Regret Bounds for Restless Markov Bandits

Ronald Ortner*, Daniil Ryabko**, Peter Auer*, Rémi Munos**

Abstract

We consider the restless Markov bandit problem, in which the state of each arm evolves according to a Markov process independently of the learner's actions. We suggest an algorithm, that first represents the setting as an MDP which exhibits some special structural properties. In order to grasp this information we introduce the notion of ε-structured MDPs, which are a generalization of concepts like (approximate) state aggregation and MDP homomorphisms. We propose a general algorithm for learning ε-structured MDPs and show regret bounds that demonstrate that additional structural information enhances learning.

Applied to the restless bandit setting, this algorithm achieves after any T steps regret of order $\tilde{O}(\sqrt{T})$ with respect to the best policy that knows the distributions of all arms. We make no assumptions on the Markov chains underlying each arm except that they are irreducible. In addition, we show that index-based policies are necessarily suboptimal for the considered problem.

Keywords:
restless bandits, Markov decision processes, regret

1. Introduction

In the bandit problem the learner has to decide at time steps $t = 1, 2, \ldots$ which of the finitely many available arms to pull. Each arm produces a reward in a stochastic manner. The goal is to maximize the reward accumulated over time.

Following [1], traditionally it is assumed that the rewards produced by each given arm are independent and identically distributed (i.i.d.). If the probability distributions of the rewards of each arm are known, the best strategy is to only pull the arm with the highest expected reward. Thus, in the i.i.d. bandit setting the *regret* is measured with respect to the best arm. An extension of this setting is to assume that the rewards generated by each arm are not i.i.d., but are governed by some more complex stochastic process. Markov chains suggest themselves as an interesting and non-trivial model. In this setting it is often natural to assume that the stochastic process (Markov chain) governing each arm does not depend on the actions

*Montanuniversitaet Leoben, A-8700 Leoben, Austria
**Inria Lille-Nord Europe, F-59650 Villeneuve d'Ascq, France
Email addresses: rortner@unileoben.ac.at (Ronald Ortner), daniil@ryabko.net (Daniil Ryabko), auer@unileoben.ac.at (Peter Auer), remi.munos@inria.fr (Rémi Munos)

of the learner. That is, the chain takes transitions independently of whether the learner pulls that arm or not (giving the name *restless bandit* to the problem). The latter property makes the problem rather challenging: since we are not observing the state of each arm, the problem becomes a partially observable Markov decision process (POMDP), rather than being a (special case of) a fully observable MDP, as in the traditional i.i.d. setting. One of the applications that motivate the restless bandit problem is the so-called *cognitive radio* problem (e.g., [2]): Each arm of the bandit is a radio channel that can be busy or available. The learner (an appliance) can only sense a certain number of channels (in the basic case only a single one) at a time, which is equivalent to pulling an arm. It is natural to assume that whether the channel is busy or not at a given time step depends on the past — so a Markov chain is the simplest realistic model — but does not depend on which channel the appliance is sensing. (See also Example 1 in Section 3 for an illustration of a simple instance of this problem.)

What makes the restless Markov bandit problem particularly interesting is that *one can do much better than pulling the best arm.* This can be seen already on simple examples with two-state Markov chains (see Section 3 below). Remarkably, this feature is often overlooked, notably by some early work on restless bandits, e.g. [3], where the regret is measured with respect to the mean reward of the best arm. This feature also makes the problem more difficult and in some sense more general than the non-stochastic bandit problem, in which the regret usually is measured with respect to the best arm in hindsight [4]. Finally, it is also this feature that makes the problem principally different from the so-called *rested* bandit problem, in which each Markov chain only takes transitions when the corresponding arm is pulled.

Thus, in the restless Markov bandit problem that we study, the regret should be measured not with respect to the best arm, but with respect to the best policy knowing the distribution of all arms. To understand what kind of regret bounds can be obtained in this setting, it is useful to compare it to the i.i.d. bandit problem and to the problem of learning an MDP. In the i.i.d. bandit problem, the minimax regret expressed in terms of the horizon T and the number of arms only is $O(\sqrt{T})$, cf. [5]. If we allow problem-dependent constants into consideration, then the regret becomes of order $\log T$ but depends also on the gap between the expected reward of the best and the second-best arm. In the problem of learning to behave optimally in an MDP, nontrivial problem-independent finite-time regret guarantees (that is, regret depending only on T and the number of states and actions) are not possible to achieve. It is possible to obtain $O(\sqrt{T})$ regret bounds that also depend on the diameter of the MDP [6] or similar related constants, such as the span of the optimal bias vector [7]. Regret bounds of order $\log T$ are only possible if one additionally allows into consideration constants expressed in terms of policies, such as the gap between the average reward obtained by the best and the second-best policy [6]. The difference between these constants and constants such as the diameter of an MDP is that one can try to estimate the latter, while estimating the former is at least as difficult as solving the original problem — finding the best policy. Turning to our restless Markov bandit problem, so far, to the best of our knowledge no regret bounds are available for the general problem. However, several special cases have been considered. Specifically, $O(\log T)$ bounds have been obtained in [8] and [9]. While the latter considers the two-armed restless bandit case, the results of [8] are constrained by some ad hoc assumptions on the transition probabilities and on the structure of the optimal policy of the problem. The algorithm proposed in [8] alternates exploration and exploitation steps, where the former shall guarantee that estimates are sufficiently precise, while in the latter an optimistic arm is chosen by a policy employing UCB-like confidence intervals. Computational aspects of the algorithm are however neglected. In addition, while the $O(\log T)$ bounds of [8] depend on the parameters of the problem (i.e., on the unknown distributions of the Markov chains), it is unclear what order the bounds assume in the worst case, that is, when one takes the supremum over the bandits satisfying the assumptions imposed by the authors.

Finally, while regret bounds for the Exp3.S algorithm [4] can be applied in the restless bandit setting, these bounds depend on the "hardness" of the reward sequences, which in the case of reward sequences generated by a Markov chain can be arbitrarily high. We refer to [10] for an overview of bandit algorithms and corresponding regret bounds.

Here we present an algorithm for which we derive $\tilde{O}(\sqrt{T})$ regret bounds, making no assumptions on the distribution of the Markov chains except that they are irreducible. The algorithm is based on constructing an approximate MDP representation of the POMDP problem, and then using a modification of the Ucrl2 algorithm of [6] to learn this approximate MDP. In addition to the horizon T and the number of arms and states, the regret bound also depends on the diameter and the mixing time (which can be eliminated however) of the Markov chains of the arms. If the regret has to be expressed only in these terms, then our lower bound shows that the dependence on T cannot be significantly improved.

A common feature of many bandit algorithms is that they look for an optimal policy in an *index* form (starting with the Gittins index [11], and including UCB [12], and, for the Markov case, [13], [9]). That is, for each arm the policy maintains an index which is a function of time, states, and rewards *of this arm only.* At each time step, the policy samples the arm that has maximal index. This idea also leads to conceptually and computationally simple algorithms. One of the results in this work is to show that, in general, for the restless Markov bandit problem, index policies are suboptimal.

The rest of the paper is organized as follows. Section 2 defines the setting, in Section 3 we give some examples of the restless bandit problem, as well as demonstrate that index-based policies are suboptimal. Section 4 presents the main results: the upper and lower bounds on the achievable regret in the considered problem; Sections 5 and 7 introduce the algorithm for which the upper bound is proven; the latter part relies on ϵ-structured MDPs, a generalization of concepts like (approximate) state aggregation in MDPs [14] and MDP homomorphism [15], introduced in Section 6. This section also presents an extension of the Ucrl2 algorithm of [6] designed to work in this setting. The (longer) proofs are given in Section 8 and 9 (with some details deferred to the appendices), while Section 10 presents some directions for further research.

2. Preliminaries

Given are K arms, where underlying each arm j there is an irreducible Markov chain with state space S_j, some initial state in S_j, and transition matrix P_j. For each state s in S_j there is a reward distribution with mean $r_j(s)$ and support in $[0,1]$. For the time being, we will assume that the learner knows the number of states for each arm and that all Markov chains are aperiodic. In Section 8, we discuss periodic chains, while in Section 10 we indicate how to deal with unknown state spaces. In any case, the learner knows neither the transition probabilities nor the mean rewards.

For each time step $t = 1, 2, \dots$ the learner chooses one of the arms, observes the current state s of the chosen arm i and receives a random reward with mean $r_i(s)$. After this, the state of each arm j changes according to the transition matrices P_j. The learner however is not able to observe the current state of the individual arms. We are interested in competing with the optimal policy π^* which knows the mean rewards and transition matrices, yet observes as the learner only the current state of the chosen arm. Thus, we are looking for algorithms which after any T steps have small regret with respect to π^*, i.e. minimize

$$T \cdot \rho^* - \textstyle\sum_{t=1}^{T} r_t,$$

where r_t denotes the (random) reward earned at step t and ρ^* is the average reward of the optimal policy π^*. It will be seen in Section 5 that we can represent the problem as an MDP, so that π^* and ρ^* are indeed well-defined. Also, while for technical reasons we consider the regret with respect to $T\rho^*$, our results also bound the regret with respect to the optimal T-step reward.

2.1. Mixing Times and Diameter

If an arm j is not selected for a large number of time steps, the distribution over states when selecting j will be close to the stationary distribution μ_j of the Markov chain underlying arm j. Let μ_s^t be the distribution after t steps when starting in state $s \in S_j$. Then setting

$$d_j(t) := \max_{s \in S_j} \|\mu_s^t - \mu_j\|_1 := \max_{s \in S_j} \sum_{s' \in S_j} |\mu_s^t(s') - \mu_j(s')|,$$

we define the *ε-mixing time* of the Markov chain as

$$T_{\text{mix}}^j(\varepsilon) := \min\{t \in \mathbb{N} \,|\, d_j(t) \leq \varepsilon\}.$$

Setting somewhat arbitrarily *the* mixing time of the chain to $T_{\text{mix}}^j := T_{\text{mix}}^j(\frac{1}{4})$, one can show (cf. eq. 4.36 in [16]) that

$$T_{\text{mix}}^j(\varepsilon) \leq \left\lceil \log_2\left(\tfrac{1}{\varepsilon}\right)\right\rceil \cdot T_{\text{mix}}^j. \tag{1}$$

Finally, let $T_j(s, s')$ be the expected time it takes in arm j to reach a state s' when starting in state s, where for $s = s'$ we set $T_j(s, s) := 1$. Then we define the *diameter* of arm j to be $D_j := \max_{s,s' \in S_j} T_j(s, s')$.

3. Examples

Next we present a few examples that give insight into the nature of the problem and the difficulties in finding solutions. In particular, the examples demonstrate that (i) the optimal reward can be (much) bigger than the average reward of the best arm, (ii) the optimal policy does not maximize the immediate reward, and (iii) the optimal policy cannot always be expressed in terms of arm indexes.

Example 1 (best arm is suboptimal)**.** In this example the average reward of each of the two arms of a bandit is $\frac{1}{2}$, but the reward of the optimal policy is close to $\frac{3}{4}$. Consider a two-armed bandit. Each arm has two possible states, 0 and 1, which are also the rewards. Underlying each of the two arms is a (two-state) Markov chain with transition matrix $\begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$, where ϵ is small. Thus, a typical trajectory of each arm looks like this:

$$000000000001111111111111111000000000\ldots,$$

and the average reward for each arm is $\frac{1}{2}$. It is easy to see that the optimal policy starts with any arm, and then switches the arm whenever the reward is 0, and otherwise sticks to the same arm. The average reward is close to $\frac{3}{4}$ — much larger than the reward of each arm.

This example has a natural interpretation in terms of *cognitive radio*: two radio channels are available, each of which can be either busy (0) or available (1). A device can only sense (and use) one channel at a time, and one wants to maximize the amount of time the channel it tries to use is available.

Example 2 (another optimal policy)**.** Consider the previous example, but with ϵ close to 1. Thus, a typical trajectory of each arm is now

$$01010101001010110\ldots.$$

Here the optimal policy switches arms if the previous reward was 1 and stays otherwise.

Example 3 (optimal policy is not myopic)**.** In this example the optimal policy does not maximize the immediate reward. Again, consider a two-armed bandit. Arm 1 is as in Example 1, and arm 2 provides Bernoulli i.i.d. rewards with probability $\frac{1}{2}$ of getting reward 1. The optimal policy (which knows the distributions) will sample arm 1 until it obtains reward 0, when it switches to arm 2. However, it will sample arm 1 again after some time t (depending on ϵ), and only switch back to arm 2 when the reward on arm 1 is 0. Note that whatever t is, the expected reward for choosing arm 1 will be strictly smaller than $\frac{1}{2}$, since the last observed reward was 0 and the limiting probability of observing reward 1 (when $t \to \infty$) is $\frac{1}{2}$. At the same time, the expected reward of the second arm is always $\frac{1}{2}$. Thus, the optimal policy will sometimes "explore" by pulling the arm with the smaller expected reward.

An intuitively appealing idea is to look for an optimal policy which is *index*-based. That is, for each arm the policy maintains an index which is a function of time, states, and rewards *of this arm only.* At

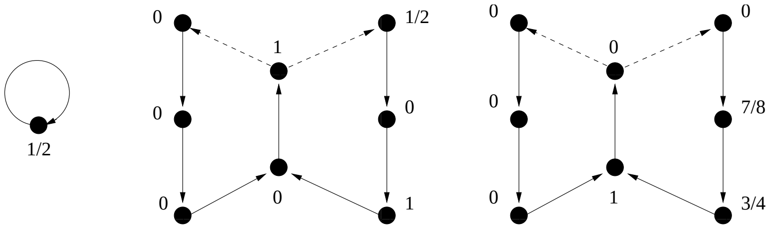

Figure 1: The example used in the proof of Theorem 4. Dashed transitions are with probability $\frac{1}{2}$, others are deterministic with probability 1. Numbers are rewards in the respective state.

each time step, the policy samples the arm that has maximal index. This seems promising for at least two reasons: First, the distributions of the arms are assumed independent, so it may seem reasonable to evaluate them independently as well; second, this works in the i.i.d. case (e.g., the Gittins index [11] or UCB [12]). This idea also motivates the setting when just one out of two arms is Markov and the other is i.i.d., see e.g. [9]. Index policies for restless Markov bandits were also studied in [13]. Despite their intuitive appeal, in general, index policies are suboptimal.

Theorem 4 (index-based policies are suboptimal)**.** *For each index-based policy π there is a restless Markov bandit problem in which π behaves suboptimally.*

Proof. Consider the three bandits L (left), C (center), and R (right) in Figure 1, where C and R start in the 1 reward state. (Arms C and R can easily be made aperiodic by adding further sufficiently small transition probabilities.) Assume that C has been observed in the $\frac{1}{2}$ reward state one step before, while R has been observed in the 1 reward state three steps ago. The optimal policy will choose arm L which gives reward $\frac{1}{2}$ with certainty (C gives reward 0 with certainty, while R gives reward $\frac{7}{8}$ with probability $\frac{1}{2}$) and subsequently arms C and R. However, if arm C was missing, in the same situation, the optimal policy would choose R: Although the immediate expected reward is smaller than when choosing L, sampling R gives also information about the current state, which can earn reward $\frac{3}{4}$ a step later. Clearly, no index based policy will behave optimally in both settings. □

4. Main Results

Theorem 5 (main upper bound on regret)**.** *Consider a restless bandit with K aperiodic arms having state spaces S_j, diameters D_j, and mixing times T^j_{mix} ($j = 1, \ldots, K$). Then with probability at least $1 - \delta$ the regret of Algorithm 2 (presented in Section 5 below) after $T > 2$ steps is upper bounded by*

$$90 \cdot S \cdot \lceil T_{\mathrm{mix}} \rceil^{3/2} \cdot \prod_{j=1}^{K} (4D_j) \cdot \left\lceil \max_i \log_2(4D_i) \right\rceil \cdot \log_2^2\left(\tfrac{T}{\delta}\right) \cdot \sqrt{T},$$

where $S := \sum_{j=1}^{K} |S_j|$ is the total number of states and $T_{\mathrm{mix}} := \max_j T^j_{\mathrm{mix}}$ the maximal mixing time. This bound also holds with a slightly worse numerical constant for the regret with respect to the best T-step policy.

Further, the dependence on T_{mix} can be eliminated to show that with probability at least $1 - \delta$ the regret is bounded by

$$O\left(S \cdot \prod_{j=1}^{K}(4D_j) \cdot \max_i \log(4D_i) \cdot \log^{7/2}\left(\tfrac{T}{\delta}\right) \cdot \sqrt{T}\right).$$

Remark 6. For periodic chains the bound of Theorem 5 has worse dependence on the state space, for details see Section 9 below.

Remark 7. Choosing $\delta = \frac{1}{T}$ in Theorem 5, it is straightforward to obtain respective upper bounds on the expected regret.

Theorem 8 (lower bound on regret)**.** *For any algorithm, any $K > 1$, and any $m \geq 1$ there is a K-armed restless bandit problem with a total number of $S := Km$ states, such that the regret after T steps is lower bounded by $\Omega(\sqrt{ST})$.*

Remark 9. While it is easy to see that lower bounds depend on the total number of states over all arms, the dependence on other parameters in our upper bound is not clear. For example, intuitively, while in the general MDP case one wrong step may cost up to D — the MDP's diameter [6] — steps to compensate for, here the Markov chains evolve independently of the learner's actions, and the upper bound's dependence on the diameter may be just an artefact of the proof.

5. Constructing the Algorithm I: MDP Representation

For the sake of simplicity, we start with the simpler case when all Markov chains are aperiodic. In Section 9, we indicate how to adapt the proofs to the periodic case.

5.1. MDP Representation

We represent the restless bandit setting as an MDP by recalling for each arm the last observed state and the number of time steps which have gone by since this last observation. Thus, each state of the MDP representation is of the form $(s_j, n_j)_{j=1}^K := (s_1, n_1, s_2, n_2, \ldots, s_K, n_K)$ with $s_j \in S_j$ and $n_j \in \mathbb{N}$, meaning that each arm j has not been chosen for n_j steps when it was in state s_j. More precisely, $(s_j, n_j)_{j=1}^K$ is a state of the considered MDP if and only if (i) all n_j are distinct and (ii) there is a j with $n_j = 1$.[1]

The action space of the MDP is $\{1, 2, \ldots, K\}$, and the transition probabilities from a state $(s_j, n_j)_{j=1}^K$ are given by the n_j-step transition probabilities $p_j^{(n_j)}(s, s')$ of the Markov chain underlying the chosen arm j (these are defined by the matrix power of the single step transition probability matrix, i.e. $P_j^{n_j}$). That is, the probability for a transition from state $(s_j, n_j)_{j=1}^K$ to $(s'_j, n'_j)_{j=1}^K$ under action j is given by $p_j^{(n_j)}(s_j, s'_j)$

[1] Actually, one would need to add for each arm j with $|S_j| > 1$ a special state for not having sampled j so far. However, for the sake of simplicity we assume that in the beginning each arm is sampled once. The respective regret is negligible.

iff (i) $n'_j = 1$, (ii) $n'_\ell = n_\ell + 1$ and $s_\ell = s'_\ell$ for all $\ell \neq j$. All other transition probabilities are 0. Finally, the mean reward for choosing arm j in state $(s_j, n_j)_{j=1}^K$ is given by $\sum_{s \in S_j} p_j^{(n_j)}(s_j, s) \cdot r_j(s)$. This MDP representation has already been considered in [8].

Obviously, within T steps any policy can reach only states with $n_j \leq T$. Correspondingly, if we are interested in the regret within T steps, it will be sufficient to consider the finite sub-MDP consisting of states with $n_j \leq T$. We call this the *T-step representation* of the problem, and the regret will be measured with respect to the optimal policy in this T-step representation.[2]

5.2. Structure of the MDP Representation

The MDP representation of our problem has some special structural properties. In particular, rewards and transition probabilities for choosing arm j only depend on the state of arm j, that is, s_j and n_j. Moreover, the support for each transition probability distribution is bounded, and for $n_j \geq T_{\text{mix}}^j(\varepsilon)$ the transition probability distribution will be close to the stationary distribution of arm j. Thus, one could reduce the T-step representation further by aggregating states[3] $(s_j, n_j)_{j=1}^K$, $(s'_j, n'_j)_{j=1}^K$ whenever $n_j, n'_j \geq T_{\text{mix}}^j(\varepsilon)$ and $s_\ell = s'_\ell$, $n_\ell = n'_\ell$ for all $\ell \neq j$. The rewards and transition probability distributions of aggregated states are ε-close, so that the error by aggregation can be bounded by results given in [17]. While this is helpful for approximating the problem when all parameters are known, it cannot be used directly when learning, since the observations in the aggregated states do not correspond to an MDP anymore. Thus, while standard reinforcement learning algorithms are still applicable, there are no theoretical guarantees for them. Instead, we will propose an algorithm which can exploit the structure information available for the MDP representation of the restless bandit setting directly. For that purpose, we first introduce the notion of *ε-structured MDPs*, which can grasp structural properties in MDPs more generally.

6. Digression: ε-structured MDPs and Colored UCRL2

ε-structured MDPs are MDPs with some additional *color* information indicating similarity of state-action pairs. Thus, state-action pairs of the same color have similar (i.e., ε-close) rewards and transition probability distributions. Concerning the latter, we allow the supports of the transition probability distributions to be different, however demand that they can be mapped to each other by a bijective translation function.

Definition 10. *An ε-structured MDP is an MDP with finite state space S, finite action space A, transition probability distributions $p(\cdot|s,a)$, mean rewards $r(s,a) \in [0,1]$, and a coloring function $c : S \times A \to \mathcal{C}$, where*

[2] An undesirable consequence of this is that the optimal average reward ρ^* which we compare to may be different for different horizons T. However, as already stated, our regret bounds also hold with respect to the more intuitive optimal T-step reward.

[3] Aggregation of states $s_1, \ldots, s_n$ means that these states are replaced by a new state s_{agg} inheriting rewards and transition probabilities from an arbitrary s_i (or averaging over all s_ℓ). Transitions to this state are set to $p(s_{\text{agg}}|s,a) := \sum_\ell p(s_\ell|s,a)$.

Algorithm 1 The colored UCRL2 algorithm for learning in ε-structured MDPs

```
Input: Confidence parameter $\delta > 0$, aggregation parameter $\varepsilon > 0$, state space $S$, action space $A$, coloring and translation functions, a bound $B$ on the size of the support of transition probability distributions.
Initialization: Set $t := 1$, and observe the initial state $s_1$.
for episodes $k = 1, 2, \ldots$ do
    Initialize episode $k$:
    Set the start time of episode $k$, $t_k := t$. Let $N_k\,(c)$ be the number of times a state-action pair of color $c$ has been visited prior to episode $k$, and $v_k(c)$ the number of times a state-action pair of color $c$ has been visited in episode $k$. Compute estimates $\hat{r}_k(s,a)$ and $\hat{p}_k(s'|s,a)$ for rewards and transition probabilities, using all samples from state-action pairs of the same color $c(s,a)$, respectively.
    Compute policy $\tilde{\pi}_k$:
    Let $\mathcal{M}_k$ be the set of plausible MDPs with rewards $\tilde{r}(s,a)$ and transition probabilities $\tilde{p}(\cdot|s,a)$ satisfying
        $\left|\tilde{r}(s,a) - \hat{r}_k(s,a)\right| \leq \varepsilon + \sqrt{\frac{7\log(2Ct_k/\delta)}{2\max\{1,N_k(c(s,a))\}}},$   (2)
        $\left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \leq \varepsilon + \sqrt{\frac{56B\log(2Ct_k/\delta)}{\max\{1,N_k(c(s,a))\}}}\,,$   (3)
    where $C$ is the number of distinct colors. Let $\rho(\pi, M)$ be the average reward of a policy $\pi : S \to A$ on an MDP $M \in \mathcal{M}_k$. Choose (e.g. by extended value iteration [6]) an optimal policy $\tilde{\pi}_k$ and an optimistic $\tilde{M}_k \in \mathcal{M}_k$ such that
        $\rho(\tilde{\pi}_k, \tilde{M}_k) = \max\left\{\rho(\pi, M)\,\middle|\,\pi : S \to A,\, M \in \mathcal{M}_k\right\}.$   (4)
    Execute policy $\tilde{\pi}_k$:
    while $v_k(c(s_t, \tilde{\pi}_k(s_t))) < \max\{1, N_k(c(s_t, \tilde{\pi}_k(s_t)))\}$ do
    ▷ Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward $r_t$, and observe next state $s_{t+1}$.
    ▷ Set $t := t + 1$.
    end while
end for
```

$\mathcal{C}$ is a set of colors. Further, for each two pairs (s,a), $(s',a') \in S \times A$ with $c(s,a) = c(s',a')$ there is a bijective translation function $\phi_{s,a,s',a'} : S \to S$ such that $\sum_{s''}\left|p(s''|s,a) - p(\phi_{s,a,s',a'}(s'')|s',a')\right| < \varepsilon$ and $|r(s,a) - r(s',a')| < \varepsilon$.

If there are states s, s' in an ε-structured MDP such that $c(s,a) = c(s',a)$ for all actions a and the associated translation function $\phi_{s,a,s',a}$ is the identity, we may aggregate the states (cf. footnote 3). We call the MDP in which all such states are aggregated the *aggregated ε-structured MDP*.

For learning in ε-structured MDPs we consider a modification of the UCRL2 algorithm of [6]. The *colored*

Ucrl2 algorithm is shown as Algorithm 1. As the original Ucrl2 algorithm it maintains confidence intervals for rewards and transition probabilities which define a set of plausible MDPs $\mathcal{M}$. Unlike the original Ucrl2 algorithm, which defines the set of plausible MDPs by confidence intervals for each single state-action pair, colored Ucrl2 calculates estimates from all samples of state-action pairs *of the same color* and works with respectively adapted confidence intervals (2), (3) for each color to determine the set $\mathcal{M}$ of plausible MDPs. Generally, the algorithm proceeds in episodes, where in each episode k an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ and an optimal policy are chosen which maximize the average reward, cf. (4). An episode ends when for some color c the number of visits in state-action pairs of color c has doubled.

We note that computation of the optimistic MDP and the respective optimal policy in (4) can be done by *extended value iteration* as introduced in [6]. This is a modification of standard value iteration where each iteration can be performed in $O(|S|^2|A|)$ computation steps. For details we refer to Section 3.1.2 of [6].

6.1. Further Applications

Although the focus of our work lies on the restless bandit problem, we'd like to note and demonstrate that ε-structured MDPs are a strong concept which is applicable to a wide range of problems.

6.1.1. MDP aggregation, MDP homomorphism, and ε-structured MDPs

First, it is easy to see that ε-structured MDPs subsume previous notions of similarity like *(approximate) state aggregation* in MDPs [14], *MDP homomorphism* [15], or *lax bisimulation* [18]. In state aggregation, one merges states to meta-states when their rewards and transition probabilities are identical or close. This corresponds to a coloring where all translation functions are the identity. MDP homomorphisms and lax bisimulation are more general in that they allow arbitrary translation functions just like ε-structured MDPs, yet they can only capture "total" similarity of two states s, s' assuming that each action in s can be mapped to an action in s' with similar rewards and transitions. Unlike that, in ε-structured MDPs two states can be similar only with respect to single actions.

Note that while this allows to grasp weaker notions of similarity, the original MDP cannot always be reduced to a smaller one. However, as we will see below, learning in structured MDPs incurs less regret.

Example 11. Consider a simple gridworld example as shown in Figure 2. The goal state g is assumed to be absorbing with reward 1. Otherwise, actions *up*, *down*, *left*, *right* lead to the respective neighbored state[4] and give reward 0. Although there is a strong topological structure in this setting, state aggregation cannot simplify the MDP. MDP homomorphisms work better, as they can exploit the symmetry along the main diagonal to reduce the state space up to a factor 2. On the other hand, the respective structured MDP only

[4] For the sake of simplicity, we assume that in border states actions that would leave the environment are simply not available. Using further colors these actions could be easily taken into account however.

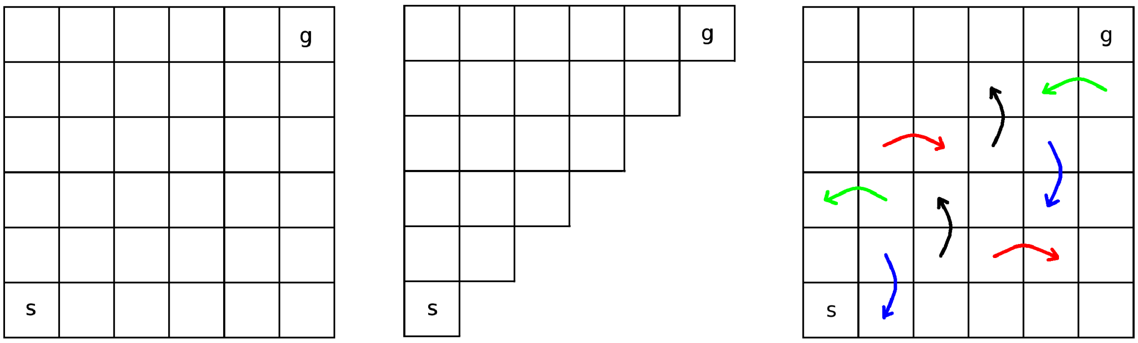

Figure 2: A simple gridworld example (left). With ordinary state aggregation no simplification is possible. An MDP homomorphism can map the states below the main diagonal to the states above it (middle). The corresponding structured MDP only needs four colors to grasp the topological structure (right).

needs four colors (one for each action) to grasp the whole topological structure (except the goal state g): Thus, for example, all state-action pairs (s, up) will obtain the same color, and the respective translation functions $\Phi_{s,up,s',up}$ will map the state above s to the state above s'. One additional color is needed for the goal state.

6.1.2. Continuous state MDPs: Discretizations as colorings

The concept of ε-structured MDPs can be straightforwardly generalized to arbitrary state spaces. Then, under the assumption that close states behave similarly according to a Lipschitz- or more generally Hölder-condition for rewards and transition probabilities, respectively, an MDP with continuous state space can be turned into a structured MDP by coloring close states with the same color. That way, a discretization of the state space also corresponds to a coloring of the state space.

6.2. Regret Bounds for Colored UCRL2

The following is a generalization of the regret bounds for Ucrl2 to ε-structured MDPs. The theorem gives improved (with respect to Ucrl2) bounds if there are only a few parameters to estimate in the MDP to learn. Recall that the *diameter* of an MDP is the maximal expected transition time between any two states (choosing an appropriate policy), cf. [6].

Theorem 12. *Let M be an ε-structured MDP with finite state space S, finite action space A, transition probability distributions $p(\cdot|s,a)$, rewards distributions with means $r(s,a)$ and support in $[0,1]$, coloring function c, and associated translation functions. Assume the learner has complete knowledge of state-action pairs $\Psi_K \subseteq S \times A$, while the state-action pairs in $\Psi_U := S \times A \setminus \Psi_K$ are unknown and have to be learned.*

However, the learner knows c and all associated translation functions as well as an upper bound B on the size of the support of each transition probability distribution in Ψ_U. Then with probability at least $1-\delta$, after any T steps colored UCRL2 [5] *gives regret upper bounded by*

$$42D_\varepsilon\sqrt{BC_U T\log\left(\tfrac{T}{\delta}\right)}+\varepsilon(D_\varepsilon+2)T,$$

where C_U is the total number of colors for states in Ψ_U, and D_ε is the diameter of the aggregated ε-structured MDP.

The proof of this theorem is given in the appendix.

Remark 13. From the proof of Theorem 12, cf. eq. (A.8), it can be seen that the accumulated reward of the best T-step policies when starting in different states cannot deviate by more than D_ε. Therefore, also the bias values of each two states differ by at most D_ε, cf. p. 339 of [19]. Application of Theorem 9.4.1a of [19] then shows that the difference between the optimal T-step reward and $T\rho^*$ is bounded by $2D_\varepsilon$. Hence, when considering the regret with respect to the best (in general non-stationary) T-step policy one obtains a bound as in Theorem 12 with an additional additive constant of $2D_\varepsilon$.

Remark 14. For $\varepsilon=0$, one can also obtain logarithmic bounds analogously to Theorem 4 of [6]. With no additional information for the learner one gets the original UCRL2 bounds (with a slightly larger constant), trivially choosing B to be the number of states and assigning each state-action pair an individual color.

Remark 15. Theorem 12 is given for finite state MDPs. However, under the mentioned Lipschitz/Hölder conditions for rewards and transition probabilities (cf. Section 6.1.2) and some additional technical assumptions an analogous result can be derived for continuous state MDPs, where ε in Theorem 12 is replaced with the precision determined by the Lipschitz/Hölder parameters. For details we refer to [20]. We note that the algorithm and the derived results in [20] differ from the ones given here in that D_ε, the diameter in the discretized MDP, is replaced with the bias span of the optimal policy. The reason for this is that the aim of [20] is to derive sublinear regret bounds by eventually choosing a suitable discretization. With that respect (an analogon of) Theorem 12 is not very satisfactory, since the regret bound depends on the chosen discretization, that is, on the respective diameter of the discretized MDP. Unlike that, as will be seen below, in the restless bandit setting we are able to bound the diameter of the respective aggregated ε-structured MDP in a satisfactory way.

7. Constructing the Algorithm II: Coloring the T-step representation

Now, we can turn the T-step representation of any restless bandit into an ε-structured MDP as follows. We assign the same color to state-action pairs where the chosen arm is in the same state, that is, we assign

[5] For the sake of simplicity the algorithm was given for the case $\Psi_K=\varnothing$. It is obvious how to extend the algorithm when some parameters are known.

colors such that $c((s_i,n_i)_{i=1}^K, j) = c((s'_i,n'_i)_{i=1}^K, j')$ iff $j = j'$, $s_j = s'_j$, and either $n_j = n'_j$ or $n_j, n'_j \geq T^j_{\text{mix}}(\varepsilon)$. The respective translation functions are chosen to map states $(s_1, n_1+1, \ldots, s_{j-1}, n_{j-1}+1, s, 1, s_{j+1}, n_{j+1}+1, \ldots, s_K, n_K+1)$ to states $(s'_1, n'_1+1, \ldots, s'_{j-1}, n'_{j-1}+1, s, 1, s'_{j+1}, n'_{j+1}+1, \ldots, s'_K, n'_K+1)$. This ε-structured MDP can be learned with colored UCRL2. This is basically our proposed restless bandits algorithm, see Algorithm 2. (The dependence on the horizon T and the mixing times T^j_{mix} as input parameters can be eliminated, cf. the proof of Theorem 5 in Section 8.)

Algorithm 2 The restless bandits algorithm

```
**Input:** Confidence parameter $\delta > 0$, the number of states $S_j$ and mixing time $T^j_{\text{mix}}$ of each arm $j$, horizon $T$.
$\triangleright$ Choose $\varepsilon = 1/\sqrt{T}$ and execute colored UCRL2 (with confidence parameter $\delta$) on the $\varepsilon$-structured MDP described in Section 7.
```

8. Proofs

8.1. Proof of the Upper Bound

We start with bounding the diameter in aggregated ε-structured MDPs corresponding to a restless bandit problem.

Lemma 16. *Consider a restless bandit with K aperiodic arms having diameters D_j and mixing times T^j_{mix} $(j = 1, \ldots, K)$. For $\varepsilon \leq 1/4$, the diameter D_ε in the respective aggregated ε-structured MDP can be upper bounded by*

$$D_\varepsilon \;\leq\; 2\left\lceil \log_2(4 \max_j D_j)\right\rceil \cdot \left\lceil T_{\text{mix}}(\varepsilon)\right\rceil \cdot \prod_{j=1}^{K}(4D_j),$$

where we set $T_{\text{mix}}(\varepsilon) := \max_j T^j_{\text{mix}}(\varepsilon)$.

Proof. Let μ_j be the stationary distribution of arm j. It is well-known that the expected *first return time* $\tau_j(s)$ in state s satisfies $\mu_j(s) = 1/\tau_j(s)$. Set $\tau_j := \max_s \tau_j(s)$, and $\tau := \max_j \tau_j$. Then, $\tau_j \leq 2D_j$.

Now consider the following scheme to reach a given state $(s_j, n_j)_{j=1}^K$: First, order the states (s_j, n_j) descendingly with respect to n_j. Thus, assume that $n_{j_1} > n_{j_2} > \ldots > n_{j_K} = 1$. Take $\lceil T_{\text{mix}}(\varepsilon)\rceil$ samples from arm j_1. (Then each arm will be ε-close to the stationary distribution, and the probability of reaching the right state s_{j_i} when sampling arm j_i afterwards is at least $\mu_{j_i}(s_{j_i}) - \varepsilon$.) Then sample each arm j_i $(i = 2, 3, \ldots, K)$ exactly $n_{j_{i-1}} - n_{j_i}$ times.

We first show the lemma for $\varepsilon \leq \mu_0 := \min_{j,s} \mu_j(s)/2$. As observed before, for each arm j_i the probability of reaching the right state s_{j_i} is at least $\mu_{j_i}(s_{j_i}) - \varepsilon \geq \mu_{j_i}(s_{j_i})/2$. Consequently, the expected number of restarts of the scheme necessary to reach a particular state $(s_j, n_j)_{j=1}^K$ is upper bounded by $\prod_{j=1}^K 2/\mu_j(s_j)$.

As each trial takes at most $2\lceil T_{\mathrm{mix}}(\varepsilon)\rceil$ steps, recalling that $1/\mu_j(s) = \tau_j(s) \leq 2D_j$ proves the bound for $\varepsilon \leq \mu_0$.

Now assume that $\varepsilon > \mu_0$. Since $D_\varepsilon \leq D_{\varepsilon'}$ for $\varepsilon > \varepsilon'$ we obtain a bound of $2\lceil T_{\mathrm{mix}}(\varepsilon')\rceil \prod_{j=1}^K (4D_j)$ with $\varepsilon' := \mu_0 = 1/2\tau$. By (1) and our assumption that $\varepsilon \leq \frac{1}{4}$, we have

$$T_{\mathrm{mix}}(\varepsilon') \;\leq\; \lceil \log_2(1/\varepsilon')\rceil \cdot T_{\mathrm{mix}}(1/4) \;\leq\; \lceil \log_2(2\tau)\rceil \cdot T_{\mathrm{mix}}(\varepsilon),$$

which proves the lemma. □

Proof of Theorem 5. First, note that in each arm j the support of the transition probability distribution is upper bounded by $|S_j|$, and that the coloring described in Section 7 uses not more than $\sum_{j=1}^K |S_j|\,\lceil T^j_{\mathrm{mix}}(\varepsilon)\rceil$ colors. Hence, Theorem 12 with $C_U = \sum_{j=1}^K |S_j|\,\lceil T^j_{\mathrm{mix}}(\varepsilon)\rceil$ and $B = \max_i |S_i|$ shows that the regret is bounded by

$$42D_\varepsilon \sqrt{\max_i |S_i| \cdot \textstyle\sum_{j=1}^K |S_j| \cdot \lceil T^j_{\mathrm{mix}}(\varepsilon)\rceil \cdot T \log\left(\frac{T}{\delta}\right)} + \varepsilon(D_\varepsilon + 2)T \tag{5}$$

with probability $\geq 1-\delta$. Since $\varepsilon = 1/\sqrt{T}$, one obtains after some minor simplifications the first bound by Lemma 16 and recalling (1). Note that when we consider regret with respect to the best T-step policy, by Remark 13 we have an additional additive constant of $2D_\varepsilon$ in (5), which only slightly increases the numerical constant of the regret bound.

If the horizon T is not known, guessing T using the doubling trick (i.e., executing the algorithm for $T = 2^i$ with confidence parameter $\delta/2^i$ in rounds $i = 1, 2, \ldots$) achieves the bound given in Theorem 5 with worse constants.

Similarly, if T_{mix} is unknown, one can perform the algorithm in rounds $i = 1, 2, \ldots$ of length 2^i with confidence parameter $\delta/2^i$, choosing an increasing function $a(t)$ to guess an upper bound on T_{mix} at the beginning t of each round. This gives a bound of order $a(T)^{3/2}\sqrt{T}$ with a corresponding additive constant. In particular, choosing $a(t) = \log t$ the regret is bounded by

$$O\big(S \prod_{j=1}^K (4D_j) \cdot \max_i \log(D_i) \cdot \log^{7/2}(T/\delta) \cdot \sqrt{T}\big)$$

with probability $\geq 1-\delta$. □

8.2. Proof of the Lower Bound

Proof of Theorem 8. Consider K arms all of which are deterministic cycles of length m and hence m-periodic. Then the learner faces m distinct ordinary bandit problems (each corresponding to states of the same period in each cycle) having K arms. By choosing suitable rewards, each of these bandit problems can be made to force regret of order $\Omega(\sqrt{KT/m})$ in the T/m steps the learner deals with the problem [4]. Overall, this gives the claimed bound of $\Omega(\sqrt{mKT}) = \Omega(\sqrt{ST})$. Adding a sufficiently small probability (with respect to the horizon T) of staying in some state of each arm, one obtains the same bounds for aperiodic arms. □

9. The Periodic Case

Now let us turn to the case where one or more arms are periodic, and let m_j be the period of arm j. Note that periodic Markov chains do not converge to a stationary distribution. However, taking into account the period of the arms, one can generalize our results to the periodic case. Considering in an m_j-periodic Markov chain the m_j-step transition probabilities given by the matrix P^{m_j}, one obtains m_j distinct *aperiodic classes* (subchains depending on the period of the initial state) each of which converges to a stationary distribution $\mu_{j,\ell}$ with respective mixing time $T^{j,\ell}_{\text{mix}}(\varepsilon)$, $\ell = 1, 2, \ldots, m_j$. The ε-mixing time $T^j_{\text{mix}}(\varepsilon)$ of the chain then can be defined as

$$T^j_{\text{mix}}(\varepsilon) := m_j \max_{\ell} T^{j,\ell}_{\text{mix}}(\varepsilon).$$

Obviously, after that many steps each aperiodic class will be ε-close to its stationary distribution when sampling in the respective period. That is, sampling after $\lceil T^j_{\text{mix}}(\varepsilon)\rceil + \ell$ steps ($\ell = 0, \ldots, m_j - 1$) one is ε-close to the stationary distribution of one of the m_j aperiodic classes. As for aperiodic chains we set $T^j_{\text{mix}} := T^j_{\text{mix}}(\frac{1}{4})$, cf. Section 2.1.

9.1. Algorithm

Due to the possible periodic nature of some arms, in general for obtaining the MDP representation we cannot simply aggregate all states (s_j, n_j), (s'_j, n'_j) with $n_j, n'_j \geq T^j_{\text{mix}}(\varepsilon)$ as in the case of aperiodic chains, but aggregate them only if additionally $n_j \equiv n'_j \mod m_j$.

If the periods m_j are not known to the learner, one can use the least common denominator (lcd) of $1, 2, \ldots, |S_j|$ as (multiple of the true) period. Since by the prime number theorem the latter is exponential in $|S_j|$ — e.g. [21] shows that the lcd of $1, 2, \ldots, n$ is between 2^n and 4^n if $n \geq 9$ — the obtained results for periodic arms show worse dependence on the number of states. Of course, in practice one can also obtain improved upper bounds by estimating the period of each arm by the greatest common divisor of the observed return times in each state. However, it is not obvious how to obtain high probability bounds for the convergence of these estimates to the true period of a Markov chain, even when assuming knowledge of the mixing time as in our setting.

9.2. Regret Bound and Proof

First, concerning (the proof of) Lemma 16 the sampling scheme has to be slightly adapted to the setting of periodic arms so that one samples in the right period when trying to reach a particular state, giving slightly worse bounds depending on the arms' periods.

Lemma 17. *Consider a restless bandit with K arms having periods m_j, diameters D_j, and mixing times T^j_{mix} ($j = 1, \ldots, K$). For $\varepsilon \leq 1/4$, the diameter D_ε in the respective aggregated ε-structured MDP can be* *upper bounded by*

$$D_\varepsilon \;\leq\; \big(2\lceil T_{\mathrm{mix}}(\varepsilon)\rceil + \mathrm{lcd}(m_1, m_2, \ldots, m_K)\big)\big\lceil \log_2(4 \max_j D_j)\big\rceil \cdot \prod_{j=1}^{K}(4D_j),$$

where we set $T_{\mathrm{mix}}(\varepsilon) := \max_j T^j_{\mathrm{mix}}(\varepsilon)$*.*

Proof. Let $\mu_{j,\ell}$ be the stationary distribution of the aperiodic class of period ℓ in arm j. As before, in each subchain the expected *first return time* $\tau_{j,\ell}(s)$ in a state s of period ℓ satisfies $\mu_{j,\ell}(s) = 1/\tau_{j,\ell}(s)$. Set $\tau_{j,\ell} := \max_s \tau_{j,\ell}(s)$, and $\tau := \max_{j,\ell} \tau_{j,\ell}$. Then, $\tau_{j,\ell} \leq 2D_j$ for $\ell = 1, 2, \ldots, m_j$. (Note that $\tau_{j,\ell}$ counts only steps in the aperiodic subchain of period ℓ and considers only states in this chain, while D_j considers all steps and all states.)

Now consider the following modified scheme to reach a given state $(s_j, n_j)_{j=1}^K$, assuming that this state is reachable (which need not be the case in the periodic setting): First, as before, order the states (s_j, n_j) descendingly with respect to n_j, so that $n_{j_1} > n_{j_2} > \ldots > n_{j_K} = 1$. Take $\lceil T_{\mathrm{mix}}(\varepsilon)\rceil$ samples from arm j_1. (Then each arm will be ε-close to the stationary distribution, and the probability of reaching the right state s_{j_i} when sampling arm j_i afterwards *in the right period* is at least $\mu_{j_i}(s_{j_i}) - \varepsilon$.) Unlike in the aperiodic case where we can continue sampling each arm j_i ($i \geq 2$) exactly $n_{j_{i-1}} - n_{j_i}$ times, here we have to take into account that we can hit the right state in each arm only if we sample it in the right period. Thus, in order to assure that we can hit each of the states by the above mentioned scheme, we continue sampling arm j_1 an appropriate (with respect to the current state) number of times to reach the right period. Obviously, this can be done within at most $\mathrm{lcd}(m_1, m_2, \ldots, m_K)$ steps. Only then we sample each arm j_i ($i = 2, \ldots, K$) exactly $n_{j_{i-1}} - n_{j_i}$ times, guaranteeing that the probability of hitting each state s_j is at least $\mu_{j,\ell(s_j)}(s_j) - \varepsilon$, where $\ell(s)$ denotes the period of state s.

The rest of the proof then is analogous to the proof of Lemma 16. Again, we first show the lemma for $\varepsilon \leq \mu_0 := \min_{j,s} \mu_{j,\ell(s)}(s)/2$. In this case $\mu_{j,\ell(s_j)}(s_j) - \varepsilon \geq \mu_{j_i,\ell(s_j)}(s_j)/2$, and the expected number of restarts of the scheme necessary to reach a particular state $(s_j, n_j)_{j=1}^K$ is upper bounded by $\prod_{j=1}^K 2/\mu_{j,\ell(s_j)}(s_j)$. As each trial takes at most $2\lceil T_{\mathrm{mix}}(\varepsilon)\rceil + \mathrm{lcd}(m_1, \ldots, m_K)$ steps, recalling that $1/\mu_{j,\ell(s)}(s) = \tau_{j,\ell(s)}(s) \leq 2D_j$ proves the bound for $\varepsilon \leq \mu_0$.

If $\varepsilon > \mu_0$, we can again use that $D_\varepsilon \leq D_{\varepsilon'}$ for $\varepsilon > \varepsilon'$. Then setting $\varepsilon' := \mu_0 = 1/2\tau$ we obtain a bound of $\big(2\lceil T_{\mathrm{mix}}(\varepsilon')\rceil + \mathrm{lcd}(m_1, \ldots, m_K)\big)\prod_{j=1}^K(4D_j)$. Application of (1) gives

$$T_{\mathrm{mix}}(\varepsilon') \;\leq\; \lceil \log_2(1/\varepsilon')\rceil\, T_{\mathrm{mix}}(1/4) \;\leq\; \lceil \log_2(4\tau)\rceil\, T_{\mathrm{mix}}(\varepsilon)$$

and proves the lemma. □

Concerning Theorem 5, the proof given in Section 8 still holds for the periodic case. In particular, in spite of the different aggregation the bound ST_{mix} on the number of needed colors used in the original proof is still valid. The only difference in the proofs is that to bound the diameter now Lemma 16 is replaced with Lemma 17, giving slightly worse bounds when periods of the single arms are known. As already discussed above, when the periods are unknown, the learner can upper bound them by $\mathrm{lcd}(1, 2, \ldots, S)$, which results in bounds exponential in the number of states. Still, these bounds have optimal dependence on the horizon T.

10. Extensions and Outlook

Unknown state space. If (the size of) the state space of the individual arms is unknown, some additional exploration of each arm will sooner or later determine the state space. Thus, we may execute our algorithm on the known state space where between two episodes we sample each arm until all known states have been sampled at least once. The additional exploration is upper bounded by $O(\log T)$, as there are only $O(\log T)$ many episodes (cf. Appendix A.4), and the time of each exploration phase can be bounded with known results. That is, the expected number of exploration steps needed until all states of an arm j have been observed is upper bounded by $D_j \log(3|S_j|)$ (cf. Theorem 11.2 of [16]), while the deviation from the expectation can be dealt with by Markov inequality or results from [22]. That way, one obtains bounds as in Theorem 5 for the case of unknown state space.

Improving the bounds. All parameters considered, there is still a large gap between the lower and the upper bound on the regret. As a first step, it would be interesting to find out whether the dependence on the diameter of the arms is necessary. Also, the current regret bounds do not make use of the interdependency of the transition probabilities in the Markov chains and treat n-step and n'-step transition probabilities independently. Finally, a related open question is how to obtain estimates and upper bounds on mixing times. Whereas it is not easy to obtain upper bounds on the mixing time in general, for *reversible* Markov chains T_{mix} can be linearly upper bounded by the diameter, cf. Lemma 15 in Chapter 4 of [23]. While it is possible to compute an upper bound on the diameter of a Markov chain from samples of the chain, we did not succeed in deriving any useful results on the quality of such bounds.

More general models. After considering bandits with i.i.d. and Markov arms, the next natural step is to consider more general time-series distributions. Generalizations are not straightforward: already for the case of Markov chains of order (or memory) 2 the MDP representation of the problem (Section 5) breaks down, and so the approach taken here cannot be easily extended. Stationary ergodic distributions are an interesting more general case, for which the first question is whether it is possible to obtain asymptotically sublinear regret. Further important generalizations include problems in which arms are dependent and possibly non-stationary. For example, if each arm is a model of the environment, and “pulling” an arm means executing a policy that is optimal for the selected model, then there is both dependence between the arms and non-stationarity that results from attempting to learn the parameters of the models; for details and some results on this problem see [24, 25, 26].

Acknowledgments.

This research was funded by the Austrian Science Fund (FWF): J 3259-N13 and P 26219-N15, the Ministry of Higher Education and Research of France, Nord-Pas-de-Calais Regional Council and FEDER (Contrat de Projets Etat Region CPER 2007-2013), and by the European Community's FP7 Program under grant agreement n° 270327 (CompLACS).

References

[1] T. L. Lai, H. Robbins, Asymptotically efficient adaptive allocation rules, Adv. in Appl. Math. 6 (1985) 4–22.

[2] I. F. Akyildiz, W.-Y. L. W.-Y. Lee, M. C. Vuran, S. Mohanty, A survey on spectrum management in cognitive radio networks, IEEE Commun. Mag. 46 (4) (2008) 40–48.

[3] V. Anantharam, P. Varaiya, J. Walrand, Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays, part II: Markovian rewards, IEEE Trans. Automat. Control 32 (11) (1987) 977–982.

[4] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, SIAM J. Comput. 32 (2002) 48–77.

[5] J.-Y. Audibert, S. Bubeck, Minimax policies for adversarial and stochastic bandits, in: colt2009. Proc. 22nd Annual Conf. on Learning Theory, 2009, pp. 217–226.

[6] T. Jaksch, R. Ortner, P. Auer, Near-optimal regret bounds for reinforcement learning, J. Mach. Learn. Res. 11 (2010) 1563–1600.

[7] P. L. Bartlett, A. Tewari, REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs, in: Proc. 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009, AUAI Press, 2009, pp. 35–42.

[8] C. Tekin, M. Liu, Adaptive learning of uncontrolled restless bandits with logarithmic regret, in: 49th Annual Allerton Conference, IEEE, 2011, pp. 983–990.

[9] S. Filippi, O. Cappe and, A. Garivier, Optimally sensing a single channel without prior information: The tiling algorithm and regret bounds, IEEE J. Sel. Topics Signal Process. 5 (1) (2011) 68–76.

[10] S. Bubeck, N. Cesa-Bianchi, Regret analysis of stochastic and nonstochastic multi-armed bandit problems, Found. Trends Mach. Learn. 5 (1) (2012) 1–122.

[11] J. C. Gittins, Bandit processes and dynamic allocation indices, J. R. Stat. Soc. Ser. B Stat. Methodol. 41 (2) (1979) 148–177.

[12] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multi-armed bandit problem, Mach. Learn. 47 (2002) 235–256.

[13] P. Whittle, Restless bandits: Activity allocation in a changing world, J. Appl. Probab. 25 (1988) 287–298.

[14] R. Givan, T. Dean, M. Greig, Equivalence notions and model minimization in Markov decision processes., Artif. Intell. 147 (1-2) (2003) 163–223.

[15] B. Ravindran, A. G. Barto, Model minimization in hierarchical reinforcement learning, in: Abstraction, Reformulation and Approximation, 5th International Symposium, SARA 2002, 2002, pp. 196–211.

[16] D. A. Levin, Y. Peres, E. L. Wilmer, Markov chains and mixing times, American Mathematical Society, 2006.

[17] R. Ortner, Pseudometrics for state aggregation in average reward Markov decision processes, in: Proc. 18th International Conf. on Algorithmic Learning Theory, ALT 2007, Springer, 2007, pp. 373–387.

[18] J. Taylor, D. Precup, P. Panangaden, Bounding performance loss in approximate MDP homomorphisms, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), Advances in Neural Information Processing Systems 21, 2009, pp. 1649–1656.

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

[20] R. Ortner, D. Ryabko, Online regret bounds for undiscounted continuous reinforcement learning, in: P. Bartlett, F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, 2012, pp. 1772–1780.

[21] M. Nair, On Chebyshev-type inequalities for primes, Amer. Math. Monthly 89 (2) (1982) 126–129.

[22] D. Aldous, Threshold limits for cover times, J. Theoret. Probab. 4 (1991) 197–211.

[23] D. Aldous, J. A. Fill, Reversible markov chains and random walks on graphs, unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/∼aldous/RWG/book.html` (2002).

[24] D. Ryabko, M. Hutter, On the possibility of learning in reactive environments with arbitrary dependence, Theoret. Comput. Sci. 405 (3) (2008) 274–284.

[25] O. Maillard, R. Munos, D. Ryabko, Selecting the state-representation in reinforcement learning, in: J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 24, 2011, pp. 2627–2635.

[26] O.-A. Maillard, P. Nguyen, R. Ortner, D. Ryabko, Optimal regret bounds for selecting the state representation in reinforcement learning, in: JMLR Workshop and Conference Proceedings Volume 28 : Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 543 – 551.

Appendix A. Proof of Theorem 12

The proof is an adaptation of the proof of the original regret bound for UCRL2, that is, Theorem 2 in [6]. We therefore follow the main steps of the proof of Theorem 2 in [6], and often refer to the original proof for technical details.

Let us first give a brief overview. In Appendix A.1 we first define the regret Δ_k of an episode k, so that we can bound the the regret by the sum over the Δ_k and another term dealing with the randomness of the observed rewards. In Appendix A.2 we handle the regret due to failing confidence intervals. This is the only part of the proof where the adaptations of the original proof are not straightforward and a more refined argument is necessary. In Appendix A.3 and Appendix A.4 we bound the difference of the optimal average reward and the mean reward of each visited state, using the confidence intervals for the rewards and the transition probabilities. In the final Appendix A.5 we conclude by summing up the individual regret terms.

Before starting, we recall some notation. Let M denote the true MDP with transition probabilities $p(\cdot|s,a)$, mean rewards $r(s,a)$, and optimal average reward ρ^*. $\mathcal{M}_k$ is the set of plausible MDPs whose transition probabilities $\tilde{p}(\cdot|s,a)$ and rewards $\tilde{r}(s,a)$ satisfy (2) and (3), where $\hat{p}(\cdot|s,a)$ and $\hat{r}(s,a)$ are the estimated transition probabilities and rewards, respectively. Further, let $\tilde{M}_k$ be the optimistic MDP chosen by the algorithm from the set $\mathcal{M}_k$, and $\tilde{\pi}_k$ the policy chosen by the algorithm in episode k.

Appendix A.1. Splitting into Episodes

Let $v_k(s,a)$ be the number of times action a has been chosen in state s in episode k, and set

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

Using Hoeffding's inequality to deal with the randomness of the observed rewards one can show (cf. Section 4.1 of [6]) that with probability at least $1 - \frac{\delta}{12T^{5/4}}$ the regret after T steps is upper bounded by

$$\sum_{k=1}^{m} \Delta_k + \sqrt{\tfrac{5}{8} T \log\left(\tfrac{8T}{\delta}\right)}\ . \tag{A.1}$$

Appendix A.2. Failing Confidence Intervals

Concerning the regret with respect to the true MDP M being not contained in the set of plausible MDPs $\mathcal{M}_k$, we cannot use the same argument (that is, Lemma 17 in Appendix C.1) as in [6], since the random variables we consider for rewards and transition probabilities are independent, yet not identically distributed.

Instead, fix a state-action pair (s,a), let $S(s,a)$ be the set of states s' with $p(s'|s,a) > 0$ and recall that $\hat{r}(s,a)$ and $\hat{p}(\cdot|s,a)$ are the estimates for rewards and transition probabilities calculated from all samples of state-action pairs of the same color $c(s,a)$. Now assume that at step t there have been $n > 0$ samples of state-action pairs of color $c(s,a)$ and that in the i-th sample action a_i has been chosen in state s_i and a transition to state s'_i has been observed ($i = 1, \ldots, n$). Then

$$\begin{aligned}
\Big\|\hat{p}(\cdot|s,a) - \mathbb{E}[\hat{p}(\cdot|s,a)]\Big\|_1 &= \sum_{s' \in S(s,a)} \Big|\hat{p}(s'|s,a) - \mathbb{E}[\hat{p}(s'|s,a)]\Big| \\
\leq \quad & \sup_{x \in \{-1,1\}^{|S(s,a)|}} \sum_{s' \in S(s,a)} \Big(\hat{p}(s'|s,a) - \mathbb{E}[\hat{p}(s'|s,a)]\Big)\, x(s') \\
= \quad & \sup_{x \in \{-1,1\}^{|S(s,a)|}} \frac{1}{n} \sum_{i=1}^{n} \Big(x(\phi_{s_i,a_i,s,a}(s'_i)) - \sum_{s'} p(s'|s_i,a_i) \cdot x(\phi_{s_i,a_i,s,a}(s')) \Big)\,. \qquad (A.2)
\end{aligned}$$

For fixed $x \in \{-1,1\}^{|S(s,a)|}$,

$$X_i := x(\phi_{s_i,a_i,s,a}(s'_i)) - \sum_{s'} p(s'|s_i,a_i) \cdot x(\phi_{s_i,a_i,s,a}(s'))$$

is a martingale difference sequence with $|X_i| \leq 2$, so that by Azuma-Hoeffding inequality (e.g., Lemma 10 in [6]), $\Pr\{\sum_{i=1}^n X_i \geq \theta\} \leq \exp(-\theta^2/8n)$ and in particular

$$\Pr\Big\{ \textstyle\sum_{i=1}^n X_i \geq \sqrt{56Bn \log\left(\frac{2tC_U}{\delta}\right)} \Big\} \leq \left(\frac{\delta}{2tC_U}\right)^{7B} < \frac{\delta}{2^B 20 t^7 C_U}.$$

Recalling that by assumption $|S(s,a)| \leq B$, a union bound over all sequences $x \in \{0,1\}^{|S(s,a)|}$ then shows from (A.2) that

$$\Pr\Big\{ \Big\|\hat{p}(\cdot|s,a) - \mathbb{E}[\hat{p}(\cdot|s,a)]\Big\|_1 \geq \sqrt{\tfrac{56B}{n} \log\left(2C_U t/\delta\right)} \Big\} \quad \leq \quad \frac{\delta}{20t^7 C_U}. \tag{A.3}$$

Concerning the rewards, as in the proof of Lemma 17 in Appendix C.1 of [6] — but now using Hoeffding's inequality for independent and not necessarily identically distributed random variables — we have that

$$\Pr\left\{\left|\hat{r}(s,a) - \mathbb{E}[\hat{r}(s,a)]\right| \geq \sqrt{\tfrac{7}{2n}\log\left(2C_U t/\delta\right)}\right\} \quad \leq \quad \tfrac{\delta}{60t^7 C_U}. \tag{A.4}$$

A union bound over all t possible values for n and all C_U colors of states in Ψ_U shows that the confidence intervals in (A.3) and (A.4) hold with probability at least $1 - \frac{\delta}{15t^6}$ for the actual counts $N(c(s,a))$ and all state-action pairs (s,a). (Note that equations (A.3) and (A.4) are the same for state-action pairs of the same color.)

By linearity of expectation, $\mathbb{E}[\hat{r}(s,a)]$ can be written as $\frac{1}{n}\sum_{i=1}^{n} r(s_i,a_i)$ for the sampled state-action pairs (s_i,a_i). Since the (s_i,a_i) are assumed to have the same color $c(s,a)$, it holds that $|r(s_i,a_i) - r(s,a)| < \varepsilon$ and hence $|\mathbb{E}[\hat{r}(s,a)] - r(s,a)| < \varepsilon$. Similarly, $\left\|\mathbb{E}[\hat{p}(\cdot|s,a)] - p(\cdot|s,a)\right\|_1 < \varepsilon$. Together with (A.3) and (A.4) this shows that with probability at least $1 - \frac{\delta}{15t^6}$ for all state-action pairs (s,a)

$$\left|\hat{r}(s,a) - r(s,a)\right| \quad < \quad \varepsilon + \sqrt{\tfrac{7\log(2C_U t/\delta)}{2N(c(s,a))}}, \tag{A.5}$$

$$\left\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\right\|_1 \quad < \quad \varepsilon + \sqrt{\tfrac{56B\log(2C_U t/\delta)}{N(c(s,a))}}. \tag{A.6}$$

Thus, the true MDP is contained in the set of plausible MDPs $\mathcal{M}(t)$ at step t with probability at least $1 - \frac{\delta}{15t^6}$, just as in Lemma 17 of [6]. Then as in Section 4.2 of [6], bounding the sum by a respective integral we have $\sum_{\lfloor T^{1/4}\rfloor+1}^{T} \frac{\delta}{15t^6} \leq \frac{\delta}{12T^{5/4}}$, so that

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M\notin\mathcal{M}_k} \;\leq\; \sum_{t=1}^{T} t\mathbb{1}_{M\notin\mathcal{M}(t)} \;\leq\; \sum_{t=1}^{\lfloor T^{1/4}\rfloor} t\mathbb{1}_{M\notin\mathcal{M}(t)} + \sum_{t=\lfloor T^{1/4}\rfloor+1}^{T} t\mathbb{1}_{M\notin\mathcal{M}(t)} \;\leq\; \sqrt{T} \tag{A.7}$$

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

Appendix A.3. Episodes with $M \in \mathcal{M}_k$

Now assuming that the true MDP M is in $\mathcal{M}_k$, we first reconsider extended value iteration [6]. In Section 4.3.1 of [6] it is shown that for the state values $u_i(s)$ in the i-th iteration it holds that $\max_s u_i(s) - \min_s u_i(s) \leq D$, where D is the diameter of the MDP. Now we want to replace D with the diameter D_ε of the aggregated MDP. For this, first note that for any two states s, s' which are aggregated we have by definition of the aggregated MDP that $u_i(s) = u_i(s')$. As it takes at most D_ε steps on average to reach any aggregated state, repeating the argument of Section 4.3.1 of [6] shows that

$$\max_s u_i(s) - \min_s u_i(s) \leq D_\varepsilon. \tag{A.8}$$

Let $\tilde{\boldsymbol{P}}_k := \big(\tilde{p}_k(s'|s,\tilde{\pi}_k(s))\big)_{s,s'}$ be the transition matrix of $\tilde{\pi}_k$ on $\tilde{M}_k$, and $\boldsymbol{v}_k := \big(v_k\big(s,\tilde{\pi}_k(s)\big)\big)_s$ the row vector of visit counts in episode k for each state and the corresponding action chosen by $\tilde{\pi}_k$. Then as shown in Section 4.3.1 of [6], we have by definition and convergence[6] of (extended) value iteration

$$\Delta_k \quad \leq \quad \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + \sum_{s,a} v_k(s,a)\big(\tilde{r}_k(s,a) - r(s,a)\big),$$

where $\boldsymbol{w}_k$ is the normalized state value vector with $w_k(s) := u(s) - (\min_s u(s) - \max_s u(s))/2$, so that $\|\boldsymbol{w}_k\| \leq \frac{D_\varepsilon}{2}$. Now for $(s,a) \in \Psi_K$ we have $\tilde{r}_k(s,a) = r(s,a)$, while for $(s,a) \in \Psi_U$ the term $\tilde{r}_k(s,a) - r(s,a) \leq |\tilde{r}_k(s,a) - \hat{r}_k(s,a)| + |r(s,a) - \hat{r}_k(s,a)|$ is bounded according to (2) and (A.5), as we assume that $\tilde{M}_k, M \in \mathcal{M}_k$. Summarizing state-action pairs of the same color we get

$$\Delta_k \quad \leq \quad \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + 2 \sum_{c \in C(\Psi_U)} v_k(c) \cdot \Big(\varepsilon + \sqrt{\tfrac{7\log(2C_U t_k/\delta)}{2\max\{1, N_k(c)\}}}\Big),$$

where $C(\Psi_U)$ is the set of colors of state-action pairs in Ψ_U. Let T_k be the length of episode k. Then noting that $N'_k(c) := \max\{1, N_k(c)\} \leq t_k \leq T$ we obtain

$$\Delta_k \quad \leq \quad \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k + 2\varepsilon T_k + \sqrt{14\log\big(\tfrac{2C_U T}{\delta}\big)} \sum_{c \in C(\Psi_U)} \frac{v_k(c)}{\sqrt{N'_k(c)}}. \tag{A.9}$$

Appendix A.4. The True Transition Matrix

Let $\boldsymbol{P}_k := \big(p(s'|s, \tilde{\pi}_k(s))\big)_{s,s'}$ be the transition matrix of $\tilde{\pi}_k$ in the true MDP M. We split

$$\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{I}\big)\boldsymbol{w}_k \quad = \quad \boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k + \boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k. \tag{A.10}$$

By assumption $\tilde{M}_k, M \in \mathcal{M}_k$, so that using (3) and (A.6) the first term in (A.10) can be bounded by (cf. Section 4.3.2 of [6])

$$\begin{aligned}
\boldsymbol{v}_k\big(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k\big)\boldsymbol{w}_k \quad &\leq \quad \sum_{s,a} v_k(s,a) \cdot \big\|\tilde{p}_k(\cdot|s,a) - p(\cdot|s,a)\big\|_1 \cdot \|\boldsymbol{w}_k\|_\infty \\
&\leq \quad 2 \sum_{c \in C(\Psi_U)} v_k(c) \cdot \Big(\varepsilon + \sqrt{\tfrac{56B\log(2C_U T/\delta)}{N'_k(c)}}\Big) \cdot \tfrac{D_\varepsilon}{2} \\
&\leq \quad \varepsilon D_\varepsilon T_k + D_\varepsilon \sqrt{56B\log\big(\tfrac{2C_U T}{\delta}\big)} \sum_{c \in C(\Psi_U)} \frac{v_k(c)}{\sqrt{N'_k(c)}},
\end{aligned} \tag{A.11}$$

since — as for the rewards — the contribution of state-action pairs in Ψ_K is 0.

Concerning the second term in (A.10), we define the sequence $X_t := \big(p(\cdot|s_t, a_t) - \mathbf{e}_{s_{t+1}}\big)\, \boldsymbol{w}_{k(t)} \mathbb{1}_{M \in \mathcal{M}_{k(t)}}$, where $\mathbf{e}_i$ denotes the unit vector with 1 in coordinate i, and $k(t)$ is the index of the episode which contains step t. Then as shown in Section 4.3.2 of [6], we can write

$$\boldsymbol{v}_k\big(\boldsymbol{P}_k - \boldsymbol{I}\big)\boldsymbol{w}_k = \sum_{t_k}^{t_{k+1}-1} X_t + w_k(s_{t_{k+1}}) - w_k(s_{t_k}) \leq \sum_{t_k}^{t_{k+1}-1} X_t + D_\varepsilon.$$

[6] For the sake of simplicity, we neglect the error by value iteration, which is explicitly considered in Section 4.3.1 of [6]. Taking into account this error only slightly deteriorates the constant in our bound.

Further, X_t is a martingale difference sequence with $|X_t| \leq D_\varepsilon$, so that application of Azuma-Hoeffding inequality (cf. Section 4.3.2 of [6]) gives that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\sum_{k=1}^{m} \boldsymbol{v}_k(\boldsymbol{P}_k - \boldsymbol{I})\boldsymbol{w}_k \mathbb{1}_{M \in \mathcal{M}_k} \quad \leq \quad D_\varepsilon \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + D_\varepsilon\, C_U \log_2\left(\tfrac{8T}{C_U}\right). \tag{A.12}$$

Here m is the number of episodes, and the bound $m \leq C_U \log_2(8T/C_U)$ used to obtain (A.12) is derived analogously to Appendix C.2 of [6].

Appendix A.5. Summing over Episodes with $M \in \mathcal{M}_k$

To conclude, we sum (A.9) over all episodes with $M \in \mathcal{M}_k$, using (A.10), (A.11), and (A.12), which yields that with probability at least $1 - \frac{\delta}{12T^{5/4}}$

$$\begin{aligned} \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} \leq{} & D_\varepsilon \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + D_\varepsilon\, C_U \log_2\left(\tfrac{8T}{C_U}\right) + \varepsilon(D_\varepsilon + 2)T \\ & + \left(D_\varepsilon \sqrt{56B \log\left(\tfrac{2C_U T}{\delta}\right)} + \sqrt{14 \log\left(\tfrac{2C_U T}{\delta}\right)} \right) \sum_{k=1}^{m} \sum_{c \in C(\Psi_U)} \frac{v_k(c)}{\sqrt{N'_k(c)}}. \end{aligned} \tag{A.13}$$

As shown in Section 4.3.3 and Appendix C.3 of [6], using that $v_k(c) \leq N'_k(c)$ for all colors c and applying Jensen's inequality, one obtains

$$\sum_{c \in C(\Psi_U)} \sum_{k} \frac{v_k(c)}{\sqrt{N'_k(c)}} \ \leq\ \left(\sqrt{2} + 1\right) \sqrt{C_U T}.$$

Thus, evaluating (A.1) by summing Δ_k over all episodes, by (A.7) and (A.13) the regret is upper bounded with probability $\geq 1 - \frac{\delta}{4T^{5/4}}$ by

$$\begin{aligned} & \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} + \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \in \mathcal{M}_k} + \sqrt{\tfrac{5}{8} T \log\left(\tfrac{8T}{\delta}\right)} \\ & \quad \leq \sqrt{\tfrac{5}{8} T \log\left(\tfrac{8T}{\delta}\right)} + \sqrt{T} + D_\varepsilon \sqrt{\tfrac{5}{2} T \log\left(\tfrac{8T}{\delta}\right)} + D_\varepsilon\, C_U \log_2\left(\frac{8T}{C_U}\right) \\ & \qquad + \varepsilon(D_\varepsilon + 2)T + 3\big(\sqrt{2} + 1\big) D_\varepsilon \sqrt{14 B C_U T \log\left(\tfrac{2C_U T}{\delta}\right)}\,. \end{aligned}$$

Further simplifications as in Appendix C.4 of [6] finish the proof. □