NEAR-OPTIMAL REGRET BOUNDS FOR REINFORCEMENT LEARNING

Peter Auer, Thomas Jaksch, Ronald Ortner

January 2009
University of Leoben
Chair of Information Technology
Technical Report No. CIT-2009-01

NEAR-OPTIMAL REGRET BOUNDS FOR REINFORCEMENT LEARNING

Peter Auer, Thomas Jaksch, Ronald Ortner

University of Leoben
Chair of Information Technology
Franz Joseph Straße 18
A-8700 Leoben
Austria
www.unileoben.ac.at/~infotech

January 2009

Technical Report No. CIT-2009-01

Acknowledgments

This work was supported in part by the Austrian Science Fund FWF (S9104-N13 SP4). The research leading to these results has 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). This publication only reflects the authors' views.

Summary

This technical report is an extended version of [1].

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. This bound holds with high probability. We also present a corresponding lower bound of $\Omega(\sqrt{DSAT})$ on the total regret of any learning algorithm.

Contents

List of Figures

1 Introduction

In a Markov decision process (MDP) M with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$, a learner in 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 with mean $\bar{r}(s,a)$ according to some distribution on $[0,1]$. 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. [2, 3] and also the discussion and references given in the introduction of [4]) — 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.

In this paper we will consider *undiscounted* rewards. The *accumulated reward* of an algorithm $\mathfrak{A}$ after T steps in an MDP M is defined as

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

where s is the initial state and r_t are the rewards received during the execution of algorithm $\mathfrak{A}$. The *average reward*

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

can be maximized by an appropriate stationary *policy* $\pi : \mathcal{S} \to \mathcal{A}$ which defines an optimal action for each state [5].

The difficulty of learning an MDP does not only depend on its 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 this pair of states s and s':

Definition 1. *Let $T(s'|M, \pi, s)$ be the first (random) time step in which state s' is reached when policy π is executed on MDP M with initial state s. Then the* diameter *of M is given by*

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

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 about D steps to reach again a "good part" of the MDP. Hence, the learner may suffer regret D for such exploration, and it is very plausible that the diameter appears in the regret bound.

For MDPs with finite diameter (which usually are called *communicating*, see e.g. [5]) the optimal average reward ρ^* does not depend on the initial state (cf. [5], 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 for a learning algorithm $\mathfrak{A}$, and we define the *total regret* of $\mathfrak{A}$ after T steps as[1]

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

In the following, we present our reinforcement learning algorithm UCRL2 (a variant of the UCRL algorithm of [6]) 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. Further, the diameter seems to be more natural than other parameters that have been proposed for various PAC and regret bounds, such as the *mixing time* [4, 7] or the *hitting time* of an optimal policy [8] (cf. the discussion below).

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

1.1 Relation to previous Work

We first compare our results to the PAC bounds for the well-known algorithms E^3 of Kearns, Singh [4], and R-Max of Brafman, Tennenholtz [7] (see also Kakade [9]). 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 $1/\varepsilon^3$, 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. 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 $T_\varepsilon^{\text{mix}} \approx D/\varepsilon$. This additional dependency 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 [4] and [7] are substantially larger than in our bound.

The MBIE algorithm of Strehl and Littman [10, 11] — similarly to our approach — applies confidence bounds to compute an optimistic policy. However, Strehl and Littman consider only a discounted reward setting, which seems to be less natural when dealing with regret. Their definition of regret measures the difference between the rewards[2] 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*.

Tewari and Bartlett [8] propose a generalization of the *index policies* of Burnetas and Katehakis [12]. 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 [12] and the OLP algorithm of [8] 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 [8] and [12] 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.

2 Results

We summarize the results achieved for our algorithm UCRL2 which is 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 parameter δ.

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

$$\Delta(M, \text{UCRL2}, s, T) \;\le\; c_1 \cdot DS\sqrt{TA \log \tfrac{T}{\delta}},$$

for a constant c_1 which is independent of M, T, and δ.

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

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

$$T \;\ge\; c_2 \frac{D^2 S^2 A}{\varepsilon^2} \log\left(\frac{DSA}{\delta\varepsilon}\right)$$

steps, where c_2 is a constant independent of the MDP.

Theorem 4. *For any initial state $s \in \mathcal{S}$, any $T \ge 1$ and any $\varepsilon > 0$ the expected regret of* UCRL2 *(with parameter $\delta := 1/(3T)$) is*

$$\mathbb{E}\left[\Delta(M, \text{UCRL2}, s, T)\right] < c_3 \frac{D^2 S^2 A \log(T)}{g},$$

[2] Actually, the state values.

where $g := \rho^*(M) - \max_{\pi,s}\{\rho(M,\pi,s) : \rho(M,\pi,s) < \rho^*(M)\}$ *is the gap between the optimal average reward and the second best average reward achievable in* M*, and* c_3 *is an MDP independent constant. Using the doubling trick to set the parameter* δ*, the same bound can be achieved without knowledge of the horizon* T*.*

These new bounds are improvements over the bounds that have been achieved in [6] 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 in [6], 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 some* $c_4 > 0$*, any algorithm* $\mathfrak{A}$*, and any natural numbers* $S, A \geq 10$*,* $D \geq 20 \log_A S$*, and* $T \geq DSA$*, there is an MDP*[3] M *with* S *states,* A *actions, and diameter* D*, such that for any initial state* $s \in \mathcal{S}$ *the expected regret of* $\mathfrak{A}$ *after* T *steps is*

$$\mathbb{E}\left[\Delta(M, \mathfrak{A}, s, T)\right] \geq c_4 \cdot \sqrt{DSAT}\,.$$

In a different setting, a modification of UCRL2 can also deal with changing MDPs.

Theorem 6. *Assume that the MDP (i.e. its transition probabilities and reward distributions) is allowed to change* ℓ *times up to step* T*, such that the diameter is always at most* D *(we assume an initial change at time* $t = 1$*). Restarting* UCRL2 *with parameter* δ/ℓ^2 *at steps* $\lceil i^3/\ell^2 \rceil$ *for* $i = 1, 2, 3 \ldots$*, the regret measured as the sum of missed rewards compared to the* ℓ *policies which are optimal after the changes of the MDP is upper bounded by*

$$c_5 \cdot \ell^{\frac{1}{3}} T^{\frac{2}{3}} DS \sqrt{A \log \tfrac{T}{\delta}}$$

with probability $1 - \delta$ *for an MDP independent constant* c_5*.*

MDPs with a different model of changing rewards have already been considered in [13]. There, the transition probabilities are assumed to be fixed and known to the learner, but the rewards are allowed to change in every step. A best possible upper bound of $O(\sqrt{T})$ on the regret against an optimal stationary policy, given all the reward changes in advance, is derived.

3 The UCRL2 Algorithm

Our algorithm is a variant of the UCRL algorithm in [6]. As its predecessor, UCRL2 implements the paradigm of "optimism in the face of uncertainty". As such, 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 (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 $\hat{p}_k(s'|s,a)$ and $\hat{r}_k(s,a)$ for the transition probabilities and mean rewards 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.[4]

[3] The diameter of any MDP with S states and A actions is at least $\log_A S$.

[4] 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 $v_k(s,a)$.

Input: A confidence parameter $\delta \in (0,1)$.

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,
$$R_k(s,a) := \sum_{\tau=1}^{t_k-1} r_\tau \mathbb{1}_{s_\tau = s, a_\tau = a},$$
$$P_k(s,a,s') := \#\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\},$$
and 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,
$$\left|\tilde{r}(s,a) - \hat{r}_k(s,a)\right| \leq \sqrt{\frac{7\log(2SAt_k/\delta)}{2\max\{1, N_k(s,a)\}}} \quad \text{and} \tag{1}$$
$$\left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \leq \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1, N_k(s,a)\}}}. \tag{2}$$
5. Use extended value iteration (Section 3.1) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that
$$\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} \rho(M', \pi, s') - \frac{1}{\sqrt{t_k}}. \tag{3}$$

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.

3.1 Extended Value Iteration

In Step 5 of the UCRL2 algorithm we need to find a near-optimal policy $\tilde{\pi}_k$ for an optimistic MDP. While value iteration typically calculates a policy for a fixed MDP, we also need to select an optimistic MDP $\tilde{M}_k$ which gives almost maximal reward among all plausible MDPs. This can be achieved by extending value iteration to search also among the plausible MDPs. Formally, this can be seen as undiscounted value iteration [5] on an MDP with extended action set. Consider an MDP $\tilde{M}^+$ with continuous action space, where each action identifies the original action, an admissible transition probability distribution and mean reward. 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 $\tilde{M}^+$.

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

$$\begin{aligned} u_0(s) &= 0, \\ u_{i+1}(s) &= \max_{a \in \mathcal{A}} \left\{ \tilde{r}_k(s,a) + \max_{p(\cdot) \in \mathcal{P}(s,a)} \left\{ \sum_{s' \in \mathcal{S}} p(s') \cdot u_i(s') \right\} \right\}, \end{aligned} \tag{4}$$

where $\tilde{r}_k(s,a)$ are the maximal rewards satisfying condition (1) in algorithm UCRL2, and $\mathcal{P}(s,a)$ is the set of transition probabilities $\tilde{p}(\cdot|s,a)$ satisfying condition (2).

While (4) 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 (4) is equivalent to value iteration on an MDP $\tilde{M}'$ with finite action set, since only the finite number of vertices of the polytope need to be considered as extended actions. Now, after computing a sequence of states sorted according to the values $\boldsymbol{u}_i$ in an iteration once, the inner maximum can be computed in $O(S)$ computation steps for each state-action pair (s,a) as follows.

The idea is to put as much transition probability as possible to the state with maximal value at the expense of transition probabilities to states with small values. That is, we start with setting $\boldsymbol{p} := \hat{p}(\cdot|s,a)$. Then we modify $\boldsymbol{p}$ by putting as much transition probability as possible from s to the state s' with maximal $u_i(s')$, i.e. $p(s') := \min\{1, \hat{p}(s'|s,a) + d(s,a)/2\}$ where $d(s,a)$ denotes the confidence interval as given in (2). In order to make $\boldsymbol{p}$ correspond to a probability distribution again, we have to reduce the transition probabilities from s to states s'' with small $u_i(s'')$ in sum by $d(s,a)/2$ as well (so that $\|\boldsymbol{p} - \hat{p}(\cdot|s,a)\|_1 = d(s,a)$). More precisely, this is done iteratively as follows: We first choose the s'' with minimal $u_i(s'')$ among those with $p(s'') > 0$. Then we reduce the transition probability from s to s'' by as much as possible or as much as necessary, respectively, i.e. $p(s'') := \max\{0, 1 - \sum_{s' \neq s''} p(s'|s,a)\}$. This is repeated until $\boldsymbol{p}$ is a probability distribution. Updating $\sum_{s' \in \mathcal{S}} p(s'|s,a)$ with every change of $\boldsymbol{p}$ for the computation of $\sum_{s' \neq s''} p(s'|s,a)$, this iterative procedure takes $O(S)$ steps. Thus computing a sorting sequence once per iteration, each iteration can be done with at most $O(S^2 A)$ computation steps.

Further, by using a fixed sorting sequence throughout an iteration, in each iteration there is some single fixed state s' which is regarded as the "best" target state. Then for each state s, in the inner maximum an action with positive transition probability to s' will be chosen. Hence, the corresponding policy is aperiodic and has state independent average reward. The aperiodicity of these policies, together with the fact that $\tilde{M}'$ is communicating, guarantees convergence of $\boldsymbol{u}_i$. This can be seen by inspecting Lemma 9.4.3 in [5], and noting that it can be restated with the set of policies E restricted to contain only aperiodic policies. Then, even though not all optimal policies are guaranteed to be aperiodic in our setting, Theorem 9.4.4. in [5] can be proved for our setting as well like in [5], since only the aperiodicity of policies in E is required in the proof. Then also Theorem 9.4.5. and Corollary 9.4.6 in [5] hold, which implies convergence of (4), since $\tilde{M}'$ is communicating. The value iteration is stopped when

$$\max_{s \in S} \{u_{i+1}(s) - u_i(s)\} - \min_{s \in S} \{u_{i+1}(s) - u_i(s)\} < \frac{1}{\sqrt{t_k}}, \tag{5}$$

which means that by Theorem 8.5.6. in [5] the greedy policy with respect to $\boldsymbol{u}_i$ is $\frac{1}{\sqrt{t_k}}$-optimal.

4 Analysis of UCRL2 (Proof of Theorem 2)

We start with a rough outline of the proof. First, in Section 4.1, the random fluctuation of the rewards is dealt with, and the regret is expressed as the sum of the regret accumulated in the individual episodes. That is, setting the *regret in episode* k to be

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

it is shown that the total regret can be bounded by

$$\sum_k \Delta_k + \sqrt{\tfrac{5}{2} T \log \tfrac{8T}{\delta}}.$$

In Section 4.2 we consider the regret that is caused by failing confidence regions. We show that this regret can be upper bounded by $\sqrt{T}$ with high probability.

After this intermezzo, the regret of episodes for which $M \in \mathcal{M}_k$ is analyzed 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 in that episode, and $\boldsymbol{w}_k$ is a vector with $\|\boldsymbol{w}_k\|_\infty \le D(M)$. The last two terms in the above expression stem from the reward confidence intervals and the approximation error of value iteration. 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 in episode k. Substituting for $\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k$ the lengths of the confidence intervals, 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 a high probability bound of $\sum_k \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \le D\sqrt{\tfrac{5}{2} T \log \tfrac{8T}{\delta}} + Dm$ concludes Section 4.3.2, where m is the number of episodes, which is shown to be logarithmic in T in Appendix A.2. Section 4.3.3 concludes the analysis of episodes with $M \in \mathcal{M}_k$ by summing the individual regret terms over all episode 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 to 4.3 and employing some further simplifications.

4.1 Splitting into Episodes

Let r_t be the (random) reward UCRL2 received at step t when started in state s_1. For given state-action counts $N(s,a)$ after T steps, the r_t are independent random variables, so that by Chernoff bounds

$$\mathbb{P}\left\{ \sum_{t=1}^{T} r_t \le \sum_{(s,a)} N(s,a)\bar{r}(s,a) - \sqrt{2T \cdot \tfrac{5}{4}\log\tfrac{8T}{\delta}} \;\middle|\; \big(N(s,a)\big)_{s,a} \right\} < \frac{\delta}{12T^{5/4}}. \tag{6}$$

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

$$\Delta(s_1,T) = T\rho^* - \sum_{t=1}^{T} r_t \;<\; T\rho^* - \sum_{(s,a)} N(s,a)\bar{r}(s,a) + \sqrt{\tfrac{5}{2} T \log \tfrac{8T}{\delta}}$$

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

$$\Delta(s_1,T) \;\le\; \sum_{k=1}^{m} \Delta_k + \sqrt{\tfrac{5}{2} T \log \tfrac{8T}{\delta}} \tag{7}$$

with probability $1 - \frac{\delta}{12T^{5/4}}$, as $\sum_{k=1}^{m} v_k(s,a) = N(s,a)$ and $\sum_{(s,a)} N(s,a) = T$.

4.2 Dealing with Failing Confidence Regions

We consider the regret of episodes in which the set of plausible MDPs does not contain the true MDP, $\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k}$. By the stopping criterion for episode k we have (except for episodes

where $v_k(s,a) = 1$ and $N_k(s,a) = 0$)

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

Hence, since $\rho^* \le 1$

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

where $\mathcal{M}(t)$ is the set of plausible MDPs as given by (1) and (2), using the estimates available at step t. Now, $\mathbb{P}\{M \notin \mathcal{M}(t)\} \le \frac{\delta}{15t^6}$ (see Appendix A.1), and since

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

it follows that $\mathbb{P}\{\exists t : T^{1/4} < t \le T : M \notin \mathcal{M}(t)\} \le \frac{\delta}{12T^{5/4}}$,

Thus, with probability at least $1 - \frac{\delta}{12T^{5/4}}$ it holds that

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} \;\le\; \sqrt{T} \tag{8}$$

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

We assume $M \in \mathcal{M}_k$ 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 $\tilde{\pi}_k$. Since $M \in \mathcal{M}_k$, and by the choice of $\tilde{\pi}_k$ and $\tilde{M}_k$ in Step 5 of UCRL2, $\tilde{\rho}_k \ge \rho^* - 1/\sqrt{t_k}$. Thus for the regret Δ_k during episode k we get

$$\Delta_k \le \sum_{(s,a)} v_k(s,a)\big(\rho^* - \bar{r}(s,a)\big) \;\le \sum_{(s,a)} v_k(s,a)\big(\tilde{\rho}_k - \bar{r}(s,a)\big) + \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{t_k}}\,. \tag{9}$$

4.3.1 Extended Value Iteration revisited

To proceed, we reconsider the extended value iteration in 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,

$$\max_s u_i(s) - \min_s u_i(s) \le D. \tag{10}$$

To see this, observe that $u_i(s)$ is the total expected reward after i steps of an optimal non-stationary i-step policy starting in state s, on the MDP with extended action set as considered for the extended value iteration. The diameter of this extended MDP is at most D as it contains the actions of the true MDP M. If there were states with $u_i(s_1) - u_i(s_0) > D$, then an improved value for $u_i(s_0)$ could be achieved by the following policy: First follow a policy which moves from s_0 to s_1 most quickly, which takes at most D steps on average. Then follow the optimal i-step policy for s_1. Since only D of the i rewards of the policy for s_1 are missed, this policy gives $u_i(s_0) \ge u_i(s_1) - D$, proving (10).

For the convergence criterion (5) it is a direct consequence of Theorem 8.5.6. in [5], that at the corresponding iteration

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

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 $\tilde{M}_k$.[5] Expanding $u_{i+1}(s)$ according to (4), 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

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

Defining $\boldsymbol{r}_k := \left(\tilde{r}_k\big(s, \tilde{\pi}_k(s)\big)\right)_s$ as the (column) vector of rewards for policy $\tilde{\pi}_k$, $\tilde{\boldsymbol{P}}_k := \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right)\right)_{s,s'}$ as the transition matrix of $\tilde{\pi}_k$ on $\tilde{M}_k$, and $\boldsymbol{v}_k := \left(v_k\big(s, \tilde{\pi}_k(s)\big)\right)_s$ as the (row) vector of visit counts for each state and the corresponding action chosen by $\tilde{\pi}_k$, we can rewrite (9) as

$$\begin{aligned}
\Delta_k &\leq \sum_{(s,a)} v_k(s,a)\big(\tilde{\rho}_k - \bar{r}(s,a)\big) + \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{t_k}} \\
&= \sum_{(s,a)} v_k(s,a)\big(\tilde{\rho}_k - \tilde{r}_k(s,a)\big) + \sum_{(s,a)} v_k(s,a)\big(\tilde{r}_k(s,a) - \bar{r}(s,a)\big) + \sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{t_k}} \\
&\leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{u}_i + \sum_{(s,a)} v_k(s,a)\big(\tilde{r}_k(s,a) - \bar{r}(s,a)\big) + 2\sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{t_k}} \qquad (11)
\end{aligned}$$

recalling that $v_k(s,a) = 0$ for $a \neq \tilde{\pi}_k(s)$. Since the rows of $\tilde{\boldsymbol{P}}_k$ sum to 1, we can replace $\boldsymbol{u}_i$ by $\boldsymbol{w}_k$ with $w_k(s) = u_i(s) - \min_s u_i(s)$ (we again use the subscript k to reference the episode). Since we consider $M \in \mathcal{M}_k$, $\tilde{r}_k(s,a) - \bar{r}(s,a)$ is bounded according to (1) to yield

$$\Delta_k \leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + 2\sum_{(s,a)} v_k(s,a)\sqrt{\tfrac{7\log(2SAt_k/\delta)}{2\max\{1,N_k(s,a)\}}} + 2\sum_{(s,a)} \frac{v_k(s,a)}{\sqrt{t_k}}, \qquad (12)$$

where $\|\boldsymbol{w}_k\|_\infty \leq D$ by (10). Noting that $\max\{1, N_k(s,a)\} \leq t_k$ we rewrite (12) as

$$\Delta_k \leq \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + \left(\sqrt{14\log\left(\tfrac{2SAT}{\delta}\right)} + 2\right) \sum_{(s,a)} \tfrac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}. \qquad (13)$$

4.3.2 The true Transition Matrix

Replacing the transition matrix $\tilde{\boldsymbol{P}}_k$ of the policy $\tilde{\pi}_k$ in the optimistic MDP $\tilde{M}_k$ by the transition matrix $\boldsymbol{P}_k$ of $\tilde{\pi}_k$ in the true MDP M, we get

$$\begin{aligned}
\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k &= \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k + \boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k \\
&= \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k + \boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k. \qquad (14)
\end{aligned}$$

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 (14) can be bounded using condition (2) in algorithm UCRL2:

$$\begin{aligned}
\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k &= \sum_s \sum_{s'} v_k\big(s, \tilde{\pi}_k(s)\big) \cdot \big(\tilde{\boldsymbol{P}}_k(s,s') - \boldsymbol{P}_k(s,s')\big) \cdot w_k(s') \\
&\leq \sum_s v_k\big(s, \tilde{\pi}_k(s)\big) \cdot \left\| \tilde{\boldsymbol{P}}_k(s,\cdot) - \boldsymbol{P}_k(s,\cdot) \right\|_1 \cdot \|\boldsymbol{w}_k\|_\infty \\
&\leq \sum_s v_k\big(s, \tilde{\pi}_k(s)\big) \cdot 2\sqrt{\tfrac{14S\log(2AT/\delta)}{\max\{1,N_k(s,\tilde{\pi}_k(s))\}}} \cdot D \\
&\leq 2D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)} \sum_{(s,a)} \tfrac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}. \qquad (15)
\end{aligned}$$

This term will turn out to mainly determine our regret bound, since it yields the dominating contribution.

[5] 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 $\tilde{\rho}_k$.

The second Term. The intuition about the second term in (14) is that the counts of the state visits $\boldsymbol{v}_k$ are relatively close to the stationary distribution of the transition matrix $\boldsymbol{P}_k$, 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 7 (Azuma-Hoeffding inequality, [14])**.** *Let $X_1, X_2, \ldots$ be a martingale difference sequence with bounded coordinates, i.e. $|X_i| \leq c$. Then for all $\varepsilon > 0$ and $n \in \mathbb{N}$,*

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

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 $X_t := \big(p\left(\cdot|s_t,a_t\right) - \boldsymbol{e}_{s_{t+1}}\big)\,\boldsymbol{w}_{k(t)} \mathbb{1}_{M \in \mathcal{M}_{k(t)}}$ for $t = 1, \ldots, T$. Then for any episode k with $M \in \mathcal{M}_k$, since $\|\boldsymbol{w}_k\|_\infty \leq D$, we have

$$\begin{aligned}
\boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \;&=\; \sum_{t=t_k}^{t_{k+1}-1} \big(p\left(\cdot|s_t,a_t\right) - \boldsymbol{e}_{s_t}\big)\boldsymbol{w}_k \\
&=\; \Bigg(\sum_{t=t_k}^{t_{k+1}-1} p\left(\cdot|s_t,a_t\right) - \sum_{t=t_k}^{t_{k+1}-1} \boldsymbol{e}_{s_{t+1}} + \boldsymbol{e}_{s_{t_{k+1}}} - \boldsymbol{e}_{s_{t_k}}\Bigg)\boldsymbol{w}_k \\
&=\; \sum_{t=t_k}^{t_{k+1}-1} X_t + w_k(s_{t_{k+1}}) - w_k(s_{t_k}) \leq \sum_{t=t_k}^{t_{k+1}-1} X_t + 2D\,.
\end{aligned}$$

Also due to $\|\boldsymbol{w}_k\|_\infty \leq D$, we have $|X_t| \leq (\|p\left(\cdot|s_t,a_t\right)\|_1 + \|\boldsymbol{e}_{s_{t+1}}\|_1)D \leq 2D$. Further, $\mathbb{E}\left[X_t \middle| s_1, a_1, \ldots, s_t, a_t\right] = 0$, so that X_t is a sequence of martingale differences, and application of Lemma 7 gives

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

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

$$\sum_{k=1}^m \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})w_k \mathbb{1}_{M\in\mathcal{M}_k} \;\leq\; \sum_{t=1}^T X_t + mD \;\leq\; D\sqrt{\tfrac{5}{2}T\log\tfrac{8T}{\delta}} + DSA\log_2\big(\tfrac{8T}{SA}\big) \qquad (16)$$

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

4.3.3 Summing over Episodes with $M \in \mathcal{M}_k$

To conclude Section 4.3, we sum (13) over all episodes with $M \in \mathcal{M}_k$, using (14), (15), and (16) which yields that with probability $1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned}
\sum_{k=1}^m \Delta_k \mathbb{1}_{M\in\mathcal{M}_{k(t)}} \;&\leq\; \sum_{k=1}^m \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k \mathbb{1}_{M\in\mathcal{M}_{k(t)}} + \sum_{k=1}^m \boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k \mathbb{1}_{M\in\mathcal{M}_{k(t)}} \\
&\quad + \sum_{k=1}^m \left(\sqrt{14\log\big(\tfrac{2SAT}{\delta}\big)} + 2\right) \sum_{(s,a)} \tfrac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \\
&\leq\; 2D\sqrt{14S\log\big(\tfrac{2AT}{\delta}\big)} \cdot \sum_{k=1}^m \sum_{(s,a)} \tfrac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}} \\
&\quad + D\sqrt{\tfrac{5}{2}T\log\big(\tfrac{8T}{\delta}\big)} + DSA\log_2\big(\tfrac{8T}{SA}\big) \\
&\quad + \left(\sqrt{14\log\big(\tfrac{2SAT}{\delta}\big)} + 2\right) \sum_{k=1}^m \sum_{(s,a)} \tfrac{v_k(s,a)}{\sqrt{\max\{1, N_k(s,a)\}}}\,. \qquad (17)
\end{aligned}$$

Recall that $N(s,a) := \sum_k v_k(s,a)$ such that $\sum_{(s,a)} N(s,a) = T$, and that $N_k(s,a) = \sum_{i<k} v_i(s,a)$. By the condition of the while-loop in Step 6 of algorithm UCRL2, we have that $v_k(s,a) \le N_k(s,a)$. Using that (see Appendix A.3)

$$\sum_{k=1}^{n} \frac{x_k}{\sqrt{X_{k-1}}} \le \left(\sqrt{2}+1\right)\sqrt{X_n}\,,$$

where $X_k = \max\left\{1, \sum_{i=1}^{k} x_i\right\}$ and $0 \le x_k \le X_{k-1}$, 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

$$\sum_{(s,a)}\sum_{k} \frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} \le \left(\sqrt{2}+1\right)\sqrt{SAT}, \tag{18}$$

and get from (17) (after minor simplifications) that with probability $1-\frac{\delta}{12T^{5/4}}$

$$\begin{aligned}\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\in\mathcal{M}_{k(t)}} \;&\le\; D\sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)} + DSA\log_2\left(\tfrac{8T}{SA}\right)\\ &\quad + \left(3D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)}+2\right)\left(\sqrt{2}+1\right)\sqrt{SAT}\,. \end{aligned}\tag{19}$$

4.4 Completing the Proof

Evaluating (7) by summing Δ_k over all episodes, using (8), (13), and (19), we get

$$\begin{aligned}\Delta(s_1,T) \;&\le\; \sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)} + \sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\notin\mathcal{M}_{k(t)}} + \sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\in\mathcal{M}_{k(t)}}\\ &\le\; \sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)} + \sqrt{T} + D\sqrt{\tfrac{5}{2}T\log\left(\tfrac{8T}{\delta}\right)} + DSA\log_2\left(\tfrac{8T}{SA}\right)\\ &\quad + \left(3D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)}+2\right)\left(\sqrt{2}+1\right)\sqrt{SAT}\end{aligned}\tag{20}$$

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

Simplifying (20) as given in Appendix A.4 yields that for any $T>1$ with probability $1-\frac{\delta}{4T^{5/4}}$

$$\Delta(s_1,T) \le 49DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}. \tag{21}$$

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

5 The logarithmic Bound (Proof of Theorem 4)

Our aim is to show a logarithmic upper bound on the expected regret. In order to achieve this, we start with a bound on the number of steps in suboptimal episodes (in the spirit of *sample complexity bounds* as given in [9]).

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[6] $\Delta_k = \sum_{t=t_k}^{t_{k+1}-1}(\rho^* - r_t)$. Then the following result gives an upper bound on the number of steps taken in ε-bad episodes.

Theorem 8. *For any initial state $s\in\mathcal{S}$, any $T>1$ and any $\varepsilon>0$, with probability $1-3\delta$, the number L_ε of steps taken in ε-bad episodes is*

$$L_\varepsilon \;\le\; 48^2\frac{D^2S^2A\log(T/\delta)}{\varepsilon^2}.$$

[6] In the following we use the same notation as in the proof of Theorem 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 $\varepsilon L_\varepsilon \le \Delta'_\varepsilon(s,T)$.

Let K_ε and J_ε be two random sets that contain the indices of the ε-bad episodes and the corresponding time steps t taken in these episodes, respectively. Analyzing the random reward fluctuations and the regret caused by failing confidence intervals we get with probability $1-2\delta$

$$\Delta'_\varepsilon(s,T) \;:=\; \sum_{k\in K_\varepsilon} \Delta_k \;\le\; 1 + \sqrt{2L_\varepsilon \log\tfrac{T}{\delta}} + \sum_{k\in K_\varepsilon}\sum_{s,a} v_k(s,a)\big(\rho^* - \bar r(s,a)\big)\mathbb{1}_{M\in\mathcal{M}_k}, \quad (22)$$

since analogously to (6) in Section 4.1, with probability $1-\delta$

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

and, similar to (8) in Section 4.2, one can show that

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

To bound the regret of a single episode with $M\in\mathcal{M}_k$ we may follow the lines of the proof of Theorem 2 in Section 4.3. By combining (13), (14), and (15) we arrive at

$$\Delta_k \le \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k + \left(3D\sqrt{14S\log\big(\tfrac{2AT}{\delta}\big)} + 2\right)\sum_{(s,a)}\tfrac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\;. \quad (23)$$

Using the observation in Appendix B.1, we get an analogon of (18), that is,

$$\sum_{k\in K_\varepsilon}\sum_{s,a}\frac{v_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} \;\le\; \left(\sqrt{2}+1\right)\sqrt{L_\varepsilon SA}\;. \quad (24)$$

From (22), (23), and (24) it follows that with probability $1-2\delta$

$$\begin{aligned}\Delta'_\varepsilon(s,T) \;\le\; & 1 + \sqrt{2L_\varepsilon\log\tfrac{T}{\delta}} + \left(3D\sqrt{14S\log\big(\tfrac{2AT}{\delta}\big)}+2\right)\cdot\left(\sqrt{2}+1\right)\cdot\sqrt{L_\varepsilon SA} \quad (25)\\ & + \sum_{k\in K_\varepsilon}\boldsymbol{v}_k(\boldsymbol{P}_k-\boldsymbol{I})\boldsymbol{w}_k\mathbb{1}_{M\in\mathcal{M}_k}\;. \quad (26)\end{aligned}$$

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 (16) in the original proof (Section 4.3.2). Here we have to consider a slightly modified martingale difference sequence

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

for $t=1,\dots,T$ to get

$$\sum_{k\in K_\varepsilon}\boldsymbol{v}_k(\boldsymbol{P}_k-\boldsymbol{I})\boldsymbol{w}_k\mathbb{1}_{M\in\mathcal{M}_k} \;\le\; \sum_{t=1}^{T(L_\varepsilon)} X_t + DSA\log_2\tfrac{8T}{SA}\;, \quad (27)$$

where $T(L) := \min\big\{t : \#\{\tau\le t, \tau\in J_\varepsilon\} = L\big\}$.

The application of the Azuma-Hoeffding inequality in the original proof is replaced with the following consequence of Bernstein's inequality for martingales [15]:

Lemma 9. *Let $X_1, X_2, \dots$ be a martingale difference sequence. Then*

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

Application of Lemma 9 with $\kappa = 2D\sqrt{L\log(T/\delta)}$ and $\gamma = D^2 L$ yields that if $L \geq \log(T/\delta)/D^2$ we have

$$\mathbb{P}\left\{ \sum_{t=1}^{T(L)} X_t > 2D\sqrt{L\log\tfrac{T}{\delta}} \,\middle|\, T(L) = \min\left\{ t : \#\{\tau \leq t, \tau \in J_\varepsilon\} = L \right\} \right\} < \frac{\delta}{T}. \tag{28}$$

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

$$\sum_{t=1}^{T(L)} X_t \leq DL \;=\; D\sqrt{L}\sqrt{L} \;<\; \sqrt{L}\sqrt{\log\tfrac{T}{\delta}} = \sqrt{L\log\tfrac{T}{\delta}} \;<\; 2D\sqrt{L\log\tfrac{T}{\delta}}. \tag{29}$$

Hence, (28) and (29) give by a union bound over all L that with probability $1-\delta$

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

which together with (27) yields that with probability $1-\delta$

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

Thus (25) yields that with probability $1-3\delta$

$$\begin{aligned}\Delta'_\varepsilon(s,T) \;\leq\; & 1 + \sqrt{2L_\varepsilon\log\tfrac{T}{\delta}} + \left(3D\sqrt{14S\log\left(\tfrac{2AT}{\delta}\right)} + 2\right)\cdot\left(\sqrt{2}+1\right)\cdot\sqrt{L_\varepsilon SA} \\ & + 2D\sqrt{L_\varepsilon\log\tfrac{T}{\delta}} + DSA\log_2\tfrac{8T}{SA}\ . \end{aligned} \tag{30}$$

As the theorem also holds trivially for $L_\varepsilon \leq 48^2 A\log\left(\frac{T}{\delta}\right)$ and also for $T \leq 48^2 A$, by similar arguments than those used in Section 4.4 we get

$$\Delta'_\varepsilon(s,T) \leq 48DS\sqrt{L_\varepsilon A\log\tfrac{T}{\delta}} \tag{31}$$

with probability $1-3\delta$. Since $\varepsilon L_\varepsilon \leq \Delta'_\varepsilon(s,T)$ we get

$$L_\varepsilon \leq 48^2\frac{D^2S^2A\log\frac{T}{\delta}}{\varepsilon^2} \tag{32}$$

the theorem follows. □

Theorem 8 can be used to obtain the claimed logarithmic upper bound on the expected regret.

Theorem 10. *For any initial state $s\in\mathcal{S}$, any $T\geq 1$ and any $\varepsilon > 0$, with probability $1-3\delta$ the regret of* UCRL2 *(with parameter δ) is*

$$\Delta(M, \textsc{UCRL2}, s, T) = 48^2\frac{D^2S^2A\log(T/\delta_0)}{\varepsilon} + \varepsilon T.$$

Moreover setting

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

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

$$\mathbb{E}\left[\Delta(M, \textsc{UCRL2}, s, T)\right] < c_3\frac{D^2S^2A\log(T)}{g}.$$

Using the doubling trick to set the parameter δ, the same bound can be achieved without knowledge of the horizon T.

Proof. Using (32) in (31) we we may bound the regret $\Delta'_\varepsilon(s,T)$ accumulated in ε-bad episodes by

$$\Delta'_\varepsilon(s,T) \leq 48^2 \frac{D^2 S^2 A \log \frac{T}{\delta}}{\varepsilon} \tag{33}$$

with probability $1-3\delta$. 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 the expected regret in $\frac{g}{2}$-bad episodes is upper bounded by $48^2 \cdot 2 \cdot D^2 S^2 A \log(AT)/g + 1$, as $\delta = 1/(3T)$. What remains to do is to consider episodes k with average regret smaller than $g/2$ in which however a non-optimal policy $\tilde{\pi}_k$ was chosen. Note that for sufficiently large episode length ℓ_k the expected ℓ_k-step return is $\frac{g}{2}$-close to the average reward, so that any policy applied in an episode that is not $\frac{g}{2}$-bad will be optimal. The regret accumulated until the episode lengths are sufficiently large will be an additive constant depending on the MDP, subsumed by c_3. □

6 The Lower Bound (Proof of Theorem 5)

We first consider the two-state MDP depicted in Figure 2. That is, there are two states, s_0 and s_1, and $A' = \lfloor (A-1)/2 \rfloor$ actions. For each action a, let the deterministic rewards be $r(s_0,a) = 0$ and $r(s_1,a) = 1$, and $p(s_0|s_1,a) = \delta$, where $\delta = 4/D$. For the rest of the proof we assume[7] $\delta \leq 1/3$. For all but a single "good" action a^* let $p(s_1|s_0,a) = \delta$ whereas $p(s_1|s_0,a^*) = \delta + \varepsilon$ for some $0 < \varepsilon < \delta$ specified later in the proof. The diameter of this MDP is $D' = 1/\delta = D/4$.

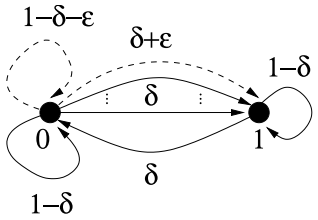

Figure 2: The MDP for the lower bound. The single action a^* with higher transition probability from state s_0 to state s_1 is shown as dashed line.

Consider $k := \lfloor S/2 \rfloor$ 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, using at most $A - A'$ additional actions. This can be done by introducing $A'+1$ additional deterministic actions per state, which do not leave the s_1-states but connect the s_0-states of the k copies by inducing an A'-ary tree structure on the s_0-states (one action for going toward the root, A' actions to go toward the leaves). The reward for each of those actions in any state is zero. The diameter of the resulting MDP is at most $2(D/4 + \lceil \log_{A'} k \rceil)$ which is twice the time 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 $\leq D$.

First note, that the problem of learning M gets easier when the additional actions (connecting the s_0 states and not leaving the s_1 states) are removed, and instead the learning algorithm is allowed to do the following: Before performing an action in any of the s_0 states any of the s_0 states may be chosen for free, and before performing an action in any of the s_1 states any of the s_1 states may be chosen for free. This is equivalent to a single MDP M' like the one in Figure 2 with kA' actions.

We prove the theorem by applying the same techniques as in the proof of the lower bound for the multi-armed bandit problem in [16]. The pair (s_0^*, a^*) identifying the copy with the better action and

[7] Otherwise we have $D < 12$, and for this to be possible $A > 2S$. In this case we use a different construction: Using $S-1$ actions, we connect all states to get an MDP with diameter 1, and with the remaining $A - S + 1$ actions we set up a bandit problem in each state as in the proof of the lower bound in [16], where only one state has a better action. This yields $\Omega\big(\sqrt{SAT}\big)$ regret, which is sufficient, since D is bounded in this case.

the better action are considered to be chosen uniformly at random from $\{1 \dots k\} \times \{1 \dots A'\}$, and we denote the expectation with respect to the random choice of (s_0^*, a^*) as $\mathbb{E}_* [\cdot]$. We show that ε can be chosen such that M' and thus M forces regret $\mathbb{E}_* [\Delta(M, \mathfrak{A}, s, T)] \geq \mathbb{E}_* [\Delta(M', \mathfrak{A}, s, T)] > 0.022\sqrt{D'kA'T}$ 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_0 to s_1 is δ for all actions), and $\mathbb{E}_a [\cdot]$ for the expectation conditioned on a being the special action a^* in M'. As argued in [16], it is sufficient to consider deterministic strategies for choosing actions. That is, we 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 in [16]. Let the random variables N_1, N_0 and N_{0a} denote the total number of visits to state s_1, the total number of visits to state s_0, and the number of times action a^* is chosen in state s_0, respectively. Since the expected number of consecutive steps spent in state s_1 after reaching it from s_0 is at most $D' = 1/\delta$ (we might reach step T first), and since choosing a^* instead of any other action in s_0 reduces the probability of staying in state s_0, the reward accumulated by any algorithm can be bounded as

$$\begin{aligned} \mathbb{E}_a [R(M, \mathfrak{A}, s, T)] &= \mathbb{E}_a [N_1] = \mathbb{E}_a [N_0 - N_{0a}] \, \delta D' + \mathbb{E}_a [N_{0a}] \, (\delta + \varepsilon) D' \\ &= \mathbb{E}_a [N_0 - N_{0a} + N_{0a}] + \mathbb{E}_a [N_{0a}] \, \varepsilon D' \\ &\leq \frac{T}{2} + \mathbb{E}_a [N_{0a}] \, \varepsilon D'. \end{aligned} \tag{34}$$

As the actions are deterministically chosen by $\mathfrak{A}$ based on the observed pairs of reward and next state, N_{0a} is a function of the observations up to step T. A slight difference to [16] is that in our setting the sequence of observations consists not just of the rewards but also of the next state, i.e. upon playing action a_t the algorithm observes s_{t+1} and r_t. Since the immediate reward is fully determined by the current state, N_{0a} is also a function of just the sequence of states, and we may bound $\mathbb{E}_a [N_{0a}]$ by the following lemma, adapted from [16].

Lemma 11. *Let $f : \{s_0, s_1\}^{T+1} \to [0, B]$ be any function defined on state sequences $\boldsymbol{s} \in \{s_0, s_1\}^{T+1}$ observed in MDP M'. Then for any $0 \leq \delta \leq \frac{1}{2}$, any $0 \leq \varepsilon \leq 1 - 2\delta$ and any $a \in \{1, \dots, kA'\}$,*

$$\mathbb{E}_a [f(\boldsymbol{s})] \leq \mathbb{E}_{\text{unif}} [f(\boldsymbol{s})] + \frac{B}{2} \cdot \frac{\varepsilon}{\sqrt{\delta}} \sqrt{2\mathbb{E}_{\text{unif}} [N_{0a}]}.$$

There are only minor modifications to the original proof in [16] to get a proof for Lemma 11, as discussed in Appendix C. Now, since N_{0a} is a function of the state sequence with $N_{0a} \in [0, T]$, and for $\varepsilon \leq \delta$ we may apply Lemma 11 to get

$$\mathbb{E}_a [N_{0a}] \leq \mathbb{E}_{\text{unif}} [N_{0a}] + \frac{T}{2} \varepsilon \sqrt{D'} \sqrt{2\mathbb{E}_{\text{unif}} [N_{0a}]}. \tag{35}$$

Using $\sum_{a=1}^{kA'} \mathbb{E}_{\text{unif}} [N_{0a}] \leq \frac{T}{2} + \frac{D'}{2}$ yields $\sum_{a=1}^{kA'} \sqrt{2\mathbb{E}_{\text{unif}} [N_{0a}]} \leq \sqrt{kA'(T + D')}$ and thus

$$\sum_{a=1}^{kA'} \mathbb{E}_a [N_{0a}] \leq \frac{T}{2} + \frac{D'}{2} + \frac{\varepsilon T}{2} \sqrt{D'} \sqrt{kA'(T + D')} \leq \frac{T}{2} + \frac{D'}{2} + \frac{\varepsilon T}{2} \sqrt{D'kA'T} + \frac{\varepsilon T D'}{2} \sqrt{kA'}.$$

Therefore, combining with (34),

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

By assumption we have $T \geq DSA \geq 16D'kA'$ and thus $D' \leq \frac{T}{16kA'}$. Further, calculating the stationary distribution, we find that the optimal average reward for the MDP M' is $\frac{\delta + \varepsilon}{2\delta + \varepsilon}$. Thus the

expected regret with respect to the random choice of a^* is at most

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

We choose $\varepsilon = c\sqrt{\frac{kA'}{TD'}}$, where $0 < c < \frac{1}{4}$. Then due to $\frac{1}{\delta} = D' \leq \frac{T}{16kA'}$ we have $\varepsilon \leq \frac{\delta}{16}$ (sufficient to get (35)), and further $\frac{1}{4\delta+2\varepsilon} \geq \frac{1}{4+1/8}D'$. Hence we get

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

Choosing $c = 0.2$ we have due to $kA' \geq 20$ that

$$\mathbb{E}_*\left[\Delta(M,\mathfrak{A},s,T)\right] > 0.022\sqrt{D'kA'T}.$$

7 Proof of Regret Bounds for Changing MDPs (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 the reward received by $\mathfrak{A}$ in step t when starting in state s.

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

Because the horizon T is usually unknown, we apply an alternative scheme for restarting which exhibits similar properties: UCRL2′ restarts UCRL2 with parameter δ/ℓ^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

$$\Delta'(\text{UCRL2}', s, T) \leq 92 \cdot \ell^{\frac{1}{3}}T^{\frac{2}{3}}DS\sqrt{A\log\tfrac{T}{\delta}}$$

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

Proof of Theorem 6. Let n be the largest natural number such that $\left\lceil \frac{n^3}{\ell^2} \right\rceil \leq T$. Then $\frac{n^3}{\ell^2} \leq \tau_n \leq T \leq \tau_{n+1} - 1 < \frac{(n+1)^3}{\ell^2}$ and thus

$$\ell^{\frac{2}{3}}T^{\frac{1}{3}} - 1 \leq n \leq \ell^{\frac{2}{3}}T^{\frac{1}{3}}. \tag{36}$$

The regret Δ_c incurred due to changes of the MDP can be bounded by the number of steps taken in periods where 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. We have

$$\tau_{n+1} - 1 - \tau_{n-\ell+1} \leq \tfrac{1}{\ell^2}(n+1)^3 - \tfrac{1}{\ell^2} - \tfrac{1}{\ell^2}(n-\ell+1)^3 = 3\tfrac{n^2}{\ell} + 6\tfrac{n}{\ell} - 3n - \tfrac{1}{\ell^2} + \ell - 3 + 3\tfrac{1}{\ell}.$$

For $\ell \geq 2$ we get

$$\tau_{n+1} - 1 - \tau_{n-\ell+1} \leq 3\,\frac{n^2}{\ell} + \ell \leq 3\,\frac{\ell^{\frac{4}{3}} T^{\frac{2}{3}}}{\ell} + \ell = 3\ell^{\frac{1}{3}} T^{\frac{2}{3}} + \ell,$$

and for $\ell = 1$ we have

$$\tau_{n+1} - 1 - \tau_{n-\ell+1} \leq 3T^{\frac{2}{3}} + 3T^{\frac{1}{3}}.$$

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

$$\Delta_c \;\leq\; 3\ell^{\frac{1}{3}} T^{\frac{2}{3}} + 3T^{\frac{1}{3}} + \ell \leq 6\ell^{\frac{1}{3}} T^{\frac{2}{3}} + \ell^{\frac{1}{3}} \ell^{\frac{2}{3}} \leq 6\ell^{\frac{1}{3}} T^{\frac{2}{3}} + \ell^{\frac{1}{3}} T^{\frac{2}{3}} \leq 7\ell^{\frac{1}{3}} T^{\frac{2}{3}}\;. \tag{37}$$

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 (21) in the proof of Theorem 2 by

$$\Delta(s_{\tau_i}, T_i) \;\leq\; 49DS\sqrt{T_i A \log \tfrac{\ell^2 T_i}{\delta}} \;\leq\; 49\sqrt{3}DS\sqrt{T_i}\sqrt{A \log \tfrac{T}{\delta}}$$

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

$$\Delta_f \;\leq\; \sum_{i=1}^n \Delta(s_{\tau_i}, T_i) \;\leq\; 49\sqrt{3}DS\sqrt{n}\sqrt{T}\sqrt{A \log \tfrac{T}{\delta}} \;\leq\; 49\sqrt{3}DS\;\ell^{\frac{1}{3}}\;T^{\frac{2}{3}}\;\sqrt{A \log \tfrac{T}{\delta}} \tag{38}$$

with probability at least $1 - \sum_{i=1}^n \frac{\delta}{4\ell^2 T_i^{5/4}}$. Using that for $\lfloor \frac{\ell^2}{3} \rfloor < i < n$

$$T_i = \left\lceil \frac{(i+1)^3}{\ell^2} \right\rceil - \left\lceil \frac{i^3}{\ell^2} \right\rceil \geq \frac{(i+1)^3}{\ell^2} - \frac{i^3}{\ell^2} - \frac{\ell^2 - 1}{\ell^2} = \frac{3i^2}{\ell^2} + \frac{3i + 2 - \ell^2}{\ell^2} > \frac{3i^2}{\ell^2},$$

and $T_i \geq 1$ we get

$$\begin{aligned} 1 - \sum_{i=1}^n \tfrac{\delta}{4\ell^2 T_i^{5/4}} \;&\geq\; 1 - \frac{\delta}{4\ell^2} - \sum_{i=1}^{\left\lfloor \frac{\ell^2}{3} \right\rfloor} \frac{\delta}{4\ell^2} - \sum_{\left\lceil \frac{\ell^2}{3} \right\rceil}^{n-1} \frac{\delta}{4i^2} \\ &>\; 1 - \frac{\delta}{4} - \frac{\ell^2}{12}\frac{\delta}{\ell^2} - \frac{\delta}{4}\sum_{i=1}^{\infty} \frac{1}{i^2} \;=\; 1 - \frac{\delta}{3} - \frac{\delta}{4}\frac{\pi^2}{6} \;>\; 1 - \delta. \end{aligned}$$

As $\Delta'(\textsc{Ucrl2}', s, T) \leq \Delta_c + \Delta_f$, using (37) and (38) yields

$$\Delta'(\textsc{Ucrl2}', s, T) \;\leq\; 7\ell^{\frac{1}{3}} T^{\frac{2}{3}} + 49\sqrt{3}DS\;\ell^{\frac{1}{3}}\;T^{\frac{2}{3}}\;\sqrt{A \log \tfrac{T}{\delta}}$$

with probability $1 - \delta$. □

References

[1] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*. 2009.

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

[3] 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*. MIT Press, 1999.

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

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

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

[7] 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.

[8] Ambuj Tewari and Peter Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems 20*, pages 1505–1512. MIT Press, 2008.

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

[10] Alexander L. Strehl and Michael L. Littman. A theoretical analysis of model-based interval estimation. In *Proc. 22nd ICML 2005*, pages 857–864, 2005.

[11] 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.

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

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

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

[15] D.A. Freedman. On tail probabilities for martingales. *Ann.Probab.*, 3:100–118, 1975.

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

Appendix

A Technical Details for the Proof of Theorem 2

A.1 Confidence Intervals

Lemma 12. *For any $t \geq 1$, the probability that the true MDP M is not contained in the set of plausible MDPs $\mathcal{M}(t)$ as used by algorithm* UCRL2 *with input parameter δ 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 $n > 0$ before step t. Denote the estimates for transition probabilities and rewards obtained from these n observations made in steps $1,\ldots,(t-1)$ by $\hat{p}\,(\cdot|s,a)$ and $\hat{r}\,(s,a)$, respectively. Let us first consider for some fixed (s,a) the probability with which a confidence interval for the transition probabilities fails. The random event observed for the transition probabilities estimates is the state to which the transition occurs. Generally, the L1 deviation of the true distribution and the empirical distribution over m distinct events from n samples is bounded by [weissman03].

$$\mathbb{P}\Big\{\left\|\hat{p}\,(\cdot) - p\,(\cdot)\right\|_1 \geq \varepsilon\Big\} \leq (2^m - 2)\exp\left(-\tfrac{n\varepsilon^2}{2}\right). \tag{39}$$

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

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

we get from (39) for each state-action pair (s,a)

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

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 $[0,1]$

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

Setting

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

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

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

Note that when there haven't been any observations, then 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 $N(s,a)$ gives

$$\begin{aligned}\mathbb{P}\left\{\left|\hat{r}\,(s,a) - \bar{r}(s,a)\right| \geq \sqrt{\frac{7\log\left(2SAt/\delta\right)}{2\max\{1,N(s,a)\}}}\right\} \;&\leq\; \sum_{n=1}^{t-1}\frac{\delta}{60t^7SA} < \frac{\delta}{60t^6SA} \quad \text{and}\\ \mathbb{P}\left\{\left\|p\,(\cdot|s,a) - \hat{p}\,(\cdot|s,a)\right\|_1 \geq \sqrt{\frac{14S\log\left(2At/\delta\right)}{\max\{1,N(s,a)\}}}\right\} \;&\leq\; \sum_{n=1}^{t-1}\frac{\delta}{20t^7SA} < \frac{\delta}{20t^6SA}.\end{aligned}$$

Summing these error probabilities for all state-action pairs we get

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

□

A.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 13. *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 $N_k(s,a) > 0$. Then if $N(s,a) > 0$ we have

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

because $v_k(s,a) = N_k(s,a)$, $N_k(s,a) > 0$ implies $N_{k+1}(s,a) = 2N_k(s,a)$. 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

$$T \;=\; \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} N(s,a) \;\geq\; \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(2^{K(s,a)} - 1\right). \tag{40}$$

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

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

Together with (40) this gives

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

which implies

$$m \leq 1 + 2SA + SA \log_2 \tfrac{T}{SA},$$

from which the claimed bound on m follows for $T \geq SA$. □

A.3 The Sum in (17)

Lemma 14. *For any sequence of numbers $x_1, \ldots, x_n$ with $0 \leq x_k \leq X_{k-1} := \max\left\{1, \sum_{i=1}^{k-1} x_i\right\}$*

$$\sum_{k=1}^{n} \frac{x_k}{\sqrt{X_{k-1}}} \leq \left(\sqrt{2} + 1\right) \sqrt{X_n}\,.$$

Proof. We prove the statement by induction over n.

Base case, $n = 1$: We have $X_0 = 1$, hence $x_1 \leq 1$ and $X_1 = 1$. Thus

$$\sum_{k=1}^{1} \frac{x_k}{\sqrt{X_{k-1}}} \leq 1 < \sqrt{2} + 1 = \left(\sqrt{2} + 1\right) \sqrt{X_1}.$$

Inductive step: By the induction hypothesis we have

$$\sum_{k=1}^{n} \frac{x_k}{\sqrt{X_{k-1}}} \leq \left(\sqrt{2}+1\right)\sqrt{X_{n-1}} + \frac{x_n}{\sqrt{X_{n-1}}}\,.$$

Since $x_n \leq X_{n-1}$ we thus have

$$\begin{aligned}
\left(\sqrt{2}+1\right)\sqrt{X_{n-1}} + \frac{x_n}{\sqrt{X_{n-1}}} &= \sqrt{\left(\sqrt{2}+1\right)^2 X_{n-1} + 2\left(\sqrt{2}+1\right)x_n + \frac{x_n^2}{X_{n-1}}} \\
&\leq \sqrt{\left(\sqrt{2}+1\right)^2 X_{n-1} + \left(2+2\sqrt{2}+1\right)x_n} \\
&= \sqrt{\left(\sqrt{2}+1\right)^2 X_{n-1} + \left(\sqrt{2}+1\right)^2 x_n} \\
&= \left(\sqrt{2}+1\right)\sqrt{X_{n-1}+x_n} = \left(\sqrt{2}+1\right)\sqrt{X_n}\,,
\end{aligned}$$

which proves the lemma. □

A.4 Simplifying (20)

Combining like terms, (20) yields that with probability $1 - \frac{\delta}{4T^{5/4}}$

$$\begin{aligned}
\Delta(s_1,T) \;\leq\; & DS\sqrt{AT}\left(\sqrt{10\log\left(\tfrac{8T}{\delta}\right)} + 3\left(\sqrt{2}+1\right)\sqrt{14\log\left(\tfrac{2AT}{\delta}\right)} + \sqrt{8} + 3\right) \\
& + DSA\log_2\left(\tfrac{8T}{SA}\right). \qquad (41)
\end{aligned}$$

For $1 < T \leq 49^2 A\log\left(\frac{T}{\delta}\right)$ we have $\Delta(s_1,T) \leq 49\sqrt{AT\log\left(\frac{T}{\delta}\right)}$ trivially. Considering only $T > 49^2 A\log\left(\frac{T}{\delta}\right)$ we have $A < \frac{1}{49\log\left(\frac{T}{\delta}\right)}\sqrt{AT\log\left(\frac{T}{\delta}\right)}$ and since $\log_2(8T) < 2\log(T)$ for the values of T considered, we get

$$DSA\log_2\left(\tfrac{8T}{SA}\right) \;<\; \tfrac{2}{49}DS\sqrt{AT\log\tfrac{T}{\delta}}\,.$$

Further, $T > 49^2 A\log\left(\frac{T}{\delta}\right)$ 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 (41), that for any $T > 1$ with probability $1 - \frac{\delta}{4T^{5/4}}$

$$\begin{aligned}
\Delta(s_1,T) &\leq DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}\left(\sqrt{20} + 3\left(\sqrt{2}+1\right)\sqrt{28} + \sqrt{8} + 3 + \frac{2}{49}\right) \\
&\leq 49DS\sqrt{AT\log\left(\tfrac{T}{\delta}\right)}.
\end{aligned}$$

B Technical Details for the Proof of Theorem 4

B.1 Proof of (24).

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

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

The idea is, to "rearrange" the sum, so that Lemma 14 becomes applicable. Indeed, when counting visits in earlier episodes than the one they actually occurred in, the inner sum can only increase (due to the smaller denominator). Consequently we may redistribute the v_k's that occur after step L_ε into "gaps" of episodes $\notin K_\varepsilon$.

To evaluate the inner sum we use the following fact. Let $\ell_\varepsilon(s,a) := \sum_{k\in K_\varepsilon} \boldsymbol{v}_k(s,a)$, 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 $d_k := \sqrt{\max\{1, N_k(s,a)\}}$, 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

$$v_{m_\varepsilon} = (N_{m_\varepsilon+1} - \ell_\varepsilon) + (\ell_\varepsilon - N_{m_\varepsilon}). \tag{42}$$

By (42) and since $N_{m_\varepsilon} = \sum_{k=1}^{m_\varepsilon-1} v_k$,

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

or equivalently,

$$\big(\ell_\varepsilon - N_{m_\varepsilon}\big)\mathbb{1}_{m_\varepsilon\notin K_\varepsilon} + \sum_{k=1}^{m_\varepsilon-1} v_k \mathbb{1}_{k\notin K_\varepsilon} \;=\; \big(N_{m_\varepsilon+1} - \ell_\varepsilon\big)\mathbb{1}_{m_\varepsilon\in K_\varepsilon} + \sum_{k=m_\varepsilon+1}^{m} v_k \mathbb{1}_{k\in K_\varepsilon}. \tag{43}$$

By (42) 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 (43), using that $d_k \leq d_{m_\varepsilon}$ for $k \leq m_\varepsilon$

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

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 14 that

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

By Jensen's inequality and as $\sum_{(s,a)} \ell_\varepsilon(s,a) = L_\varepsilon$ we finally get the claimed

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

C Proof of Lemma 11

To denote the probability conditioned on a being the "good" action we write $\mathbb{P}_a[\cdot]$. The probability with respect to a setting where all actions in state s_0 are equivalent (i.e. $\varepsilon = 0$) is denoted by $\mathbb{P}_{\text{unif}}[\cdot]$. For convenience we abbreviate $\mathcal{S} := \{s_0, s_1\}$. We denote the state observed at step τ by S_τ, the state-sequence up to step τ by $\boldsymbol{s}^\tau = \langle S_1, \ldots, S_\tau \rangle$. We go into detail only where there are differences to the proof of Lemma A.1 in [16]. 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 in [16], using the notation of [16], that

$$\mathbb{E}_a[f(\boldsymbol{s})] - \mathbb{E}_{\text{unif}}[f(\boldsymbol{s})] \leq \frac{B}{2}\sqrt{2\log(2)\,\text{KL}\left(\mathbb{P}_{\text{unif}}\middle\|\mathbb{P}_a\right)}, \tag{44}$$

and we also have

$$\text{KL}\left(\mathbb{P}_a\middle\|\mathbb{P}_{\text{unif}}\right) = \sum_{t=1}^{T}\text{KL}\left(\mathbb{P}_{\text{unif}}\left[S_{t+1}\middle|\boldsymbol{s}^t\right]\middle\|\mathbb{P}_a\left[S_{t+1}\middle|\boldsymbol{s}^t\right]\right). \tag{45}$$

Like in [16] and by the Markov property we have

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

Since for the transition probabilities in MDP M' we have $\log_2\frac{\mathbb{P}_{\text{unif}}\left[s''|s',a'\right]}{\mathbb{P}_a\left[s''|s',a'\right]} \neq 0$ only for $s' = s_0$ and a' being the special action we get

$$\begin{aligned}
&\text{KL}\left(\mathbb{P}_{\text{unif}}\left[S_{t+1}\middle|\boldsymbol{s}^t\right]\middle\|\mathbb{P}_a\left[S_{t+1}\middle|\boldsymbol{s}^t\right]\right) = \\
&= \sum_{\boldsymbol{s}^{t-1}\in\mathcal{S}^{t-1}}\mathbb{P}_{\text{unif}}\left[\boldsymbol{s}^{t-1}\right]\mathbb{P}_{\text{unif}}\left[S_t = s_0, a_t = a|\boldsymbol{s}^{t-1}\right] \\
&\quad\cdot\sum_{s'\in\mathcal{S}}\mathbb{P}_{\text{unif}}\left[s'|s_0,a\right]\log_2\frac{\mathbb{P}_{\text{unif}}\left[s'|s_0,a\right]}{\mathbb{P}_a\left[s'|s_0,a\right]} \\
&= \mathbb{P}_{\text{unif}}\left[S_t = s_0, a_t = a\right]\sum_{s'\in\mathcal{S}}\mathbb{P}_{\text{unif}}\left[s'|s_0,a\right]\log_2\frac{\mathbb{P}_{\text{unif}}\left[s'|s_0,a\right]}{\mathbb{P}_a\left[s'|s_0,a\right]} \\
&= \mathbb{P}_{\text{unif}}\left[S_t = s_0, a_t = a\right]\left(\delta\log_2\frac{\delta}{\delta+\varepsilon} + (1-\delta)\log_2\frac{1-\delta}{1-\delta-\varepsilon}\right).
\end{aligned} \tag{46}$$

To complete the proof we use the following Lemma (proved below).

Lemma 15. *For any $0 \leq \delta \leq \frac{1}{2}$ and $\varepsilon \leq 1-2\delta$ we have*

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

Applying Lemma 15, by (45) and (46) we have that

$$\begin{aligned}
\text{KL}\left(\mathbb{P}_a\middle\|\mathbb{P}_{\text{unif}}\right) &= \sum_{t=1}^{T}\text{KL}\left(\mathbb{P}_{\text{unif}}\left[S_{t+1}\middle|\boldsymbol{s}^t\right]\middle\|\mathbb{P}_a\left[S_{t+1}\middle|\boldsymbol{s}^t\right]\right) \\
&\leq \sum_{t=1}^{T}\mathbb{P}_{\text{unif}}\left[S_t = s_0, a_t = a\right]\frac{\varepsilon^2}{\delta\log(2)} = \mathbb{E}_{\text{unif}}\left[N_{0a}\right]\frac{\varepsilon^2}{\delta\log(2)},
\end{aligned}$$

which together with (44) yields

$$\mathbb{E}_a\left[f(\boldsymbol{s})\right] - \mathbb{E}_{\text{unif}}\left[f(\boldsymbol{s})\right] \leq \frac{B}{2} \cdot \frac{\varepsilon}{\sqrt{\delta}} \sqrt{2\mathbb{E}_{\text{unif}}\left[N_{0a}\right]},$$

as claimed by Lemma 11.

Proof of Lemma 15. Consider

$$f_\delta(\varepsilon) := \frac{\varepsilon^2}{\delta} - \delta \log \frac{\delta}{\delta+\varepsilon} - (1-\delta) \log \frac{1-\delta}{1-\delta-\varepsilon}$$

and note that $f_\delta(0) = 0$ for all δ. For the first derivative

$$f'_\delta(\varepsilon) := \frac{\partial}{\partial\varepsilon} f_\delta(\varepsilon) = 2\frac{\varepsilon}{\delta} + \frac{\delta}{\delta+\varepsilon} - \frac{1-\delta}{1-\delta-\varepsilon}$$

we have $f'_\delta(\varepsilon) \geq 0$ for $\delta \leq \frac{1}{2}$ and $0 \leq \varepsilon \leq \varepsilon_0$, where

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

It is sufficient to show $\varepsilon \leq \varepsilon_0$ for $\delta < \frac{1}{2}$ and $\varepsilon \leq 1-2\delta$. Then we have $\varepsilon_0 > 0$,

$$\big(\varepsilon_0 - \varepsilon\big) \cdot \varepsilon_0 \geq \big(\varepsilon_0 - (1-2\delta)\big) \cdot \varepsilon_0 = -\left(\frac{1}{2} - \delta\right)^2 + \frac{1}{4}(1-2\delta) = \frac{1}{2}\delta - \delta^2 \geq 0,$$

and thus $\varepsilon \leq \varepsilon_0$. This implies $f_\delta(\varepsilon) \geq 0$ for all $0 \leq \delta \leq \frac{1}{2}$ and $0 \leq \varepsilon \leq 1-2\delta$. □