Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-13 (2016) 247–252

A Generalization of Robust Normal Two-Armed Bandit ⋆

Alexander V. Kolnogorov *

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

Abstract: We consider Normal two-armed bandit problem with a priori known variances and unknown mathematical expectations of incomes in robust (minimax) setting. This setup naturally arises in group control of data processing. We show that one can solve the problem using the main theorem of the theory of games, i.e. determine minimax strategy and minimax risk as Bayesian corresponding to the worst-case prior distribution. We obtain recursive invariant Bellman-type equation for calculation appropriate Bayesian risk and Bayesian strategy. The requirement of a priori known variances of incomes may be omitted because they may be estimated at the initial stage of control.

Keywords: Two-armed bandit problem, control in a random environment, minimax and Bayesian approaches, group 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 choice of alternatives (see, e.g. Sragovich (2006), Nazin and Poznyak (1986)) 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 currently chosen actions y_n and are normally distributed with probability densities $f_{D_\ell}(x|m_\ell)$ if $y_n = \ell$ $(\ell = 1, 2)$ where

$$f_D(x|m) = (2\pi D)^{-1/2} \exp\left\{-(x-m)^2/(2D)\right\},$$

We assume that D_1, D_2 are a priori known variances and m_1, m_2 are a priori unknown mathematical expectations. Such two-armed bandit can be described by a vector parameter $\theta = (m_1, m_2)$. The goal is to maximize (in some sense) the total expected income. Control strategy σ at the point of time $n = n_1 + n_2$ is a function of the current statistics (X_1, n_1, X_2, n_2), where n_1, n_2 are current total numbers of both actions' applications, X_1, X_2 are corresponding total incomes. Thus

$$\sigma_\ell(X_1, n_1, X_2, n_2) = \Pr(y_n = \ell | X_1, n_1, X_2, n_2),$$

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

If parameter θ is known then the optimal strategy should always apply the action corresponding to the larger value of m_1, m_2. The total expected income would thus be equal to $N(m_1 \vee m_2)$. If parameter is unknown then the function

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

describes expected losses of total income due to the 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 total expected losses on the set of parameters Θ should be minimized on the set of strategies Σ. The value

$$R_N^M(\Theta) = \inf_{\Sigma} \sup_{\Theta} L_N(\sigma, \theta) \qquad (2)$$

is called the minimax risk and corresponding strategy is called the minimax strategy. The minimax approach to the problem was proposed 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

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

$\ell = 1, 2$. Such bandit is described by parameter $\theta = (p_1, p_2)$ with the set of 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 in Vogel (1960) which states that minimax risk has the order $N^{1/2}$ as $N \to \infty$ and provides the estimates of the factor. This theorem holds true for the Normal two-armed bandit as well.

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 Cesa-Bianchi (2006); Juditsky et al (2008); Gasnikov et al (2015). In these articles stochastic approximation method and mirror descent algorithm are used for the control. The

⋆ 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.

10.1016/j.ifacol.2016.07.959

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 group control of processing a large amount of data. Let $T = NM$ data be given that can be processed using either of the two alternative methods. The result of processing of the t-th item of data is ζ_t. For example, processing may be successful ($\zeta_t = 1$) or unsuccessful ($\zeta_t = 0$). Or $\{\zeta_t\}$ may be equal to durations of processing; in this case the goal is to minimize the total expected duration. Distributions of $\{\zeta_t\}$ depend only on the selected methods (actions). We partition the data into N packets of M data in each packet and use the same method for data processing in the same packet. For the control, we use the values of the process $\xi_n = Z^{-1}\sum_{t=(n-1)M+1}^{nM}\zeta_t$, $n = 1,\dots,N$ with Z being some normalizing factor. According to the central limit theorem, distributions of ξ_n, $n = 1,\dots,N$ are close to normal as in considered setting. If $\{\zeta_t\}$ count successfully processed data then the data in the same packet may be processed in parallel. And if $\{\zeta_t\}$ are durations of processing then the same method may be applied to successively incoming data of the packet (the duration of parallel data processing is equal to the longest duration but not to their sum).

It is important that parallel control almost does not increase the value of the minimax risk if the number of packets is large enough (see, Kolnogorov (2012, 2014)). In more details it is discussed in section 2.

Remark 2. Note that parallel control for the two-armed and multi-armed bandit problems was first suggested for the problem of treating a large group of patients by either of the two drugs with different unknown efficiencies. 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 on the set Θ. The value

$$R_N^B(\lambda) = \inf_{\Sigma}\int_{\Theta} L_N(\sigma,\theta)\lambda(\theta)d\theta \qquad (3)$$

is called the Bayesian risk and corresponding strategy is called the Bayesian strategy. Bayesian approach is very popular because it allows to write recursive Bellman-type equation for determination of both Bayesian strategy and Bayesian risk by a dynamic programming technique. On the other hand, it was often criticized (see, e.g. Presman and Sonin (1990), Berry and Fristedt (1985)) for the lack of clear criteria how to choose an appropriate prior distribution. Minimax and Bayesian approaches are integrated by the main theorem of the theory of games. According to this theorem the minimax risk (2) is equal to the Bayesian risk (3) calculated over the worst-case prior distribution corresponding to the maximum of the Bayesian risk. And the minimax strategy is equal to corresponding Bayesian strategy as well.

Determination of the minimax strategy and minimax risk as Bayesian ones corresponding to the worst-case prior distribution is considered below. The structure of the paper is the following. In section 2 we explain the choice of normal distributions of incomes and show that the requirement of a priori known variances may be omitted. In section 3 we specify asymptotically the worst-case prior distribution. In section 4 we provide recursive equations for determination of the Bayesian strategy, Bayesian risk and expected losses over this prior distribution. In section 5 numerical results are presented.

2. MOTIVATION

In this section we discuss how normal distributions of incomes with equal or different variances may occur in considered problem. We also show that the requirement of a priori known variances may be omitted.

2.1 Parallel Processing of Data. Close and Distant Expectations

Let's recall parallel processing considered in section 1. Of course, it is more convenient to control the aggregated process $\{\xi_n\}$ than the original process $\{\zeta_t\}$. However, the following question naturally arises. How much are the losses of the control quality due to such aggregation? The answer to this question depends on how close are mathematical expectations m_1, m_2.

First, let's make the following remark. The maximal expected losses in the two-armed bandit problem have the order $N^{1/2}$ and are attained for close expectations $|m_1 - m_2| \le cN^{-1/2}$ with $c > 0$ large enough. For distant expectations $|m_1 - m_2| \ge \delta > 0$, the maximal expected losses have the order $\log(N)$. These estimates follow from the results of Vogel (1960) and Lai et al (1980).

Let's give a short explanation of this result. If $|m_1 - m_2| \le cN^{-1/2}$ then probability of the error to determine the largest value of m_1, m_2 always is not less than some $p_e \ge \alpha > 0$. Therefore, maximal total expected losses are not less than $cN^{-1/2}p_eN \ge c\alpha N^{1/2}$. For distant expectations $|m_1 - m_2| > \delta > 0$ the largest value of m_1, m_2 may be confidently determined on the time horizon of the order $\log(N)$. This specifies the order of expected losses in this case.

Second, in case of close expectations grouping of data as it is described in section 1 almost does not affect the maximal expected losses if the number of groups is large enough, e.g. if the number of groups is 30 or larger. Therefore, say 30,000 items of data may 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. In case of distant expectations another strategies should be used. In more details it is discussed in Kolnogorov (2016) where the adaptive strategy is proposed which at initial comparatively short stage checks the closeness of expectations and applies the appropriate strategy at the final stage of control.

Remark 3. In section 4 we present the invariant Bellman-type recursive equation for determination of the Bayesian risk and Bayesian strategy. Since the maximal expected losses are attained for close expectations, the invariant recursive equation is valid in the domain of close expectations as well.