Theoretical Computer Science 558 (2014) 62–76

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Regret bounds for restless Markov bandits

Ronald Ortner [a,*], Daniil Ryabko [b], Peter Auer [a], Rémi Munos [b]

[a] Montanuniversitaet Leoben, A-8700 Leoben, Austria
[b] Inria Lille-Nord Europe, F-59650 Villeneuve d'Ascq, France

ARTICLE INFO

Available online 21 September 2014

Keywords:
Restless bandits
Markov decision processes
Regret

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.

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 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

* Corresponding author.
E-mail addresses: rortner@unileoben.ac.at (R. Ortner), daniil@ryabko.net (D. Ryabko), auer@unileoben.ac.at (P. Auer), remi.munos@inria.fr (R. Munos).

http://dx.doi.org/10.1016/j.tcs.2014.09.026
0304-3975/

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 Sections 8 and 9 (with some details deferred to Appendix A), 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 specified 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.