BioSystems 135 (2015) 55–65

Contents lists available at ScienceDirect

BioSystems

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

Overtaking method based on sand-sifter mechanism: Why do optimistic value functions find optimal solutions in multi-armed bandit problems?

Kento Ochi [a], Moto Kamiura [a,b,c,*]

[a] Graduate School of Science and Engineering, Tokyo Denki University, Japan
[b] School of Science and Engineering, Tokyo Denki University, Japan
[c] Research Institute of Electrical Communication, Tohoku University, Japan

ARTICLE INFO

Article history:
Received 6 June 2015
Received in revised form 25 June 2015
Accepted 26 June 2015
Available online 10 July 2015

Keywords:
Exploration–exploitation dilemma
Multi-armed bandit problem
Confidence interval
UCB algorithm
Optimism

ABSTRACT

A multi-armed bandit problem is a search problem on which a learning agent must select the optimal arm among multiple slot machines generating random rewards. UCB algorithm is one of the most popular methods to solve multi-armed bandit problems. It achieves logarithmic regret performance by coordinating balance between exploration and exploitation. Since UCB algorithms, researchers have empirically known that *optimistic* value functions exhibit good performance in multi-armed bandit problems. The terms *optimistic* or *optimism* might suggest that the value function is sufficiently larger than the sample mean of rewards. The first definition of UCB algorithm is focused on the optimization of regret, and it is not directly based on the optimism of a value function. We need to think the reason why the optimism derives good performance in multi-armed bandit problems. In the present article, we propose a new method, which is called *Overtaking method*, to solve multi-armed bandit problems. The value function of the proposed method is defined as an upper bound of a confidence interval with respect to an estimator of expected value of reward: the value function asymptotically approaches to the expected value of reward from the upper bound. If the value function is larger than the expected value under the asymptote, then the learning agent is almost sure to be able to obtain the optimal arm. This structure is called *sand-sifter mechanism*, which has no regrowth of value function of suboptimal arms. It means that the learning agent can play only the current best arm in each time step. Consequently the proposed method achieves high accuracy rate and low regret and some value functions of it can outperform UCB algorithms. This study suggests the advantage of optimism of agents in uncertain environment by one of the simplest frameworks.

1. Introduction

Reinforcement learning is a type of machine learning which is based on maximization of total rewards (Sutton and Barto, 1998). A reinforcement learning agent adapts to environment through trial and error based on a policy of the agent. It can know only rewards which are implicit information for the agent on the environment, although a supervised learning agent is given true/false as supervisory signals which are explicit instruction for agent behavior. Such uncertainty of the environment derives the exploration–exploitation dilemma which is a decision tradeoff between searching for a better action (i.e. exploration) and taking a temporally selected action as the current optimal solution (i.e. exploitation). Most of the reinforcement learning methods attempt to provide an optimal balance between exploration and exploitation, to achieve quick and accurate learning.

A multi-armed bandit problem, the importance of which was realized by Robbins (1952), is one of simple tasks on reinforcement learning. It is illustrated as a search problem on which an agent must select the best arm among multiple slot machines generating random rewards. Selecting an arm and playing it imply an action in reinforcement learning. The exploration and exploitation in the picture of slot machines corresponds to searching for the best arm and playing the current best arm. Multi-armed bandit problems are used in a wide range of applications: e.g. as formulation of pay-per-click auctions for internet advertising (Babaioff et al., 2009; Devanur and Kakade, 2009) or of Go game situations (Gelly and

* Corresponding author at: School of Science and Engineering, Tokyo Denki University, Ishizaka, Hatoyama, Hiki, Saitama, Japan. Tel.: +81 80 1236 2612.
E-mail addresses: kamiura@mail.dendai.ac.jp, moto@goo.jp (M. Kamiura).

http://dx.doi.org/10.1016/j.biosystems.2015.06.009

Wang, 2006; Gelly et al., 2006), as a task for cognitively inspired heuristics (Oyo and Takahashi, 2013) or for models of human cognition (Shinohara et al., 2007; Takahashi et al., 2010), etc.

The seminal paper of Lai and Robbins (1985) proposed using minimization of the regret instead of maximization of the total rewards in a multi-armed bandit problem. Besides they show that the lower bound of the regret grows at logarithmic order of n when the regret grows at polynomial order of n for each arm, where n is the total number of plays (i.e. THEOREM1 in Lai and Robbins (1985)). UCB algorithm which was proposed by Auer et al. (2002) is one of the most popular methods to solve multi-armed bandit problems. Auer et al. define a value function of UCB algorithm by the sum of a sample mean of rewards and an additional term which depends on $\sqrt{\ln n}$. They also propose some variants of UCB, the differences of that appear in their additional terms which are based on distributions of rewards (Auer et al., 2002). UCB algorithm is extended and is used in many applied problems (Gelly and Wang, 2006; Gelly et al., 2006; Sturtevant, 2008).

One of the most important features of UCB algorithm is asymptotical achievability of logarithmic regret. UCB algorithm achieves the order of the lower bound of the regret growth shown by Lai and Robbins (1985). Inheriting the feature of Lai and Robbins (1985) and Auer et al. (2002), some researchers advanced mathematically rigorous analyses of UCB algorithms (Audibert et al., 2009; Bubeck et al., 2011; Bubeck and Cesa-Bianchi, 2012; Salomon and Audibert, 2014).

These solutions based on regret analyses might have the following blind spot as the approaches to multi-armed bandit problems. The results provided by the UCB algorithms seem to be robust since they underlay the THEOREM1 in Lai and Robbins (1985). If we attempt to provide a solver which can outperform UCB algorithms in multi-armed bandit problems, then we might have to pay attention to the assumption of the THEOREM1 in Lai and Robbins (1985) which says that the regret grows at polynomial order of n for each slot machine. This assumption implies that suboptimal solutions (i.e. selecting ineffectual machines) are never ignored. UCB algorithm achieves the logarithmic regret performance through balance between exploration and exploitation. It does not forsake suboptimal solutions and does not stop exploring.

Since UCB algorithms, researchers have empirically known that *optimistic* value functions exhibit good performance in multi-armed bandit problems. The terms *optimistic* or *optimism* might suggest that the value function is sufficiently larger than the sample mean of rewards although they were primarily non-technical terms. The first definition of UCB algorithm is focused on the optimization of regret, as the above, and it is not directly based on the optimism of a value function, although UCB algorithm implicitly includes the aspect of the optimistic value function. We need to think the reason why the optimism derives good performance in multi-armed bandit problems.

In the present study, we propose an alternative methodology for a multi-armed bandit problem. We call the new multi-armed bandit algorithm *Overtaking method*. The learning agent of Overtaking method can eventually find an optimal arm with a given probability which we call *guarantee probability*. Overtaking method is based on a mechanism to discard losing arms and to prune worse choices for a learning agent: i.e. this method is not based on the THEOREM1 in Lai and Robbins (1985). This structure is called *sand-sifter mechanism*, in which regrowth of value function of suboptimal arms does not arise. The learning agent can play only the current best arm in each time step. Some of the proposed methods can averagely outperform UCB algorithms in terms of regret and accuracy rate. In the previous paper (Ochi and Kamiura, 2013), we have already shown brief ideas of Overtaking method and numerical experiments for some problems on normally-distributed rewards. In the present paper, we show the detail of Overtaking method and test the performance of this method numerically. We consider two cases for multi-armed bandit problem, differing in the range of reward value; one is *normal case* which implies a task with normally-distributed rewards, the other is *binary case* which implies a task with Bernoulli-distributed rewards. In Section 2, we explain the outline and the general formula of Overtaking method. In addition, we define a class of multi-armed bandit problems and describe a main theorem expressing sand-sifter mechanism which provides a basis for Overtaking method. In Section 3, we define the concrete value functions of Overtaking method, to use the numerical experiments which consist of the normal cases and the binary cases. In Section 4, the configurations of the numerical experiments are given. The results are shown with analyses of them. Finally the conclusions are given.

Biologically, the optimism bias is known as a cognitive bias in higher brain functions (Fox, 2012; Shepperd et al., 2002). Our study suggests that "optimism" has important effects not only in a higher brain function but also in a multi-armed bandit agent as a primitive learning mechanism. In multi-armed bandit problems, the learning agent must select an arm with maximum expected reward as soon as possible, to maximize the total reward which the agent obtains. The expected reward estimated optimistically as the value function enhances the compatibility between exploration and exploitation: i.e. optimism of a multi-armed bandit agent has substance as a statistical mechanism. Optimism might have importance in multiple layer of cognition and learning.

2. Overtaking method

2.1. Outline of method

The theorem described in the following section implies that the learning agent can find the optimal arm if it makes the value function approach from the upper side of the expected value. A value function for Overtaking method consists of the following two terms to hold the above features: the first term is a sample mean of rewards and the second term is a buffering factor to keep the value function on the upper side of the expected value. It means that the value function is optimistic.

The Convergence Condition in the theorem requires that the value function almost surely converge to the expected value. Based on the condition, the value function is configured as an estimator of the expected value of the rewards. According to the strong law of large numbers, the first term of the value function which is a sample mean of iid random variables almost surely converges to the expected value of rewards. We need to configure the second term which almost surely converges to zero.

Moreover, with respect to the Overtaking Condition in the theorem, the value function of the optimal arm a^* has to be on the upper side of the expected value of the arm a^*. The learning agent does not know which arm is the optimal, thus we make every value function be on the upper side of each expected value in our method. We say "a value function overtakes an expected value" for such a situation which probabilistically occurs.

The Overtaking Condition is clearly distinguished from the optimism of the value function. The former implies that the value function is probabilistically larger than the expected value of rewards, and the latter implies that the value function is larger than the sample mean of rewards.

Probability on a value function Overtaking an expected value is called *guarantee probability*. We, the user of Overtaking method, can quantitatively control the guarantee probability under the above conditions. The guarantee probability is connected to probability to be eventually able to select an optimal solution.

Readers grasping outline of features of the proposed method, we suggest a metaphors for it as a tower of piled sand-sifters: i.e.