

Near-optimal Regret Bounds for Reinforcement Learning

Peter Auer Thomas Jaksch Ronald Ortner
University of Leoben, Franz-Josef-Strasse 18,
8700 Leoben, Austria
{auer,tjaksch,rortner}@unileoben.ac.at

December 3, 2007

Abstract

For undiscounted reinforcement learning we consider the *total regret* of a learning algorithm in respect to an optimal policy. We present a reinforcement learning algorithm with total regret $\tilde{O}(DS\sqrt{AT})$ after T steps for any unknown Markov decision process (MDP) with S states, A actions per state, and *diameter* D . The diameter of an MDP is at most D if for any pair of states s_1, s_2 there is a policy which moves from s_1 to s_2 in at most D steps (on average). Our upper bound holds with high probability and it can be converted into a logarithmic regret bound, if a fixed difference between the average reward of the optimal policy and the second optimal policy is assumed.

We also present a corresponding lower bound $\Omega(\sqrt{DSAT})$ on the worst case total regret of any learning algorithm.

Acknowledgments: This work was supported in part by the the Austrian Science Fund FWF (S9104-N04 SP4) and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. We also acknowledge support by the PASCAL pump priming projects “Sequential Forecasting and Partial Feedback: Applications to Machine Learning” and “Online Performance of Reinforcement Learning with Internal Reward Functions”. This publication only reflects the authors’ views.

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 , he receives a random reward r according to a distribution $q(r|s, a)$ on $[0, 1]$. Further, according to the transition probabilities $p(s'|s, a)$, he will observe a random transition to a state $s' \in \mathcal{S}$.

Reinforcement learning of MDPs is considered as a standard model for learning with delayed feedback. In contrast to most work on reinforcement learning — where the performance of the *learned* policy is considered [12, 7] — 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*

$$\rho(M, \mathfrak{A}, s) := \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}[R(M, \mathfrak{A}, s, T)]$$

can be maximized by an appropriate stationary *policy* $\pi : \mathcal{S} \rightarrow \mathcal{A}$ which defines an optimal action for each state [9].

We will consider only MDPs with finite *diameter* D , where D is the minimal average time to move from any state s_1 to any other state s_2 , using appropriate policies. Let $T(s_2|M, \pi, s_1)$ be the first (random) time, that state s_2 is reached when policy π is executed on MDP M with initial state s_1 . Then the diameter is given by

$$D(M) := \max_{s_1, s_2 \in \mathcal{S}} \min_{\pi: \mathcal{S} \rightarrow \mathcal{A}} \mathbb{E}[T(s_2|M, \pi, s_1)].$$

A finite diameter seems necessary for interesting regret bounds, since otherwise some parts of the MDP might not be reachable.¹

For MDPs with finite diameter the optimal average reward ρ^* does not depend on the initial state (cf. [9]), 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* after time T as

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

(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.)

¹Alternatively, one can achieve regret bounds when starting in one of the strongly connected components of the MDP, induced by the relation $s_1 \rightarrow s_2$ iff $\min_{\pi} T(s_2|M, \pi, s_1) < \infty$.

1.1 Results

We summarize the results achieved for our new reinforcement learning algorithm UCRL2 (described in Section 2) and also state the corresponding lower bound. In the following we assume a given MDP M to be learned, with finite diameter $D := D(M)$, a number of states $S := |\mathcal{S}|$, and $A := |\mathcal{A}|$ actions. We omit explicit references to M and UCRL2 when they are clear from the context.

Theorem 1. *For any initial state $s \in \mathcal{S}$ and any $T \geq 1$, with probability $1 - \delta/T$ the regret of UCRL2 is*

$$\Delta(M, \text{UCRL2}, s, T) = O\left(DS\sqrt{AT \log(AT/\delta)}\right).$$

Theorem 2. *For any initial state $s \in \mathcal{S}$, any $T \geq 1$ and any $\varepsilon > 0$, with probability $1 - \delta$, the regret of UCRL2 is*

$$\Delta(M, \text{UCRL2}, s, T) = O\left(\frac{D^2 S^2 A \log(AT/\delta)}{\varepsilon^2}\right) + \varepsilon T.$$

Corollary 3. *Let $g := \rho^*(M) - \max_{\pi, s} \{\rho(M, \pi, s) : \rho(M, \pi, s) < \rho^*(M)\}$ be the gap in average reward between best and second best policy in M . Then for any initial state $s \in \mathcal{S}$ and any $T \geq 1$, with probability $1 - \delta$, the regret of UCRL2 is*

$$\Delta(M, \text{UCRL2}, s, T) = O\left(\frac{D^2 S^2 A \log(AT/\delta)}{g^2}\right).$$

Theorem 4. *For any algorithm \mathfrak{A} and any natural numbers T , S and $A > 1$ there is an MDP M with S states, A actions, a diameter D , and a suitable initial state $s \in \mathcal{S}$ such that the expected regret after T steps is*

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

The new bounds given here are improvements over the bounds that have been achieved in [1] (for a slightly different version of the algorithm presented here) 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 [1], and finally, the ergodicity assumption is replaced by the much weaker and more natural assumption that the MDP has finite diameter.

The fact that the diameter D appears in the bounds can be explained as follows. In the process of learning the MDP, the learner will also explore suboptimal actions. Such a suboptimal action may take the learner into a “bad part” of the MDP from which it may take D steps to reach again a “good part” of the MDP. Thus, the learner may suffer regret D for such exploration.

1.2 Relation to Previous Work

We first compare our results to the well-known PAC bounds for the algorithms E^3 of Kearns, Singh [8] and R-Max of Brafman, Tennenholtz [3]. These algorithms achieve ε -optimal average rewards with probability $1 - \delta$ after time polynomial in $\frac{1}{\delta}$, $\frac{1}{\varepsilon}$, S , A , and the mixing time $T_\varepsilon^{\text{mix}}$ (see below). As this polynomial dependence on ε is of order $1/\varepsilon^3$, the PAC bounds translate into $T^{2/3}$ regret bounds at the best. Both algorithms need the ε -return mixing time $T_\varepsilon^{\text{mix}}$ of an optimal policy π^*

as input parameter². $T_\varepsilon^{\text{mix}}$ 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 regret bounds for E³ and R-max. Moreover, the exponents of the parameters S , A , and $T_\varepsilon^{\text{mix}}$ in the PAC bounds are substantially larger than in our bounds.

An important precursor of our paper is the work by Burnetas and Katehakis [4]. They prove regret bounds that are *asymptotically* logarithmic in T for their *index policies*. The bounds of Burnetas and Katehakis have recently been generalized by Tewari and Bartlett [13]. Both papers assume that the MDP is *ergodic*, such that *any* policy will reach every state after a sufficient number of steps. The asymptotic bounds also hide an additive term which is proportional to the number of policies, A^S .

While the index policies of Burnetas and Katehakis choose actions optimistically by using confidence bounds only for the estimates in the current state, the MBIE algorithm of Strehl and Littman [10, 11] — similarly to our approach — applies confidence bounds for the whole MDP 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 rewards³ 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 received *along the trajectory of an optimal policy*⁴.

2 The UCRL2 Algorithm

Our algorithm UCRL2 (Figure 1) proceeds in episodes and computes a new policy at the beginning of each episode. The lengths of the episodes are not fixed a priori, but depend on the observations made. Episode k ends (see Step 8), when in the next step a state-action pair (s, a) would be observed *in* episode k equally often as *before* episode k . The count $v_k(s, a)$ keeps track of the visits to (s, a) in episode k .⁵

In Steps 3–5 UCRL2 computes estimates $\hat{p}_k(s'|s, a)$ of the transition probabilities from the transitions observed before episode k . Here $N_k(s, a)$ counts the state-action pairs observed before episode k , such that $N_k(s, a) = \sum_{i=1}^{k-1} v_i(s, a)$. In Step 6, a set \mathcal{M}_k of plausible MDPs is defined through confidence regions around the estimated transition probabilities $\hat{p}_k(s'|s, a)$. This guarantees that with high probability the true MDP $M \in \mathcal{M}_k$. An optimistic policy $\tilde{\pi}_k$ is calculated in Step 7 (details about policy $\tilde{\pi}_k$ and its bias are given below in Section 2.1) and executed in Step 8.

To keep the exposition simple, we are making the following assumption:

²The knowledge of this parameter could be eliminated by guessing $T_\varepsilon^{\text{mix}}$ to be $1, 2, \dots$, so that sooner or later the correct $T_\varepsilon^{\text{mix}}$ will be reached (cf. [8, 3]). However, since there is no condition on when to stop increasing $T_\varepsilon^{\text{mix}}$, the assumed mixing time might grow exponentially large, such that the PAC bounds become exponential in the true $T_\varepsilon^{\text{mix}}$ (cf. [3]).

³Actually, the state values.

⁴Indeed, 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.

⁵Since 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 constant $\delta \in (0, 1]$.

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

For episodes $k = 1, 2, \dots$ **do**

Initialize episode k :

1. Set the start time of episode k , $t_k := t$.
2. For all s, a initialize the state-action counts for episode k , $v_k(s, a) := 0$.
3. For all s, a set the state-action counts prior to episode k ,

$$N_k(s, a) := N(s, a; t_k) := \#\{\tau < t_k : s_\tau = s, a_\tau = a\}.$$

4. For all s, a, s' set the transition counts prior to episode k ,

$$P_k(s, a, s') := P(s, a, s'; t_k) := \#\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}$$

5. Compute estimates $\hat{p}_k(s'|s, a) := \hat{p}(s'|s, a; t_k) := P(s, a, s'; t_k) / \max\{1, N(s, a; t_k)\}$.

Compute policy $\tilde{\pi}_k$:

6. Let $\mathcal{M}_k := \mathcal{M}(t_k)$ be the set of all MDPs with states, actions and rewards as in M , and with transition probabilities $\tilde{p}(\cdot|s, a)$ close to $\hat{p}(\cdot|s, a; t_k)$,

$$\|\tilde{p}(\cdot|s, a) - \hat{p}(\cdot|s, a; t_k)\|_1 \leq \sqrt{\frac{12S \log(2At_k/\delta)}{\max\{1, N(s, a; t_k)\}}}.$$

7. Choose optimistically an MDP $\tilde{M}_k \in \mathcal{M}_k$ and a policy $\tilde{\pi}_k$ with state independent maximal average reward $\tilde{\rho}_k = \rho(\tilde{M}_k, \tilde{\pi}_k)$. If such an optimal pair $(\tilde{M}_k, \tilde{\pi}_k)$ is not unique, then choose an optimal pair with minimal $\|\tilde{\lambda}_k\|_\infty$, where $\tilde{\lambda}_k$ solves the Poisson equation (1) (see Section 2.1).

Execute policy $\tilde{\pi}_k$:

8. **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.

Assumption 5. For the unknown MDP M , the mean rewards

$$\bar{r}(s, a) := \mathbb{E}[r|s, a]$$

are known for each state-action pair (s, a) .

This assumption can be easily removed by extending the algorithm to also keep estimates and confidence intervals for the mean rewards $\bar{r}(s, a)$. The analysis of the algorithm can be extended to this case, since dealing with uncertain transition probabilities is much harder than dealing with uncertain rewards. The regret bounds are not affected by removing this assumption, except for a small change of the constants.

2.1 The optimistic policy $\tilde{\pi}_k$ and its bias

A central tool in our analysis is the *Poisson equation*, a result from MDP theory (see e.g. [9]). For a policy π it gives the relation between the *bias vector* $\boldsymbol{\lambda}$, the average reward ρ , the mean reward vector $\mathbf{r} = (\bar{r}(s, \pi(s)))_s$, and the transition matrix $\mathbf{P} = (p(s'|s, \pi(s)))_{s,s'}$. The bias $\lambda(s)$ can be understood as the expected advantage in total reward, for $T \rightarrow \infty$, of starting (with policy π) in state s , over starting in the stationary distribution of π . Starting with the stationary distribution the expected reward in each step equals ρ . With this interpretation of the bias the Poisson equation

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

is quite intuitive ($\mathbf{1}$ is the vector of all 1's): the advantage $\lambda(s)$ of state s is the immediate reward $r(s, \pi(s))$ minus ρ plus the expected advantage of the subsequent state s' , $\sum_{s'} p(s'|s, \pi(s)) \lambda(s')$. As main fact in our analysis we use that the Poisson equation for the chosen policy $\tilde{\pi}_k$ in \tilde{M}_k ,

$$\tilde{\boldsymbol{\lambda}}_k = \mathbf{r}_k - \tilde{\rho}_k \mathbf{1} + \tilde{\mathbf{P}}_k \tilde{\boldsymbol{\lambda}}_k, \quad (1)$$

has a small solution $\tilde{\boldsymbol{\lambda}}_k$, where $\mathbf{r}_k = (\bar{r}(s, \tilde{\pi}_k(s)))_s$ and $\tilde{\mathbf{P}}_k = (\tilde{p}_k(s'|s, \tilde{\pi}_k(s)))_{s,s'}$, with \tilde{p}_k the transition probabilities in \tilde{M}_k .

Lemma 6. *If $M \in \mathcal{M}_k$, then for policy $\tilde{\pi}_k$ in MDP \tilde{M}_k , as chosen by algorithm UCRL2, there is a vector $\tilde{\boldsymbol{\lambda}}_k$ satisfying the Poisson equation (1) with*

$$\|\tilde{\boldsymbol{\lambda}}_k\|_\infty \leq D(M)/2.$$

Proof idea. Since the Poisson equation holds for any $\boldsymbol{\lambda}' = \boldsymbol{\lambda} + c\mathbf{1}$ if it holds for $\boldsymbol{\lambda}$, we only need to show that (1) has a solution $\tilde{\boldsymbol{\lambda}}_k$ with $d := \max_s \tilde{\lambda}_k(s) - \min_s \tilde{\lambda}_k(s) \leq D := D(M)$.

Assume that for policy $\tilde{\pi}_k$ in \tilde{M}_k , $d > D$. Let $s_1 = \arg \min_s \tilde{\lambda}_k(s)$ and $s_0 = \arg \max_s \tilde{\lambda}_k(s)$. There is a policy π_0 which quickly moves from s_1 to s_0 in M , i.e. $\mathbb{E}[T(s_0|M, \pi_0, s_1)] \leq D$. Modifying policy $\tilde{\pi}_k$ (and \tilde{M}_k) by using π_0 to move quickly from s_1 to s_0 , would improve $\tilde{\lambda}_k(s_1)$ to $\tilde{\lambda}_k(s_0) - D$, which contradicts the choice of $\tilde{\pi}_k$. Details are given in the appendix. \square

3 Analysis of UCRL2

3.1 Bounding UCRL2's regret

3.1.1 Basic properties of UCRL2

We start with some simple observations about algorithm UCRL2. By the stopping criterion for episode k we have (except for the episode where $v_k(s, a) = 1$ and $N_k(s, a) = 0$)

$$\sum_{s,a} v_k(s, a) \leq \sum_{s,a} N_k(s, a) = t_k - 1. \quad (2)$$

Let m be the number of episodes started by algorithm UCRL2 up to step T , and let $N(s, a) := N(s, a; T)$ be the total number of observations of the state-action pair (s, a) up to step T . In each episode $k < m$ there are 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 for $N(s, a) > 0$ 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 $N_{k+1}(s, a) = 2N_k(s, a)$. Thus

$$T = \sum_{s, a} N(s, a) \geq \sum_{s, a} 2^{K(s, a)} \geq (SA)2^{m/(SA)-1},$$

since $m \leq (SA) + \sum_{s, a} K(s, a)$. Thus the number of episodes is bounded by

$$m \leq (SA) \log_2 \frac{2T}{SA}. \quad (3)$$

Let r_t be the reward received at time t when executing UCRL2. For given $N(s, a)$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, the r_t are independent random variables, such that by Chernoff bounds

$$\mathbb{P} \left\{ \sum_{t=1}^T r_t \leq \sum_{s, a} N(s, a) \bar{r}(s, a) + \sqrt{2T \log \frac{T}{\delta}} \left| N(s, a), s \in \mathcal{S}, a \in \mathcal{A} \right. \right\} \leq \frac{\delta}{T}.$$

Hence for any initial state s_1 we have with probability $1 - \delta/T$ that

$$\Delta(s_1, T) = T\rho^* - \sum_{t=1}^T r_t \leq T\rho^* - \sum_{s, a} N(s, a) \bar{r}(s, a) + \sqrt{2T \log \frac{T}{\delta}} = \sum_{k=1}^m \Delta_k + \sqrt{2T \log \frac{T}{\delta}}, \quad (4)$$

where

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

since $\sum_{k=1}^m v_k(s, a) = N(s, a)$ and $\sum_{s, a} N(s, a) = T$.

3.1.2 Dealing with failing confidence regions

We consider the regret of episodes when $M \notin \mathcal{M}_k$. By (2) and since $\rho^* \leq 1$ we have

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

where $\mathcal{M}(t)$ is the set of plausible MDPs using the estimates available at time t as defined in algorithm UCRL2. Using the following Lemma 7, we find that $\mathbb{P}\{\exists t : T^{1/4} < t \leq T : M \notin \mathcal{M}(t)\} \leq \delta/T$. Thus,

$$\mathbb{P} \left\{ \sum_{k=1}^m \Delta_k \mathbf{1}_{M \notin \mathcal{M}_k} > \sqrt{T} \right\} \leq \frac{\delta}{T}. \quad (5)$$

Lemma 7. *Using the notation of algorithm UCRL2, $\mathbb{P}\{M \notin \mathcal{M}(t)\} \leq \delta/t^5$.*

Proof. The result follows from Chernoff bounds or an improved bound by Weissman et al. [14]. Details are given in the appendix. \square

3.1.3 Bounding the regret of an episode with $M \in \mathcal{M}_k$

Let $\tilde{\pi}_k$ be the optimistic policy chosen for episode k and let \tilde{M}_k be the corresponding plausible MDP. When π^* is an optimal policy for M and $M \in \mathcal{M}_k$, then

$$\tilde{\rho}_k := \rho(\tilde{M}_k, \tilde{\pi}_k) \geq \rho(M, \pi^*) = \rho^*.$$

Since $v_k(s, a) = 0$ if $a \neq \tilde{\pi}_k(s)$, we simplify notation and define $a_s := \tilde{\pi}_k(s)$, row vector $\mathbf{v}_k := (v_k(s, a_s))_s$, column vector $\mathbf{r}_k := (\bar{r}(s, a_s))_s$, and the transition matrices $\tilde{\mathbf{P}}_k := (\tilde{p}_k(s'|s, a_s))_{s, s'}$, $\mathbf{P}_k := (p(s'|s, a_s))_{s, s'}$. Applying the Poisson equation (1) for policy $\tilde{\pi}_k$ with bias vector $\tilde{\boldsymbol{\lambda}}_k$ in \tilde{M}_k , we get

$$\begin{aligned} \Delta_k &= \sum_{s, a} v_k(s, a) [\rho^* - \bar{r}_k(s, a)] \leq \sum_{s, a} v_k(s, a) [\tilde{\rho}_k - \bar{r}_k(s, a)] \\ &= \mathbf{v}_k(\tilde{\rho}_k \mathbf{1} - \mathbf{r}_k) = \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k = \mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k + \mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \\ &\leq \|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 \|\tilde{\boldsymbol{\lambda}}_k\|_\infty + \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \\ &\leq \frac{D}{2} \|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 + \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \end{aligned} \quad (6)$$

The last inequality follows from Lemma 6 since $M \in \mathcal{M}_k$. Since $\tilde{M}_k \in \mathcal{M}_k$, too,

$$\|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 \leq \sum_s v_k(s, a_s) \|\tilde{p}_k(\cdot|s, a_s) - p(\cdot|s, a_s)\|_1 \leq 2 \sum_{s, a} v_k(s, a) \sqrt{\frac{12S \log(2At_k/\delta)}{\max\{1, N_k(s, a)\}}}. \quad (7)$$

3.1.4 Bounding $\mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k$

It remains to bound the second term in (6). We derive a bound for the sum over all episodes k with $M \in \mathcal{M}_k$. Let $s_1, a_1, s_2, \dots, a_T, s_{T+1}$ be the sequence of states and actions, and let $k(t)$ be the episode which contains step t . Denote the standard basis vectors by \mathbf{e}_i , where each \mathbf{e}_i has 1 for its i -th component and 0 for every other component. Let $X_t = (p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}}) \tilde{\boldsymbol{\lambda}}_{k(t)} \mathbf{1}_{M \in \mathcal{M}_{k(t)}}$ for $t = 1, \dots, T$. Since $\mathbb{E}[X_t | s_1, a_1, \dots, s_t, a_t] = 0$, X_t is a sequence of martingale differences.

Lemma 8 (Azuma-Hoeffding inequality, see [2] or [6]). *Let X_1, X_2, \dots be a martingale difference sequence with bounded components $|X_i| \leq c$. Then for all $\varepsilon > 0$ and $n > 0$,*

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

By Lemma 6, $\|\tilde{\boldsymbol{\lambda}}_k\|_\infty \leq D/2$ if $M \in \mathcal{M}_k$, so that $|X_t| \leq \|p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}}\|_1 \|\tilde{\boldsymbol{\lambda}}_k\|_\infty \leq (\|p(\cdot|s_t, a_t)\|_1 + \|\mathbf{e}_{s_{t+1}}\|_1)D/2 \leq D$. The Azuma-Hoeffding inequality gives

$$\mathbb{P}\left\{\sum_{t=1}^T X_t \geq D\sqrt{2T \log \frac{T}{\delta}}\right\} \leq \frac{\delta}{T}.$$

For any episode k with $M \in \mathcal{M}_k$ we have

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

Summing over all episodes with $M \in \mathcal{M}_k$ yields

$$\sum_{k=1}^m \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \mathbb{1}_{M \in \mathcal{M}_k} \leq \sum_{t=1}^T X_t + Dm.$$

Hence, with probability $1 - \delta/T$,

$$\sum_{k=1}^m \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \mathbb{1}_{M \in \mathcal{M}_k} \leq D \left(\sqrt{2T \log \frac{T}{\delta}} + m \right). \quad (8)$$

3.1.5 Putting things together: Proof of Theorem 1

Combining inequalities (3)–(8) we get

$$\begin{aligned} \Delta(s, T) &\leq \sqrt{2T \log \frac{T}{\delta}} + \sqrt{T} + D \sqrt{12S \log(2A \frac{T}{\delta})} \sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \\ &\quad + D \left(\sqrt{2T \log \frac{T}{\delta}} + SA \log_2 \left(\frac{2T}{SA} \right) \right) \end{aligned} \quad (9)$$

with probability $1 - 3\delta/T$.

Since $v_k(s, a) = N_{k+1}(s, a) - N_k(s, a)$, $v_k(s, a) \leq N_k(s, a)$, and $\sum_{k,s,a} v_k(s, a) = T$, we have

$$\sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \leq \sqrt{8TSA} \quad (10)$$

(a derivation is given in the Appendix), which gives the theorem.

3.2 The logarithmic upper bound

An episode k is said to be 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 Δ_k as defined in Section 3.1.3. Our aim is to show an upper bound on the number of steps L_ε taken in ε -bad episodes that holds with high probability. We define the random sets K_ε and J_ε that contain the indices of the ε -bad episodes and the corresponding time steps t taken in these episodes, respectively.

Similar to Section 3.1.2, we have

$$\mathbb{P} \left\{ \sum_{k \in K_\varepsilon} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} > 1 \right\} \leq \delta. \quad (11)$$

For the regret term of $\sum_{k \in K_\varepsilon} \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \mathbf{1}_{M \in \mathcal{M}_k}$ (cf. Section 3.1.4) we need to consider a slightly modified martingale difference sequence $X_t = (p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}}) \tilde{\boldsymbol{\lambda}}_{k(t)} \mathbf{1}_{M \in \mathcal{M}_{k(t)}} \mathbf{1}_{t \in J_\varepsilon}$ for $t = 1, \dots, T$. The following bound is a consequence of Bernstein's inequality for martingales [5].

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

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

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

$$\mathbb{P} \left\{ \sum_{t=1}^{T(L)} X_t > 2D\sqrt{2L \log \frac{T}{\delta}} \mid T(L) = \min\{t : \#\{\tau \leq t, \tau \in J_\varepsilon\} = L\} \right\} < \frac{\delta}{T^2}. \quad (12)$$

On the other hand, if $L < \log(T/\delta)/D^2$, for $T(L) = \min\{t : \#\{\tau \leq t, \tau \in J_\varepsilon\} = L$ we have

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

Then a union bound over all L gives together with a similar argument as in Section 3.1.4 that with probability $1 - \delta/T$

$$\sum_{k \in K_\varepsilon} \mathbf{v}_k(\mathbf{P}_k - \mathbf{I})\tilde{\boldsymbol{\lambda}}_k \mathbf{1}_{M \in \mathcal{M}_k} \leq D \left(SA \log_2 \left(\frac{T}{SA} \right) + 2\sqrt{2L_\varepsilon \log \frac{T}{\delta}} \right). \quad (13)$$

The final regret term of $\sum_{k \in K_\varepsilon} \|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 \|\tilde{\boldsymbol{\lambda}}_k\|_\infty$ can be upper bounded just as (7) to be

$$\sum_{k \in K_\varepsilon} \|\mathbf{v}_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)\|_1 \|\tilde{\boldsymbol{\lambda}}_k\|_\infty = O(DS\sqrt{L_\varepsilon A \log(2AT/\delta)}). \quad (14)$$

Combining (11), (13), and (14) with an analogon of (4) yields that with probability $1 - \delta - \delta/T$

$$\varepsilon L_\varepsilon \leq \Delta'(s, T) \leq O\left(DS\sqrt{L_\varepsilon A \log(2AT/\delta)}\right) + O\left(DSA \log_2 \left(\frac{2T}{SA}\right)\right),$$

where $\Delta'(s, T)$ is the regret in ε -bad episodes. This implies

$$L_\varepsilon = O\left(\frac{D^2 S^2 A \log(AT/\delta)}{\varepsilon^2}\right)$$

with probability $1 - \delta - \delta/T$, proving Theorem 2.

4 Proof Sketch for the Lower Bound

We consider an MDP with two states, $\mathcal{S} = \{0, 1\}$, and A actions. For all $a \in \mathcal{A}$, let $r(0, a) = 0$, $r(1, a) = 1$, and $p(0|1, a) = \delta$. For all but a single action $a^* \in \mathcal{A}$ let $p(1|0, a) = \delta$ whereas

$p(1|0, a^*) = \delta + \varepsilon$, $\varepsilon < \delta$. Such an MDP is depicted in Figure 2, where the single action a^* with

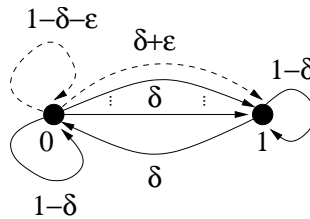


Figure 2: An MDP.

higher transition probability from state 0 to state 1 is shown as dashed line. The diameter of this MDP is $D = 1/\delta$. The average reward of a policy which chooses action a^* in state 0 is $\frac{\delta+\varepsilon}{2\delta+\varepsilon}$, while the average reward of any other policy is $\frac{1}{2}$. Thus the regret of a suboptimal policy in T steps is $\Omega(\varepsilon T/\delta)$.

To detect the better action a^* reliably, any action in state 0 needs to be probed $\Omega(\delta/\varepsilon^2)$ times. If we consider $k := \lceil S/2 \rceil$ copies of this MDP, where only one copy has such a better action a^* , then all actions in all 0-states need to be probed $\Omega(\delta/\varepsilon^2)$ times. Setting $\varepsilon = \sqrt{\delta k A/T}$ this takes $\Omega(T)$ time and yields $\Omega(\sqrt{S A T/\delta}) = \Omega(\sqrt{D S A T})$ regret.

Finally, connecting the 0-states of the S copies by $A + 1$ additional deterministic actions per state, gives an MDP with diameter $D + 2 \log_A S$ (e.g. by using an A -ary tree). Since any MDP with A actions and S states has diameter $\Omega(\log_A S)$, the theorem follows.

5 Outlook

Concerning the gap between the upper and the lower bound, we conjecture that the lower bound gives the right exponents for the parameters S and D . However, taking a look at the proof of the upper bound it seems to be hard to get rid of the additional factors \sqrt{S} and \sqrt{D} .

References

- [1] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for reinforcement learning. In *Proc. 19th NIPS*, pages 49–56. MIT Press, 2006.
- [2] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. J. (2)*, 19:357–367, 1967.
- [3] 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.
- [4] Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for Markov decision processes. *Math. Oper. Res.*, 22(1):222–255, 1997.
- [5] D.A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3:100–118, 1975.
- [6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

- [7] Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Proc. 11th NIPS*. MIT Press, 1999.
- [8] Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.*, 49:209–232, 2002.
- [9] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.
- [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. preprint, 2006.
- [12] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- [13] Ambuj Tewari and Peter L. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. In *Proc. 20th NIPS*, 2007. to appear.
- [14] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marco L. Weinberger. Inequalities for the L1 deviation of the empirical distribution. online, 2003.

A Proof of Lemma 6.

Lemma 6. *If $M \in \mathcal{M}_k$ then for policy $\tilde{\pi}_k$ in MDP \tilde{M}_k as chosen by algorithm UCRL2, there is a vector $\tilde{\lambda}_k$ satisfying the Poisson equation (1) with*

$$\|\tilde{\lambda}_k\|_\infty \leq D(M)/2.$$

Proof. We first show that $\tilde{\pi}_k$ is also average reward optimal in the following modified MDP M_k^+ in which $D(M_k^+) \leq D(M)$, if $M \in \mathcal{M}_k$. In order to obtain M_k^+ we take M and add for each action $a \in \mathcal{A}$ an action a' with rewards $r(s, a') := r(s, a)$ and transition probabilities $p_k^+(s'|s, a') := \tilde{p}_k(s'|s, a)$. Clearly, $D(M_k^+) \leq D(M)$, and since $M \in \mathcal{M}_k$, $\tilde{\pi}_k$ is also average reward optimal and minimizes the max-norm of the bias in M_k^+ . The average reward and bias of $\tilde{\pi}_k$ in M_k^+ are $\tilde{\rho}_k$ and $\tilde{\lambda}_k$, as in \tilde{M}_k . Since the Poisson equation (1) is solved by $\lambda + c\mathbf{1}$ if it is solved by λ , it is sufficient to show that there exists an average reward optimal policy π^+ for MDP M_k^+ with bias λ^+ such that $d := \max_s \lambda^+(s) - \min_s \lambda^+(s) \leq D(M) := D$.

Existence of π^+ : Let π^* be any average reward optimal policy in M_k^+ , ρ^* and λ^* be the average reward and bias of π^* . Let $s_0 = \arg \max_{s \in \mathcal{S}} \{\lambda^*(s)\}$ be a state with maximal bias. Assume that there is a state $s \in \mathcal{S}$ with $\lambda^*(s_0) - \lambda^*(s) > D$, for otherwise we use $\pi^+ = \pi^*$. We show how to construct a policy π^+ for M_k^+ with bias λ^+ such that $\max_s \lambda^+(s) - \min_s \lambda^+(s) \leq D$. Let $T_{s,s'}^\pi$ be the expected time to reach state s' from state s in M_k^+ under policy π . We write π_0 for the policy moving as fast as possible from every state to s_0 in M_k^+ and thus realizing⁶

$$T_{s,s_0}^{\pi_0} = \arg \min_{\pi} \left\{ T_{s,s_0}^\pi \right\}.$$

Now we identify a set of states \mathcal{S}' with small bias under π^* and define the policy π^+ and a vector z as follows:

$$\begin{aligned} \mathcal{S}' &:= \left\{ s : \lambda^*(s_0) - \lambda^*(s) > T_{s,s_0}^{\pi_0} \right\}, \\ \pi^+(s) &:= \begin{cases} \pi^*(s) & \text{if } s \notin \mathcal{S}', \\ \pi_0(s) & \text{if } s \in \mathcal{S}', \end{cases} \\ z(s) &:= \begin{cases} \lambda^*(s) - \lambda^*(s_0) & \text{if } s \notin \mathcal{S}', \\ -T_{s,s_0}^{\pi_0} & \text{if } s \in \mathcal{S}'. \end{cases} \end{aligned}$$

Note that this guarantees that for all s

$$z(s) \geq -T_{s,s_0}^{\pi_0} \tag{15}$$

and

$$z(s) + \lambda^*(s_0) \geq \lambda^*(s). \tag{16}$$

We claim that $r(s, \pi^+(s)) - \rho^* + p(\cdot|s, \pi^+(s))z \geq z(s)$ for all s . Indeed, if $s \notin \mathcal{S}'$ we have by

⁶Such a policy can be computed by constructing an MDP such that an average reward optimal policy for it has that property.

(16) and the Poisson equation (1)

$$\begin{aligned} r(s, \pi^+(s)) - \rho^* + p(\cdot|s, \pi^+(s)) \mathbf{z} &= r(s, \pi^*(s)) - \rho^* + p(\cdot|s, \pi^*(s)) (\mathbf{z} + \boldsymbol{\lambda}^*(s_0)\mathbf{1}) - \boldsymbol{\lambda}^*(s_0) \\ &\geq r(s, \pi^*(s)) - \rho^* + p(\cdot|s, \pi^*(s)) \boldsymbol{\lambda}^* - \boldsymbol{\lambda}^*(s_0) = \boldsymbol{\lambda}^*(s) - \boldsymbol{\lambda}^*(s_0) = \mathbf{z}(s). \end{aligned}$$

On the other hand, if $s \in \mathcal{S}'$ we get by (15) and since $\rho^*, r(s, \pi^+(s)) \in [0, 1]$

$$r(s, \pi^+(s)) - \rho^* + p(\cdot|s, \pi^+(s)) \mathbf{z} \geq -1 - \sum_{s' \in \mathcal{S}} p(s'|s, \pi_0(s)) \cdot T_{s', s_0}^{\pi_0} = -T_{s, s_0}^{\pi_0} = \mathbf{z}(s).$$

Thus there is a vector \mathbf{x} with no negative entry such that

$$\mathbf{z} = -\mathbf{x} + \mathbf{r}^+ - \rho^* \mathbf{1} + \mathbf{P}^+ \mathbf{z}, \quad (17)$$

where \mathbf{P}^+ and \mathbf{r}^+ are the transition matrix and the reward vector of policy π^+ in M_k^+ . Let \mathcal{R}^+ denote the set of recurrent states in M_k^+ under π^+ . Since the transition probability from states in \mathcal{R}^+ to any state not in \mathcal{R}^+ is zero under π^+ the restriction of (17) to \mathcal{R}^+ holds. So we have

$$\mathbf{z}_{\mathcal{R}^+} = -\mathbf{x}_{\mathcal{R}^+} + \mathbf{r}_{\mathcal{R}^+}^+ - \rho^* \mathbf{1}_{\mathcal{R}^+} + \mathbf{P}_{\mathcal{R}^+}^+ \mathbf{z}_{\mathcal{R}^+}, \quad (18)$$

where \mathcal{R}^+ in the subscript means that only the rows (and for a matrix also columns) corresponding to a state in \mathcal{R}^+ are considered. Bias $\boldsymbol{\lambda}'$ and average reward⁷ ρ' of an MDP with transition matrix \mathbf{P}^+ and rewards \mathbf{x} satisfy the equation

$$\boldsymbol{\lambda}'_{\mathcal{R}^+} = \mathbf{x}_{\mathcal{R}^+} - \rho' \mathbf{1}_{\mathcal{R}^+} + \mathbf{P}_{\mathcal{R}^+}^+ \boldsymbol{\lambda}'_{\mathcal{R}^+}. \quad (19)$$

Adding (18) and (19) we get

$$\mathbf{z}_{\mathcal{R}^+} + \boldsymbol{\lambda}'_{\mathcal{R}^+} = \mathbf{r}_{\mathcal{R}^+}^+ - \rho' \mathbf{1}_{\mathcal{R}^+} + \rho^* \mathbf{1}_{\mathcal{R}^+} + \mathbf{P}_{\mathcal{R}^+}^+ (\mathbf{z}_{\mathcal{R}^+} + \boldsymbol{\lambda}'_{\mathcal{R}^+}). \quad (20)$$

Case A: There is a state $s \in \mathcal{R}^+$ with $\mathbf{x}(s) > 0$:

Then we have $\rho'_{\mathcal{R}^+} \geq \rho^* \mathbf{1}_{\mathcal{R}^+}$ for some $\rho' > 0$. Since the Poisson equation (1) and thus (20) uniquely determines the average reward of \mathcal{R}^+ under π^+ , this result immediately implies that there is a policy with unichain transition matrix and average reward $\rho' + \rho^* > \rho^*$, which contradicts average reward optimality of π^* in M_k^+ .

Case B: For the recurrent states we have $\mathbf{x}_{\mathcal{R}^+} = \mathbf{0}$:

We first show that $\mathcal{R}^+ \cap \mathcal{S}' = \emptyset$. Note that \mathcal{S}' is not empty, and assume that $\mathcal{R}^+ \cap \mathcal{S}'$ is not empty as well. Further note that $\mathcal{R}^+ \setminus \mathcal{S}' \neq \emptyset$: If $\mathcal{R}^+ \subseteq \mathcal{S}'$ then the policy was changed to π_0 in all states of \mathcal{R}^+ and, since $s_0 \notin \mathcal{S}'$ by definition, \mathcal{R}^+ is not recurrent under π^+ . Thus consider any state $s_u \in \mathcal{R}^+ \setminus \mathcal{S}'$ where the policy is left unchanged. From the Poisson equation for π^* we have

$$r(s_u, \pi^*(s_u)) - \rho^* + \sum_{s \in \mathcal{S}} p(s|s_u, \pi^*(s_u)) (\boldsymbol{\lambda}^*(s) - \boldsymbol{\lambda}^*(s_0)) = \boldsymbol{\lambda}^*(s_u) - \boldsymbol{\lambda}^*(s_0). \quad (21)$$

⁷The average reward ρ' may depend on the recurrence class for reward vector \mathbf{x} .

Since s_u is in the recurrent component \mathcal{R}^+ , we have

$$\sum_{s \in \mathcal{S} \setminus \mathcal{R}^+} p(s|s_u, \pi^*(s_u)) = 0.$$

By definition we have $\lambda^*(s_u) - \lambda^*(s_0) = \mathbf{z}(s_u)$, so that (21) and (18) give

$$\begin{aligned} r(s_u, \pi^*(s_u)) - \rho^* + \sum_{s \in \mathcal{R}^+} p(s|s_u, \pi^*(s_u)) (\lambda^*(s) - \lambda^*(s_0)) &= \lambda^*(s_u) - \lambda^*(s_0) \\ &= \mathbf{z}(s_u) = r(s_u, \pi^+(s_u)) - \rho^* + \sum_{s \in \mathcal{R}^+} p(s|s_u, \pi^+(s_u)) \mathbf{z}(s). \end{aligned}$$

Hence, because of $\pi^*(s_u) = \pi^+(s_u)$,

$$\sum_{s \in \mathcal{R}^+} p(s|s_u, \pi^*(s_u)) (\lambda^*(s) - \lambda^*(s_0)) = \sum_{s \in \mathcal{R}^+} p(s|s_u, \pi^*(s_u)) \mathbf{z}(s), \quad (22)$$

and subtracting

$$\sum_{s \in \mathcal{R}^+ \setminus \mathcal{S}'} p(s|s_u, \pi^*(s_u)) (\lambda^*(s) - \lambda^*(s_0)) = \sum_{s \in \mathcal{R}^+ \setminus \mathcal{S}'} p(s|s_u, \pi^*(s_u)) \mathbf{z}(s) \quad (23)$$

from both sides gives

$$\sum_{s \in \mathcal{R}^+ \cap \mathcal{S}'} p(s|s_u, \pi^*(s_u)) (\lambda^*(s) - \lambda^*(s_0)) = \sum_{s \in \mathcal{R}^+ \cap \mathcal{S}'} p(s|s_u, \pi^*(s_u)) \mathbf{z}(s).$$

Now since $\mathbf{z}(s) > \lambda^*(s) - \lambda^*(s_0)$ for all $s \in \mathcal{R}^+ \cap \mathcal{S}'$, it follows that $p(s|s_u, \pi^*(s_u)) = 0$ for all $s \in \mathcal{R}^+ \cap \mathcal{S}'$. Because this holds for any $s_u \in \mathcal{R}^+ \setminus \mathcal{S}'$ there is no transition from $\mathcal{R}^+ \setminus \mathcal{S}'$ to $\mathcal{R}^+ \cap \mathcal{S}'$, which contradicts our assumption that \mathcal{R}^+ is a recurrent component. This completes the proof of $\mathcal{R}^+ \cap \mathcal{S}' = \emptyset$.

Since the policy in all states in \mathcal{R}^+ is left unchanged and the bias of the recurrent states does not depend on the transient states, the bias of all states in \mathcal{R}^+ is the same under policies π^* and π^+ . Further, due to the definition of \mathcal{S}' , $\lambda^+ \geq \lambda^*(s_0) - D$ for all $s \notin \mathcal{S}'$.

Because of $\mathbf{x}_{\mathcal{R}^+} = \mathbf{0}$ we have $\rho' = \mathbf{0}$ and $\lambda'_{\mathcal{R}^+} = \mathbf{0}$, and thus π^+ is average reward optimal due to (20). It is easy to see that the bias of any state $s \notin \mathcal{R}^+$ under policy π^+ is $\lambda^+(s) = \mathbf{z}(s) + \lambda'(s) + \lambda^*(s_0)$. Since $\mathbf{x}_{\mathcal{R}^+} = \mathbf{0}$ implies $\lambda'(s) \geq 0$ for all states s , this shows due to (15) that $\lambda^+(s) \geq \lambda^*(s_0) - D$. On the other hand, if for some state $s \notin \mathcal{R}^+$ we have $\lambda^+(s) \geq \lambda^*(s_0)$ we get by a further iteration of our argument—modifying π^+ once more—a contradiction as in case A. Thus we get $d = \max_s \{\lambda^+(s)\} - \min_s \{\lambda^+(s)\} \leq D$ and the claim of the lemma follows. \square

B Proof of Lemma 7

Lemma 7. *Using the notation of algorithm UCRL2, $\mathbb{P}\{M \notin \mathcal{M}(t)\} \leq \delta/t^5$.*

Proof. The L1 deviation of the empirical distribution over m distinct events from n samples is

bounded by (cf. Weissman et al. [14])

$$\mathbb{P} \left\{ \left\| \hat{p}(\cdot) - p(\cdot) \right\|_1 \geq \varepsilon \right\} \leq [2^m - 2] \exp \left(-\frac{n\varepsilon^2}{2} \right).$$

Setting

$$\varepsilon = \sqrt{\frac{2}{n} \log(2^S S A t^6 / \delta)} \leq \sqrt{\frac{12S}{n} \log(2At/\delta)}$$

we get for every state-action pair

$$\mathbb{P} \left\{ \left\| p(\cdot|s, a) - \tilde{p}_t(\cdot|s, a) \right\|_1 \geq \sqrt{\frac{12S}{n} \log(2At/\delta)} \right\} \leq 2^S \exp \left(-\frac{n}{2} \frac{2}{n} \log(2^S S A t^6 / \delta) \right) \leq \frac{\delta}{t^6 S A}.$$

Summing up over all n from 1 to t and over all state-action pairs proves the lemma. \square

C Proof of (10)

For every state-action pair denote $k(s, a)$ the episode where action a is used for the first time in state s . Then we have

$$\sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} \leq \sum_{s,a} \left(1 + \sum_{k=k(s,a)+1}^m \frac{N_{k+1}(s, a) - N_k(s, a)}{\sqrt{N_k(s, a)}} \right).$$

Using $2N_m(s, a) \geq N_{m+1}(s, a)$ and $2N_k(s, a) \geq N_{k+1}(s, a)$ for the first and $N_m(s, a) \leq N_{m+1}(s, a)$ for the second inequality, we find for every state-action pair

$$\begin{aligned} & 1 + \sum_{k=k(s,a)+1}^m \frac{N_{k+1}(s, a) - N_k(s, a)}{\sqrt{N_k(s, a)}} = \\ & = 1 + \frac{N_{m+1}(s, a)}{\sqrt{N_m(s, a)}} + \sum_{k=k(s,a)+1}^{m-1} \left[\frac{N_{k+1}(s, a)}{\sqrt{N_k(s, a)}} - \frac{N_{k+1}(s, a)}{\sqrt{N_{k+1}(s, a)}} \right] - \frac{N_1(s, a)}{\sqrt{N_1(s, a)}} \\ & = 1 + \frac{\sqrt{2}N_{m+1}(s, a)}{\sqrt{2N_m(s, a)}} + \sum_{k=k(s,a)+1}^{m-1} N_{k+1}(s, a) \sqrt{2} \frac{\sqrt{N_{k+1}(s, a)} - \sqrt{N_k(s, a)}}{\sqrt{2N_k(s, a)}\sqrt{N_{k+1}(s, a)}} - 1 \\ & \leq \sqrt{2N_{m+1}(s, a)} + \sqrt{2} \sum_{k=k(s,a)+1}^{m-1} \left[\sqrt{N_{k+1}(s, a)} - \sqrt{N_k(s, a)} \right] \\ & = \sqrt{2N_{m+1}(s, a)} + \sqrt{2N_m(s, a)} - \sqrt{2} \leq \sqrt{8N_{m+1}(s, a)}. \end{aligned}$$

Since $\forall x \in \mathbb{R}^n : \|x\|_1 \leq \sqrt{n}\|x\|_2$ and $\sum_{s,a} N_{m+1}(s, a) = T$ we get

$$\sum_{k=1}^m \sum_{s,a} \frac{v_k(s, a)}{\sqrt{\max\{1, N_k(s, a)\}}} = \sqrt{8} \sum_{s,a} \sqrt{N_{m+1}(s, a)} \leq \sqrt{8} \sqrt{SA} \cdot \sqrt{T}.$$