Online Regret Bounds for Markov Decision Processes with Deterministic Transitions[☆]

Ronald Ortner

Department Mathematik und Informationstechnologie, Montanuniversität Leoben, A-8700 Leoben, Austria

Abstract

We consider an upper confidence bound algorithm for learning in Markov decision processes with deterministic transitions. For this algorithm we derive upper bounds on the *online* regret with respect to an (ε-)optimal policy that are logarithmic in the number of steps taken. We also present a corresponding lower bound. As an application, multi-armed bandits with switching cost are considered.

Key words:
Markov decision process, regret, labeled digraph

1. Introduction

This paper considers learning in Markov decision processes (MDPs) with deterministic transitions. Unlike in general MDPs, a learner can easily determine the MDP's transition structure. After that, the remaining problem is to deal with the exploitation-exploration problem concerning the rewards. Thus, the situation is similar to a multi-armed bandit problem. However, dealing with deterministic transition MDPs that way does not give any satisfying bounds, as in general the number of different policies to consider is exponential in the number of states. In the following, we present an algorithm (a simple generalization of the UCB1 algorithm of Auer et al. [2]) that achieves logarithmic regret in the number of steps taken. More precisely, after T steps the regret is $O\big(\frac{|A|\log T}{\Delta}\big)$ for MDPs with action space A and a gap of Δ between the optimal and the second-best average reward of a deterministic policy. We point out that unlike in the general MDP setting where A usually is the set of actions available in each single state, here we assume that each state s has an individual set $A(s)$ of available actions, and A is the union of these disjoint sets. Thus, $|A|$ in our setting corresponds to $|S||A|$ in the more usual setting with S being the state

[☆] A preliminary version of this paper appeared as [1].

Email addresses: ronald.ortner@unileoben.ac.at (Ronald Ortner)

space. Note that our bound corresponds to the bound in the original bandit setting as given by Auer et al. [2].

There are also logarithmic regret bounds for general (average reward) MDPs with state space S and a set of actions A available in each state. The first of these bounds due to Burnetas and Katehakis [3] was recently generalized by Tewari and Bartlett [4] (at the cost of a worse constant in the bound). This latter bound is of order $O\big(\frac{\kappa^2|A||S|\log T}{\Delta}\big)$ for an MDP dependent parameter κ, but — as the original bound of Burnetas and Katehakis [3] — it holds only asymptotically and makes the assumption that the MDP is *ergodic*, i.e., any two states are connected by *any* policy. We do not make this assumption in our setting.

Finite horizon bounds have been achieved by Auer and Ortner [5] and have further been improved by Auer et al. [6]. This improved bound of $O\big(\frac{D^2|A||S|^2\log T}{\Delta}\big)$ for an MDP dependent parameter D has slightly worse dependence on the parameters than the bound of Tewari and Bartlett [4], yet it holds more generally in *communicating* MDPs, where each two states are connected by a suitable policy. Recently, modifying algorithm and methods of Auer et al. [6], Bartlett and Tewari [7] managed to replace the parameter D in the mentioned regret bound with a smaller parameter $D_1 \leq D$. Moreover, their bound also holds when the MDP has some *transient* states that are not reachable under any policy. However, this bound is only obtained when the learner knows an upper bound on the parameter D_1. In case the learner has no such upper bound, a doubling trick can be applied which however deteriorates the bound's dependence on the number of states from $|S|$ to $|S|^{3/2}$.

The MDP dependent parameters in the mentioned logarithmic bounds are transition parameters (roughly, the expected time needed to connect either any two states [4, 6], or any state with some particular state [7], respectively). In the general MDP setting such a parameter is necessary as the lower bounds given by Auer et al. [6] and Bartlett and Tewari [7] show. In the deterministic transition case we achieve finite horizon bounds whose main term is not dependent on any similar parameter. The diameter of the MDP's underlying transition graph only appears in an additional term stemming from the costs incurred by switching policies. Thus, MDPs with deterministic transitions resemble more the multi-armed bandit case (with some kind of switching cost) than the general MDP case.

The cost of deriving finite horizon bounds instead of asymptotic bounds is usually that optimality is lost. Thus, while the asymptotic bound of Burnetas and Katehakis [3] was shown to be optimal, there is still a gap between the lower and upper bound on the finite horizon regret given by Auer et al. [6]. In our case, it is possible to come at least quite close to optimality. We give a lower bound on the regret that matches the main term of the upper bounds. Concerning the term for switching policies, we will indicate that such a term is necessary as well. However, this lower bound on the switching cost term does not quite match the switching cost term of the upper bound obtained for our algorithm.

1.1. Outline

We proceed with definitions and some basic observations concerning MDPs with deterministic transitions. Section 3 then introduces the upper confidence bound algorithm UCycle for the considered deterministic transition MDP setting. In Section 4, we prove a logarithmic bound on the expected regret of UCycle and complement it with a bound that holds with high probability. A lower bound is derived as well. Finally, in Section 5 we consider the setting of multi-armed bandits with switching cost as a special case of deterministic transition MDPs.

2. MDPs with Deterministic Transitions

A *Markov decision process* (*MDP*) [8] can be specified as follows. There is a finite set of *states* S and a finite set of *actions* A such that for each state s there is a nonempty set $A(s) \subset A$ of actions that are available in s. We assume that $A(s) \cap A(s') = \varnothing$ for $s \neq s'$, and $A = \bigcup_{s \in S} A(s)$. Actually, it is more usual to assume that the sets $A(s)$ coincide for all states s, yet for our purposes it is more useful to consider distinct action sets. For a state $s \in S$ and an action $a \in A(s)$ *transition probability* distributions $p(\cdot|s,a)$ determine the probability $p(s'|s,a)$ that choosing a in s leads to state s'. Further, a *reward* function r gives the mean $r(s,a)$ of the random reward obtained for choosing action a in state s. We assume that successively choosing action a in state s gives random rewards $r_1(s,a), r_2(s,a), \ldots$, which are independent and identically distributed according to an unknown probability distribution with support in $[0,1]$. Generally, the rewards $r_i(s,a)$ and $r_j(s',a')$ shall be independent for all states s, s', all actions $a \in A(s)$, $a' \in A(s')$, and all $i, j \in \mathbb{N}$.

A policy is a function $\pi : S \to A$ that assigns each state s a fixed action $\pi(s) \in A(s)$. The *average reward of a policy* π is defined as

$$\rho_\pi(s_0) := \lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} r\big(s_t, \pi(s_t)\big),$$

where the process starts in s_0, and generally, s_t is a random variable for the state at step t.

In MDPs with *deterministic* transitions, for all states s and all $a \in A(s)$ we assume that $p(s'|s,a) = 1$ for a unique $s' \in S$, while $p(s''|s,a) = 0$ for all $s'' \neq s'$. Thus each action leads deterministically from one state to another (or the same) state, so that the transition structure may be considered as a directed graph (loops allowed) with vertex set S and edge set $\bigcup_{s \in S} A(s) = A$. Accordingly, in the following we will use the terms *action* and *edge* synonymously. As we assume that the action sets $A(s)$ are pairwise disjoint, the mean reward $r(s,a)$ depends only on the edge a in this *transition (di)graph*, so that we will write $r(a)$ for the mean reward of edge a. Similarly, $r_i(a)$ will denote the random reward for the i-th visit in edge a. Summarizing, a deterministic transition MDP may be considered as a directed graph where the edges are labeled with the respective mean rewards.

We introduce some terminology from graph theory. Given a graph with vertex set V and a set $E \subseteq V^2$ of directed edges, an edge $(v, v') \in E$ is said to *start* in its *initial vertex* v and *end* in its *terminal vertex* v'. We also say that (v, v') is an *outgoing edge of* v. A (directed) *path* is a sequence of edges $e_1, e_2, \ldots, e_\ell$ such that for $2 \leq i \leq \ell$ the edge e_i starts in the same vertex in which edge e_{i-1} ends. Such a path is called a (directed) *cycle*, if the initial vertex of e_1 is identical to the terminal vertex of e_ℓ. Paths and cycles are called *simple*, if the initial vertices of all edges are pairwise distinct. In the following, we will often sloppily identify a simple cycle with the set of its edges.

As we assume that $A(s) \neq \varnothing$ for all $s \in S$, each state has at least one outgoing edge, so that playing an arbitrary but fixed policy π eventually leads into a directed simple cycle $a_1^\pi, a_2^\pi, \ldots, a_\ell^\pi$. A policy may induce more than one such cycle, and the cycle that is eventually reached depends on the initial state. Generally, any policy π will partition the edge set A into one or more cycles and a (possible empty) set of *transient* edges not contained in any cycle. However, starting in such a transient edge leads to a cycle, so that each edge can be uniquely assigned to an induced cycle. Consequently, depending on the initial state s_0, the average reward ρ_π of a policy π can be written as

$$\rho_\pi(s_0) = \frac{1}{\ell} \sum_{i=1}^{\ell} r(a_i^\pi),$$

where $a_1^\pi, a_2^\pi, \ldots, a_\ell^\pi$ is the respective cycle induced by π and s_0. We are interested in the optimal policy π^* that gives maximal reward ρ^*,[1] which basically means that we are looking for a cycle with maximal mean reward.

As one step in our suggested algorithm for the learning setting is to determine an optimal cycle (in an optimistic estimate of the MDP), we briefly point out possibilities how to deal with this task. The first algorithm for finding the optimal cycle mean in a labeled digraph has been suggested by Karp [9]. His algorithm is based on a formula which expresses the optimal cycle mean via optimal weights $w_k(v)$ of paths of length k from a fixed source vertex to the vertex v. The weights $w_k(v)$ can be calculated via a recurrence relation, which results in an algorithm with run-time $O(|A||S|)$ and $\Omega(|A||S|)$. For our purposes Karp's algorithm is in principle sufficient. Still, some refinements are possible [10] which improve the run-time in some cases. For an overview of algorithms (some dealing with more general problems [11, 12]) and their experimental evaluation see [13]. Finally, note that as for general MDPs standard value iteration may be used to find an optimal cycle. The run-time behavior of value iteration on MDPs with deterministic transitions has been studied by Madani [14].

We consider the learning setting when the MDP is not known and a learner can only observe her current state and the actions she may choose in this state.

[1] It can be shown that allowing time-dependent policies does not increase the achievable maximal reward. This also holds in the general MDP setting (see Puterman [8]).

As a measure how well a learning algorithm works, we consider its *regret* after a finite number of T steps with respect to an optimal policy, defined as

$$R_T := T\rho^* - \sum_{t=1}^{T} r_t,$$

where r_t is the random reward received by the algorithm at step t. We also consider the *regret $R_{T,\varepsilon}$ with respect to an ε-optimal policy*, when the learner does not compete with the optimal average reward ρ^* but only with $\rho^* - \varepsilon$ for some $\varepsilon > 0$.

Note that if the transition graph of the MDP is not *strongly connected*[2], the achievable optimal reward ρ^* will depend on the initial state (as the optimal cycle may not be reachable from each initial state). Even if the learner may in principle reach an optimal cycle from her initial state, as she first has to explore the transition structure of the MDP, choosing a wrong action may lead into a suboptimal part of the state space that cannot be left anymore. In this case it seems fair to compete at each step with the optimal reward achievable in the strongly connected part containing the learner's current state.[3] As we assume deterministic transitions, any learner that explores all actions (which in general is obviously necessary) will eventually reach a strongly connected part that cannot be left anymore. Since our proposed learning algorithm will have explored all actions after at most $|S||A|$ steps (see Proposition 2 below), in the following we may simply assume that the transition graph is strongly connected, so that ρ^* depends only on the MDP, and we may sloppily identify optimal policies with optimal cycles. The additional regret in the general case is at most $|S||A|$.

3. An Upper Confidence Bound Algorithm

As algorithm for the deterministic transition MDP setting we suggest a simple adaptation of known upper confidence bound algorithms such as UCB1 [2] (for multi-armed bandits) or UCRL2 [6] (for communicating MDPs). The common idea of such algorithms is to choose an optimal policy in an optimistic but plausible model of the situation, where plausibility is specified by confidence intervals for the estimated parameters (rewards, transition probabilities) of the system.

[2] A digraph is called *strongly connected* if there is a directed path between any two vertices. Note that a deterministic transition MDP is communicating iff the transition graph is strongly connected.

[3] This basically has been suggested as one possible approach for learning in general (i.e., not necessarily communicating) MDPs by Kearns and Singh [15]. By the way, the alternative suggestion of Kearns and Singh [15] to compete with $\min_s \rho^*(s)$, where $\rho^*(s)$ is the highest achievable reward when starting in s, seems to be too weak. A lucky learner may reach a part of the MDP in which the reward is larger than $\min_s \rho^*(s)$ for *any* policy. In that case, it seems to be more natural to compete with the highest reward achievable *in that part*.

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

Initialization and determination of transition structure:

While some action has not been sampled yet **do**:

▷ **If** there is an unsampled action $a \in A(s)$ in the current state s, choose a.

▷ **Otherwise** choose shortest path to a state s' that has already been visited before and in which there is an unsampled action.[4]

For episodes $i = 1, 2, \ldots$ **do**

▷ *Calculate optimal cycle for episode i:*
Determine an optimal cycle $\tilde{C}_i$ in the MDP with transition structure as determined and whose rewards for any action a are given by

$$\tilde{r}_t(a) := \hat{r}_t(a) + \sqrt{\frac{\log \frac{|A|t^4}{\delta}}{2n_t(a)}}, \tag{1}$$

where t is the overall number of steps taken so far, $\hat{r}_t(a)$ is the average reward obtained from action a, and $n_t(a)$ is the number of times action a was chosen.

▷ *Transition phase:* Take a shortest path to a state in the cycle $\tilde{C}_i$.

▷ *Cycle phase:* Run through the whole cycle $\tilde{C}_i$ for $\min_{a \in \tilde{C}_i} n_t(a)$ times,[5]when episode i is terminated.

Figure 1: The UCycle algorithm.

In the case of deterministic transition MDPs, the upper confidence bound strategy will be applied only to the rewards. As the transitions are assumed to be deterministic (and the learner is aware of this fact), they can easily be determined with certainty. Thus, our suggested algorithm UCycle first investigates the transition structure of the MDP by playing each available action in each state once. Then an upper confidence bound strategy is applied to the rewards associated with each action in order to determine the cycle $\tilde{C}$ with the highest average plausible reward. As indicated above, plausibility means that the reward is contained in some suitable confidence interval. The optimal cycle can be computed efficiently by any of the algorithms from the literature mentioned in the introduction. After computing the optimal cycle $\tilde{C}$, the algorithm chooses the shortest route to a state in $\tilde{C}$ and remains in $\tilde{C}$ for an appropriate number of time steps (cf. discussion below). The algorithm is depicted in Figure 1.

Note that UCycle proceeds in episodes of increasing length. In fact, it is

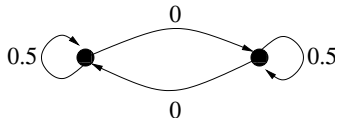

Figure 2: Switching policies too often may lead to linear regret.

a tempting but bad idea to switch the cycle whenever another cycle looks more promising. The following example demonstrates that there are very simple cases where this strategy leads to linear regret.

Example 1. *Consider the MDP shown in Figure 2, where not only the transitions but also all the rewards are assumed to be deterministic. There are obviously two optimal cycles, viz. the loops in each of the two states with optimal average reward of* $\frac{1}{2}$*. If we would take our upper confidence bound approach and choose the more promising loop at each step, then each loop would be played only twice, before the other loop has a higher upper confidence bound (due to the larger confidence interval). As switching (which hence happens each third step) gives no reward, the average reward after* T *steps will be at most* $\frac{2}{3} \cdot \frac{1}{2}T = \frac{1}{3}T$*, so that the regret of this strategy is* $\Omega(T)$*. Note that our* UCycle *algorithm also keeps switching between the two optimal loops, but the number of switches is* $O(\log T)$*.*

4. Online Regret Bounds

4.1. Logarithmic Upper Bounds

The bounds in this section improve the respective bounds of the previous version of this paper [1] which contained an additional factor λ, the largest simple cycle length in the transition graph. The main idea of the proof of the original bounds [1] was to determine a sufficient precision for the estimates of the rewards in order to guarantee the optimality of the chosen cycle $\tilde{C}_i$. Unlike that, for the bounds given below the intuition is that the suffered regret is upper bounded by the sum of the lengths of the confidence intervals.

We start with some basic properties of UCycle. First, we bound the number of episodes and the time spent on the determination of the transition graph. Then we bound the probability that at a given step t a confidence interval fails.

Proposition 2. *It takes at most* $|A||S|$ *steps until* UCycle *has determined the transition structure of any deterministic transition MDP.*

[4]The first condition guarantees that the learner need not know the state space in advance. Note that due to the condition of strong connectivity, the transition graph will be completely determined as soon as there is no such state s'. That way, only unsampled actions in the current and already visited states need to be considered in the loop, so that it is not necessary that the learner knows the number of actions in advance either.

[5]That way, each action in the selected cycle $\tilde{C}_i$ is chosen the same number of times.

Proof. It takes at most $|S|-1$ steps to reach a state in which there is an action to explore, and playing that action takes another step. Thus, $|A|$ distinct actions can be explored in at most $|A||S|$ steps. □

While this bound is not sharp for arbitrary $|S|$, it is easy to give examples where it is sharp at least asymptotically (for arbitrary $|S|$ and $|A| \to \infty$).

Proposition 3. *The number of episodes up to some step $T > |A|$ is upper bounded by $|A| \log_2 \frac{2T}{|A|}$.*

Proof. First note that a new episode starts only after the number of visits in one edge $a \in A$ has been doubled. Let M_a be the number of episodes which ended after the number of visits in a has been doubled, and let T_a be the number of steps in these episodes. As it may happen that in an episode the number of visits is doubled in more than one edge, we assume that M_a and T_a count only the episodes/steps where a is the first edge for which this happens. It is easy to see that $M_a \leq 1 + \log_2 T_a = \log_2 2T_a$ for $T_a > 0$. Further, $\sum_a \log_2 2T_a$ is maximal under the constraint $\sum_a T_a = T$ when $T_a = \frac{T}{|A|}$ for all a, which gives the claimed bound. □

Lemma 4. *(i) At each step t, the probability that some true mean reward is larger than the upper confidence bound value given in* (1) *is at most $\frac{\delta}{t^3}$, that is,*

$$\mathbb{P}\big\{\tilde{r}_t(a) < r(a) \text{ for some } a \in A\big\} \;\leq\; \frac{\delta}{t^3}.$$

(ii) Moreover, for each step t, it holds that

$$\mathbb{P}\Big\{\tilde{r}_t(a) - r(a) > 2\sqrt{\tfrac{\log(|A|t^4/\delta)}{2n_t(a)}} \text{ for some } a \in A\Big\} \;\leq\; \frac{\delta}{t^3}.$$

For the proof we apply the following special case of Hoeffding's inequality, which we will also need further down below.

Lemma 5 (Hoeffding's inequality [16])**.** *Let $X_1,\ X_2, \ldots, X_n$ be independent random variables with values in the unit interval $[0,1]$, and let S_n be the sum $X_1 + \ldots + X_n$. Then*

$$\begin{aligned} \mathbb{P}\big\{S_n - \mathbb{E}(S_n) > n\epsilon\big\} &\leq \exp(-2\epsilon^2 n), \\ \textit{and} \quad \mathbb{P}\big\{\mathbb{E}(S_n) - S_n > n\epsilon\big\} &\leq \exp(-2\epsilon^2 n). \end{aligned}$$

Proof of Lemma 4. For given $a \in A$ we have by Lemma 5 for all $t \in \mathbb{N}$ and all $n \leq t$,

$$\mathbb{P}\left\{ \tfrac{1}{n_t(a)} \sum_{i=1}^{n_t(a)} r_i(a) + \sqrt{\tfrac{\log(|A|t^4/\delta)}{2n_t(a)}} < r(a) \;\middle|\; n_t(a) = n \right\} \;\leq\; \frac{\delta}{|A|t^4}.$$

Because $\hat{r}_t(a) = \frac{1}{n_t(a)} \sum_{i=1}^{n_t(a)} r_i(a)$, a union bound over all possible values of $n_t(a)$ and all actions in A proves claim (i). The second statement follows analogously from the symmetric

$$\mathbb{P}\Big\{\hat{r}_t(a) - \sqrt{\tfrac{\log(|A|t^4/\delta)}{2n_t(a)}} > r(a) \;\Big|\; n_t(a) = n\Big\} \;\leq\; \frac{\delta}{|A|t^4}. \qquad \square$$

The error bounds of Lemma 4 allow to derive the following sample complexity bound on the number of steps taken in suboptimal cycles. The bound is logarithmic in the total number of steps taken and grows linearly with the *total* number of actions (i.e., the number of edges in the transition graph – recall that this corresponds to $|S||A|$ in the standard MDP setting).

Theorem 6. *The number of steps up to step T which* UCycle *(with input parameter δ) in the cycle phase spends in cycles whose average reward is smaller than $\rho^* - \varepsilon$ is upper bounded by*

$$\frac{12\,|A| \log \frac{|A|T^4}{\delta}}{\varepsilon^2},$$

provided that the confidence intervals given in Lemma 4 hold at the beginning of each episode. This latter condition holds with probability at least $1 - \frac{5}{2}\delta$.

Proof. Our assumption is that the confidence intervals given in Lemma 4 (i) and (ii) hold at the beginning of each episode. Note that this is assumption is guaranteed in particular when these confidence intervals hold for all t, which is true with probability at least $1 - 2\sum_t \frac{\delta}{t^3} > 1 - \frac{5}{2}\delta$.

Let $\mathcal{M}_\varepsilon$ be the set of all indices i of *ε-bad* episodes where UCycle chooses a cycle $\tilde{C}_i$ with expected reward $< \rho^* - \varepsilon$. Further, write τ_i for the length of the cycle phase of episode i. Finally, denote the average reward of a cycle C in the original MDP (with the real rewards) with $\rho(C)$ and its average reward in the optimistic MDP (with rewards $\tilde{r}$) with $\tilde{\rho}(C)$. We are interested in the value

$$\Delta_\varepsilon \;:=\; \sum_{i \in \mathcal{M}_\varepsilon} \Big(\rho^* - \rho(\tilde{C}_i)\Big)\,\tau_i, \tag{2}$$

which is basically the expected regret accumulated in these ε-bad episodes. Writing $N_\varepsilon := \sum_{i \in \mathcal{M}_\varepsilon} \tau_i$ for the total number of steps taken in the cycle phases of ε-suboptimal episodes gives the lower bound

$$\Delta_\varepsilon \;\geq\; \varepsilon N_\varepsilon. \tag{3}$$

In the rest of the proof we are going to derive also an upper bound on Δ_ε in terms of N_ε, which together with (3) will allow us to derive the claimed sample complexity bound.

Let C^* be an optimal cycle in the MDP. Then by our assumption on the confidence intervals we have by Lemma 4 (i)

$$\rho^* \;=\; \rho(C^*) \;\leq\; \tilde{\rho}(C^*) \;\leq\; \tilde{\rho}(\tilde{C}_i). \tag{4}$$

Further, writing t_i for the last step before episode i, due to (4), Δ_ε can be upper bounded as

$$\begin{aligned}\Delta_\varepsilon &\leq \sum_{i\in\mathcal{M}_\varepsilon}\Big(\tilde{\rho}(\tilde{C}_i)-\rho(\tilde{C}_i)\Big)\,\tau_i\\ &= \sum_{i\in\mathcal{M}_\varepsilon}\left(\frac{1}{|\tilde{C}_i|}\sum_{a\in\tilde{C}_i}\tilde{r}_{t_i}(a)-\frac{1}{|\tilde{C}_i|}\sum_{a\in\tilde{C}_i}r(a)\right)\tau_i\\ &= \sum_{i\in\mathcal{M}_\varepsilon}\sum_{a\in\tilde{C}_i}\frac{\tau_i}{|\tilde{C}_i|}\big(\tilde{r}_{t_i}(a)-r(a)\big). \qquad (5)\end{aligned}$$

Now let $\tau_i(a)$ denote the number of times edge a is visited in the cycle phase of episode i. Then we may rewrite (5) as

$$\Delta_\varepsilon \leq \sum_{i\in\mathcal{M}_\varepsilon}\sum_{a\in\tilde{C}_i}\tau_i(a)\big(\tilde{r}_{t_i}(a)-r(a)\big) = \sum_{a\in A}\sum_{i\in\mathcal{M}_\varepsilon}\tau_i(a)\big(\tilde{r}_{t_i}(a)-r(a)\big),$$

because $\tau_i(a)=0$, if $a\notin\tilde{C}_i$. Application of Lemma 4 (ii) shows that

$$\begin{aligned}\Delta_\varepsilon &\leq \sum_{a\in A}\sum_{i\in\mathcal{M}_\varepsilon}2\tau_i(a)\sqrt{\frac{\log(|A|t_i^4/\delta)}{2n_{t_i}(a)}}\\ &\leq \sqrt{2\log\tfrac{|A|T^4}{\delta}}\sum_{a\in A}\sum_{i\in\mathcal{M}_\varepsilon}\frac{\tau_i(a)}{\sqrt{n_{t_i}(a)}}. \qquad (6)\end{aligned}$$

Now one can show that (see Lemma 14 and its proof in Appendix A)

$$\sum_{i\in\mathcal{M}_\varepsilon}\frac{\tau_i(a)}{\sqrt{n_{t_i}(a)}} \leq \big(1+\sqrt{2}\big)\sqrt{n_\varepsilon(a)},$$

where $n_\varepsilon(a)$ is the total number of visits (up to the final step T) in edge a in the cycle phases of episodes with index in $\mathcal{M}_\varepsilon$. This yields from (6)

$$\Delta_\varepsilon < \sqrt{12\log\tfrac{|A|T^4}{\delta}}\sum_{a\in A}\sqrt{n_\varepsilon(a)}.$$

Since the term $\sum_{a\in A}\sqrt{n_\varepsilon(a)}$ under the constraint $\sum_{a\in A}n_\varepsilon(a)=N_\varepsilon$ is maximal when $n_\varepsilon(a)=N_\varepsilon/|A|$ for each $a\in A$ (so that $\sum_{a\in A}\sqrt{n_\varepsilon(a)}=\sqrt{|A|N_\varepsilon}$), it follows that

$$\Delta_\varepsilon < \sqrt{12|A|N_\varepsilon\log\tfrac{|A|T^4}{\delta}}. \qquad (7)$$

Combining (3) and (7) gives

$$\varepsilon N_\varepsilon < \sqrt{12|A|N_\varepsilon\log\tfrac{|A|T^4}{\delta}}.$$

Calculating N_ε then gives the claimed

$$N_\varepsilon < \frac{12|A|\log\frac{|A|T^4}{\delta}}{\varepsilon^2}. \qquad \square$$

Together with Proposition 3, Theorem 6 is sufficient to yield a *high probability* bound on the regret (Theorem 8 below). For the following bound on the *expected* regret we deal with the error probabilities in a slightly different way.

Theorem 7. *The expected regret of* UCycle *(with input parameter δ) after T steps with respect to an ε-optimal policy is upper bounded as*

$$\mathbb{E}(R_{T,\varepsilon}) \;\leq\; \frac{12\,|A|\log\frac{|A|T^4}{\delta}}{\varepsilon} + \big(D + \tfrac{20+15D}{12}\delta\big)|A|\log_2\tfrac{2T}{|A|} + |S||A|,$$

where D is the diameter of the transition graph, i.e. the length of the longest among all shortest simple paths between any pair of vertices.

Proof. The way in which we derive the bound on the expected regret from Theorem 6 is different from the one employed in the previous version of this paper [1] and gives slightly better constants.

First, according to Proposition 3, the regret accumulated in the transition phases caused by switching from one cycle to another one can be upper bounded by $D|A|\log_2\frac{2T}{|A|}$, using that by assumption at each step we suffer a loss of at most $\rho^* \leq 1$.

For the analysis of the cycle phases, we use the notation introduced in the proof of Theorem 6. Then the expected regret $R^{\circ}_{T,\varepsilon}$ accumulated in the cycle phases can be written as

$$\begin{aligned} R^{\circ}_{T,\varepsilon} &= \sum_{i\in\mathcal{M}_\varepsilon}\Big(\rho^* - \varepsilon - \rho(\tilde{C}_i)\Big)\,\tau_i \\ &= \sum_{i\in\mathcal{M}_\varepsilon}\Big(\rho^* - \rho(\tilde{C}_i)\Big)\,\tau_i - \sum_{i\in\mathcal{M}_\varepsilon}\varepsilon\tau_i, \\ &\leq \Delta_\varepsilon \end{aligned} \tag{8}$$

according to the definition (2) of Δ_ε. Note that the expectation here is only taken with respect to the random fluctuations of the rewards obtained in each episode. There are still the random values $\mathcal{M}_\varepsilon$, C_i, and τ_i in $R^{\circ}_{T,\varepsilon}$. However, we will bound $R^{\circ}_{T,\varepsilon}$ independent of these random values to obtain a bound on $\mathbb{E}(R_{T,\varepsilon})$. By (8) and (7) we obtain

$$R^{\circ}_{T,\varepsilon} \;\leq\; \sqrt{12|A|N_\varepsilon\log\tfrac{|A|T^4}{\delta}}. \tag{9}$$

Now from Theorem 6 we know that

$$N_\varepsilon \;\leq\; \frac{12|A|\log\frac{|A|T^4}{\delta}}{\varepsilon^2},$$

provided the confidence intervals for the rewards hold at the beginning of each episode. Together with (9) this gives

$$R^{\circ}_{T,\varepsilon} \;\leq\; \frac{12|A|\log\frac{|A|T^4}{\delta}}{\varepsilon}.$$

This holds for all possible values of $\mathcal{M}_\varepsilon$, C_i, and τ_i, under the assumption that the confidence intervals for the rewards according to Lemma 4 hold at the beginning of each episode. Thus to complete the proof, we have to consider the error probability with which these confidence intervals do not hold. The regret for a failing confidence interval at t_i is upper bounded by the length of the episode's cycle phase (the regret of the respective transition phase has already been considered above). Due to the episode termination criterion, if an episode's cycle phase starts at step t_i°, the length of the cycle phase is at most t_i°, which in turn can be bounded by $t_i + D$. Hence, by Lemma 4 and the bound on the number M of episodes of Proposition 3, the expected regret accumulated due to failing confidence intervals is at most

$$\begin{aligned}
&\sum_{i=1}^{M}(t_i + D) \cdot \mathbb{P}\{\text{confidence interval fails at } t_i\} \\
&\quad\leq \sum_{i=1}^{M}\sum_{t=1}^{T}(t + D) \cdot \mathbb{P}\{t_i = t \text{ and confidence interval fails at } t\} \\
&\quad\leq \sum_{i=1}^{M}\sum_{t=1}^{T}(t + D) \cdot \frac{2\delta}{t^3} \;=\; \sum_{i=1}^{M}\Big(\sum_{t=1}^{T}\frac{2}{t^2} + D\sum_{t=1}^{T}\frac{2}{t^3}\Big)\delta \;<\; \sum_{i=1}^{M}\big(\tfrac{10}{3} + \tfrac{5D}{2}\big)\delta \\
&\quad\leq \tfrac{20+15D}{6}\delta|A|\log_2\tfrac{2T}{|A|}.
\end{aligned}$$

Summarizing we obtain

$$\mathbb{E}(R_{T,\varepsilon}) \;\leq\; \frac{12\,|A|\log\frac{|A|T^4}{\delta}}{\varepsilon} + \big(D + \tfrac{20+15D}{6}\delta\big)|A|\log_2\tfrac{2T}{|A|} + |S||A|,$$

now also taking into account the regret caused in the exploration phase of the transition structure according to Proposition 2. □

In order to obtain a high probability bound on the regret from Theorem 6, we also have to consider deviations from the average reward in each cycle, which will be handled by Lemma 5.

Theorem 8. *With probability $1 - \frac{9}{2}\delta$, the regret of* UCycle *(with input parameter δ) with respect to an ε-optimal policy after T steps can be upper bounded as*

$$R_{T,\varepsilon} \quad\leq\quad \frac{24\,|A|\log\frac{|A|T^4}{\delta}}{\varepsilon} + D|A|\log_2\tfrac{2T}{|A|} + |S||A| + \frac{16\lambda|A|\log\frac{|A|}{\delta}}{\varepsilon^2},$$

where λ is the largest simple cycle length and D the diameter of the transition graph.

Proof. We basically repeat the proof of Theorem 7, but in order to achieve high probability bounds on the regret with respect to an ε-optimal cycle, we consider $\frac{\varepsilon}{2}$-optimal cycles and reserve $\frac{\varepsilon}{2}$ for the deviation from the average reward. Thus,

Lemma 5 shows that in a cycle phase of length θ the probability that the random average reward is worse than the expected average reward minus $\frac{\varepsilon}{2}$ is upper bounded by $\exp\big(-\frac{\varepsilon^2\theta}{2}\big)$.

Now we book all episodes with cycle phases shorter than $\theta_0 := \frac{2\log(|A|/\delta)}{\varepsilon^2}$ (which corresponds to error probability $\frac{\delta}{|A|}$) as having maximal possible regret. Similarly to Proposition 3, the number of episodes with cycle phase of length $< \theta_0$ in which the number of visits of a fixed action a is doubled can be upper bounded by $\log_2 2\theta_0$. By the criterion for episode termination (first, visits in an action are doubled, then the cycle is completed), we may upper bound the total number of steps taken in the cycle phases of these short episodes (and consequently also the respective regret) by

$$|A| \sum_{i=0}^{\lceil \log_2 2\theta_0 \rceil} \lambda \cdot 2^i \;\le\; 8\lambda|A|\theta_0 \;=\; \frac{16\lambda|A|\log\frac{|A|}{\delta}}{\varepsilon^2}, \tag{10}$$

where λ is the largest simple cycle length in the transition graph.

Concerning the episodes with longer cycle phases, consider for a fixed action a all episodes with cycle phase of length $\ge \theta_0$ in which the number of visits in a is doubled. The history and hence the corresponding cycle phase lengths $\theta_1(a) < \theta_2(a) < \ldots$ are random, however by the doubling criterion for episode termination we certainly have $\theta_i(a) \ge 2^{i-1}\theta_0$. Consequently, the probability that the average reward of the i-th episode's cycle phase is more than $\frac{\varepsilon}{2}$ below expectation is at most $\exp\big(-\frac{\varepsilon^2 2^{i-1}\theta_0}{2}\big)$. Thus, summing up over all actions a, each episode is covered and the total error probability can be bounded by

$$|A| \sum_{i=0}^{\lceil \log_2 \frac{2T}{|A|} \rceil} \exp\Big(-\frac{\varepsilon^2 2^i\theta_0}{2}\Big) \;=\; |A| \sum_{i=0}^{\lceil \log_2 \frac{2T}{|A|} \rceil} \Big(\frac{\delta}{|A|}\Big)^{2^i} \;<\; 2\delta.$$

The rest of the proof is as for Theorem 7, only with ε replaced with $\frac{\varepsilon}{2}$ and without the error term, so that one obtains including (10) the claimed regret bound, which holds with probability $1 - \frac{5}{2}\delta - 2\delta = 1 - \frac{9}{2}\delta$. □

Note that due to the different handling of the error probabilities in the proofs of Theorems 7 and 8, Theorem 8 only makes sense for sufficiently small $\delta < \frac{2}{9}$, while Theorem 7 remains sensible also for larger values of δ.

For sufficiently small ε, any ε-optimal policy will be optimal, which yields the following corollary from Theorems 7 and 8.

Corollary 9. *Let $\Delta := \rho^* - \max_{\pi:\rho_\pi<\rho^*} \rho_\pi$ be the difference between the average reward of an optimal cycle and the average reward of the best suboptimal cycle. Then for the regret of* UCycle *(with input parameter δ)*

$$\begin{aligned} \mathbb{E}(R_T) \;&\le\; \frac{12|A|\log\frac{|A|T^4}{\delta}}{\Delta} + \big(D + \tfrac{20+15D}{6}\delta\big)|A|\log_2\tfrac{2T}{|A|} + |S||A|, \qquad \textit{and} \\ R_T \;&\le\; \frac{24|A|\log\frac{|A|T^4}{\delta}}{\Delta} + D|A|\log_2\tfrac{2T}{|A|} + |S||A| + \frac{16\lambda|A|\log\frac{|A|}{\delta}}{\Delta^2}, \end{aligned}$$

the latter with probability $1 - \frac{13}{2}\delta$.

Proof. The bound on the expected regret is straightforward from Theorem 7. For the high probability bound one also has to consider episodes whose cycle is Δ-good without being optimal (which causes additional regret with respect to an *optimal* policy). This may happen if the random reward the learner obtains for a suboptimal cycle is higher than the expected reward. However, this problem can be handled using a similar strategy as in the proof of Theorem 8. We consider $\frac{\Delta}{2}$-optimal cycles and reserve $\frac{\Delta}{2}$ for the confidence interval of the random average reward of a suboptimal cycle. Note that this is different from what we did in the proof of Theorem 8. There we had to deal with episodes in which the performance in the cycle phase was *below* the average reward of the played cycle, while here we have to consider episodes where the performance is *above* the average reward.

Still, the argument is symmetric to the one given in the proof of Theorem 8. We assume that episodes with cycle phase shorter than $\frac{2\log(|A|/\delta)}{\Delta^2}$ have maximal possible regret, while by Lemma 5 the random reward of all longer cycle phases is larger than the expected reward by at most $\frac{\Delta}{2}$ with a total error probability $< 2\delta$. Together with Theorem 8, this results in the claimed bound, which holds with probability at least $1 - \frac{13}{2}\delta$. □

Remark 10. *Although there are MDPs in which the distance* Δ *is so large that the second term in the regret bounds of Corollary 9 is the larger one, in general* Δ *can be arbitrarily small, while the diameter* D *is always upper bounded by the number of states. Thus, the first term can essentially be considered to be the main term in our regret bounds.*

4.2. Lower Bounds

There are two kinds of lower bounds on the expected regret in the multi-armed bandit setting (cf. Section 5.1 below). First, there is a lower bound due to Mannor and Tsitsiklis [17] of $\Omega\big(\frac{|B|\log T}{\Delta}\big)$, where B is the set of given arms and Δ is the difference between the best and the second-best average reward of an arm. For the case where the reward distribution is allowed to depend on the horizon T, a lower bound of $\Omega(\sqrt{|B|T})$ has been derived by Auer et al. [18].

While for $|S| = 1$ one has an ordinary multi-armed bandit problem, for $|S| \geq 2$ it is easy to reproduce these bounds for the deterministic transition MDP scenario with $|B|$ being replaced with $|A|$, provided that there are $|A| \geq 2|S|$ actions (i.e., edges in the transition graph). This is done simply by inflating the respective multi-armed bandit problem as follows. The transition graph is chosen to consist of a directed cycle of length $|S|$ (containing each state in S) with (one or more) loops added in each state (cf. Figure 3). The rewards in the loops are chosen as for the arms in the multi-armed bandit problems that give the lower bounds mentioned above. All other rewards (for the transitions to different states in the cycle) are set to 0. Obviously, learning such a deterministic transition MDP is at least as hard as learning the corresponding bandit, while the regret is actually larger due to the 0-reward transitions. As the total number

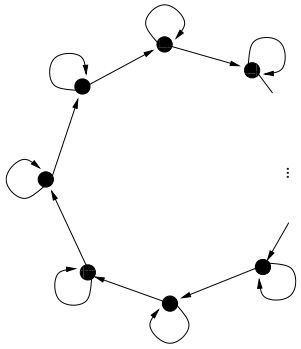

Figure 3: The transition graph for the lower bound deterministic transition MDP.

of edges $|A|$ is $|B| + |S|$ (where $|B|$ corresponds to the number of arms in the bandit setting), this gives the following lower bounds for deterministic transition MDPs.

Theorem 11. *For any algorithm and any natural numbers $|S|, |A|$ with $|A| > |S|$ there is a deterministic transition MDP with $|S|$ states and $|A|$ actions, such that the algorithm's expected regret after T steps is*

$$\mathbb{E}(R_T) = \Omega\left(\frac{(|A| - |S|)\log T}{\Delta}\right). \tag{11}$$

If the MDP is allowed to depend on T, a lower bound of $\Omega(\sqrt{(|A| - |S|)\, T})$ holds.

Note that $|A| - |S| \geq \frac{|A|}{2}$ when $|A| \geq 2|S|$, so that in this case the lower bound of (11) meets the main term of the upper bounds of Corollary 9.

Remark 12. *The MDP in Figure 3 also indicates that a switching cost term of $\Omega(D \log T)$ is necessary for each learner that wants to achieve logarithmic regret. Indeed, partition the T steps into episodes of length $T_0, 2T_0, 4T_0, \ldots, 2^i T_0$. Then, by the lower bound of Mannor and Tsitsiklis [17] for a suitable T_0 the expected number of choices of a suboptimal arm/loop has to be at least 1 in each episode. On the other hand, the algorithm cannot afford to keep playing a suboptimal choice, as the regret would not be logarithmic anymore. Therefore, one has at least one switch per episode and since the number of episodes is $\Theta(\log T)$, this shows that $\Omega(\log T)$ switches are necessary. As in the MDP of Figure 3 each switch costs D (until one gets back to the optimal loop again), the switching costs are $\Omega(D \log T)$. This does not quite match the term of $D|A| \log T$ in our upper bounds, which we however conjecture to be necessary.*

5. An Application: Bandits with Switching Cost

5.1. Setting

A special case of practical relevance is the setting of stochastic multi-armed bandits with switching cost. In ordinary multi-armed bandit problems, a learner

has to choose an *arm* from a (usually finite) set B. Choosing $b \in B$ gives random reward $\in [0,1]$ with mean $r(b)$, where we make the same independence assumptions as in the MDP setting. The learner tries to maximize her accumulated rewards. This corresponds to a single state MDP (whose transitions are trivially deterministic).

It is a natural constraint to assume that the learner may switch arms not for free but has to pay a fixed cost of $\gamma > 0$ when switching from an arm b to an arm $b' \neq b$. This can be interpreted as a negative reward of $-\gamma$ for switching arms.

Bandit settings with switching cost have mainly been considered in the economics literature (for an overview see Jun [19]). Even though most of this literature deals with the optimization problem when the whole setting is known, there is also some work on the problem when the learner has no primary knowledge of the payoffs of the individual arms. In the wake of the seminal paper of Lai and Robbins [20], which dealt with the ordinary multi-armed bandit problem, there was an adaptation of their approach to the setting with switching costs by Agrawal et al. [21]. Their bounds later were improved by Brezzi and Lai [22]. However, as the original (optimal) bounds of Lai and Robbins [20], the bounds given by Agrawal et al. [21] and Brezzi and Lai [22] are *asymptotic* in the number of steps. From our results for deterministic transition MDPs it is easy to obtain logarithmic bounds that hold uniformly over time.

5.2. Bandits with Switching Cost as Deterministic MDPs

Translated into the deterministic transition MDP setting a multi-armed bandit problem with arm set B and switching cost γ corresponds to a complete digraph with $|B|$ vertices, each with loop. These loops have mean rewards according to the actions in B, while all other edges in the graph have deterministic negative reward of $-\gamma$. Note that the situation in Example 1 is an MDP corresponding to a bandit problem with switching cost 0. Hence, switching arms too often is also harmful in the simpler setting of bandits with switching cost.

In fact, the situation is a little bit different to the deterministic transition MDP setting, as in the bandit setting it is assumed that the learner knows the cost for switching. With this knowledge, it is obviously disadvantageous to choose a cycle that is not a loop in some state. Hence, a sensible adaptation of UCycle would choose the loop in the state that has the highest upper confidence bound value. This corresponds to the UCB1 algorithm of Auer et al. [2] with the only difference being that increasing episodes are used (which is necessary to obtain sublinear regret as Example 1 shows). Indeed, Auer et al. [2] have already proposed an algorithm called UCB2 that also works in episodes and whose regret (including switching costs) is also logarithmic in T.

Although due to the negative switching costs, the rewards are not in $[0,1]$, it is easy to adapt the bounds we have derived in the deterministic transition MDP setting. Indeed, we have $D = 1$, while, as switching costs γ, the transition term in the bounds has to be multiplied by γ. This yields the following bounds.

Corollary 13. *The regret of the adapted* UCycle *algorithm (with input parameter δ) in the multi-armed bandit setting with $|B|$ arms and switching cost γ is upper bounded as*

$$\begin{aligned} \mathbb{E}(R_T) &\leq \frac{12|B|\log\frac{|B|T^4}{\delta}}{\Delta} + \big(\gamma + 6\delta\big)|B|\log_2\tfrac{2T}{|B|}, \quad \textit{and} \\ R_T &\leq \frac{24|B|\log\frac{|B|T^4}{\delta}}{\Delta} + \gamma|B|\log_2\tfrac{2T}{|B|} + \frac{16|B|\log\frac{|B|}{\delta}}{\Delta}, \end{aligned}$$

the latter with probability $1 - \frac{13}{2}\delta$.

Indeed, in the bandit setting a more refined analysis is possible, so that one easily achieves bounds of the form

$$\sum_{b\in B: r(b)<r^*} \frac{const\cdot\log\frac{T}{\delta}}{r^* - r(b)} + \gamma|B|\log_2\tfrac{2T}{|B|}, \quad \text{where } r^* := \max_{b\in B} r(b),$$

as given by Auer et al. [2] (apart from the switching cost term, cf. Remark 12) by adapting the proof of Theorem 1 of Auer et al. [2] to the episode setting. This gives slightly worse constants in the main term than in the bound given by Auer et al. [2], since a suboptimal arm will be played till the end of an episode. As all this is straightforward, we neither bother to give the precise bounds nor further details concerning the proof.

Of course, the deterministic transition MDP setting also allows to deal with settings with individual switching costs or where switching between certain arms is not allowed. In these more general settings one trivially obtains corresponding bounds with γ replaced by the cost of the most expensive switch between any two arms. This switch need not be performed in a single step, as it may be cheaper to switch from b to b' via a sequence of other arms. Note however, that when not switching directly, the learner not only has to pay switching costs but also loses time and reward by choosing the probably suboptimal intermediate arms. There is a similar problem in the original UCycle algorithm, as taking the *shortest* path to the assumed best cycle may not be optimal. Generally, in order to solve this problem one has to consider *bias-* or *Blackwell-optimal* policies [8]. However, as this has no influence on the regret bounds, we do not consider this further.

Finally, we would like to remark that the episode strategy also works well in more general bandit settings, such as continuous bandits with Lipschitz condition. Such settings were considered e.g. by Kleinberg [23] or Auer et al. [24], and it is easy to modify e.g. the proposed algorithm CAB of Kleinberg [23] to achieve bounds when switching costs are present. As in the settings mentioned above, the main term of the bounds remains basically the same with slightly worse constants. We note that these bounds are not logarithmic anymore and neither is the switching cost term.

6. Conclusion

We have shown that unlike in the general MDP case, in regret bounds for MDPs with deterministic transitions the transition parameter only appears in the term incurred by switching policies. Our bounds are close to optimal, and the only open question in that respect is whether the factor $|A|$ in the switching term is necessary.

As in the deterministic transition MDP setting the transition structure is more or less given, a related open question is whether our results can be generalized to general MDPs with known transition probabilities. A similar scenario has already been considered by Even-Dar et al. [25]. However, Even-Dar et al. [25] consider rewards that are allowed to change over time, which makes learning more difficult, so that the achieved $O(\sqrt{T})$ bound (including a transition parameter) is best possible.

Acknowledgments.

The author would like to thank the ALT08 reviewers for pointing out some errors and for other valuable comments on the previous version of this paper, the TCS reviewers for help on improving the presentation of the paper, and Thomas Jaksch for proof-reading. This work was supported in part by the Austrian Science Fund FWF (S9104-N13 SP4). The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 216886 (PASCAL2 Network of Excellence), and n° 216529 (Personal Information Navigator Adapting Through Viewing, PinView). This publication only reflects the author's views.

A. An Inequality and its Proof

Lemma 14. *Let $\mathcal{M}_\varepsilon$ be the set of all indices i of ε-bad episodes where* UCYCLE *chooses a cycle $\tilde{C}_i$ with expected reward $< \rho^* - \varepsilon$. Further, let $\tau_i(a)$ denote the number of times edge a is visited in the cycle phase of episode i, and let $n_{t_i}(a)$ be the number of times a was chosen before episode i. Finally, set $n_\varepsilon(a) := \sum_{i\in\mathcal{M}_\varepsilon} \tau_i(a)$ to be the total number of visits in a in cycle phases of episodes in $\mathcal{M}_\varepsilon$. Then*

$$\sum_{i\in\mathcal{M}_\varepsilon} \frac{\tau_i(a)}{\sqrt{n_{t_i}(a)}} \;\le\; \big(1+\sqrt{2}\big)\sqrt{n_\varepsilon(a)}.$$

Although our algorithm and its termination criterion differ slightly from the UCRL2 algorithm of Auer et al. [6], the proof of Lemma 14 is basically the same as for an analogous result given in Appendices A.3 and B.1 of Auer et al. [26]. We reproduce the proof here for the sake of completeness, starting with the following preliminary lemma.

Lemma 15. *For any sequence of numbers $z_1, \dots, z_n$ with $1 \le z_k \le Z_{k-1} := \sum_{i=1}^{k-1} z_i$ for $k \ge 1$ and $Z_0 \ge 1$,*

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

Proof. We give a proof by induction over n. For $n = 1$ we have $Z_1 = z_1 \le Z_0$, so that

$$\frac{z_1}{\sqrt{Z_0}} \;=\; \frac{Z_1}{\sqrt{Z_0}} \;\le\; \sqrt{Z_1} \;<\; \big(1+\sqrt{2}\big)Z_1.$$

For the induction step, we have by the induction hypothesis and as $z_n \le Z_{n-1}$,

$$\begin{aligned}
\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} &\le \big(1+\sqrt{2}\big)\sqrt{Z_{n-1}} + \frac{z_n}{\sqrt{Z_{n-1}}} \\
&= \sqrt{\big(1+\sqrt{2}\big)^2 Z_{n-1} + 2\big(1+\sqrt{2}\big)z_n + \frac{z_n^2}{Z_{n-1}}} \\
&\le \sqrt{\big(1+\sqrt{2}\big)^2 Z_{n-1} + \Big(2+2\sqrt{2}+1\Big)\, z_n} \\
&= \sqrt{\big(1+\sqrt{2}\big)^2 Z_{n-1} + \big(1+\sqrt{2}\big)^2 z_n} = \big(1+\sqrt{2}\big)\sqrt{Z_{n-1}+z_n} \\
&= \big(1+\sqrt{2}\big)\sqrt{Z_n}\,.
\end{aligned}$$

□

Proof of Lemma 14. For the sake of readability, in the following we skip references to the action a from the notation and abbreviate $\sqrt{n_{t_i}}$ with d_i. Let j_ε be the episode that contains step n_ε. Note that by definition of n_ε the number of steps *up to* (and including) step n_ε that are in episodes $\notin \mathcal{M}_\varepsilon$ equals the number of steps *after* n_ε that are in episodes $\in \mathcal{M}_\varepsilon$. Consequently,

$$\sum_{i<j_\varepsilon} \tau_i \mathbb{1}_{i\notin\mathcal{M}_\varepsilon} + \big(n_\varepsilon - \sum_{i<j_\varepsilon} \tau_i\big)\mathbb{1}_{j_\varepsilon\notin\mathcal{M}_\varepsilon} = \big(\sum_{i\le j_\varepsilon} \tau_i - n_\varepsilon\big)\mathbb{1}_{j_\varepsilon\in\mathcal{M}_\varepsilon} + \sum_{i>j_\varepsilon} \tau_i \mathbb{1}_{i\in\mathcal{M}_\varepsilon}.$$

Now, since $d_{j_\varepsilon} \le d_i$ for $i \ge j_\varepsilon$ and $d_i \le d_{j_\varepsilon}$ for $i \le j_\varepsilon$ this observation gives

$$\begin{aligned}
\sum_{i\in\mathcal{M}_\varepsilon} \frac{\tau_i}{d_i} &\le \sum_{i<j_\varepsilon} \frac{\tau_i}{d_i}\mathbb{1}_{i\in\mathcal{M}_\varepsilon} + \frac{\tau_{j_\varepsilon}}{d_{j_\varepsilon}}\mathbb{1}_{j_\varepsilon\in\mathcal{M}_\varepsilon} + \frac{1}{d_{j_\varepsilon}}\sum_{i>j_\varepsilon} \tau_i \mathbb{1}_{i\in\mathcal{M}_\varepsilon} \\
&\le \sum_{i<j_\varepsilon} \frac{\tau_i}{d_i}\mathbb{1}_{i\in\mathcal{M}_\varepsilon} + \frac{1}{d_{j_\varepsilon}}\sum_{i<j_\varepsilon} \tau_i \mathbb{1}_{i\notin\mathcal{M}_\varepsilon} + \frac{1}{d_{j_\varepsilon}}\big(n_\varepsilon - \sum_{i<j_\varepsilon} \tau_i\big) \\
&\le \sum_{i<j_\varepsilon} \frac{\tau_i}{d_i} + \frac{1}{d_{j_\varepsilon}}\big(n_\varepsilon - \sum_{i<j_\varepsilon} \tau_i\big).
\end{aligned}$$

Since $d_i = \sqrt{n_{t_i}} \geq \sqrt{\sum_{k=1}^{i-1} \tau_k}$ we may apply Lemma 15 to obtain the claimed

$$\begin{aligned} \sum_{i \in \mathcal{M}_\varepsilon} \frac{\tau_i}{d_i} &\leq \sum_{i<j_\varepsilon} \frac{\tau_i}{\sqrt{\sum_{k=1}^{i-1} \tau_k}} + \frac{n_\varepsilon - \sum_{i<j_\varepsilon} \tau_i}{\sqrt{\sum_{k=1}^{j_\varepsilon - 1} \tau_k}} \\ &\leq \big(1+\sqrt{2}\big)\sqrt{n_\varepsilon}. \end{aligned}$$

□

References

[1] Ronald Ortner. Online regret bounds for Markov decision processes with deterministic transitions. In *Algorithmic Learning Theory, 19th International Conference, ALT 2008, Proceedings*, pages 123–137, 2008.

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

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

[4] 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.

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

[6] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 21*, 2009. to appear.

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

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

[9] Richard M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Math.*, 23(3):309–311, 1978.

[10] Ali Dasdan and Rajesh Gupta. Faster maximum and minimum mean cycle algorithms for system performance analysis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 17(10):889–899, 1998.

[11] Neal E. Young, Robert E. Tarjan, and James B. Orlin. Faster parametric shortest path and minimum-balance algorithms. *Networks*, 21(2):205–221, 1991.

[12] Mark Hartmann and James B. Orlin. Finding minimum cost to time ratio cycles with small integral transit times. *Networks*, 23(6):567–574, 1993.

[13] Ali Dasdan, Sandy S. Irani, and Rajesh K. Gupta. Efficient algorithms for optimum cycle mean and optimum cost to time ratio problems. In *DAC '99: Proceedings of the 36th ACM/IEEE Conference on Design Automation*, pages 37–42, New York, NY, USA, 1999. ACM.

[14] Omid Madani. Polynomial value iteration algorithms for deterministic MDPs. In Adnan Darwiche and Nir Friedman, editors, *UAI '02, Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*, pages 311–318. Morgan Kaufmann, 2002.

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

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

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

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

[19] Tackseung Jun. A survey on the bandit problem with switching costs. *De Economist*, 152:513–541, 2004.

[20] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.*, 6:4–22, 1985.

[21] Rajeev Agrawal, Manjunath V. Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Trans. Automat. Control*, 33(10): 899–906, 1988.

[22] Monica Brezzi and Tze Leung Lai. Optimal learning and experimentation in bandit problems. *J. Econom. Dynam. Control*, 27:87–108, 2002.

[23] Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704. MIT Press, 2005.

[24] Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, Proceedings*, pages 454–468. Springer, 2007.

[25] 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.

[26] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. Technical Report CIT-2009-01, University of Leoben, Chair for Information Technology, 2009. URL `http://www.unileoben.ac.at/~infotech/publications/TR/CIT-2009-01.pdf`. extended version of [6].