Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-13 (2016) 241–246

Adaptive Normal Two-Armed Bandit and Data Processing Optimization [★]

Alexander V. Kolnogorov [*]

[*] *Yaroslav-the-Wise Novgorod State University, Velikiy Novgorod, 173003 Russia, (e-mail: Alexander.Kolnogorov@novsu.ru).*

Abstract: We consider the two-armed bandit problem as applied to data processing provided that there are two alternative processing methods with different a priori unknown efficiencies. One should determine more efficient method and ensure its preferable application. Normal two-armed bandit is a generalization which allows to process data in parallel and almost without loss of the control performance, i.e. without increasing of the minimax risk. However, it requires that methods must have close efficiencies. Below we propose the adaptive modification of the algorithm which works properly with methods which efficiencies are not obligatory close.

Keywords: Two-armed bandit problem, control in a random environment, minimax and Bayesian approaches, parallel processing, an asymptotic minimax theorem.

1. INTRODUCTION

We consider the two-armed bandit problem (see, e.g. Berry and Fristedt (1985), Presman and Sonin (1990)) which is also well-known as the problem of expedient behavior in a random environment (see, e.g. Tsetlin (1973), Varshavsky (1973)) and the problem of adaptive control (see, e.g. Nazin and Poznyak (1986), Sragovich (2006)) in the following setting. Let ξ_n, $n = 1, \dots, N$ be a controlled random process which values are interpreted as incomes, depend only on a currently chosen actions y_n ($y_n \in \{1, 2\}$) and are normally distributed with probability densities

$$f(x|m_\ell) = (2\pi)^{-1/2} \exp\left\{-(x - m_\ell)^2/2\right\}$$

if $y_n = \ell$ ($\ell = 1, 2$). So, this is Normal (or Gaussian) two-armed bandit. It can be described by a vector parameter $\theta = (m_1, m_2)$. Control strategy σ at a point of time n assigns a random choice of the action y_n depending on the current history of the process, i.e. replies $x^{n-1} = x_1, \dots, x_{n-1}$ to applied actions $y^{n-1} = y_1, \dots, y_{n-1}$:

$$\Pr(y_n = \ell|y^{n-1}, x^{n-1}) = \sigma_\ell(y^{n-1}, x^{n-1}),$$

$\ell = 1, 2$. The set of strategies is denoted by Σ.

The goal is to maximize (in some sense) the total expected income. So, if parameter θ is known then the optimal strategy should always choose the action corresponding to the larger value of m_1, m_2. The total expected income would thus be equal to $N(m_1 \vee m_2)$ where $\vee$ stands for maximum. If parameter is unknown then the loss function

$$L_N(\sigma, \theta) = N(m_1 \vee m_2) - E_{\sigma,\theta}\left(\sum_{n=1}^{N} \xi_n\right)$$

describes expected losses of total income with respect to its maximal possible value due to incomplete information. Here $E_{\sigma,\theta}$ denotes the mathematical expectation calculated with respect to the measure generated by strategy σ and parameter θ. The set of parameters is assumed to be the following

$$\Theta = \{\theta : |m_1 - m_2| \le 2C\},$$

where $0 < C < \infty$. Restriction $C < \infty$ ensures the boundedness of the loss function on Θ.

According to the minimax approach the maximal value of the loss function on the set of parameters Θ should be minimized on the set of strategies Σ. The value

$$R_N^M(\Theta) = \inf_{\Sigma} \sup_{\Theta} L_N(\sigma, \theta) \qquad (1)$$

is called the minimax risk and corresponding strategy σ^M (if it exists) is called the minimax strategy. Note that if strategy σ^M is applied then inequality

$$L_N(\sigma^M, \theta) \le R_N^M(\Theta)$$

holds for all $\theta \in \Theta$ and this means robustness of the control.

The minimax approach to the problem was proposed by H. Robbins in Robbins (1952). This article caused a significant interest to considered problem. The classic object of the most of arisen articles was Bernoulli two-armed bandit which can be described by distribution

$$\Pr(\xi_n = 1|y_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0|y_n = \ell) = q_\ell,$$

where $p_\ell + q_\ell = 1$, $\ell = 1, 2$. Such two-armed bandit is described by a parameter $\theta = (p_1, p_2)$ with the set of possible values $\Theta = \{\theta : 0 \le p_\ell \le 1; \ell = 1, 2\}$. It was shown in Fabius and van Zwet (1970) that explicit determination of the minimax strategy and minimax risk is practically impossible already for $N > 4$. However, the asymptotic minimax theorem was proved by W. Vogel in Vogel (1960) by indirect techniques. It states that minimax risk has the order $N^{1/2}$ as $N \to \infty$. In more details it is presented in section 2.

Remark 1. There are some different approaches to robust control in the two-armed and multi-armed bandit problems, see, e.g. Nazin and Poznyak (1986); Lugosi and

[★] This work was supported in part by the Project Part of the State Assignment in the Field of Scientific Activity by the Ministry of Education and Science of the Russian Federation, project no. 1.949.2014/K.

2405-8963 © 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.07.958

Cesa-Bianchi (2006); Juditsky et al (2008); Gasnikov et al (2015). In these articles stochastic approximation method, mirror descent algorithm and some other techniques are used for the control. Instead of minimax risk, the authors often consider the equivalent attitude called the guaranteed rate of convergency. The order of the minimax risk for these algorithms is $N^{1/2}$ or close to $N^{1/2}$.

Let's explain the choice of the normal distribution of incomes. We consider the problem as applied to control of processing a large amount of data in a comparatively small number of stages by grouping them and then processing in parallel. Let $T = NM$ data be given that can be processed using either of the two alternative methods. Processing can be successful ($\zeta_t = 1$) or unsuccessful ($\zeta_t = 0$). The probabilities of successful and unsuccessful processing depend only on the selected methods (actions), that is, $\Pr(\zeta_t = 1|y_t = \ell) = p_\ell$ and $\Pr(\zeta_t = 0|y_t = \ell) = q_\ell$, $\ell = 1, 2$. Let p_1 and p_2 be known to be close to p $(0 < p < 1)$. We partition the data into N packets of M data in each packet and use the same method for parallel data processing in the same packet. For the control, we use the values of the process $\xi_n = (DM)^{-1/2} \sum_{t=(n-1)M+1}^{nM} \zeta_t$, $n = 1, \dots, N$, where $D = p(1-p)$. According to the central limit theorem, distributions of ξ_n, $n = 1, \dots, N$ are close to normal, and their variances are close to unity as in considered setting.

Certainly, there is a question of losses in the control performance as the result of such aggregation. It was shown in Kolnogorov (2012) that $N^{-1/2} R_N^M(\Theta_N)$ remains almost unchanged already for $N \ge 30$. Therefore, say 30,000 items of data can be processed in 30 steps by packets of 1,000 data with almost the same maximal losses as if the data were processed optimally one-by-one.

Remark 2. Parallel control for the two-armed bandit problem was first suggested for the problem of treating a large group of patients by either of the two alternative drugs with different unknown efficiencies. Clearly, the doctor cannot treat the patients sequentially one by one. Say, if the result of the treatment will be manifest in a week and there is a thousand of patients, then one-by-one treatment would take almost twenty years. Therefore, it was proposed to give both drugs to sufficiently large groups of patients, and then the more effective one to give to the rest of them. As the result, the entire treatment will take two weeks. The discussion and bibliography of the problem can be found, for example, in Lai et al (1980)

Another well-known approach to the problem is a Bayesian one. Denote by λ a prior distribution density of the parameter θ on the set Θ. The value

$$R_N^B(\lambda) = \inf_{\Sigma} \int_{\Theta} L_N(\sigma, \theta)\lambda(\theta)d\theta \tag{2}$$

is called the Bayesian risk and corresponding strategy σ^B is called the Bayesian strategy. Bayesian approach is very popular because it allows to write recursive equation for determination of both Bayesian strategy and Bayesian risk by a dynamic programming technique. An adaptive nature of Bayesian formalism was recognized by many researchers. For example, Berry and Fristedt write in Berry and Fristedt (1985): "It is not that researchers in bandit problems tend to Bayesians; rather Bayes's theorem provides a convenient mathematical formalism that allows for *adaptive learning* and so is an ideal tool in sequential decision problems".

Both minimax and Bayesian approaches are integrated by the main theorem of the theory of games. According to this theorem the minimax risk (1) is equal to the Bayesian risk (2) calculated over the worst-case prior distribution corresponding to the maximum of the Bayesian risk, i.e.

$$R_N^M(\Theta) = R_N^B(\lambda_0) = \sup_{\lambda} R_N^B(\lambda). \tag{3}$$

And the minimax strategy is equal to corresponding Bayesian strategy as well.

Below we use the main theorem of the theory of games for finding minimax risk and minimax strategy. We propose a recursive equation which allows to determine Bayesian risk and Bayesian strategy for the parallel control of packets of data. However, this method works well only for close mathematical expectations m_1, m_2, for which expected losses have the order $N^{1/2}$. For distant expectations, such that maximal value of m_1, m_2 can be quickly detected, expected losses have the order $\log(N)$ and another strategy should be used for the control. For expectations, which are not obligatory close, we propose adaptive modification of the strategy which checks the closeness of m_1, m_2 at the initial stage of control and then applies the proper strategy at the final stage.

The structure of the paper is the following. In section 2 we discuss the paradoxical situation with maximal expected losses for close and distant expectations m_1, m_2. Namely, it may be surprisingly that maximal expected losses are attained for close expectations where no control at all seems to be necessary. A solution to the paradox is grouping of data. In section 3 we describe the approach based on the main theorem of the theory of games. This approach allows to determine minimax strategy and minimax risk for the case of close expectations. In section 4 we propose an adaptive algorithm which works properly for all expectations by detecting close and distant those ones at the initial stage of the control. Section 5 contains conclusion.

2. MAXIMAL EXPECTED LOSSES

2.1 An Asymptotic Minimax Theorem

Maximal expected losses are described by the asymptotic minimax theorem which was proved in Vogel (1960).

Theorem 1. The following estimates hold as $N \to \infty$ for Bernoulli two-armed bandit:

$$0.612 \le (DN)^{-1/2} R_N^M(\Theta) \le 0.752 \tag{4}$$

with $D = 0.25$ being the maximal variance of one-step income. The lower estimate, presented here, was obtained in Bather (1983). The upper estimate was obtained in Vogel (1960) for the following strategy.

Thresholding strategy. *Use actions turn-by-turn until the absolute difference of total incomes for their applications exceeds the value of the threshold $\alpha(DN)^{1/2}$ or the control time expires. If the threshold has been achieved and the control time has not expired then at the rest of the control horizon use only the action corresponding to the larger value of total initial income.*