Approximations of the Restless Bandit Problem

Steffen Grünewälder S.GRUNEWALDER@LANCASTER.AC.UK

Azadeh Khaleghi A.KHALEGHI@LANCASTER.AC.UK

Department of Mathematics & Statistics
Lancaster University, Lancaster, UK

Editor: EWRL 2016

Abstract

The multi-armed restless bandit problem is studied in the case where the pay-offs are not necessarily independent over time nor across the arms, but the joint distribution over the arms is φ-mixing. Even though this version of the problem provides a more realistic model for most real-world applications, it cannot be optimally solved in practice since it is known to be PSPACE-hard. In this paper, a special setting is characterised where *good* approximate solutions can indeed be found via simple UCB-type algorithms. In particular, it is shown that under some conditions on the φ-mixing coefficients, a modified version of UCB recovers the best achievable regret of the i.i.d. setting (up to some function of the φ-mixing coefficients).

1. Introduction

Multi-armed bandit problems are typically studied under the assumption that the pay-offs are independently and identically distributed (i.i.d.), and the arms are independent. Under this assumption, the empirical mean accurately estimates the true mean of the arm and confidence intervals can easily be obtained around the estimates. However, in practice reward distributions are not necessarily i.i.d. and as such, both the mean estimates and the confidence intervals may deviate greatly from their true values. One of the most popular applications of multi-armed bandits concerns Online Advertisement in which the aim is to garner as many clicks as possible from a user. Grouping adverts into categories and associating with each category an arm, this problem turns into that of multi-armed bandit's. There is dependence over time and across the arms since, for example, we expect a user to be more likely to click on adverts that are related to her selections in the recent past.

In this paper, we are concerned with developing algorithms that are robust with respect to such dependencies. More specifically, we consider the multi-armed bandit problem in the case where the distribution of the arms are φ-mixing, and each arm evolves over time regardless of whether or not it is played. This is an instance of the so-called *restless* bandit problem (Whittle, 1988; Guha, Munagala, and Shi, 2010; Ortner, Ryabko, Auer, and Munos, 2014). Note that, since in this setting an optimal policy can leverage the inter-dependencies between the samples and switch between the arms at appropriate times, it can obtain an overall pay-off much higher than that given by playing the *best arm*, i.e. the distribution with the highest expected pay-off, see Example 1 in (Ortner et al., 2014). However, finding the best such *switching strategy* is PSPACE-hard, even in the case where the process distributions are Markovian with known dynamics (Papadimitriou and Tsitsiklis, 1999).

We consider a relaxation of the problem in the case where the φ-mixing coefficients are small, and develop an algorithm to approximate the optimal policy of the relaxed problem. The proposed algorithm recovers the best achievable regret of the i.i.d. setting (up to some function of the φ-mixing coefficients). Furthermore, we show that in this case, the optimum of the relaxed problem is close to that given by the best switching strategy. Note that even this relaxed version of the problem is not straightforward. The main challenge lies in obtaining confidence intervals around empirical estimates of the stationary means. Since Hoeffding-type concentration bounds exist for φ-mixing processes, it may be tempting to use such inequalities directly with standard UCB algorithms designed for the i.i.d. setting to find the best arm; in fact, this seems to be the approach taken by Audiffren and Ralaivola (2015). However, as we show through an example in this paper, a sequence of random variables obtained by sampling a stationary φ-mixing process at random times, may not necessarily be φ-mixing. As a result, in order for Hoeffding-type concentration results (for φ-mixing processes) to be applicable in this setting, the sampling policy must be designed appropriately.

2. Preliminaries

Let $\mathbb{N}_+ := \{1, 2, \ldots\}$ and $\overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$ denote the set and extended set of natural numbers respectively. We introduce the abbreviation $\mathbf{a}_{m..n},\ m, n \in \mathbb{N}_+, m \leq n$, for sequences $a_m, a_{m+1}, \ldots, a_n$. Given a finite subset $C \subset \mathbb{N}_+$ and a sequence $\mathbf{a}$, we let $\mathbf{a}_C := \{a_i : i \in C\}$ denote the set of elements of $\mathbf{a}$ indexed by C. If X_C is a sequence of random variables indexed by $C \subset \mathbb{N}_+$, we denote by $\sigma(X_C)$ the smallest σ-algebra generated by X_C.

Stochastic processes & φ-mixing. Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ be a measurable space; we let $\mathcal{X} \subset [0, 1]$ [1] and denote by $\mathcal{B}_\mathcal{X}^{(m)}$ the Borel σ-algebra on $\mathcal{X}^m,\ m \in \mathbb{N}_+$. A stochastic process is a probability measure over the space $(\mathcal{X}^\infty, \mathcal{B})$ where $\mathcal{B}$ denotes the σ-algebra on $\mathcal{X}^\infty$ generated by the cylinder sets. A process distribution ρ is stationary if $\rho(X_{1..m} \in B) = \rho(X_{i+1..i+m} \in B)$ for all Borel sets $B \in \mathcal{B}_\mathcal{X}^{(m)}$, $i \in \mathbb{N}_+, m \in \mathbb{N}_+$. The focus of this work is on stationary φ-mixing processes which may be defined as follows (Doukhan, 1994). Let $(\mathcal{X}^\infty, \mathcal{B}, \rho)$ be a stochastic process as defined above. The φ-dependence between X_A and X_B is defined as $\varphi(X_A, X_B) := \sup\{|\rho(V) - \rho(U \cap V)/\rho(U)| : U \in \sigma(X_A), \rho(U) > 0, V \in \sigma(X_B)\}$. A process $(X_i)_{i \in \mathbb{N}_+}$ is φ-mixing if $\lim_{n \to \infty} \varphi_n(u, v) = 0$, for all $u, v \in \mathbb{N}_+$, where $\varphi_n(u, v) := \sup\{\varphi(X_A, X_B) : A = 1..u, B = a..a + v - 1, a \geq u + n\}$, $n \in \mathbb{N}, u, v \geq 1$.

Stationary φ-mixing Bandits. We are concerned with $k < \infty$ stochastic processes. We define the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega := \Omega_1 \times \ldots \times \Omega_k$ with $\Omega_i := \mathcal{X}^\infty,\ i \in 1..k$, P a probability measure and $\mathcal{A}$ obtained via the cylinder sets as above. In much the same way as for a single process, we say that the joint process is φ-mixing if we have $\lim_{n \to \infty} \varphi_n(u, v) = 0$ for all $u,\ v \in \mathbb{N}_+$ where $\varphi_n(u, v) := \{\varphi(X_{A,1..k}, X_{B,1..k}) : A = 1..u,\ B = a..a + v - 1, a \geq u + n\},\ n \in \mathbb{N}_+,\ u, v \geq 1$ and $\varphi(X_{A,1..k}, X_{B,1..k}) := \sup\{|P(V) - P(U \cap V)/P(U)| : U \in \sigma(X_{A,1}, \ldots, X_{A,k}), P(U) > 0, V \in \sigma(X_{B,1}, \ldots, X_{B,k})\}$. Note that under this assumption there could be dependence between the arms; we only require for the joint process to be φ-mixing. An inherent part of the problem involves working with *pay-offs* $X_{\tau,i},\ i \leq k$, obtained by the player at random times $\tau : \Omega \to \overline{\mathbb{N}}_+$ when arm i is played. We define these as $X_{\tau,i} := \sum_{t \in \mathbb{N}_+} X_{t,i} \times \chi\{\tau = t\}$, where χ denotes the indicator function.

[1] More generally $\mathcal{X}$ can be a finite set or a closed interval $[a, b]$, for $a < b$, $a, b \in \mathbb{R}$.

3. Problem Formulation

We consider the restless bandit problem in the following setting. A total number of k bandit arms are given, where for each $i \in 1..k$, arm i corresponds to a stationary process that generates a time series of pay-offs $X_{1,i}, X_{2,i}, \ldots$. The joint process over the k arms is φ-mixing in the sense defined in Section 2, and $\|\varphi\| := \sum_{i=1}^{\infty} \varphi_i < \infty$. At every time-step $t \in \mathbb{N}_+$, a player chooses one of k arms according to a *policy* π_t and receives a reward X_{t,π_t}. The player's objective is to maximize the sum of the pay-offs received. The policy has access only to the pay-offs gained at earlier stages and to the arms it has chosen. More formally, a policy is a sequence of mappings $\pi_t : \Omega \to \{1, \ldots, k\}$, $t \geq 1$, each of which is measurable with respect to $\mathcal{F}_{t-1}$ where $\langle \mathcal{F}_t \rangle_{t \geq 0}$ is a filtration that tracks the pay-offs obtained in the past t rounds, i.e. $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and for $t \geq 1$ we let $\mathcal{F}_t = \sigma(X_{1,\pi_1}, \ldots, X_{t,\pi_t})$. This assumption is equivalent to the assumption that the policy can be written as a function of the past pay-offs and chosen arms Shiryaev (1991)[Thm. 3, p.174]. Let $\Pi = \{\langle \pi_t \rangle_{t \geq 1} : \pi_t \text{ is } \mathcal{F}_{t-1}\text{-measurable for all } t \geq 1\}$ denote the space of all possible policies. We define the maximal value that can be achieved in T rounds as

$$v_T^* = \sup_{\pi \in \Pi} \sum_{t=1}^{T} \mathbb{E}[X_{t,\pi_t}].$$

Our objective is to devise policies that achieve an expected pay-off (summed over T rounds) close to v_T^*. As discussed in the Introduction, it is easy to see that the optimal pay-off in this setting cannot be obtained by identifying the *best arm*, i.e. the distribution with the highest expected pay-off, see Example 1 in (Ortner et al., 2014). The optimal strategy is to switch between the arms at appropriate time steps. However, obtaining the best switching strategy is PSPACE-hard, even in the case where each arm follows a Markov chain with known dynamics. In this paper, we consider a relaxation of the problem in the case where the φ-mixing coefficients are small, and develop an algorithm to approximate the optimal policy of the relaxed problem. Furthermore, we show that in this case, the optimum of the relaxed problem is close to v_T^*.

4. Main Results

In this section we outline our findings. Due to space constraints these are provided without proof.

4.1 Approximation Error

We start by translating φ-mixing properties to those of expectations in order to control the difference between what a switching strategy can achieve as compared to the best stationary mean. This is established by Proposition 2, which shows that if we have a random variable X that takes values in $[0, 1]$ and depends only weakly on some collected information, then the conditional expectation of X given that information is close to the expected value of X.

Proposition 1 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, let X be a real-valued random variable taking values in $\mathcal{X}$ and denote by $\mathcal{G}$ some σ-subalgebra of $\mathcal{A}$. If there exists $\varphi \geq 0$ such that*

$$|P(A)P(B) - P(A \cap B)| \leq \varphi P(B)$$

for all $A \in \sigma(X), B \in \mathcal{G}$, then for any $B \in \mathcal{G}$ we have

$$\int_B |E(X|\mathcal{G}) - E(X)|\, dP \leq 4\varphi P(B), \text{ and } \|E(X|\mathcal{G}) - E(X)\|_{\mathcal{L}^1(P)} \leq 4\varphi.$$

The result is tight in the sense that for any $0 < \varphi < 1/2$ there exists a probability space $(\Omega, \mathcal{A}, P)$, a σ-subalgebra $\mathcal{G} \subset \mathcal{A}$, and a random variable X, such that $|P(A)P(B') - P(A \cap B')| \leq \varphi P(B')$ for all $B' \in \mathcal{G}$ and there exists a set $B \in \mathcal{G}$, $P(B) > 0$, with

$$\varphi P(B) \leq \int_B |E(X|\mathcal{G}) - E(X)| \; dP.$$

Using Proposition 1 we have the following statement.

Proposition 2 *Consider a k-armed stationary φ-mixing bandit problem. Let $\mu_1, \ldots, \mu_k$ be the means of the stationary distributions and let $\mu^* = \max\{\mu_1, \ldots, \mu_k\}$. Then for every $T \geq 1$ we have*

$$v_T^* - T\mu^* \leq T\varphi_1.$$

In particular, if the process has a weak dependence such that $\varphi_1 \leq \epsilon$ for some small ϵ, then we are guaranteed to only lose a factor of $T\epsilon$ if we settle for the arm with the highest mean instead of using the best possible switching policy.

4.2 Policies and the φ-mixing Property

The policies we consider have information about the whole (observed) past which may lead to stronger couplings between past and future pay-offs. As a result, depending on the policy used, the pay-off sequences obtained by playing a set of jointly φ-mixing bandit arms are not guaranteed to be φ-mixing. This is demonstrated by the following example.

Example 1 *Consider a two-armed bandit problem, where the first arm is deterministically set to 0, i.e. $X_{t,1} = 0,\ t \in \mathbb{N}_+$ and the second arm has a process distribution described by a two state Markov chain with the following transition matrix,*

$$T = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}, \quad \textit{with some } \epsilon \in (0,1).$$

Observe that for this process, if ϵ is small, with high probability the Markov chain stays in its current state. It is easy to check that the arms are jointly φ-mixing. Now consider a policy π, and denote by $\tau_1, \tau_2, \ldots$ the sequence of random times at which π samples the second arm according to the following simple rule. Set $\tau_1 = 1$. For subsequent random times, if $X_{\tau_n,2} = X_{1,2}$ for $n \in \mathbb{N}_+$ then $\tau_{n+1} = \tau_n + 1$. Otherwise, τ_{n+1} is set to be significantly larger than τ_n to guarantee that the distribution of $X_{\tau_{n+1},2}$ given $X_{\tau_n,2}$ is close to the stationary distribution of the Markov chain, during which time the first arm is sampled. The sequence $X_{\tau_1,2}, X_{\tau_2,2}, \ldots$ so generated is highly dependent on $X_{1,2}$, does not have a stationary distribution, and is not φ-mixing. In fact the expectations $\mathbb{E}[X_{\tau_n,2}],\ n \in \mathbb{N}_+$ are very different from the stationary mean $\mathbb{E}[X_{1,2}]$.

4.3 Proposed Method

In this section we describe a UCB-type algorithm to identify the arm with the highest stationary mean in a jointly φ-mixing bandit problem. The main challenge in achieving this objective lies in building confidence intervals around empirical estimates of the stationary means. Indeed, as shown in Example 1, the sampling process may introduce long range dependencies in such a way that the resulting pay-off sequence may neither be stationary nor φ-mixing. This is the reason why a standard UCB algorithm designed for the i.i.d. setting may not be suitable here, even when equipped with a Hoeffding-type concentration bound for φ-mixing processes. Thus, care must be taken when devising a sampling policy in order to allow for Hoeffding-type concentration results (for φ-mixing processes) to be applicable.

```
Input: $\|\varphi\| := \sum_{i=1}^{\infty} \varphi_i$

Initialization: Play each arm once in order

  – $t_i \leftarrow i,\ \overline{X}_i \leftarrow X_{t_i},\ s_i \leftarrow 1,$ for $i = 1..k$

Loop $t = k + 1..\infty$

  – Select arm $j \in 1..k$ that maximises

        $\overline{X}_j + \sqrt{\frac{8\xi(\frac{1}{8} + \ln t)}{2^{s_j}}} + \frac{\|\varphi\|}{2^{s_j - 1}}$

    where $\xi := 1 + 4\|\varphi\|$

  – Update: $t_j \leftarrow t$, $t \leftarrow t + 2^{s_j}$, $\overline{X}_j \leftarrow \frac{1}{2^{s_j}} \sum_{t'=t_j}^{t_j + 2^{s_j} - 1} X_{t'}$, $s_j \leftarrow s_j + 1$
```

Algorithm 1: A UCB-type Algorithm for φ-mixing bandits.

Our approach is based on two key observations. First, consecutive samples $X_{\tau,i}, \dots, X_{\tau+\ell,i}$ where $\tau \in \overline{\mathbb{N}}_+$ is a random starting time at which arm $i \in 1..k$ is sampled in a batch of length ℓ, form a φ-mixing *though not necessarily stationary* sequence. Second, for a long enough batch the average expectations $n^{-1} \sum_{j=0}^{n-1} \mathbb{E}[X_{\tau+j,i}]$ converge to the stationary mean μ_i and the empirical mean of the batch is concentrated around its expectation. We thus propose Algorithm 1, which given an upper-bound on $\|\varphi\|$ works as follows. First, each arm is sampled once for initialisation. Next, from $t = k+1$ on, arms are played in batches of exponentially growing length. Specifically, at each round, arm j with the highest upper-confidence on its empirical mean is selected, and played for 2^{s_j} consecutive time-steps, where s_j denotes the number of times that arm j has been selected so far. The 2^{s_j} samples obtained by playing the selected arm are used in turn to calculate (from scratch) the arm's empirical mean. The upper confidence is calculated based on a Hoeffding-type bound for φ-mixing processes given by Corollary 2.1 of Rio (1999). Theorem 1 gives an upper bound on the algorithm's regret.

Theorem 1 (Regret Bound.) *Let $\mathcal{R}(T) := T\mu^* - \sum_{t=1}^{T} \mathbb{E}[X_{t,\pi_t}]$ be the expected regret of Algorithm 1 compared to the best stationary mean μ^* after T rounds of play. We have,*

$$\mathcal{R}(T) \leq \sum_{\substack{i=1 \\ \mu_i \neq \mu^*}}^{k} \frac{32(1 + 4\|\varphi\|)\ln T}{\Delta_i} + (1 + 2\pi^2/3)(\sum_{i=1}^{k} \Delta_i) + \|\varphi\| \log T$$

where $\Delta_i := \mu^* - \mu_i$.

References

J. Audiffren and L. Ralaivola. Cornering stationary and restless mixing bandits with remix-ucb. In *Advances in Neural Information Processing Systems*, 2015.

P. Doukhan. *Mixing: Properties and Examples*. Springer Lecture Notes, 1994.

S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *J. ACM*, 2010.

R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless markov bandits. *Theoretical Computer Science*, 2014.

C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res.*, 1999.

E Rio. Théorie asymptotique des processus aléatoires faiblement dépendants. 1999.

A. N. Shiryaev. *Probability*. Springer, 1991.

P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 1988.