Online Regret Bounds for Markov Decision Processes with Deterministic Transitions

Ronald Ortner

University of Leoben, A-8700 Leoben, Austria
ronald.ortner@unileoben.ac.at

Abstract. We consider an upper confidence bound algorithm for Markov decision processes (MDPs) 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. These bounds also match known *asymptotic* bounds for the general MDP setting. We also present corresponding lower bounds. As an application, multi-armed bandits with switching cost are considered.

1 Introduction

1.1 MDPs with Deterministic Transitions

A *Markov decision process* (*MDP*) can be specified as follows. First, there is a finite set S of *states* 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 that $A = \bigcup_{s \in S} A(s)$.[1] For a state $s \in S$ and an action $a \in A(s)$, a *reward* function r gives the mean $r(s,a)$ of the random rewards for choosing a in s. We assume that these random rewards are bounded in $[0,1]$. Further, *transition probability* distributions $p(\cdot|s,a)$ determine the probability $p(s'|s,a)$ that choosing an action a in state s leads to state s'.

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) one, so that the transition structure may be considered as a directed graph (loops allowed) with vertex set S and edge set $\bigcup_{s \in S} A(s) = A$. As we

[1] 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.

assume that the action sets $A(s)$ are pairwise disjoint, the mean reward depends only on the action (or edge[2]) in this *transition (di)graph*, so that we will write $r(a)$ for the mean reward of edge a. Thus summarizing, a deterministic 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 assumed 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. Starting in a transient edge a leads to a cycle, so that each edge can uniquely be 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 induced cycle of π. We are interested in the optimal policy π^* that gives maximal reward ρ^*,[3] which basically means that we are looking for a cycle with maximal mean reward.

The first algorithm for finding the optimal cycle mean has been suggested by Karp [2]. His algorithm has run-time $O(|A||S|)$. As the run-time of Karp's algorithm is also $\Omega(|A||S|)$, other algorithms have been proposed [3–6] which in some cases are faster. Value iteration on deterministic MDPs has been studied as well [7].

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. 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 at step t. When the learner does not compete with the optimal average reward ρ^* but only with $\rho^* - \varepsilon$ for some $\varepsilon > 0$, one considers the *regret* R_T^ε *with respect to an* ε*-optimal policy.*

[2] In the following, we will use the terms *action* and *edge* synonymously.

[3] 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 [1]).

Note that if the transition graph of the MDP is not *strongly connected*[4], 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 has first 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.[5] As we assume deterministic transitions, any learner that explores all actions (which obviously is 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 1 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 $|S||A|$.

1.2 General Remarks

After exploring the transition structure, the remaining problem is to deal with the exploitation-exploration problem concerning the rewards. The situation is similar to a multi-armed bandit problem. However, dealing with deterministic MDPs that way does not give any satisfying bounds, as in general the number of cycles is exponential in $|S|$. In the following, we present an algorithm (a simple generalization of the UCB1 algorithm of Auer et al. [9]) that achieves logarithmic regret in the number of steps taken. More precisely, after T steps the regret is $O\big(\frac{\lambda|A|\log T}{\Delta}\big)$ for an MDP dependent parameter $\lambda \leq |S|$ and a gap of Δ between ρ^* and the second-best average reward of a cycle. Apart from the parameter λ, this bound corresponds to the bound in the original bandit setting as given in [9].

On the other hand, there are logarithmic regret bounds for the general (average reward) MDP setting as well. These bounds usually hold under the assumption that the MDP is *ergodic*, i.e., any two states are connected by *any* policy.[6] The first of these bounds due to Burnetas and Katehakis [10] was recently generalized by Tewari and Bartlett [11]. This latter bound is of order[7] $O\big(\frac{\kappa_1^2|A||S|\log T}{\Delta}\big)$ for an MDP dependent parameter κ_1, but — as the original bound of [10] —

[4] A digraph is called *strongly connected* if there is a directed path between any two vertices.

[5] This basically has been suggested as one possible approach for learning in multichain MDPs in [8]. By the way, the alternative suggestion of [8] 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*.

[6] Note that this assumption does not hold in our setting.

[7] In these bounds for general MDPs, A is the set of actions available in each state, so that the $|A|$ in our bound corresponds rather to $|S||A|$ in the general MDP setting.

it holds only asymptotically. Finite horizon bounds have been achieved in [12]. However, as the bound is $O\big(\frac{\kappa_2 \kappa_3^2 |A||S|^5 \log T}{\Delta^2}\big)$ with MDP dependent parameters $\kappa_2 > \kappa_1$ and $\kappa_3 < \kappa_2$, the dependence on the parameters is worse than in the bounds of [11].[8] Here we achieve finite horizon bounds that basically correspond to the bounds of [11] in the simpler setting of deterministic MDPs, yet without the ergodicity assumption.

1.3 Outline

We proceed by introducing the upper confidence bound algorithm UCycle for the deterministic MDP setting. In Section 3, we prove a logarithmic bound on the expected regret of UCycle and complement it with a bound that holds with high probability. Lower bounds are derived as well. Finally, in Section 4 we consider the setting of multi-armed bandits with switching cost as a special case of deterministic MDPs.

2 An Upper Confidence Bound Algorithm

As algorithm for the deterministic MDP setting we suggest a simple adaptation of known upper confidence bound algorithms such as UCB1 [9] (for multi-armed bandits) or UCRL [12] (for ergodic 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 represented by confidence intervals for the estimated parameters (rewards, transition probabilities) of the system.

In the case of deterministic 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. Here again, 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 in detail.

Note that UCycle proceeds in episodes of increasing length. In fact, it is a tempting but bad idea to switch the cycle whenever another cycle looks more

[8] In fact, the exponent of $|S|$ in the bounds of [12] can be reduced by using the perturbation bounds of [13] (as applied e.g. in [14]) instead of those given in [15]. Moreover the exponent of the "gap" in the denominator can be reduced as well by using the "fillet" technique demonstrated in the proof of Theorem 2 below. Still, the improved bound of $O\big(\frac{\kappa_2 \kappa_3^2 |A||S|^3 \log T}{\Delta}\big)$ usually remains worse than that of [11].

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.[9]

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

▷ *Calculate optimal cycle for episode m:*
Determine an optimal cycle $\tilde{C}_m$ in the deterministic MDP with transition structure as determined and whose rewards for an 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 the shortest path to a state in the cycle $\tilde{C}_m$.

▷ *Cycle phase:* Play each action in $\tilde{C}_m$ for the next $\min_{a \in \tilde{C}_m} n_t(a)$ time steps,[10] when episode m is terminated.

Fig. 1. The UCycle algorithm.

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 better 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 $O(\log T)$.

[9] 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.

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

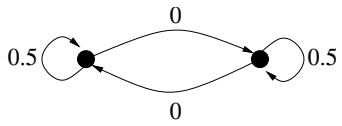

Fig. 2. Switching policies too often may lead to linear regret.

3 Online Regret Bounds

3.1 Logarithmic Upper Bounds

We start with some 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 1. *It takes at most $|A||S|$ steps until* UCycle *has determined the transition structure of any deterministic MDP.*

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 2. *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,[11] and let T_a be the number of steps in these episodes. It is easy to see that $M_a \le 1 + \log_2 T_a = \log_2 2T_a$. 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 1. *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,*

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

Proof. For fixed $a \in A$ and $n(a) \le t$, a standard Chernoff-Hoeffding bound (see e.g. Fact 1 in [9]) shows that

$$\mathbb{P}\Big\{\hat{r}_t(a) + \sqrt{\frac{\log \frac{|A|t^4}{\delta}}{2n_t(a)}} \le r(a)\Big\} \;<\; \frac{\delta}{|A|t^4}.$$

[11] Actually, it may happen that in an episode the number of visits is doubled in more than one edge. We assume that M_a counts only the episodes where a is the first edge for which this happens.

A union bound over all actions in A and all possible values of $n(a)$ proves the lemma. □

The error bound of Lemma 1 allows to derive the following sample complexity bound on the number of steps taken in suboptimal cycles.

Theorem 1. *The number of steps up to step T which* UCycle *in the cycle phase spends in cycles whose average reward is smaller than $\rho^* - \varepsilon$ is upper bounded by*

$$\frac{6\lambda|A|\log\frac{|A|T^4}{\delta}}{\varepsilon^2}$$

with probability at least $1 - \frac{5}{2}\delta$, where λ is the length of the largest simple cycle in the transition graph of the MDP.

Proof. Let C^* be an optimal cycle in the MDP, and let $\tilde{C}_m$ be the cycle chosen by UCycle in the current episode m. 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 assume throughout the proof that the confidence intervals given in (1) hold for all t, which by Lemma 1 is true with probability at least $1 - \sum_t \frac{\delta}{t^3} > 1 - \frac{5}{4}\delta$. Then

$$\rho^* - \rho(\tilde{C}_m) \;=\; \rho(C^*) - \rho(\tilde{C}_m) \;\leq\; \tilde{\rho}(C^*) - \rho(\tilde{C}_m) \;\leq\; \tilde{\rho}(\tilde{C}_m) - \rho(\tilde{C}_m)$$

with probability $1 - \frac{5}{4}\delta$. Thus, if at the beginning of an episode the estimation error $\tilde{\rho}(\tilde{C}_m) - \rho(\tilde{C}_m)$ is upper bounded by ε, UCycle will choose an ε-optimal cycle. This will happen in particular if $\tilde{r}_t(a) - r(a) < \varepsilon$ for all actions $a \in A$. On the other hand, this means that whenever UCycle chooses a cycle $\tilde{C}_m$ for which $\rho(\tilde{C}_m) < \rho^* - \varepsilon$, then there is an edge a in $\tilde{C}_m$ for which $\tilde{r}_t(a) - r(a) \geq \varepsilon$ at the initial step t of episode m. Now when each edge a was visited sufficiently often, that is, when for all $a \in A$

$$n_t(a) \;>\; \frac{2\log\frac{|A|T^4}{\delta}}{\varepsilon^2}, \tag{2}$$

then $\tilde{r}_t(a) - \hat{r}_t(a) < \frac{\varepsilon}{2}$ and (by a Chernoff-Hoeffding bound analogously to Lemma 1) also $\hat{r}_t(a) - r(a) < \frac{\varepsilon}{2}$ for all a, each with error probability at most $\frac{5}{4}\delta$. Hence, in this case $\tilde{r}_t(a) - r(a) < \varepsilon$ with probability $1 - \frac{5}{2}\delta$, and the chosen cycle is ε-optimal. We are going to determine how many steps in ε-suboptimal cycles are taken at most, until (2) holds for all actions a.

If UCycle chooses an ε-suboptimal cycle of length $|\tilde{C}_m|$ in an episode m, then each edge is visited exactly $\frac{\tau_m}{|\tilde{C}_m|}$ times, where τ_m is the length of episode m. For a fixed action a, let $M(a)$ be the number of episodes i in which a is part of the chosen, ε-suboptimal cycle $\tilde{C}_i(a)$, and (2) does not hold for a at the beginning of the episode. Further, let $\tau_i(a)$ be the length of the i-th respective episode. Then denoting the largest simple cycle length in the MDP by λ, we

have that after the last step t' of episode $M(a)-1$,

$$\sum_{i=1}^{M(a)-1} \frac{\tau_i(a)}{\lambda} \;\leq\; \sum_{i=1}^{M(a)-1} \frac{\tau_i(a)}{|\tilde{C}_i(a)|} \;\leq\; n_{t'}(a) \;\leq\; \frac{2\log\frac{|A|T^4}{\delta}}{\varepsilon^2}. \tag{3}$$

Within the $M(a)$-th episode, a is finally visited sufficiently often so that (2) holds. Thus at the first step t'' of this final episode (note that a may have been played in the meantime in an optimal episode or in a transition phase)

$$n_{t''}(a) \;\leq\; \frac{2\log\frac{|A|T^4}{\delta}}{\varepsilon^2}.$$

By the criterion when an episode ends, the number of visits in a (and indeed in all other edges as well) in this final episode can be upper bounded by $2\cdot\frac{2\log(|A|T^4/\delta)}{\varepsilon^2}$, so that together with (3),

$$\sum_{i=1}^{M(a)} \frac{\tau_i(a)}{\lambda} \;\leq\; \sum_{i=1}^{M(a)-1} \frac{\tau_i(a)}{\lambda} + \frac{4\log\frac{|A|T^4}{\delta}}{\varepsilon^2} \;\leq\; \frac{6\log\frac{|A|T^4}{\delta}}{\varepsilon^2}.$$

Consequently,

$$\sum_{i=1}^{M(a)} \tau_i(a) \;\leq\; \frac{6\lambda\log\frac{|A|T^4}{\delta}}{\varepsilon^2},$$

and summing over all actions $a \in A$ finishes the proof. □

Together with Proposition 2, Theorem 1 is sufficient to yield a *high probability* bound on the regret (Theorem 3 below). For the following bound on the *expected* regret we deal with the error probabilities in a slightly more sophisticated way.

Theorem 2. *The expected regret of* UCycle *after T steps with respect to an ε-optimal policy can be upper bounded as*

$$\mathbb{E}(R_T^\varepsilon) \;\leq\; \frac{48\lambda|A|\log\frac{|A|T^4}{\delta}}{\varepsilon} + \big(D+\tfrac{10}{3}\delta\big)|A|\log_2\tfrac{2T}{|A|} + |S||A|,$$

where λ is the largest simple cycle length and D the diameter of the transition graph, i.e. the length of the longest simple path between two vertices.

Proof. First, according to Proposition 2, 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 cycle phases, Theorem 1 bounds the number of steps taken in ε-suboptimal cycles with high probability. Note that the expected regret accumulated in a cycle phase of length τ when $\tilde{C}_m$ is an ε-optimal cycle is at most $\tau\varepsilon$ (this is due to the fact that episodes end only after all edges in the cycle have

been visited equally often). Now we fix an $\varepsilon > 0$ and partition all suboptimal episodes[12] with respect to their expected regret:[13] we summarize all episodes whose expected regret in the cycle phase is in the same interval $[2^{-i}, 2^{-i+1})$. For each $\varepsilon' \in [2^{-i}, 2^{-i+1})$, the number of steps in ε'-suboptimal cycles is upper bounded by $\frac{6\lambda|A|\log(|A|T^4/\delta)}{\varepsilon'^2}$ according to Theorem 1,[14] so that the expected regret in the cycle phases of these episodes is upper bounded by

$$\frac{6\lambda|A|\log\frac{|A|T^4}{\delta}}{2^{-2i}} \cdot 2^{-i+1}.$$

Let $k \in \mathbb{N}$ be such that $2^{-k} \leq \varepsilon < 2^{-k+1}$. Then summing up over all $i = 0, \ldots, k$ allows to upper bound the regret by

$$\sum_{i=0}^{k} \frac{6\lambda|A|\log\frac{|A|T^4}{\delta}}{2^{-2i}} \cdot 2^{-i+1} \;<\; 12\lambda|A|\log\left(\frac{|A|T^4}{\delta}\right)2^{k+1} \;\leq\; \frac{48\lambda|A|\log\frac{|A|T^4}{\delta}}{\varepsilon}.$$

Finally, we have to consider the error probability with which the confidence intervals do not hold. Writing t_m for the beginning of the m-th episode, the regret for a failing confidence interval at t_m is at most $(t_{m+1} - t_m) \leq 2t_m$ (this inequality holds due to the episode termination criterion). Hence, by Lemma 1 and the bound on the number M of episodes of Proposition 2, the expected regret accumulated due to failing confidence intervals is at most

$$\begin{aligned}
&2\sum_{m=1}^{M} t_m \cdot \mathbb{P}\{\text{confidence interval fails at } t_m\} \\
&\leq\; 2\sum_{m=1}^{M}\sum_{t=1}^{T} t \cdot \mathbb{P}\{t_m = t \text{ and confidence interval fails at } t\} \\
&\leq\; 2\sum_{m=1}^{M}\sum_{t=1}^{T} t \cdot \frac{\delta}{t^3} \;=\; 2\sum_{m=1}^{M}\sum_{t=1}^{T} \frac{\delta}{t^2} \;<\; 2\sum_{m=1}^{M} \tfrac{5}{3}\delta \;\leq\; \tfrac{10}{3}\delta|A|\log_2\tfrac{2T}{|A|}.
\end{aligned}$$

Summarizing we obtain

$$\mathbb{E}(R_T^{\varepsilon}) \;\leq\; \frac{48\lambda|A|\log\frac{|A|T^4}{\delta}}{\varepsilon} + \big(D + \tfrac{10}{3}\delta\big)|A|\log_2\tfrac{2T}{|A|} + |S||A|,$$

now also taking into account the regret caused in the exploration phase of the transition structure according to Proposition 1. □

In order to obtain high probability bounds on the regret from Theorem 1, we have to consider deviations from the average reward in each cycle.

[12] When speaking of an (ε-)suboptimal episode m we mean that the respective chosen cycle $\tilde{C}_m$ is (ε-)suboptimal.

[13] The following technique was suggested to me by Peter Auer.

[14] Actually, we do not use Theorem 1 itself, but rather refer to its proof, as we deal slightly differently with the error probabilities here.

Theorem 3. *With probability* $1-\frac{9}{2}\delta$*, the regret of* UCYCLE *with respect to an* ε*-optimal policy after* T *steps can be upper bounded as*

$$R_T^\varepsilon \;\le\; \frac{96\lambda|A|\log\frac{|A|T^4}{\delta}}{\varepsilon} + D|A|\log_2\tfrac{2T}{|A|} + |S||A| + \frac{16\lambda|A|\log\frac{|A|}{\delta}}{\varepsilon}.$$

Proof. We basically repeat the proof of Theorem 2, 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, another application of Chernoff-Hoeffding shows that in a cycle phase of length τ the probability that the random average reward is worse than the expected average reward minus $\frac{\varepsilon}{2}$ can be upper bounded by $\exp\big(-\frac{\varepsilon\tau}{2}\big)$. Now we book all episodes that are shorter than $\tau_0 := \frac{2\log(|A|/\delta)}{\varepsilon}$ (which corresponds to error probability $\frac{\delta}{|A|}$) as having maximal possible regret. Similarly to Proposition 2, the number of episodes of length $< \tau_0$ in which the number of visits of a fixed action a is doubled (cf. footnote 11) can be upper bounded by $\log_2 2\tau_0$. 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 these short episodes (and consequently also the respective regret) by

$$|A| \sum_{i=0}^{\lceil\log_2 2\tau_0\rceil} \lambda\cdot 2^i \;\le\; 8\lambda|A|\tau_0 \;=\; \frac{16\lambda|A|\log\frac{|A|}{\delta}}{\varepsilon}. \tag{4}$$

Similarly, the error probabilities of all longer episodes can be (by the doubling criterion for episode termination) summed up and bounded by

$$|A| \sum_{i=0}^{\lceil\log_2\frac{2T}{|A|}\rceil} \exp\Big(-\frac{\varepsilon 2^i\tau_0}{2}\Big) \;=\; |A| \sum_{i=0}^{\lceil\log_2\frac{2T}{|A|}\rceil} \Big(\frac{\delta}{|A|}\Big)^{2^i} \;<\; 2\delta.$$

The rest of the proof is as for Theorem 2, only with ε replaced with $\frac{\varepsilon}{2}$ and without the error term, so that one obtains including (4) the claimed regret bound, which holds with probability $1-\frac{9}{2}\delta$. □

Note that due to the different handling of the error probabilities in the proofs of Theorems 2 and 3, Theorem 3 only makes sense for sufficiently small $\delta < \frac{2}{9}$, while Theorem 2 remains sensible also for larger values of δ.

When ε is chosen sufficiently small, any ε-optimal policy will be optimal, which yields the following corollary from Theorems 2 and 3.

Corollary 1. *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*

$$\mathbb{E}(R_T) \;\le\; \frac{48\lambda|A|\log\frac{|A|T^4}{\delta}}{\Delta} + \big(D+\tfrac{10}{3}\delta\big)|A|\log_2\tfrac{2T}{|A|} + |S||A|, \qquad \textit{and}$$

$$R_T \;\le\; \frac{96\lambda|A|\log\frac{|A|T^4}{\delta}}{\Delta} + D|A|\log_2\tfrac{2T}{|A|} + |S||A| + \frac{16\lambda|A|\log\frac{|A|}{\delta}}{\Delta},$$

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

Proof. The bound on the expected regret is straightforward from Theorem 2. For the high probability bound one also has to consider episodes that are 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 solved using a similar strategy as in the proof of Theorem 3. We consider $\frac{\Delta}{2}$-optimal episodes 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 3. There we had to deal with episodes in which the performance 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 3. We consider that episodes shorter than $\frac{2\log(|A|/\delta)}{\Delta^2}$ have maximal possible regret, while the random reward of all longer episodes is larger than their expected reward by at most $\frac{\Delta}{2}$ with a total error probability $< 2\delta$. Together with Theorem 3, this results in the claimed bound which holds with probability at least $1 - \frac{13}{2}\delta$. □

3.2 Lower Bounds

There are two kinds of lower bounds (on the expected regret) in the multi-armed bandit setting (cf. Section 4.1 below). First, there is a lower bound due to Mannor and Tsitsiklis [16] of $\Omega(\frac{|B|\log T}{\Delta})$ 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 in [17].

It is easy to reproduce these bounds for the deterministic MDP scenario with $|B|$ being replaced with $|A|$, when there are $|S| \geq 3$ states[15] and $|A| \geq 3(|S|-1)$ actions (i.e., edges in the transition graph). This is done simply by inflating the respective multi-armed bandit problem. Figure 3 shows the basic construction of the transition graph with $|A| = 3(|S|-1)$. Further actions may be added in each of the states. The rewards for 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) are set to 0. Obviously, learning such a deterministic MDP is equally hard as learning the corresponding bandit, while the regret is actually larger due to the 0-reward transitions. As the total number of edges $|A|$ is three times the number of loops $|B|$ (corresponding to the number of arms in the bandit setting), this gives the claimed lower bounds for deterministic MDPs.

[15] For $|S| = 1$ one has an ordinary multi-armed bandit problem, while for $|S| = 2$ a deterministic MDP with transitions as in Figure 2 works instead of the construction given here.

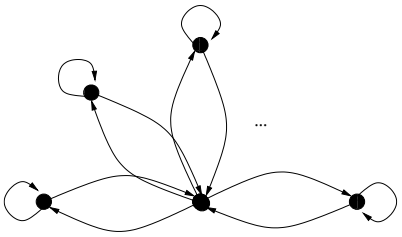

Fig. 3. The transition graph for the lower bound deterministic MDPs.

4 An Application: Bandits with Switching Cost

4.1 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 $r(b)$, and the learner tries to maximize her accumulated rewards. This corresponds to a (trivially deterministic) single state MDP.

It is a natural constraint to assume that the learner may switch arms not for free, but that she has to pay a fixed cost of $\gamma > 0$ when switching from an arm a to an arm $a' \neq a$. 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 [18]). 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 [19], 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. [20]. Their bounds later were improved by Brezzi and Lai [21]. However, as the original bounds of [19], the bounds given in [20, 21] are only *asymptotic* in the number of steps. From our results for deterministic MDPs it is easy to obtain logarithmic bounds that hold uniformly over time.

4.2 Bandits with Switching Cost as Deterministic MDPs

Translated into the deterministic 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. 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 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. [9] 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. [9] 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 MDP setting. We have already argued that it is sufficient to look for optimal cycles among the loops in each state, so that λ can be chosen to be 1. Moreover, $D = 1$. However, as switching costs γ, the transition term in the bounds has to be multiplied by γ. This yields the following bounds.

Corollary 2. *The regret of* UCycle *in the multi-armed bandit setting with $|B|$ arms and switching cost γ can be upper bounded as*

$$\begin{aligned}\mathbb{E}(R_T) &\leq \frac{48|B|\log\frac{|B|T^4}{\delta}}{\Delta} + \big(\gamma + \tfrac{10}{3}\delta\big)|B|\log_2\tfrac{2T}{|B|}, \quad \textit{and}\\ R_T &\leq \frac{96|B|\log\frac{|B|T^4}{\delta}}{\Delta} + \gamma|B|\log_2\tfrac{2T}{|B|} + \frac{16|B|\log\frac{|B|}{\delta}}{\Delta},\end{aligned}$$

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

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

as given in [9] (apart from the switching cost term) by adapting the proof to the episode setting (which gives slightly worse constants in the main term than in [9]). As all this is straightforward, we neither bother to give the precise bounds nor further details concerning the proof.

Of course, the deterministic 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.[16]

[16] 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 [1]. 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. in [22, 23], and it is easy to modify e.g. the proposed algorithm CAB of [22] 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.

5 Conclusion

Although usually there is some kind of transition (or mixing time) parameter in regret bounds for general MDPs (e.g. the κ_i in the bounds of [12, 11] mentioned above), it is not clear whether the largest simple cycle length parameter λ is necessary in regret bounds for deterministic MDPs. Interestingly, the parameter λ and the diameter D (which may be considered as an alternative transition parameter of the MDP) are in general not comparable to each other. On the one hand, complete graphs have largest possible $\lambda = |S|$ and smallest possible diameter $D = 1$. On the other hand, there are also graphs with large diameter and small λ as Figure 4 shows.

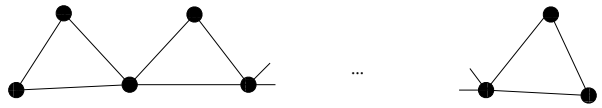

Fig. 4. Graph with $\lambda = 3$ and diameter $D = \frac{|S|-1}{2}$.

A related question is what optimal bounds look like in the case of general MDPs with known transition probabilities. In particular, also in this setting it is an interesting question whether in such bounds the appearance of a transition parameter is necessary. A similar scenario has already been considered in [24]. However, in [24] the rewards are allowed to change over time, which makes learning more difficult, so that the achieved $O(\sqrt{T})$ bounds (including a transition parameter) are best possible.

Acknowledgements. The author would like to thank the reviewers for pointing out some errors and for other valuable comments. This work was supported in part by the Austrian Science Fund FWF (S9104-N13 SP4). The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 216529 and n° 216886. This publication only reflects the authors' views.

References

1. Puterman, M.L.: Markov Decision Processes. Wiley, New York (1994)

2. Karp, R.M.: A characterization of the minimum cycle mean in a digraph. Discrete Math. 23(3), 309–311 (1978)
3. Dasdan, A., Gupta, R.: Faster maximum and minimum mean cycle algorithms for system performance analysis. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. 17(10), 889–899 (1998)
4. Dasdan, A., Irani, S.S., Gupta, R.K.: Efficient algorithms for optimum cycle mean and optimum cost to time ratio problems. In: Proc. 36th DAC, pp. 37–42. ACM, New York (1999)
5. Hartmann, M., Orlin, J.B.: Finding minimum cost to time ratio cycles with small integral transit times. Networks 23(6), 567–574 (1993)
6. Young, N.E., Tarjan, R.E., Orlin, J.B.: Faster parametric shortest path and minimum-balance algorithms. Networks 21(2), 205–221 (1991)
7. Madani, O.: Polynomial value iteration algorithms for deterministic MDPs. In: Proc. 18th UAI, pp. 311–318. Morgan Kaufmann (2002)
8. Kearns, M.J., Singh, S.P.: Near-optimal reinforcement learning in polynomial time. Mach. Learn. 49, 209–232 (2002)
9. Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multi-armed bandit problem. Mach. Learn. 47, 235–256 (2002)
10. Burnetas, A.N., Katehakis, M.N.: Optimal adaptive policies for Markov decision processes. Math. Oper. Res. 22(1), 222–255 (1997)
11. Tewari, A., Bartlett, P.L.: Optimistic linear programming gives logarithmic regret for irreducible MDPs. In: Proc. 20th NIPS, to appear.
12. Auer, P., Ortner, R.: Logarithmic online regret bounds for undiscounted reinforcement learning. In: Proc. 19th NIPS, pp. 49–56. MIT Press (2006)
13. Hunter, J.J.: Mixing times with applications to perturbed Markov chains. Linear Algebra Appl. 417, 108–123 (2006)
14. Ortner, R.: Pseudometrics for state aggregation in average reward Markov decision processes. In: Proc. 18th ALT, pp. 373–387. Springer (2007)
15. Cho, G.E., Meyer, C.D.: Markov chain sensitivity measured by mean first passage times. Linear Algebra Appl. 316, 21–28 (2000)
16. Mannor, S., Tsitsiklis, J.N.: The sample complexity of exploration in the multi-armed bandit problem. J. Mach. Learn. Res. 5, 623–648 (2004)
17. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multi-armed bandit problem. SIAM J. Comput. 32, 48–77 (2002)
18. Jun, T.: A survey on the bandit problem with switching costs. De Economist 152, 513–541 (2004)
19. Lai, T., Robbins, H.: Asymptotically efficient adaptive allocation rules. Adv. in Appl. Math. 6, 4–22 (1985)
20. Agrawal, R., Hedge, M.V., Teneketzis, D.: Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. IEEE Trans. Automat. Control 33(10), 899–906 (1988)
21. Brezzi, M., Lai, T.L.: Optimal learning and experimentation in bandit problems. J. Econom. Dynam. Control 27, 87–108 (2002)
22. Kleinberg, R.D.: Nearly tight bounds for the continuum-armed bandit problem. In: Proc. 17th NIPS, pp. 697–704. MIT Press (2004)
23. Auer, P., Ortner, R., Szepesvári, C.: Improved rates for the stochastic continuum-armed bandit problem. In: Proc. 20th COLT, pp. 454–468. Springer (2007)
24. Even-Dar, E., Kakade, S.M., Mansour, Y.: Experts in a Markov decision process. In: Proc. 17th NIPS, pp. 401–408. MIT Press (2004)