Neurocomputing 205 (2016) 16–21

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Brief Papers

Multi-armed bandit problem with known trend

Djallel Bouneffouf[a,*], Raphael Féraud[b,*]

[a] Canada's Michael Smith Genome Sciences Centre, University of British Columbia, Vancouver, British Columbia, Canada
[b] Orange Labs, 2 Avenue Pierre Marzin, 22300 Lannion, France

ARTICLE INFO

Article history:
Received 24 August 2015
Accepted 14 February 2016
Available online 24 March 2016

Keywords:
Multi-armed bandit
Online learning
Recommender systems

ABSTRACT

We consider a variant of the multi-armed bandit model, which we call multi-armed bandit problem with known trend, where the gambler knows the shape of the reward function of each arm but not its distribution. This new problem is motivated by different on-line problems like active learning, music and interface recommendation applications, where when an arm is sampled by the model the received reward change according to a known trend. By adapting the standard multi-armed bandit algorithm UCB1 to take advantage of this setting, we propose the new algorithm named Adjusted Upper Confidence Bound (A-UCB) that assumes a stochastic model. We provide upper bounds of the regret which compare favorably with the ones of UCB1. We also confirm that experimentally with different simulations.

1. Introduction

The basic formulation of the Multi-Armed Bandit (MAB) problem can be described as follows: there are K arms, each having a fixed, unknown and independent probability-distribution of reward. At each step, a player chooses an arm and receives a reward. This reward is drawn according to the selected arm's distribution and it is independent of previous actions. Under this assumption, many policies have been proposed to optimize the long-term accumulated reward.

A challenging variant of the MAB problem is the non-stationary bandit problem where the player must decide which arm to play while facing the possibility of a changing environment. We study here a special case of this model where the rewards of each arm of the bandit follow a known function. In this setting, is it possible to adapt standard bandit algorithms to take advantage of this new setting?

The answer of this question is interesting by itself: it could open new doors from a theoretical point of view. But the real motivation is operational: knowing the shape of the reward function assumption is realistic for several real-world problems like on-line active learning, A/B testing and music recommendation. For instance, in [1], the analysis of the active learning problem led the authors to model the active learning problem as a MAB problem. They cluster at first the input space: each cluster is considered as an arm. In this setting the authors find that the more an area is sampled by the model the less is the received reward. In [2], the authors study the recommendation of music where they observe that the interest of a user to a music follows the inverse of an exponential function called forgetting curve, where the more a music is heard the lesser it is interesting. Another model that follows a reward with known function was studied in [3]. The authors observe that when they propose a new interface to a user, at the beginning, the user dislikes it, but after using it several times, the user begin to like it, which means that if we model this problem as a bandit where the interface is an arm, we can say that the reward of the arm start to be bad at the beginning and it increases by time.

From the three above examples, we can say that all these problems can be modeled as new bandit problem called "Multi-armed Bandit Problem with Known Trend" where each arm follow a known trend reward function. For instance, in the first two examples the rewards follow a decreasing function and the third one follows a sigmoid function. In this setting we propose to study this new model derived from this problem, by adapting the existing algorithm to the new setting and analyzing their regret. Finally, we evaluate the proposed algorithms through different simulations.

The remaining of the paper is organized as follows. Section 2 reviews related works. Section 3 describes the setting MAB model with known trend reward function and the proposed algorithm A-UCB. Then we proof its regret in Section 3.2. The experimental evaluation through different simulations is illustrated in Section 4.

[*] Corresponding authors.
E-mail addresses: dbouneffouf@bcgsc.ca (D. Bouneffouf), Raphael.feraud@orange.com (R. Feraud).

http://dx.doi.org/10.1016/j.neucom.2016.02.052
0925-2312/© 2016 Elsevier B.V. All rights reserved.

The last section concludes the paper and points out possible directions for future works.

2. Related work

This section provides an overview on the MAB problem related to our work. In the bandit problem, each arm delivers rewards that are independently drawn from an unknown distribution. An efficient solution based on optimism in the face of uncertainty principle has been proposed by Lai and Robbins [4] who computed an index for each arm and they choose the arm with the highest index.

Our work is an adaptation of these classes of policies for MAB problem with known trend reward function. Our work is mostly related to the study of dynamic versions of the MAB where either the set of arms or their expected reward may change over time. There are several applications, including active learning, music and interface recommendation, where the rewards are far from being stationary random sequences. A solution to cope with non-stationary is to drop the stochastic reward assumption and assume the reward sequences to be chosen by an adversary. Even with this adversarial formulation of the MAB problem, a randomized strategy like EXP3 provides the guarantee of a minimal regret [5,6].

Another work done in [7] considers the situation where the distributions of rewards remain constant over epochs and change at unknown time instants. They analyze two algorithms: the discounted UCB and the sliding-window UCB and they establish for these two algorithms an upper-bound for the expected regret by upper-bounding the expectation of the number of times a suboptimal arm is played. They establish a lower-bound for the regret in the presence of abrupt changes in the arms' reward distributions.

Similar to [7], the authors in [8] propose a Thompson Sampling strategy equipped with a Bayesian change point mechanism to tackle this problem. They develop algorithms for a variety of cases with constant switching rate: when switching occurs all arms change (Global Switching), switching occurs independently for each arm (Per-Arm Switching), when the switching rate is known and when it must be inferred from data.

Motivated by task scheduling, the author in [9] proposed a policy where only the state of the arm currently selected can change in a given step, and proved its optimality for time discounting. This result gave rise to a rich line of work. For example, the authors [10,11] studied the restless bandits, where the states of all arms can change in each step according to an arbitrary stochastic transition function.

To deal with the partial information nature of the bandit problem, in Adapt-Eve [12] the mean reward of the estimated best arm is monitored. The drawback of this approach is that it does not tackle the case of a suboptimal arm becoming the best arm.

In [13] the authors study specific classes of drifting restless bandits selected for their relevance to modelling an online website optimization process. The contribution was a feasible weighted least squares technique capable of utilizing contextual arm parameters while considering the parameter space drifting non-stationary within reasonable bounds.

Another line of work studies the non-stationary reward of arms by considering that each arm has a finite lifetime. In this mortal bandits' setting, each disappearing arm changes the set of available arms. Several algorithms were proposed and analyzed in [14] for mortal bandits under stochastic reward assumptions. In sleeping bandits' problem [15], the set of strategies is fixed but only a subset of them available in each step. In their model they study the mixture-of-experts paradigm, where a set of experts is specified in each time period. The goal of the algorithm is to choose one expert in each time period to minimize regret against the best mixture of experts.

Our new model can be considered an extension of the work done in [14,15], the main difference is in the fact that, in our case the reward function of each arm can follow any function not specially a decreasing function. Our model can also be a specification of the general model of restless bandits with a known shape of the reward function.

3. Problem statement

In this section, we present the algorithm and the main theorem that bounds its regret. Before that, we first provide the setting of our problem. In the MAB setting, to maximize his gain the player has to find the best arm as soon as possible, and then exploit it. In our setting, the rewards follow a known function. When the player has found the best arm, he knows that this arm will be the best just for a certain period of time. The player needs to re-explore at each time to find the next best arm. In the following, we define our setting.

Let $r_i(1), \ldots, r_i(n)$ be a sequence of independent draws of the random variable $r_i \in [0,1]$ with n being the number of trials and let $\mu_i = E[r_i]$ be its mean reward. At each time t, the player chooses an arm $i \in \{1, \ldots, K\}$ to play according to a (deterministic or random) policy ϕ based on sequence of plays and reward, and obtains a non-stationary reward $z(t)$ where $z(t) = r_{i_t}(t) \cdot D(n_{i_t}(t))$, where $D(n_{i_t}(t))$ is a trend reward function assumed to be known, $n_{i_t}(t)$ is the number of times i is played and $r_{i_t}(t)$ is the stationary reward for arm i at time t.

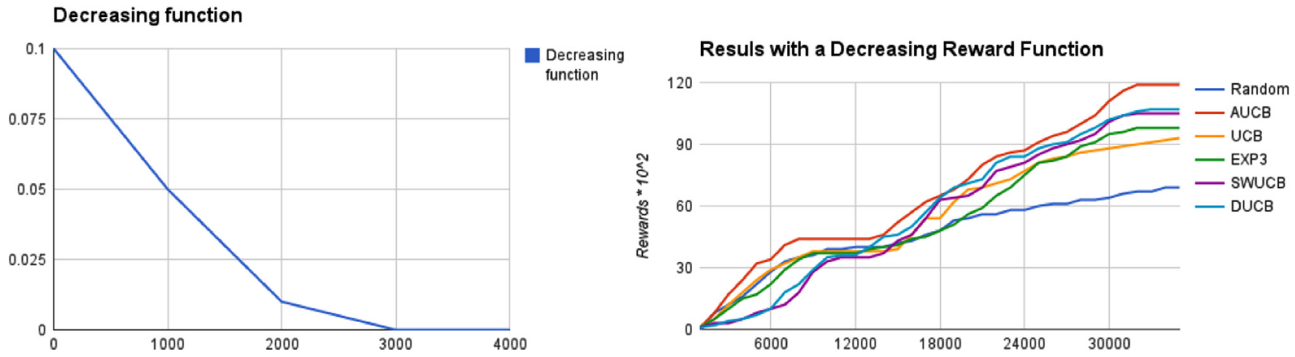

Fig. 1. Decreasing reward function.