Neurocomputing 000 (2018) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

# Possibilistic reward methods for the multi-armed bandit problem☆

Miguel Martín, Antonio Jiménez-Martín*, Alfonso Mateos

*Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Campus de Montegancedo S/N, Boadilla del Monte Madrid 28660, Spain*

ARTICLE INFO

*Article history:*
Received 28 July 2017
Revised 1 December 2017
Accepted 30 April 2018
Available online xxx

Communicated by Dr Kee-Eung Kim

*Keywords:*
Multi-armed bandit problem
Possibilistic reward
Numerical study

ABSTRACT

In this paper, we propose a set of allocation strategies to deal with the multi-armed bandit problem, the *possibilistic reward* (PR) methods. First, we use possibilistic reward distributions to model the uncertainty about the expected rewards from the arm, derived from a set of infinite confidence intervals nested around the expected value. Depending on the inequality used to compute the confidence intervals, there are three possible PR methods with different features. Next, we use a *pignistic probability transformation* to convert these possibilistic functions into probability distributions following the *insufficient reason principle*. Finally, Thompson sampling techniques are used to identify the arm with the higher expected reward and play that arm. A numerical study analyses the performance of the proposed methods with respect to other policies in the literature. Two PR methods perform well in all representative scenarios under consideration, and are the best allocation strategies if truncated poisson or exponential distributions in [0,10] are considered for the arms.



## 1. Introduction

The name *bandit* comes from imagining a gambler playing with $K$ slot machines. The gambler can pull the arm of any of the machines, which produces a reward payoff. The *multi-armed bandit problem* has been at great depth studied in statistics [10], becoming fundamental in different areas of economics, statistics or artificial intelligence [3,22,26,33,35].

A $K$-armed bandit problem can be defined by random variables $X_{i,n}$ for $1 \le i \le K$ and $n \ge 1$, where each $i$ is the index of an arm of a bandit and $n$ refers to the round of play. Successive plays of arm $i$ yield rewards $X_{i,1}, X_{i,2}, \ldots$ which are independent and identically distributed according to an unknown law with unknown expectation $\mu_i$. Other variants of the multi-armed bandit problem (bandits with side information, bandits with no stochastic rewards, bandits with a budgeted cost allocations...) can be found in the literature, see for example [11,28,29,36,37].

A *policy*, or *allocation strategy*, $A$, is an algorithm that chooses the next arm to play based on the sequence of previous plays and obtained rewards.

The goal is to maximize the sum of the rewards received, or equivalently, to minimize the regret, which is defined as the loss compared to the total reward that can be achieved given full knowledge of the problem. The *regret* of $A$ after $n$ plays can be computed as

$$\mu^* n - \sum_{i=1}^{K} \mu_i E[n_i], \quad \text{where} \quad \mu^* = \max_{1 \le i \le K} \{\mu_i\},$$

$E[\cdot]$ denotes expectation and $n_i$ is the number of times arm $i$ has been played by $A$ during the first $n$ plays.

As pointed out in [18], two families of bandit settings can be distinguished. In the first, the distribution of $X_{it}$ is assumed to belong to a family of probability distributions $\{p_\theta,\ \theta \in \Theta_i\}$, whereas in the second, the rewards are only assumed to be bounded (say, between 0 and 1), and policies rely directly on the estimates of the expected rewards for each arm.

Almost all the policies or allocation strategies in the literature focus on the first family and they can be separated, as cited in [24], in two distinct approaches: the frequentist view and the Bayesian approach. In the *frequentist view*, the expected mean rewards corresponding to all arms are considered as unknown deterministic quantities and the goal of the algorithm is to reach the best parameter-dependent performance.

Lai and Robbins [27] first constructed a theoretical framework for determining optimal policies. For specific families of reward distributions, they found that the optimal arm is played exponentially more often than any other arm, at least asymptotically. They also proved that this regret is the best one.

These policies work by associating a quantity called *upper confidence index to each arm*, which relies on the entire sequence of

☆ The paper was supported by the Spanish Ministry of Economy and Competitiveness MTM2014-56949-C3-2-R and MTM2017-86875-C3-3R.

* Corresponding author.
*E-mail addresses:* miguel.martin@alumnos.upm.es (M. Martín), antonio.jimenez@upm.es (A. Jiménez-Martín), alfonso.mateos@upm.es (A. Mateos).

https://doi.org/10.1016/j.neucom.2018.04.078
0925-2312/

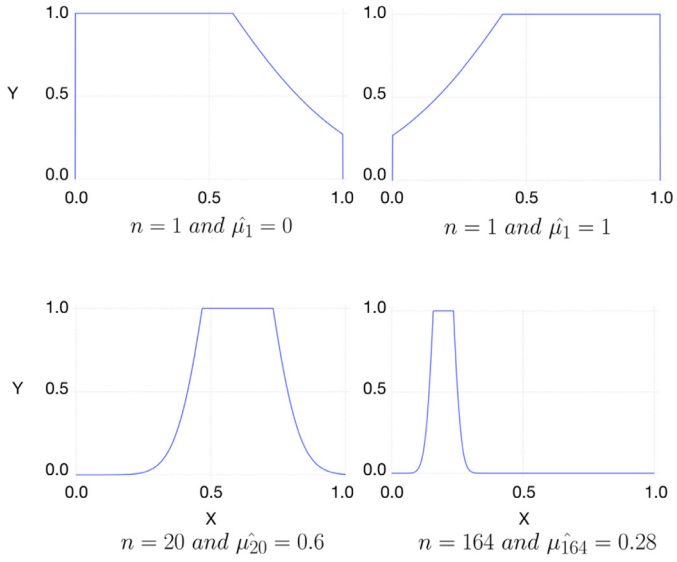


**Fig. 1.** Possibilistic rewards distributions in PR-1.

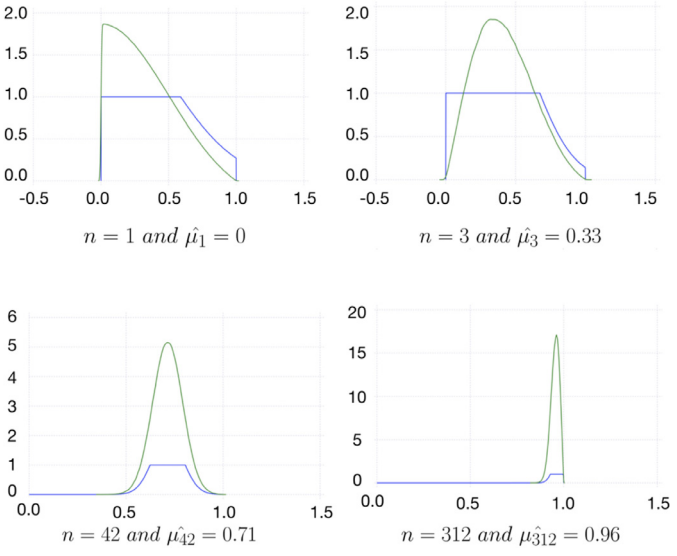


**Fig. 2.** Pignistic probability transformation examples for PR-1.

rewards obtained so far from a given arm. Burnetas and Katehakis [12] proposed an extension to multiparameter or non-parametric models that facilitated the computation of the *upper confidence index*.

Later, Agrawal [1] introduced a generic class of index policies termed *upper confidence bounds* (UCB), where the index can be expressed as simple function of the total reward obtained so far from the arm. These policies are thus much easier to compute than Lai and Robbins', yet their regret retains the optimal logarithmic behavior.

From then, different policies based on UCB can be found in the literature. First, Auer et al. [6] strengthen previous results by showing simple to implement and computationally efficient policies (UCB1, UCB2 and UCB-Tuned) that achieve logarithmic regret uniformly over time, rather than only asymptotically.

Later, Audibert et al. [5] proposed the UCB-V policy, which uses an empirical version of the Bernshtein bound to obtain refined upper confidence bounds. In [7] the UCB method of Auer et al. [6] was modified, leading to the improved-UCB method, whereas an improved UCB1 algorithm, MOSS, was proposed by Audibert & Bubeck [4], which achieved the distribution-free optimal rate while still having a distribution-dependent rate logarithmic in the number of plays.

Another class of policies under the frequentist perspective are the Kullback–Leibler (KL)-based algorithms, including DMED, $K_{inf}$, KL-UCB and kl-UCB.

The *deterministic minimum empirical divergence* (DMED) policy was proposed by Honda & Takemura [23] motivated by a Bayesian viewpoint for the problem (although a Bayesian framework is not used for theoretical analyses).

In [30], the $K_{inf}$-based algorithm was analyzed by Maillard et al. It is inspired by the ones studied in [12,27], taking also into account the full empirical distribution of the observed rewards. Later, the KL-UCB algorithm and its variant KL-UCB+ were introduced by Garivier & Cappé [18]. KL-UCB satisfied a uniformly better regret bound than UCB and its variants for arbitrary bounded rewards, whereas it reached the lower bound of Lai and Robbins when Bernoulli rewards are considered.

New algorithms were proposed by Cappé et al. [13] based on upper confidence bounds of the arm rewards computed using different divergence functions. The kl-UCB uses the Kullback–Leibler divergence; whereas the kl-poisson-UCB and the kl-exp-UCB account for families of poisson and exponential distributions, respectively.

Finally, the BESA algorithm was proposed by Baransi et al. [8]. It is not based on the computation of an empirical confidence bounds, nor can it be classified as a KL-based algorithm. BESA is fully non-parametric.

Stochastic bandit problems have been analyzed from a *Bayesian perspective*, i.e. the parameter is drawn from a prior distribution instead of considering a deterministic unknown quantity. The Bayesian performance is then defined as the average performance over all possible problem instances weighted by the prior on the parameters.

The origin of this perspective is in the work by Gittins [19,20]. Gittins' index based policies are a family of Bayesian-optimal policies based on indices that fully characterize each arm given the current history of the game, and at each time step the arm with the highest index will be pulled. In [25], Gittins' indices for the arms a related to ladder variables for associated random walks.

Another family of algorithms to solve bandit problems is the so-called *Thompson sampling* (TS), consisting of randomly drawing each arm according to its probability of being optimal. The algorithm assumes that the arms' distributions belong to a parametric family of distributions $P = \{p(.|\theta), \theta \in \Theta\}$ where $\Theta \subseteq R$, it starts by putting a prior distribution on each one of the arms parameters, and at each time step a posterior distribution is maintained according to the rewards observed so far.

Finally, Bayes-UCB was proposed by Kaufmann et al. [24] inspired by the Bayesian interpretation of the problem but retaining the simplicity of UCB-like algorithms.

Table 1 shows the main features of the allocation strategies mentioned throughout this section. *Regret bound* refers to whether or not there is a theoretical analysis proving a regret bound, *Optimality* points out if there is a reward distribution whose performance is optimal or near optimal, *Parametric* refers to whether the reward distribution family (Bernoulli, exponential, Gaussian...) or only the upper and lower bound values have to be specified, *Delayed* denotes whether or not there are experiments testing strategy performance for delayed reward in the literature, and *Complexity* refers to the computational resources needed to compute the next action.

In this paper, we propose *possibilistic reward* (PR) methods. PR methods combine the best of upper confidence index policies, where the only available information about the reward distributions is that they are bounded, and the best of Thompson

Please cite this article as: M. Martín et al., Possibilistic reward methods for the multi-armed bandit problem, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.04.078