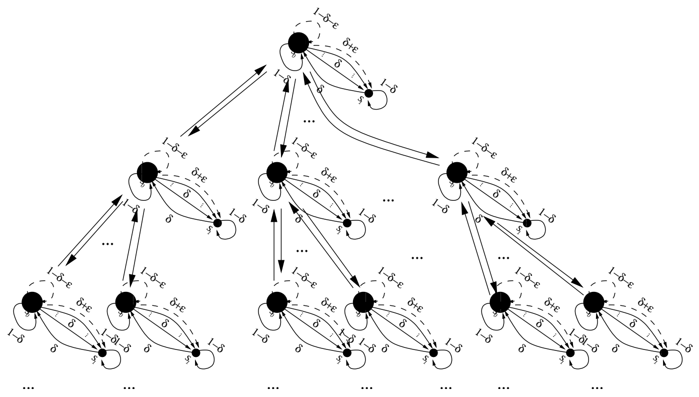

Figure 4: The composite MDP for the lower bound. Copies of the MDP of Figure 3 are arranged in an A'-ary tree, where the $s_\circ$-states are connected.

using at most $A-A'$ additional actions. This can be done by introducing $A'+1$ additional actions per state with deterministic transitions which do not leave the $s_{\shortmid}$-states and connect the $s_\circ$-states of the k copies by inducing an A'-ary tree structure on the $s_\circ$-states (one action for going toward the root, A' actions going toward the leaves—see Figure 4 for a schematic representation of the composite MDP). The reward for each of those actions is zero in any state. The diameter of the resulting MDP is at most $2(\frac{D}{4}+\lceil\log_{A'}k\rceil)$, which is twice the time it takes to travel to or from the root for any state in the MDP. Thus we have constructed an MDP M with $\leq S$ states, $\leq A$ actions, and diameter $\leq D$, for which we will show the claimed lower bound on the regret.

Actually, in the analysis we will consider the simpler MDP where all $s_\circ$-states are identified. We set this state to be the initial state. This MDP is equivalent to a single MDP M' like the one in Figure 3 with kA' actions which we assume in the following to be taken from $\{1,\ldots,kA'\}$. Note that learning this MDP is easier (as the learner is allowed to switch between different $s_\circ$-states without any cost for transition), while its optimal average reward is the same.

We prove the theorem by applying the same techniques as in the proof of the lower bound for the multi-armed bandit problem of Auer et al. (2002b). The pair $(s_\circ^*,a^*)$ identifying the copy with the better action and the better action is considered to be chosen uniformly at random from $\{1,\ldots,k\}\times\{1,\ldots,A'\}$, and we denote the expectation with respect to the random choice of $(s_\circ^*,a^*)$ as $\mathbb{E}_*[\cdot]$. We show that ε can be chosen such that M' and consequently also the composite MDP M forces regret $\mathbb{E}_*[\Delta(M,\mathfrak{A},s_\circ,T)]\geq\mathbb{E}_*[\Delta(M',\mathfrak{A},s_\circ^*,T)]>0.015\sqrt{D'kA'T}$ on any algorithm $\mathfrak{A}$.

We write $\mathbb{E}_{\text{unif}}[\cdot]$ for the expectation when there is no special action (i.e., the transition probability from $s_\circ$ to $s_{\shortmid}$ is δ for all actions), and $\mathbb{E}_a[\cdot]$ for the expectation conditioned on a being the special action a^* in M'. As already argued by Auer et al. (2002b), it is sufficient to consider deterministic strategies for choosing actions. Indeed, any randomized strategy is equivalent to an (apriori) random choice from the set of all deterministic strategies. Thus, we may assume that any algorithm $\mathfrak{A}$ maps the sequence of observations up to step t to an action a_t.

Now we follow the lines of the proof of Theorem A.2 as given by Auer et al. (2002b). Let the random variables $N_{\shortmid}$, $N_\circ$ and $N_\circ^*$ denote the total number of visits to state $s_{\shortmid}$, the total number of visits to state $s_\circ$, and the number of times action a^* is chosen in state $s_\circ$, respectively. Further, write s_t as usual for the state observed at step t. Then since $s_\circ$ is assumed to be the initial state, we have

$$\begin{aligned}
\mathbb{E}_a[N_{\shortmid}] &= \sum_{t=1}^{T}\mathbb{P}_a[s_t=s_{\shortmid}] = \sum_{t=2}^{T}\mathbb{P}_a[s_t=s_{\shortmid}] = \\
&= \sum_{t=2}^{T}\Big(\mathbb{P}_a[s_t=s_{\shortmid}|s_{t-1}=s_\circ]\,\mathbb{P}_a[s_{t-1}=s_\circ]+\mathbb{P}_a[s_t=s_{\shortmid}|s_{t-1}=s_{\shortmid}]\,\mathbb{P}_a[s_{t-1}=s_{\shortmid}]\Big) \\
&\leq \delta\sum_{t=2}^{T}\mathbb{P}_a[s_{t-1}=s_\circ,a_t\neq a^*]+(\delta+\varepsilon)\sum_{t=2}^{T}\mathbb{P}_a[s_{t-1}=s_\circ,a_t=a^*]+(1-\delta)\mathbb{E}_a[N_{\shortmid}] \\
&\leq \delta\mathbb{E}_a[N_\circ-N_\circ^*]+(\delta+\varepsilon)\mathbb{E}_a[N_\circ^*]+(1-\delta)\mathbb{E}_a[N_{\shortmid}]\,.
\end{aligned}$$

Taking into account that choosing a^* instead of any other action in $s_\circ$ reduces the probability of staying in state $s_\circ$, it follows that (using $D'=\frac{1}{\delta}$)

$$\begin{aligned}
\mathbb{E}_a\big[R(M',\mathfrak{A},s,T)\big] \leq \mathbb{E}_a[N_{\shortmid}] &\leq \mathbb{E}_a[N_\circ-N_\circ^*]+\tfrac{\delta+\varepsilon}{\delta}\mathbb{E}_a[N_\circ^*] \\
&= \mathbb{E}_a[N_\circ]+\mathbb{E}_a[N_\circ^*]\,\varepsilon D' \\
&\leq \mathbb{E}_{\mathrm{unif}}[N_\circ]+\mathbb{E}_a[N_\circ^*]\,\varepsilon D' \\
&= \mathbb{E}_{\mathrm{unif}}[T-N_{\shortmid}]+\mathbb{E}_a[N_\circ^*]\,\varepsilon D' \\
&= T-\mathbb{E}_{\mathrm{unif}}[N_{\shortmid}]+\mathbb{E}_a[N_\circ^*]\,\varepsilon D'. \qquad (34)
\end{aligned}$$

Now denoting the step where the first transition from $s_\circ$ to $s_{\shortmid}$ occurs by $\tau_{\circ\shortmid}$, we may lower bound $\mathbb{E}_{\mathrm{unif}}[N_{\shortmid}]$ by the law of total expectation as

$$\begin{aligned}
\mathbb{E}_{\mathrm{unif}}[N_{\shortmid}] &= \sum_{t=1}^{T}\mathbb{E}_{\mathrm{unif}}[N_{\shortmid}|\tau_{\circ\shortmid}=t]\,\mathbb{P}_{\mathrm{unif}}[\tau_{\circ\shortmid}=t] = \sum_{t=1}^{T}\mathbb{E}_{\mathrm{unif}}[N_{\shortmid}|\tau_{\circ\shortmid}=t]\,(1-\delta)^{t-1}\delta \\
&\geq \sum_{t=1}^{T-1}\frac{T-t}{2}(1-\delta)^{t-1}\delta = \frac{\delta T}{2}\sum_{t=1}^{T-1}(1-\delta)^{t-1}-\frac{\delta}{2}\sum_{t=1}^{T-1}t(1-\delta)^{t-1} \\
&= \frac{\delta T}{2}\cdot\frac{1-(1-\delta)^{T-1}}{\delta}-\frac{\delta}{2}\left(\frac{1-(1-\delta)^{T}}{\delta^2}-\frac{T(1-\delta)^{T-1}}{\delta}\right) \\
&= \frac{T-T(1-\delta)^{T-1}}{2}-\frac{1}{2\delta}+\frac{(1-\delta)^{T}}{2\delta}+\frac{T(1-\delta)^{T-1}}{2} \\
&= \frac{T}{2}-\frac{1}{2\delta}+\frac{(1-\delta)^{T}}{2\delta} \geq \frac{T}{2}-\frac{1}{2\delta} = \frac{T}{2}-\frac{D'}{2}. \qquad (35)
\end{aligned}$$

Therefore, combining (34) and (35) we obtain

$$\mathbb{E}_a\big[R(M',\mathfrak{A},s,T)\big] \leq \frac{T}{2}+\mathbb{E}_a[N_\circ^*]\,\varepsilon D'+\frac{D'}{2}. \qquad (36)$$

As $\mathfrak{A}$ chooses its actions deterministically based on the observations so far, $N_\circ^*$ is a function of the observations up to step T, too. A slight difference to Auer et al. (2002b) is that in our setting the sequence of observations consists not just of the rewards but also of the next state, that is, upon playing action a_t the algorithm observes s_{t+1} and r_t. However, since the immediate reward is fully determined by the current state, $N_\circ^*$ is also a function of just the state sequence, and we may bound $\mathbb{E}_a[N_\circ^*]$ by the following lemma, adapted from Auer et al. (2002b).

Lemma 13 *Let $f : \{s_\circ, s_{|}\}^{T+1} \to [0,B]$ be any function defined on state sequences $\boldsymbol{s} \in \{s_\circ, s_{|}\}^{T+1}$ observed in the MDP M'. Then for any $0 \leq \delta \leq \frac{1}{2}$, any $0 \leq \varepsilon \leq 1-2\delta$, and any $a \in \{1,\ldots,kA'\}$,*

$$\mathbb{E}_a\left[f(\boldsymbol{s})\right] \;\leq\; \mathbb{E}_{\mathrm{unif}}\left[f(\boldsymbol{s})\right] + \frac{B}{2}\cdot\frac{\varepsilon}{\sqrt{\delta}}\sqrt{2\mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right]}.$$

The proof of Lemma 13 is a straightforward modification of the respective proof given by Auer et al. (2002b). For details we refer to Appendix E.

Now let us assume that $\varepsilon \leq \delta$. (Our final choice of ε below will satisfy this requirement.) By our assumption of $\delta \leq \frac{1}{3}$ this yields that $\varepsilon \leq \delta \leq \frac{1}{3} \leq 1-2\delta$. Then, since $N_\circ^*$ is a function of the state sequence with $N_\circ^* \in [0,T]$, we may apply Lemma 13 to obtain

$$\mathbb{E}_a\left[N_\circ^*\right] \;\leq\; \mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right] + \frac{T}{2}\varepsilon\sqrt{D'}\sqrt{2\mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right]}. \tag{37}$$

An immediate consequence of (35) is that $\sum_{a=1}^{kA'}\mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right] \leq \frac{T}{2}+\frac{D'}{2}$, which yields by Jensen's inequality that $\sum_{a=1}^{kA'}\sqrt{2\mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right]} \leq \sqrt{kA'(T+D')}$. Thus we have from (37)

$$\begin{aligned}
\sum_{a=1}^{kA'}\mathbb{E}_a\left[N_\circ^*\right] &\leq \frac{T}{2}+\frac{D'}{2}+\frac{\varepsilon T}{2}\sqrt{D'}\sqrt{kA'(T+D')}\\
&\leq \frac{T}{2}+\frac{D'}{2}+\frac{\varepsilon T}{2}\sqrt{D'kA'T}+\frac{\varepsilon TD'}{2}\sqrt{kA'}.
\end{aligned}$$

Together with (36) this gives

$$\begin{aligned}
\mathbb{E}_*\left[R(M',\mathfrak{A},s,T)\right] &= \frac{1}{kA'}\sum_{a=1}^{kA'}\mathbb{E}_a\left[R(M,\mathfrak{A},s,T)\right]\\
&\leq \frac{T}{2}+\frac{\varepsilon TD'}{2kA'}+\frac{\varepsilon D'^2}{2kA'}+\frac{\varepsilon^2 TD'}{2kA'}\sqrt{D'kA'T}+\frac{\varepsilon^2 TD'^2}{2kA'}\sqrt{kA'}+\frac{D'}{2}.
\end{aligned}$$

Calculating the stationary distribution, we find that the optimal average reward for the MDP M' is $\frac{\delta+\varepsilon}{2\delta+\varepsilon}$. Hence, the expected regret with respect to the random choice of a^* is at least

$$\begin{aligned}
&\mathbb{E}_*\left[\Delta(M',\mathfrak{A},s,T)\right] = \frac{\delta+\varepsilon}{2\delta+\varepsilon}T - \mathbb{E}_*\left[R(M,\mathfrak{A},s,T)\right]\\
&\quad\geq \frac{\delta+\varepsilon}{2\delta+\varepsilon}T-\frac{T}{2}-\frac{\varepsilon TD'}{2kA'}-\frac{\varepsilon D'}{2kA'}\cdot D'-\frac{\varepsilon^2 TD'}{2kA'}\sqrt{D'kA'T}-\frac{\varepsilon^2 TD'}{2kA'}\sqrt{D'kA'}\cdot\sqrt{D'}-\frac{D'}{2}.
\end{aligned}$$

Since by assumption we have $T \geq DSA \geq 16D'kA'$ and thus $D' \leq \frac{T}{16kA'}$, it follows that

$$\begin{aligned}
&\mathbb{E}_*\left[\Delta(M',\mathfrak{A},s,T)\right]\\
&\quad\geq \frac{\varepsilon T}{4\delta+2\varepsilon}-\frac{\varepsilon TD'}{2kA'}-\frac{\varepsilon D'}{2kA'}\cdot\frac{T}{16kA'}-\frac{\varepsilon^2 TD'}{2kA'}\sqrt{D'kA'T}-\frac{\varepsilon^2 TD'}{2kA'}\sqrt{D'kA'}\sqrt{\frac{T}{16kA'}}-\frac{D'}{2}\\
&\quad= \frac{\varepsilon T}{4\delta+2\varepsilon}-\varepsilon TD'\left(\frac{1}{2kA'}+\frac{1}{32k^2A'^2}\right)-\frac{\varepsilon^2 TD'}{kA'}\sqrt{D'kA'T}\left(\frac{1}{2}+\frac{1}{8\sqrt{kA'}}\right)-\frac{D'}{2}.
\end{aligned}$$

Now we choose $\varepsilon := c\sqrt{\frac{kA'}{TD'}}$, where $c := \frac{1}{5}$. Then because of $\frac{1}{\delta} = D' \leq \frac{T}{16kA'}$ it follows that $\varepsilon \leq c\frac{1}{4D'} = \frac{\delta}{20}$ (so that also $\varepsilon \leq \delta$ as needed to get (37)), and further $\frac{1}{4\delta+2\varepsilon} \geq \frac{1}{4+1/8}D'$. Hence we obtain

$$\mathbb{E}_*\left[\Delta(M',\mathfrak{A},s,T)\right] \;\geq\; \left(\frac{c}{4+\frac{1}{8}} - \frac{c}{2kA'} - \frac{c}{32k^2A'^2} - \frac{c^2}{2} - \frac{c^2}{8\sqrt{kA'}}\right)\sqrt{D'kA'T} - \frac{D'}{2}.$$

Finally, we note that

$$\frac{D'}{2} \leq \tfrac{1}{2}\sqrt{D'}\sqrt{\frac{T}{16kA'}} = \frac{1}{8kA'}\sqrt{D'kA'T},$$

and since by assumption $S,A \geq 10$ so that $kA' \geq 20$, it follows that

$$\mathbb{E}_*\left[\Delta(M',\mathfrak{A},s,T)\right] \;>\; 0.015\sqrt{D'kA'T},$$

which concludes the proof. ■

7. Regret Bounds for Changing MDPs (Proof of Theorem 6)

Consider the learner operates in a setting where the MDP is allowed to change ℓ times, such that the diameter never exceeds D (we assume an initial change at time $t=1$). For this task we define the regret of an algorithm $\mathfrak{A}$ up to step T with respect to the average reward $\rho^*(t)$ of an optimal policy at step t as

$$\Delta'(\mathfrak{A},s,T) := \sum_{t=1}^{T} \rho^*(t) - r_t,$$

where r_t is as usual the reward received by $\mathfrak{A}$ at step t when starting in state s.

The intuition behind our approach is the following: When restarting UCRL2 every $\left(\frac{T}{\ell}\right)^{2/3}$ steps, the total regret for periods in which the MDP changes is at most $\ell^{1/3}T^{2/3}$. For each other period we have regret of $\tilde{O}\big((\frac{T}{\ell})^{1/3}\big)$ by Theorem 2. Since UCRL2 is restarted only $T^{1/3}\ell^{2/3}$ times, the total regret is $\tilde{O}\big(\ell^{1/3}T^{2/3}\big)$.

Because the horizon T is usually unknown, we use an alternative approach for restarting which however exhibits similar properties: UCRL2$'$ restarts UCRL2 with parameter $\frac{\delta}{\ell^2}$ at steps $\tau_i = \left\lceil \frac{i^3}{\ell^2} \right\rceil$ for $i = 1,2,3,\ldots$. Now we prove Theorem 6, which states that the regret of UCRL2$'$ is bounded by

$$\Delta'(\text{UCRL2}',s,T) \leq 65 \cdot \ell^{1/3}T^{2/3}DS\sqrt{A\log\left(\tfrac{T}{\delta}\right)}$$

with probability at least $1-\delta$ in the considered setting.

Let n be the largest natural number such that $\left\lceil \frac{n^3}{\ell^2} \right\rceil \leq T$, that is, n is the number of restarts up to step T. Then $\frac{n^3}{\ell^2} \leq \tau_n \leq T \leq \tau_{n+1} - 1 < \frac{(n+1)^3}{\ell^2}$ and consequently

$$\ell^{2/3}T^{1/3} - 1 \;\leq\; n \;\leq\; \ell^{2/3}T^{1/3}. \tag{38}$$

The regret Δ_c incurred due to changes of the MDP can be bounded by the number of steps taken in periods in which the MDP changes. This is maximized when the changes occur during the ℓ

longest periods, which contain at most $\tau_{n+1} - 1 - \tau_{n-\ell+1}$ steps. Hence we have

$$\begin{aligned}
\Delta_c &\leq \tau_{n+1} - 1 - \tau_{n-\ell+1} \\
&\leq \tfrac{1}{\ell^2}\left(n+1\right)^3 - \tfrac{1}{\ell^2} - \tfrac{1}{\ell^2}\left(n-\ell+1\right)^3 \\
&= 3\tfrac{n^2}{\ell} + 6\tfrac{n}{\ell} - 3n - \tfrac{1}{\ell^2} + \ell - 3 + 3\tfrac{1}{\ell}. \qquad (39)
\end{aligned}$$

For $\ell \geq 2$ we get by (39) and (38) that

$$\Delta_c \;\leq\; 3\,\frac{n^2}{\ell} + \ell \;\leq\; 3\,\frac{\ell^{4/3}T^{2/3}}{\ell} + \ell \;=\; 3\ell^{1/3}T^{2/3} + \ell,$$

while for $\ell = 1$ we obtain also from (39) and (38) that

$$\Delta_c \;\leq\; 3n^2 + 3n \;\leq\; 3T^{2/3} + 3T^{1/3}.$$

Thus the contribution to the regret from changes of the MDP is at most

$$\begin{aligned}
\Delta_c &\leq 3\ell^{1/3}T^{2/3} + 3T^{1/3} + \ell \\
&\leq 6\ell^{1/3}T^{2/3} + \ell^{1/3}\ell^{2/3} \\
&\leq 6\ell^{1/3}T^{2/3} + \ell^{1/3}T^{2/3} \\
&\leq 7\ell^{1/3}T^{2/3}\,. \qquad (40)
\end{aligned}$$

On the other hand, if the MDP does not change between the steps τ_i and $\min\{T, \tau_{i+1}\}$, the regret $\Delta(s_{\tau_i}, T_i)$ for these $T_i := \min\{T, \tau_{i+1}\} - \tau_i$ steps is bounded according to Theorem 2 (or more precisely (23)). Therefore, recalling that the confidence parameter is chosen to be $\frac{\delta}{\ell^2}$, this gives

$$\Delta(s_{\tau_i}, T_i) \;\leq\; 34DS\sqrt{T_i A \log\left(\tfrac{\ell^2 T_i}{\delta}\right)} \;\leq\; 34\sqrt{3}DS\sqrt{T_i}\sqrt{A\log\left(\tfrac{T}{\delta}\right)}$$

with probability $1 - \frac{\delta}{4\ell^2 T_i^{5/4}}$. As $\sum_{i=1}^n T_i = T$, we have by Jensen's inequality $\sum_{i=1}^n \sqrt{T_i} \leq \sqrt{n}\sqrt{T}$. Thus, summing over all $i = 1, \ldots, n$, the regret Δ_f in periods in which the MDP does not change is by (38)

$$\begin{aligned}
\Delta_f &\leq \sum_{i=1}^{n} \Delta(s_{\tau_i}, T_i) \;\leq\; 34\sqrt{3}DS\sqrt{n}\sqrt{T}\sqrt{A\log\left(\tfrac{T}{\delta}\right)} \\
&\leq 34\sqrt{3}DS\,\ell^{1/3}\,T^{2/3}\,\sqrt{A\log\left(\tfrac{T}{\delta}\right)} \qquad (41)
\end{aligned}$$

with probability at least $1 - \sum_{i=1}^n \frac{\delta}{4\ell^2 T_i^{5/4}}$. We conclude the proof by bounding this latter probability. For $\left\lfloor \frac{\ell^2}{3} \right\rfloor < i < n$,

$$\begin{aligned}
T_i &= \left\lceil \frac{(i+1)^3}{\ell^2} \right\rceil - \left\lceil \frac{i^3}{\ell^2} \right\rceil \\
&\geq \frac{(i+1)^3}{\ell^2} - \frac{i^3}{\ell^2} - \frac{\ell^2 - 1}{\ell^2} \\
&= \frac{3i^2}{\ell^2} + \frac{3i + 2 - \ell^2}{\ell^2} \;\geq\; \frac{3i^2}{\ell^2},
\end{aligned}$$

and consequently $\frac{1}{\ell^2 T_i^{5/4}} \leq \frac{1}{i^2}$. This together with $T_i \geq 1$ then yields

$$\begin{aligned} 1-\sum_{i=1}^{n}\frac{\delta}{4\ell^2 T_i^{5/4}} &\geq 1-\sum_{i=1}^{\lfloor \ell^2/3 \rfloor}\frac{\delta}{4\ell^2}-\sum_{i=\lfloor \ell^2/3 \rfloor+1}^{n-1}\frac{\delta}{4i^2}-\frac{\delta}{4\ell^2} \\ &> 1-\frac{\ell^2}{3}\cdot\frac{\delta}{4\ell^2}-\frac{\delta}{4}\sum_{i=1}^{\infty}\frac{1}{i^2}-\frac{\delta}{4} \\ &= 1-\frac{\delta}{3}-\frac{\delta}{4}\cdot\frac{\pi^2}{6} > 1-\delta. \end{aligned}$$

As $\Delta'(\text{UCRL2}', s, T) \leq \Delta_c + \Delta_f$, combining (40) and (41) yields

$$\Delta'(\text{UCRL2}', s, T) \leq 7\ell^{1/3}T^{2/3} + 34\sqrt{3}DS\,\ell^{1/3}\,T^{2/3}\sqrt{A\log\left(\tfrac{T}{\delta}\right)}$$

with probability at least $1-\delta$, and Theorem 6 follows, since the claimed bound holds trivially for $A\log\left(\frac{T}{\delta}\right) < \log 4$. ■

8. Open Problems

There is still a gap between the upper bound on the regret of Theorem 2 and the lower bound of Theorem 5. We conjecture that the lower bound gives the right exponents for the parameters S and D (concerning the dependence on S compare also the sample complexity bounds of Strehl et al., 2006). The recent research of Bartlett and Tewari (2009) also poses the question whether the diameter in our bounds can be replaced by a smaller parameter, that is, by the span of the bias of an optimal policy. As the algorithm REGAL.C of Bartlett and Tewari (2009) demonstrates, this is at least true when this value is known to the learner. However, in the case of ignorance, currently this replacement of the diameter D can only be achieved at the cost of an additional factor of $\sqrt{S}$ in the regret bounds (Bartlett and Tewari, 2009). The difficulty in the proof is that while the span of an optimal policy's bias vector in the *assumed* optimistic MDP can be upper bounded by the diameter of the *true* MDP (cf. Remark 8), it is not clear how the spans of optimal policies in the assumed and the true MDP relate to each other.

A somehow related question is that of *transient* states, that is, the possibility that some of the states are not reachable under any policy. In this case the diameter is unbounded, so that our bounds become vacuous. Indeed, our algorithm cannot handle transient states: for any time step and any transient state s, UCRL2 optimistically assumes maximal possible reward in s and a very small but still positive transition probability to s from any other state. Thus insisting on the possibility of a transition to s, the algorithm fails to detect an optimal policy.[12] The assumption of having an upper bound on an optimal policy's bias resolves this problem, as this bound indirectly also gives some information on what the learner may expect from a state that has not been reached so far and thus may be transient. Consequently, with the assumed knowledge of such an upper bound, the REGAL.C algorithm of Bartlett and Tewari (2009) is also able to deal with transient states.

[12] Actually, one can modify UCRL2 to deal with transient states by assuming transition probability 0 for all transitions not observed so far. This is complemented by an additional exploration phase between episodes where, for example, the state-action pair with the fewest number of visits is probed. While this algorithm gives asymptotically the same bounds, these however contain a large additive constant for all the episodes that occur before the transition structure assumed by the algorithm is correct.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work was supported in part by the Austrian Science Fund FWF (S9104-N13 SP4). The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 216886 (PASCAL2 Network of Excellence), and n° 216529 (Personal Information Navigator Adapting Through Viewing, PinView), and the Austrian Federal Ministry of Science and Research. This publication only reflects the authors' views.

Appendix A. A Lower Bound on the Diameter

We are going to show a more general result, from which the bound on the diameter follows. For a given MDP, let $T^*(s|s_0)$ be the minimal expected time it takes to move from state s_0 to state s.

Theorem 14 *Consider an MDP with state space $\mathcal{S}$ and A states. Let d_0 be an arbitrary distribution over $\mathcal{S}$, and $\mathcal{U} \subseteq \mathcal{S}$ be any subset of states. Then the sum of the minimal expected transition times to states in $\mathcal{U}$ when starting in an initial state distributed according to d_0 is bounded as follows:*

$$\mathcal{T}(\mathcal{U}|d_0) := \sum_{s\in\mathcal{U}} \sum_{s_0\in\mathcal{S}} d_0(s_0)\, T^*(s|s_0) \;\geq\; \min_{\substack{0\leq n_k\leq A^k, k\geq 0,\\ \sum_k n_k = |\mathcal{U}|}} \sum_k k\cdot n_k .$$

We think this bound is tight. The minimum on the right hand side is attained when the n_k are maximized for small k until $|\mathcal{U}|$ is exhausted. For $A \geq 2$, this gives an average (over the states in $\mathcal{U}$) expected transition time of at least $\log_A |\mathcal{U}| - 3$ to states in $\mathcal{U}$. Indeed, for $|\mathcal{U}| = \sum_{k=0}^{m-1} A^k + n_m$ we have $\frac{A^{m+1}-A}{(A-1)^2} < |\mathcal{U}|\big(1+\frac{1}{A-1}\big)$ as well as $m \geq \log_A\big(\frac{|\mathcal{U}|}{2}\big)$, so that

$$\begin{aligned}
\sum_{k=0}^{m-1} kA^k + m\cdot n_m &= m|\mathcal{U}| + \sum_{k=0}^{m-1}(k-m)A^k \\
&= m|\mathcal{U}| + \frac{m}{A-1} - \frac{A^{m+1}-A}{(A-1)^2} \\
&> |\mathcal{U}|\big(m-1-\tfrac{1}{A-1}\big) \\
&\geq |\mathcal{U}|\left(\log_A\left(\tfrac{|\mathcal{U}|}{2}\right) - 1 - \tfrac{1}{A-1}\right) \\
&\geq |\mathcal{U}|\left(\log_A|\mathcal{U}| - 3\right).
\end{aligned}$$

In particular, choosing $\mathcal{U} = \mathcal{S}$ gives the claimed lower bound on the diameter.

Corollary 15 *In any MDP with S states and $A \geq 2$ actions, the diameter D is lower bounded by $\log_A S - 3$.*

Remark 16 *For given S, A the minimal diameter is not always assumed by an MDP with deterministic transitions. Consider for example $S = 4$ and $A = 2$. Any deterministic MDP with four states and two actions has diameter at least 2. However, Figure 5 shows a corresponding MDP whose diameter is $\frac{3}{2}$.*

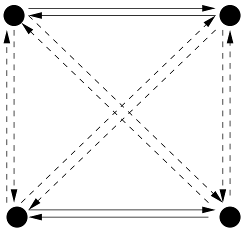

Figure 5: An MDP with four states and two actions whose diameter is $\frac{3}{2}$. In each state two actions are available. One action leads to another state deterministically, while the other action causes a random transition to each of the two other states with probability $\frac{1}{2}$ (indicated as dashed lines).

Proof of Theorem 14 Let $a^*(s_0,s)$ be the optimal action in state s_0 for reaching state s, and let $p(s|s_0,a)$ be the transition probability to state s when choosing action a in state s_0.

The proof is by induction on the size of $\mathcal{U}$. For $|\mathcal{U}|=0,1$ the statement holds.

For $|\mathcal{U}|>1$ we have

$$\begin{aligned}
\mathcal{T}(\mathcal{U}|d_0) &= \sum_{s_0\in\mathcal{S}}\sum_{s\in\mathcal{U}} d_0(s_0)\,T^*(s|s_0)\\
&= \sum_{s_0\in\mathcal{S}}\sum_{s\in\mathcal{U}\setminus\{s_0\}} d_0(s_0)\,T^*(s|s_0)\\
&= \sum_{s_0\in\mathcal{S}}\sum_{s\in\mathcal{U}\setminus\{s_0\}} d_0(s_0)\Big(\sum_{s_1\in\mathcal{S}} p\big(s_1|s_0,a^*(s_0,s)\big)\,T^*(s|s_1)+1\Big)\\
&= \sum_{s_0\in\mathcal{S}} d_0(s_0)\sum_a \sum_{\substack{s\in\mathcal{U}\setminus\{s_0\},\\ a^*(s_0,s)=a}}\Big(\sum_{s_1\in\mathcal{S}} p(s_1|s_0,a)\,T^*(s|s_1)+1\Big)\\
&= \sum_{s_0\in\mathcal{S}} d_0(s_0)\sum_a\Big(\sum_{s_1\in\mathcal{S}}\sum_{s\in\mathcal{U}_{s_0,a}} p(s_1|s_0,a)\,T^*(s|s_1)+|\mathcal{U}_{s_0,a}|\Big),
\end{aligned}$$

where $\mathcal{U}_{s_0,a} := \big\{s\in\mathcal{U}\setminus\{s_0\} : a^*(s_0,s)=a\big\}$.

If all $\mathcal{U}_{s_0,a}\subset\mathcal{U}$, we apply the induction hypothesis and obtain for suitable $n_k(s_0,a)$

$$\begin{aligned}
&\sum_{s_0\in\mathcal{S}} d_0(s_0)\sum_a\Big(\sum_{s_1\in\mathcal{S}}\sum_{s\in\mathcal{U}_{s_0,a}} p(s_1|s_0,a)\,T^*(s|s_1)+|\mathcal{U}_{s_0,a}|\Big)\\
&\quad\geq \sum_{s_0\in\mathcal{S}} d_0(s_0)\sum_a\Big(\sum_k k\cdot n_k(s_0,a)+|\mathcal{U}_{s_0,a}|\Big)\\
&\quad= \sum_{s_0\in\mathcal{S}} d_0(s_0)\sum_a\sum_k (k+1)\cdot n_k(s_0,a),
\end{aligned}$$

since $\sum_k n_k(s_0,a)=|\mathcal{U}_{s_0,a}|$. Furthermore, $n_k(s_0,a)\leq A^k$ and $|\mathcal{U}|-1\leq\sum_a|\mathcal{U}_{s_0,a}|\leq|\mathcal{U}|$. Thus setting $n'_k=\sum_{s_0}d_0(s_0)\sum_a n_{k-1}(s_0,a)$ for $k\geq1$ and $n'_0=|\mathcal{U}|-\sum_{k\geq1}n'_k$ satisfies the conditions of the statement. This completes the induction step for this case.

If $\mathcal{U}_{s_0,a} = \mathcal{U}$ for some pair (s_0,a) (i.e., for all target states $s \in \mathcal{U}$ the same action is optimal in s_0), then we construct a modified MDP with shorter transition times. This is achieved by modifying one of the actions to give a deterministic transition from s_0 to some state in $\mathcal{U}$ (which is not reached deterministically by choosing action a). For the modified MDP the induction step works and the lower bound can be proven, which then also holds for the original MDP. ■

Appendix B. Convergence of Value Iteration (Proof of Theorem 7)

As sufficient condition for convergence of value iteration, Puterman (1994) assumes only that all optimal policies have aperiodic transition matrices. Actually, the proof of Theorem 9.4.4 of Puterman (1994)—the main result on convergence of value iteration—needs this assumption only at one step, that is, to guarantee that the optimal policy identified at the end of the proof has aperiodic transition matrix. In the following we give a proof sketch of Theorem 7 that concentrates on the differences to the convergence proof given by Puterman (1994).

Lemma 9.4.3 of Puterman (1994) shows that value iteration eventually chooses only policies π that satisfy $\boldsymbol{P}_\pi \boldsymbol{\rho}^* = \boldsymbol{\rho}^*$, where $\boldsymbol{P}_\pi$ is the transition matrix of π and $\boldsymbol{\rho}^*$ is the optimal average reward vector. More precisely, there is an i_0 such that for all $i \geq i_0$

$$\max_{\pi:\mathcal{S}\to\mathcal{A}} \{\boldsymbol{r}_\pi + \boldsymbol{P}_\pi \boldsymbol{u}_i\} = \max_{\pi\in E} \{\boldsymbol{r}_\pi + \boldsymbol{P}_\pi \boldsymbol{u}_i\},$$

where $\boldsymbol{r}_\pi$ is the reward vector of the policy π, and $E := \{\pi : \mathcal{S} \to \mathcal{A} \,|\, \boldsymbol{P}_\pi \boldsymbol{\rho}^* = \boldsymbol{\rho}^*\}$.

Unlike standard value iteration, extended value iteration always chooses policies with aperiodic transition matrix (cf. the discussion in Section 3.1.3). Thus when considering only aperiodic policies $F := \{\pi : \mathcal{S} \to \mathcal{A} |\, \boldsymbol{P}_\pi \text{ is aperiodic}\}$ in the proof of Lemma 9.4.3, the same argument shows that there is an i_0' such that for all $i \geq i_0'$

$$\max_{\pi\in F} \{\boldsymbol{r}_\pi + \boldsymbol{P}_\pi \boldsymbol{u}_i\} = \max_{\pi\in E\cap F} \{\boldsymbol{r}_\pi + \boldsymbol{P}_\pi \boldsymbol{u}_i\}. \tag{42}$$

Intuitively, (42) shows that extended value iteration eventually chooses only policies from $E \cap F$.

With (42) accomplished, the proof of Theorem 9.4.4, the main result on convergence of value iteration, can be rewritten word by word from Puterman (1994), with E replaced with $E \cap F$ and using (42) instead of Lemma 9.4.3. Thus, unlike in the original proof where the optimal policy π^* identified at the end of the proof is in E, in our case π^* is in $E \cap F$. Here Puterman (1994) uses the assumption that *all* optimal policies have aperiodic transition matrices to guarantee that π^* has aperiodic transition matrix. In our case, π^* has aperiodic transition matrix by definition, as it is in $E \cap F$.

Then by the aperiodicity of $\boldsymbol{P}_{\pi^*}$, the result of Theorem 9.4.4 follows, and one obtains analogously to Theorem 9.4.5 (a) of Puterman (1994) that

$$\lim_{i\to\infty} (\boldsymbol{u}_{i+1} - \boldsymbol{u}_i) = \boldsymbol{\rho}^*. \tag{43}$$

As the underlying MDP $\tilde{M}^+$ is assumed to be communicating (so that ρ^* is state-independent), analogously to Corollary 9.4.6 of Puterman (1994) convergence of extended value iteration follows from (43). Finally, with the convergence of extended value iteration established, the error bound for the greedy policy follows from Theorem 8.5.6 of Puterman (1994). ■

Appendix C. Technical Details for the Proof of Theorem 2

This appendix collects some technical details, starting with an error bound for our confidence intervals.

C.1 Confidence Intervals

Lemma 17 *For any $t \geq 1$, the probability that the true MDP M is not contained in the set of plausible MDPs $\mathcal{M}(t)$ at time t (as given by the confidence intervals in* (3) *and* (4)*) is at most $\frac{\delta}{15t^6}$, that is*

$$\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} < \frac{\delta}{15t^6}.$$

Proof Consider a fixed state-action pair (s,a) and assume some given number of visits $n > 0$ in (s,a) before step t. Denote the estimates for transition probabilities and rewards obtained from these n observations by $\hat{p}(\cdot|s,a)$ and $\hat{r}(s,a)$, respectively. Let us first consider the probability with which a confidence interval for the transition probabilities fails. The random event observed for the transition probability estimates is the state to which the transition occurs. Generally, the L^1-deviation of the true distribution and the empirical distribution over m distinct events from n samples is bounded according to Weissman et al. (2003) by

$$\mathbb{P}\Big\{ \big\|\hat{p}(\cdot) - p(\cdot)\big\|_1 \geq \varepsilon \Big\} \leq (2^m - 2)\exp\left(-\tfrac{n\varepsilon^2}{2}\right). \tag{44}$$

Thus, in our case we have $m = S$ (for each possible transition there is a respective event), so that setting

$$\varepsilon = \sqrt{\frac{2}{n}\log\left(\tfrac{2^S 20SAt^7}{\delta}\right)} \leq \sqrt{\frac{14S}{n}\log\left(\tfrac{2At}{\delta}\right)},$$

we get from (44)

$$\begin{aligned}\mathbb{P}\left\{\Big\|p(\cdot|s,a) - \hat{p}(\cdot|s,a)\Big\|_1 \geq \sqrt{\frac{14S}{n}\log\left(\tfrac{2At}{\delta}\right)}\right\} \quad &\leq \quad 2^S \exp\left(-\frac{n}{2}\cdot\frac{2}{n}\log\left(\tfrac{2^S 20SAt^7}{\delta}\right)\right) \\ &= \quad \frac{\delta}{20t^7 SA}.\end{aligned}$$

For the rewards we observe real-valued, independent identically distributed (i.i.d.) random variables with support in $[0,1]$. Hoeffding's inequality gives for the deviation between the true mean $\bar{r}$ and the empirical mean $\hat{r}$ from n i.i.d. samples with support in $[0,1]$

$$\mathbb{P}\Big\{ \big|\hat{r} - \bar{r}\big| \geq \varepsilon_r \Big\} \leq 2\exp\Big(-2n\varepsilon_r^2\Big).$$

Setting

$$\varepsilon_r = \sqrt{\frac{1}{2n}\log\left(\tfrac{120SAt^7}{\delta}\right)} \leq \sqrt{\frac{7}{2n}\log\left(\tfrac{2SAt}{\delta}\right)},$$

we get for state-action pair (s,a)

$$\begin{aligned}\mathbb{P}\left\{\big|\hat{r}(s,a) - \bar{r}(s,a)\big| \geq \sqrt{\frac{7}{2n}\log\left(\tfrac{2SAt}{\delta}\right)}\right\} \quad &\leq \quad 2\exp\left(-2n\cdot\frac{1}{2n}\log\left(\tfrac{120SAt^7}{\delta}\right)\right) \\ &= \quad \frac{\delta}{60t^7 SA}.\end{aligned}$$

Note that when there haven't been any observations, the confidence intervals trivially hold with probability 1 (for transition probabilities as well as for rewards). Hence a union bound over all possible values of $n = 1, \ldots, t-1$ gives (now writing $N(s,a)$ for the number of visits in (s,a))

$$\mathbb{P}\left\{ \left|\hat{r}(s,a) - \bar{r}(s,a)\right| \geq \sqrt{\frac{7\log\left(\frac{2SAt}{\delta}\right)}{2\max\{1,N(s,a)\}}} \right\} \quad \leq \quad \sum_{n=1}^{t-1} \frac{\delta}{60t^7SA} < \frac{\delta}{60t^6SA} \quad \text{and}$$

$$\mathbb{P}\left\{ \left\|p(\cdot|s,a) - \hat{p}(\cdot|s,a)\right\|_1 \geq \sqrt{\frac{14S\log\left(\frac{2At}{\delta}\right)}{\max\{1,N(s,a)\}}} \right\} \quad \leq \quad \sum_{n=1}^{t-1} \frac{\delta}{20t^7SA} < \frac{\delta}{20t^6SA}.$$

Summing these error probabilities over all state-action pairs we obtain the claimed bound $\mathbb{P}\{M \notin \mathcal{M}(t)\} < \frac{\delta}{15t^6}$. ■

C.2 A Bound on the Number of Episodes

Since in each episode the total number of visits to at least one state-action pair doubles, the number of episodes m is logarithmic in T. Actually, the number of episodes becomes maximal when all state-action pairs are visited equally often, which results in the following bound.

Proposition 18 *The number m of episodes of* UCRL2 *up to step $T \geq SA$ is upper bounded as*

$$m \leq SA\log_2\left(\tfrac{8T}{SA}\right).$$

Proof Let $N(s,a) := \#\{\tau < T+1 : s_\tau = s, a_\tau = a\}$ be the total number of observations of the state-action pair (s,a) up to step T. In each episode $k < m$ there is a state-action pair (s,a) with $v_k(s,a) = N_k(s,a)$ (or $v_k(s,a) = 1$, $N_k(s,a) = 0$). Let $K(s,a)$ be the number of episodes with $v_k(s,a) = N_k(s,a)$ and $N_k(s,a) > 0$. If $N(s,a) > 0$, then $v_k(s,a) = N_k(s,a)$ implies $N_{k+1}(s,a) = 2N_k(s,a)$, so that

$$N(s,a) \;=\; \sum_{k=1}^{m} v_k(s,a) \;\geq\; 1 + \sum_{k:v_k(s,a)=N_k(s,a)} N_k(s,a) \;\geq\; 1 + \sum_{i=1}^{K(s,a)} 2^{i-1} \;=\; 2^{K(s,a)}.$$

On the other hand, if $N(s,a) = 0$, then obviously $K(s,a) = 0$, so that generally, $N(s,a) \geq 2^{K(s,a)} - 1$ for any state-action pair (s,a). It follows that

$$T \;=\; \sum_{s,a} N(s,a) \geq \sum_{s,a}\left(2^{K(s,a)} - 1\right). \tag{45}$$

Now, in each episode a state-action pair (s,a) is visited for which either $N_k(s,a) = 0$ or $N_k(s,a) = v_k(s,a)$. Hence, $m \leq 1 + SA + \sum_{s,a} K(s,a)$, or equivalently $\sum_{s,a} K(s,a) \geq m - 1 - SA$. This implies

$$\sum_{s,a} 2^{K(s,a)} \;\geq\; SA\, 2^{\sum_{s,a} K(s,a)/SA} \;\geq\; SA\, 2^{\frac{m-1}{SA} - 1}.$$

Together with (45) this gives

$$T \geq SA\left(2^{\frac{m-1}{SA}-1} - 1\right),$$

which yields

$$m \leq 1 + 2SA + SA\log_2\left(\tfrac{T}{SA}\right),$$

and the claimed bound on m follows for $T \geq SA$. ■

C.3 The Sum in (19)

Lemma 19 *For any sequence of numbers* $z_1, \dots, z_n$ *with* $0 \le z_k \le Z_{k-1} := \max\left\{1, \sum_{i=1}^{k-1} z_i\right\}$

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \le \left(\sqrt{2}+1\right)\sqrt{Z_n}\,.$$

Proof We prove the statement by induction over n.
Base case: We first show that the lemma holds for all n with $\sum_{k=1}^{n-1} z_k \le 1$. Indeed, in this case $Z_k = 1$ for $k \le n-1$ and hence $z_n \le 1$. It follows that

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} = \sum_{k=1}^{n-1} z_k + z_n \le 1+1 < \left(\sqrt{2}+1\right) Z_n.$$

Note that this also shows that the lemma holds for $n = 1$, since $\sum_{k=1}^{0} z_k = 0 \le 1$.
Inductive step: Now let us consider natural numbers n such that $\sum_{k=1}^{n-1} z_k > 1$. By the induction hypothesis we have

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \le \left(\sqrt{2}+1\right)\sqrt{Z_{n-1}} + \frac{z_n}{\sqrt{Z_{n-1}}}\,.$$

Since $z_n \le Z_{n-1} = \sum_{k=1}^{n-1} z_k$ and $Z_{n-1} + z_n = Z_n$, we further have

$$\begin{aligned}
\left(\sqrt{2}+1\right)\sqrt{Z_{n-1}} + \frac{z_n}{\sqrt{Z_{n-1}}} &= \sqrt{\left(\sqrt{2}+1\right)^2 Z_{n-1} + 2\left(\sqrt{2}+1\right) z_n + \frac{z_n^2}{Z_{n-1}}} \\
&\le \sqrt{\left(\sqrt{2}+1\right)^2 Z_{n-1} + \left(2+2\sqrt{2}+1\right) z_n} \\
&= \sqrt{\left(\sqrt{2}+1\right)^2 Z_{n-1} + \left(\sqrt{2}+1\right)^2 z_n} \\
&= \left(\sqrt{2}+1\right)\sqrt{Z_{n-1}+z_n} = \left(\sqrt{2}+1\right)\sqrt{Z_n}\,,
\end{aligned}$$

which proves the lemma. ■

C.4 Simplifying (22)

Combining similar terms, (22) yields that with probability at least $1 - \frac{\delta}{4T^{5/4}}$

$$\begin{aligned}
\Delta(s_1,T) \;\le\; & DS\sqrt{AT}\left(\tfrac{3}{2}\sqrt{\tfrac{1}{A}\cdot\tfrac{5}{2}\log\left(\tfrac{8T}{\delta}\right)} + 2\left(\sqrt{2}+1\right)\sqrt{14\log\left(\tfrac{2AT}{\delta}\right)} + \sqrt{8} + 2 + \tfrac{1}{\sqrt{A}}\right) \\
& + DSA\log_2\left(\tfrac{8T}{SA}\right). \qquad (46)
\end{aligned}$$

We assume $A \ge 2$, since the bound is trivial otherwise. Also, for $1 < T \le 34^2 A\log\left(\frac{T}{\delta}\right)$ we have $\Delta(s_1,T) \le 34\sqrt{AT\log\left(\frac{T}{\delta}\right)}$ trivially. Considering $T > 34A\log\left(\frac{T}{\delta}\right)$ we have $A < \frac{1}{34\log\left(\frac{T}{\delta}\right)}\sqrt{AT\log\left(\frac{T}{\delta}\right)}$ and also $\log_2(8T) < 2\log(T)$, so that

$$DSA\log_2\left(\tfrac{8T}{SA}\right) \;<\; \tfrac{2}{34}DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}\,.$$

Further, $T > 34A\log\left(\frac{T}{\delta}\right)$ also implies $\log\left(\frac{2AT}{\delta}\right) \leq 2\log\left(\frac{T}{\delta}\right)$ and $\log\left(\frac{8T}{\delta}\right) \leq 2\log\left(\frac{T}{\delta}\right)$. Thus, we have by (46) that for any $T > 1$ with probability at least $1 - \frac{\delta}{4T^{5/4}}$

$$\begin{aligned}\Delta(s_1,T) &\leq DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}\left(\tfrac{3}{2}\sqrt{\tfrac{5}{2}}+2\left(\sqrt{2}+1\right)\sqrt{28}+\sqrt{8}+2+\tfrac{1}{\sqrt{2}}+\tfrac{2}{34}\right)\\ &\leq 34DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}.\end{aligned}$$

Appendix D. Technical Details for the Proof of Theorem 4: Proof of (27)

For a given index set K_ε of episodes we would like to bound the sum

$$\sum_{k\in K_\varepsilon}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} = \sum_{s,a}\sum_{k=1}^{m}\frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\mathbb{1}_{k\in K_\varepsilon}.$$

We will do this by modifying the sum so that Lemma 19 becomes applicable. Compared to the setting of Lemma 19 there are some "gaps" in the sum caused by episodes $\notin K_\varepsilon$. In the following we show that the contribution of episodes that occur after step $L_\varepsilon := \sum_{k\in K_\varepsilon}\sum_{s,a} v_k(s,a)$ is not larger than the missing contributions of the episodes $\notin K_\varepsilon$. Intuitively speaking, one may fill the episodes that occur after step L_ε into the gaps of episodes $\notin K_\varepsilon$ as Figure 6 suggests.

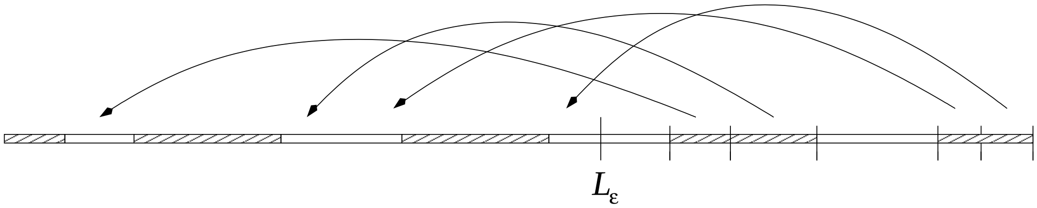

Figure 6: Illustration of the proof idea. Shaded boxes stand for episodes $\in K_\varepsilon$, empty boxes for episodes $\notin K_\varepsilon$. The contribution of episodes after step L_ε can be "filled into the gaps" of episodes $\notin K_\varepsilon$ before step L_ε.

Let $\ell_\varepsilon(s,a) := \sum_{k\in K_\varepsilon} v_k(s,a)$, so that $\sum_{s,a}\ell_\varepsilon(s,a) = L_\varepsilon$. We consider a fixed state-action pair (s,a) and skip the reference to it for ease of reading, so that N_k refers to the number of visits to (s,a) up to episode k, and N denotes the total number of visits to (s,a). Further, we abbreviate $d_k := \sqrt{\max\{1,N_k\}}$, and let $m_\varepsilon := \max\{k : N_k < \ell_\varepsilon\}$ be the episode containing the ℓ_ε-th visit to (s,a). Due to $v_k = N_{k+1} - N_k$ we have

$$v_{m_\varepsilon} = (N_{m_\varepsilon+1} - \ell_\varepsilon) + (\ell_\varepsilon - N_{m_\varepsilon}). \tag{47}$$

Since $N_{m_\varepsilon} = \sum_{k=1}^{m_\varepsilon-1} v_k$, this yields

$$\begin{aligned}\ell_\varepsilon - N_{m_\varepsilon} + \sum_{k=1}^{m_\varepsilon-1} v_k &= \ell_\varepsilon = \sum_{k=1}^{m} v_k \mathbb{1}_{k\in K_\varepsilon}\\ &= \sum_{k=1}^{m_\varepsilon-1} v_k\mathbb{1}_{k\in K_\varepsilon} + \left(N_{m_\varepsilon+1}-\ell_\varepsilon\right)\mathbb{1}_{m_\varepsilon\in K_\varepsilon} + \left(\ell_\varepsilon - N_{m_\varepsilon}\right)\mathbb{1}_{m_\varepsilon\in K_\varepsilon} + \sum_{k=m_\varepsilon+1}^{m} v_k\mathbb{1}_{k\in K_\varepsilon},\end{aligned}$$

or equivalently,

$$\big(N_{m_\varepsilon+1}-\ell_\varepsilon\big)\mathbb{1}_{m_\varepsilon\in K_\varepsilon}+\sum_{k=m_\varepsilon+1}^{m} v_k\mathbb{1}_{k\in K_\varepsilon} \;=\; \big(\ell_\varepsilon-N_{m_\varepsilon}\big)\mathbb{1}_{m_\varepsilon\notin K_\varepsilon}+\sum_{k=1}^{m_\varepsilon-1} v_k\mathbb{1}_{k\notin K_\varepsilon}. \tag{48}$$

By (47) and due to $d_k\geq d_{m_\varepsilon}$ for $k\geq m_\varepsilon$ we have

$$\begin{aligned}\sum_{k=1}^{m}\frac{v_k}{d_k}\mathbb{1}_{k\in K_\varepsilon} \;\leq\;& \sum_{k=1}^{m_\varepsilon-1}\frac{v_k}{d_k}\mathbb{1}_{k\in K_\varepsilon}+\frac{\ell_\varepsilon-N_{m_\varepsilon}}{d_{m_\varepsilon}}\mathbb{1}_{m_\varepsilon\in K_\varepsilon}\\ &+\frac{1}{d_{m_\varepsilon}}\bigg(\big(N_{m_\varepsilon+1}-\ell_\varepsilon\big)\mathbb{1}_{m_\varepsilon\in K_\varepsilon}+\sum_{k=m_\varepsilon+1}^{m} v_k\mathbb{1}_{k\in K_\varepsilon}\bigg).\end{aligned}$$

Hence, we get together with (48), using that $d_k\leq d_{m_\varepsilon}$ for $k\leq m_\varepsilon$

$$\begin{aligned}\sum_{k=1}^{m}\frac{v_k}{d_k}\mathbb{1}_{k\in K_\varepsilon} \;\leq\;& \sum_{k=1}^{m_\varepsilon-1}\frac{v_k}{d_k}\mathbb{1}_{k\in K_\varepsilon}+\frac{\ell_\varepsilon-N_{m_\varepsilon}}{d_{m_\varepsilon}}\mathbb{1}_{m_\varepsilon\in K_\varepsilon}\\ &+\frac{1}{d_{m_\varepsilon}}\bigg(\big(\ell_\varepsilon-N_{m_\varepsilon}\big)\mathbb{1}_{m_\varepsilon\notin K_\varepsilon}+\sum_{k=1}^{m_\varepsilon-1} v_k\mathbb{1}_{k\notin K_\varepsilon}\bigg)\\ \leq\;& \sum_{k=1}^{m_\varepsilon-1}\frac{v_k}{d_k}\mathbb{1}_{k\in K_\varepsilon}+\frac{\ell_\varepsilon-N_{m_\varepsilon}}{d_{m_\varepsilon}}\mathbb{1}_{m_\varepsilon\in K_\varepsilon}+\frac{\ell_\varepsilon-N_{m_\varepsilon}}{d_{m_\varepsilon}}\mathbb{1}_{m_\varepsilon\notin K_\varepsilon}+\sum_{k=1}^{m_\varepsilon-1}\frac{v_k}{d_k}\mathbb{1}_{k\notin K_\varepsilon}\\ =\;& \sum_{k=1}^{m_\varepsilon-1}\frac{v_k}{d_k}+\frac{\ell_\varepsilon-N_{m_\varepsilon}}{d_{m_\varepsilon}}.\end{aligned}$$

Now define v'_k as follows: let $v'_k:=v_k$ for $k<m_\varepsilon$ and $v'_{m_\varepsilon}:=\ell_\varepsilon-N_{m_\varepsilon}$. Then we have just seen that

$$\sum_{k=1}^{m}\frac{v_k}{d_k}\mathbb{1}_{k\in K_\varepsilon}\;\leq\;\sum_{k=1}^{m_\varepsilon}\frac{v'_k}{d_k}.$$

Since further $\sum_{k=1}^{m_\varepsilon} v'_k=\ell_\varepsilon$ we get by Lemma 19 that

$$\sum_{k=1}^{m_\varepsilon}\frac{v'_k}{d_k}\;\leq\;\Big(\sqrt{2}+1\Big)\sqrt{\ell_\varepsilon}\,.$$

As $\sum_{s,a}\ell_\varepsilon(s,a)=L_\varepsilon$, we finally obtain by Jensen's inequality

$$\sum_{k\in K_\varepsilon}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\;\leq\;\Big(\sqrt{2}+1\Big)\sqrt{L_\varepsilon SA}\,,$$

as claimed. ■

Appendix E. Proof of Lemma 13

Let us first recall some notation from Section 6. Thus $\mathbb{P}_a[\cdot]$ denotes the probability conditioned on a being the "good" action, while the probability with respect to a setting where all actions in state $s_\circ$ are equivalent (i.e., $\varepsilon=0$) is denoted by $\mathbb{P}_{\mathrm{unif}}[\cdot]$. Let $\mathcal{S}:=\{s_\circ,s_\shortmid\}$ and denote the state observed at step τ by s_τ and the state-sequence up to step τ by $\boldsymbol{s}^\tau = s_1, \dots, s_\tau$. Basically, the proof follows along the lines of the proof of Lemma A.1 of Auer et al. (2002b). The first difference is that our observations now consist of the sequence of $T+1$ states instead of a sequence of T observed rewards. Still it is straightforward to get analogously to the proof of Auer et al. (2002b), borrowing the notation, that for any function f from $\{s_\circ, s_{\shortmid}\}^{T+1}$ to $[0,B]$,

$$\mathbb{E}_a\left[f(\boldsymbol{s})\right] - \mathbb{E}_{\mathrm{unif}}\left[f(\boldsymbol{s})\right] \;\le\; \frac{B}{2}\sqrt{2\log(2)\,\mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}}\middle\|\mathbb{P}_a\right)}, \tag{49}$$

where $\mathrm{KL}(P\|Q)$ denotes for two distributions P,Q the *Kullback-Leibler divergence* defined as $\mathrm{KL}(P\|Q) := \sum_{\boldsymbol{s}\in\mathcal{S}^{T+1}} P\{\boldsymbol{s}\}\log_2\left(\frac{P\{\boldsymbol{s}\}}{Q\{\boldsymbol{s}\}}\right)$. It holds that (cf. Auer et al., 2002b)

$$\mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}}\middle\|\mathbb{P}_a\right) = \sum_{t=1}^{T}\mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\middle\|\mathbb{P}_a\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\right), \tag{50}$$

where $\mathrm{KL}(P\{s_{t+1}|\boldsymbol{s}^t\}\|Q\{s_{t+1}|\boldsymbol{s}^t\}) := \sum_{\boldsymbol{s}^{t+1}\in\mathcal{S}^{t+1}} P\{\boldsymbol{s}^{t+1}\}\log_2\left(\frac{P\{s_{t+1}|\boldsymbol{s}^t\}}{Q\{s_{t+1}|\boldsymbol{s}^t\}}\right)$. By the Markov property and the fact that the action a_t is determined by a sequence $\boldsymbol{s}^t \in \mathcal{S}^t$ we have (similar to Auer et al., 2002b)

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\middle\|\mathbb{P}_a\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\right) = \sum_{\boldsymbol{s}^{t+1}\in\mathcal{S}^{t+1}}\mathbb{P}_{\mathrm{unif}}\left[\boldsymbol{s}^{t+1}\right]\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}|\boldsymbol{s}^t\right]}{\mathbb{P}_a\left[s_{t+1}|\boldsymbol{s}^t\right]}\right)\\
&= \sum_{\boldsymbol{s}^t\in\mathcal{S}^t}\mathbb{P}_{\mathrm{unif}}\left[\boldsymbol{s}^t\right]\sum_{s'\in\mathcal{S}}\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}=s'|\boldsymbol{s}^t\right]\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}\left[s'|\boldsymbol{s}^t\right]}{\mathbb{P}_a\left[s'|\boldsymbol{s}^t\right]}\right)\\
&= \sum_{\boldsymbol{s}^{t-1}\in\mathcal{S}^{t-1}}\mathbb{P}_{\mathrm{unif}}\left[\boldsymbol{s}^{t-1}\right]\sum_{(s'',a')\in\mathcal{S}\times\mathcal{A}}\mathbb{P}_{\mathrm{unif}}\left[s_t=s'',a_t=a'|\boldsymbol{s}^{t-1}\right]\\
&\qquad\cdot\sum_{s'\in\mathcal{S}}\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}=s'|\boldsymbol{s}^{t-1},s_t=s'',a_t=a'\right]\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}\left[s'|\boldsymbol{s}^{t-1},s_t=s'',a_t=a'\right]}{\mathbb{P}_a\left[s'|\boldsymbol{s}^{t-1},s_t=s'',a_t=a'\right]}\right)\\
&= \sum_{\boldsymbol{s}^{t-1}\in\mathcal{S}^{t-1}}\mathbb{P}_{\mathrm{unif}}\left[\boldsymbol{s}^{t-1}\right]\sum_{a'=1}^{kA'}\sum_{s''\in\mathcal{S}}\mathbb{P}_{\mathrm{unif}}\left[s_t=s'',a_t=a'|\boldsymbol{s}^{t-1}\right]\\
&\qquad\qquad\qquad\cdot\sum_{s'\in\mathcal{S}}\mathbb{P}_{\mathrm{unif}}\left[s'|s'',a'\right]\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}\left[s'|s'',a'\right]}{\mathbb{P}_a\left[s'|s'',a'\right]}\right).
\end{aligned}$$

Since $\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}[s'|s'',a']}{\mathbb{P}_a[s'|s'',a']}\right) \neq 0$ only for $s'' = s_\circ$ and a' being the special action a, we get

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\middle\|\mathbb{P}_a\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\right) =\\
&= \sum_{\boldsymbol{s}^{t-1}\in\mathcal{S}^{t-1}}\mathbb{P}_{\mathrm{unif}}\left[\boldsymbol{s}^{t-1}\right]\mathbb{P}_{\mathrm{unif}}\left[s_t=s_\circ,a_t=a|\boldsymbol{s}^{t-1}\right]\cdot\sum_{s'\in\mathcal{S}}\mathbb{P}_{\mathrm{unif}}\left[s'|s_\circ,a\right]\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}\left[s'|s_\circ,a\right]}{\mathbb{P}_a\left[s'|s_\circ,a\right]}\right)\\
&= \mathbb{P}_{\mathrm{unif}}\left[s_t=s_\circ,a_t=a\right]\sum_{s'\in\mathcal{S}}\mathbb{P}_{\mathrm{unif}}\left[s'|s_\circ,a\right]\log_2\left(\frac{\mathbb{P}_{\mathrm{unif}}\left[s'|s_\circ,a\right]}{\mathbb{P}_a\left[s'|s_\circ,a\right]}\right)\\
&= \mathbb{P}_{\mathrm{unif}}\left[s_t=s_\circ,a_t=a\right]\left(\delta\log_2\left(\frac{\delta}{\delta+\varepsilon}\right)+(1-\delta)\log_2\left(\frac{1-\delta}{1-\delta-\varepsilon}\right)\right).
\end{aligned}\tag{51}$$

To complete the proof we use the following lemma.

Lemma 20 *For any* $0 \leq \delta \leq \frac{1}{2}$ *and* $\varepsilon \leq 1-2\delta$ *we have*

$$\delta \log_2\left(\frac{\delta}{\delta+\varepsilon}\right) + (1-\delta)\log_2\left(\frac{1-\delta}{1-\delta-\varepsilon}\right) \;\leq\; \frac{\varepsilon^2}{\delta\log(2)}\,.$$

Indeed, application of Lemma 20 together with (50) and (51) gives that

$$\begin{aligned} \mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}} \middle\| \mathbb{P}_a\right) &= \sum_{t=1}^{T} \mathrm{KL}\left(\mathbb{P}_{\mathrm{unif}}\left[s_{t+1}\middle|\boldsymbol{s}^t\right] \middle\| \mathbb{P}_a\left[s_{t+1}\middle|\boldsymbol{s}^t\right]\right) \\ &\leq \sum_{t=1}^{T} \mathbb{P}_{\mathrm{unif}}\left[s_t = s_\circ, a_t = a\right] \frac{\varepsilon^2}{\delta\log(2)} \;=\; \mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right] \frac{\varepsilon^2}{\delta\log(2)}, \end{aligned}$$

which together with (49) yields

$$\mathbb{E}_a\left[f(\boldsymbol{s})\right] - \mathbb{E}_{\mathrm{unif}}\left[f(\boldsymbol{s})\right] \leq \frac{B}{2}\cdot\frac{\varepsilon}{\sqrt{\delta}}\sqrt{2\mathbb{E}_{\mathrm{unif}}\left[N_\circ^*\right]},$$

as claimed by Lemma 13.

Proof of Lemma 20 Consider

$$h_\delta(\varepsilon) := \frac{\varepsilon^2}{\delta} - \delta\log\left(\frac{\delta}{\delta+\varepsilon}\right) - (1-\delta)\log\left(\frac{1-\delta}{1-\delta-\varepsilon}\right).$$

We show that $h_\delta(\varepsilon) \geq 0$ for $\delta \leq \frac{1}{2}$ and $0 \leq \varepsilon \leq \varepsilon_0$, where

$$\varepsilon_0 := \frac{1}{2} - \delta + \frac{1}{2}\sqrt{1-2\delta}.$$

Indeed, $h_\delta(0) = 0$ for all δ, while for the first derivative

$$h'_\delta(\varepsilon) := \frac{\partial}{\partial\varepsilon}h_\delta(\varepsilon) \;=\; 2\cdot\frac{\varepsilon}{\delta} + \frac{\delta}{\delta+\varepsilon} - \frac{1-\delta}{1-\delta-\varepsilon}$$

we have $h'_\delta(\varepsilon) \geq 0$ for $\delta \leq \frac{1}{2}$ and $0 \leq \varepsilon \leq \varepsilon_0$. It remains to show that $\delta \leq \frac{1}{2}$ and $\varepsilon \leq 1-2\delta$ imply $\varepsilon \leq \varepsilon_0$. Indeed, for $\delta \leq \frac{1}{2}$ and $\varepsilon \leq 1-2\delta$ we have

$$\varepsilon - \varepsilon_0 \;\leq\; 1-2\delta-\varepsilon_0 = \frac{1}{2} - \delta - \frac{1}{2}\sqrt{1-2\delta} \;=\; \frac{1}{2}\left((1-2\delta) - \sqrt{1-2\delta}\right) \;\leq\; 0.$$

■

References

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 49–56. MIT Press, 2007.

Peter Auer and Ronald Ortner. Online regret bounds for a new reinforcement learning algorithm. In *Proceedings 1st Austrian Cognitive Vision Workshop (ACVW 2005)*, pages 35–42. ÖCG, 2005.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47:235–256, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32:48–77, 2002b.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, 2009.

Ronen I. Brafman and Moshe Tennenholtz. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for Markov decision processes. *Math. Oper. Res.*, 22(1):222–255, 1997.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Experts in a Markov decision process. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 401–408. MIT Press, 2005.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online Markov decision processes. *Math. Oper. Res.*, 34(3):726–736, 2009.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory (COLT 1994)*, pages 88–97. ACM, 1994.

David A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3:100–118, 1975.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. Preprint, 2008. URL `http://arxiv.org/pdf/0805.3415`.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

Sham M. Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49:209–232, 2002.

Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 11 (NIPS 1998)*, pages 996–1002. MIT Press, 1999.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.*, 5:623–648, 2004.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

Alexander L. Strehl and Michael L. Littman. A theoretical analysis of model-based interval estimation. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005)*, pages 857–864. ACM, 2005.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for Markov decision processes. *J. Comput. System Sci.*, 74(8):1309–1331, 2008.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006)*, pages 881–888. ACM, 2006.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1505–1512. MIT Press, 2008.

Ambuj Tewari and Peter L. Bartlett. Bounded parameter Markov decision processes with average reward criterion. In *Learning Theory, 20th Annual Conference on Learning Theory (COLT 2007)*, pages 263–277, 2007.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marco L. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97, HP Laboratories Palo Alto, 2003. URL `www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf`.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009)*, pages 1177–1184, 2009.

Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Math. Oper. Res.*, 34(3):737–757, 2009.