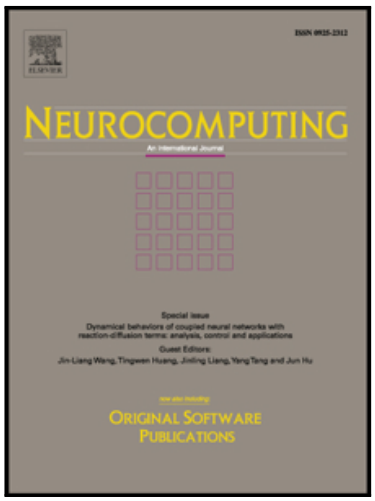

Finite Budget Analysis of Multi-armed Bandit Problems

Yingce Xia, Tao Qin, Wenkui Ding, Haifang Li, Xudong Zhang, Nenghai Yu, Tie-Yan Liu

PII: S0925-2312(17)30421-6
DOI: 10.1016/j.neucom.2016.12.079
Reference: NEUCOM 18171

To appear in: *Neurocomputing*

Received date: 25 May 2016
Revised date: 4 September 2016
Accepted date: 30 December 2016

Please cite this article as: Yingce Xia, Tao Qin, Wenkui Ding, Haifang Li, Xudong Zhang, Nenghai Yu, Tie-Yan Liu, Finite Budget Analysis of Multi-armed Bandit Problems, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2016.12.079

Finite Budget Analysis of Multi-armed Bandit Problems

Yingce Xia[1], Tao Qin[2], Wenkui Ding[3], Haifang Li[4], Xudong Zhang[5], Nenghai Yu[1], Tie-Yan Liu[2]

[1]*University of Science and Technology of China* [2]*Microsoft Research* [3]*Hulu LLC.* [4]*Chinese Academy of Sciences* [5]*Tsinghua University*
[1]*yingce.xia@gmail.com;*[1]*ynh@ustc.edu.cn;*[2]*{taoqin,tie-yan.liu}@microsoft.com;*[3]*wenkui.ding@hulu.com;*[4]*haifang.li@ia.ac.cn;*[5]*zhangxd@tsinghua.edu.cn*

Abstract

In the budgeted multi-armed bandit (MAB) problem, a player receives a random reward and needs to pay a cost after pulling an arm, and he cannot pull any more arm after running out of budget. In this paper, we give an extensive study of the upper confidence bound based algorithms and a greedy algorithm for budgeted MABs. We perform theoretical analysis on the proposed algorithms, and show that they all enjoy sublinear regret bounds with respect to the budget B. Furthermore, by carefully choosing the parameters, they can even achieve log linear regret bounds. We also prove that the asymptotic lower bound for budgeted Bernoulli bandits is $\Omega(\ln B)$. Our proof technique can be used to improve the theoretical results for fractional KUBE [26] and Budgeted Thompson Sampling [30].

Keywords: Budgeted Multi-Armed Bandits, UCB Algorithms, Regret Analysis

1. Introduction

Multi-armed bandits (MAB) correspond to a typical sequential decision problem, in which a player receives a random reward by playing one of K arms from a slot machine at each round and wants to maximize his cumulated reward. Many real world applications can be modeled as MAB problems, such as auction mechanism design [24], search advertising [27], UGC mechanism design [17] and personalized recommendation [22]. Many algorithms have been designed for MAB problems and studied from both theoretical and empirical perspectives, like UCB1, ϵ_n-GREEDY [6], UCB-V [4], LinRel [5], DMED [18], and KL-UCB [16]. A survey on MAB can be found in [11].

Most of the aforementioned works assume that playing an arm is costless, however, in many real applications including the real-time bidding problem in ad exchange [12], the bid optimization problem in sponsored search [10], the on-spot instance bidding problem in Amazon EC2 [9], and the cloud service provider selection problem in IaaS [2], one needs to pay some cost to play an arm and the number of plays is constrained by a budget. To model these applications, a new kind of MAB problems, called budgeted MAB, have been proposed and studied in recent years, in which the play of an arm is associated with both a random reward and a cost. According to different settings of budgeted MAB, the cost could be either deterministic or random, either discrete or continuous.

In the literature, a few algorithms have been developed for some particular settings of the budgeted MAB problem. The setting of deterministic cost was studied in [26], and two algorithms named KUBE and fractional KUBE were proposed, which learn the probability of pulling an arm by solving an integer programming problem. It has been proven that these two algorithms can lead to a regret bound of $O(\ln B)$. The setting of random discrete cost was studied in [13], and two upper confidence bound (UCB) based algorithms with specifically designed (complex) indexes were proposed and the log linear regret bounds were derived for the two algorithms.

The above algorithms for the budgeted MAB problem could only address some (but not all) settings of budgeted MAB. Given the tight connection between budgeted MAB and standard MAB problems, an interesting question to ask is whether some extensions of the algorithms originally designed for standard MAB (without budget constraint) could be good enough to fulfill the budgeted MAB tasks, and perhaps in a more general way. [30] shows that a simple extension of the Thompson sampling algorithm, which is designed for standard MAB, works quite well for budgeted MAB with very general settings. Inspired by that work, we are interested in whether we can handle budgeted MAB by extending other algorithms designed for standard MAB.

In order to answer the question, we study the following natural extensions of the UCB1 and ϵ_n-GREEDY algorithms [6] in this paper (these extensions do not need to know the budget B in advance). We first propose four basic algorithms for budgeted MABs: (1) i-UCB, which replaces the average reward of an arm in the exploitation term of UCB1 by the average reward-to-cost ratio; (2) c-UCB, which further incorporates the average cost of an arm into the exploration term of UCB1 (c-UCB can be regarded as an adaptive version of i-UCB); (3) m-UCB, which mixes the upper confidence bound of reward and the lower confidence bound of cost; (4) b-GREEDY, which replaces the average reward in ϵ_n-GREEDY with the average reward-to-cost ratio.

We conduct theoretical analysis on these algorithms, and show that they all enjoy sublinear regret bounds with respect to B. By carefully setting the hyperparameter in each algorithm, we show that the regret bounds can be further improved to be log linear. Although the basic idea of regret analysis for the