Minimax Normal Two-Armed Bandit with Indefinite Control Horizon

Alexander Kolnogorov[1,⋆]

[1] *Yaroslav-the-Wise Novgorod State University*
B.St-Petersburgskaya Str, 41, Velikiy Novgorod, Russia, 173003

Abstract. We consider the two-armed bandit problem as applied to data processing if there are two alternative processing methods available with different a priori unknown efficiencies. On should determine the most effective method and provide its dominating application. The total number of data, which is interpreted as a control horizon, is assumed to have a priori known probability distribution.
The problem is considered in minimax (robust) setting. According to the main theorem of the theory of games minimax risk and minimax strategy are sought for as Bayesian ones corresponding to the worst-case prior distribution. We describe the properties of the worst-case prior and present a recursive Bellman-type equation for determination of both minimax strategy and minimax risk. Numerical results illustrating the proposed algorithm are given. The algorithm can be applied to optimization of parallel data processing if the number of processed data is not definitely known in advance.

1 Introduction

We consider the two-armed bandit problem (see, e.g. [1], [2]) which is also well-known as the problem of expedient behavior in a random environment (see, e.g. [3], [4]) and the problem of adaptive control (see, e.g. [5], [6]) 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 ($y_n \in \{1, 2\}$) and have Normal probability distribution densities

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

if $y_n = \ell$ ($\ell = 1, 2$). Normal two-armed bandit can be described by a vector parameter $\theta = (m_1, m_2)$. The set of parameters is assumed to be the following

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

where $0 < C < \infty$.

Control strategy σ at the 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 $y^{n-1} = y_1, \dots, y_{n-1}$:

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

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

Generally, the goal is to maximize (in some sense) the total expected income. In this article, we consider the minimax approach. 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 loss 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 with respect to its maximal possible value due to incomplete information. Here $E_{\sigma,\theta}$ denotes the mathematical expectation calculated with respect to measure generated by a strategy σ and a parameter 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

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

is called the minimax risk and corresponding optimal strategy σ^M is called the minimax strategy. Note that application of the minimax strategy ensures that inequality holds

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

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

The minimax approach to the problem was proposed by H. Robbins in [7] for the so-called Bernoulli two-armed bandit. It is described by binary incomes $\{0, 1\}$ and probability distribution

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

$p_\ell + q_\ell = 1$, $\ell = 1, 2$. Bernoulli two-armed bandit is described by a parameter $\theta = (p_1, p_2)$ with the set of values $\Theta = \{\theta : 0 \le p_\ell \le 1; \ell = 1, 2\}$. The article [7] caused a significant interest to considered problem. It was shown in [8] that explicit determination of the minimax strategy and

⋆e-mail: Alexander.Kolnogorov@novsu.ru

minimax risk is practically impossible already for $N > 4$. However, the asymptotic minimax theorem was proved by W. Vogel in [9] which states that minimax risk has the order $N^{1/2}$ and provides lower and upper estimates for a factor. This theorem holds true for the Normal two-armed bandit as well.

A very popular approach to the problem is a Bayesian one. Let $\lambda(\theta)$ be a prior probability density. The value

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

is called Bayesian risk and corresponding optimal strategy is called Bayesian strategy. Bayesian approach allows to find Bayesian strategy and risk by solving a recursive Bellman-type equation.

According to the main theorem of the theory of games the minimax risk (2) is equal to Bayesian risk (3) calculated with respect to the worst-case prior distribution, i.e.

$$R_N^M(\Theta) = R_N^B(\lambda_0) = \sup_{\lambda} R_N^B(\lambda) \tag{4}$$

This approach allows to obtain the following asymptotic estimate for the Normal two-armed bandit (see [10]–[13]):

$$\lim_{N\to\infty} N^{-1/2} R_N^M(\Theta) = R_0 \tag{5}$$

with $R_0 \approx 0.637$.

Let's explain the choice of 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 $\zeta_t = 1$ if processing is successful and $\zeta_t = 0$ if it is unsuccessful. Probabilities $\Pr(\zeta_t = 1|y_t = \ell) = p_\ell$ ($\ell = 1, 2$) depend only on selected methods (actions). Let's assume that p_1, p_2 are 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 data processing in the same packet. For the control, we use the values of the process $\xi_n = (D_p M)^{-1/2} \sum_{n=(n-1)M+1}^{nM} \zeta_t$, $n = 1, \dots, N$ with $D_p = p(1-p)$. 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.

Note that data in the same packet may be processed in parallel. However, there is a question of losses in the control performance as the result of such aggregation. It was shown in [10]–[13] that if N is large enough (e.g. $N \geq 30$) then parallel control is close to optimal. Therefore, say 30000 items of data can be processed in 30 steps by packets of 1000 data with almost the same maximal losses as if the data were processed optimally one-by-one.

Remark 1. There are some different approaches to robust control in the two-armed bandit problem, see, e.g. [6, 14–16]. In these articles stochastic approximation method and mirror descent algorithm 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 $N^{1/2}$. However, more precise estimates were not presented for these algorithms. The versions for parallel processing were not proposed as well.

Remark 2. Parallel control for the two-armed bandit problem was first proposed 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 in [17].

The goal of this paper is to investigate the robust control of the Normal two-armed bandit with indefinite horizon. The structure of the paper is the following. In Section 2 we set the control problem, i.e. define the loss function and corresponding minimax and Bayesian risks and strategies if there is a priori known probability distribution on the set of horizons. We state the main theorem of the theory of games in this case. In Section 3 we analyze the properties of the worst-case prior distribution and present the recursive Bellman-type equation for calculation of the Bayesian risk with respect to this worst-case prior. We also present recursive Bellman-type equation for calculation of the expected losses. In Section 4 the results of numerical experiments are presented. Section 5 contains a conclusion.

2 Setup of the Control Problem with Indefinite Horizon

The disadvantage of approach considered in Section 1 is that the control horizon N is fixed. However, it is often more likely to consider the problem with indefinite control horizon. Let's consider $\{N_i\} = \{N_i; i = 1, \dots, I\}$ a finite set of control horizons s.t. $2 \leq N_1 < \dots < N_I = N$. Let's define appropriate loss function as follows

$$L(\sigma, \theta, \{N_i\}) = \sum_{i=1}^{I} \beta_i L_{N_i}(\sigma, \theta), \tag{6}$$

where $L_{N_i}(\sigma, \theta)$, $i = 1, \dots, I$ are defined in (1). Factors β_i, $i = 1, \dots, I$ may be arbitrary chosen. However, it is natural to choose them so that all $\{\beta_i L_{N_i}(\sigma, \theta)\}$ have approximately equal maximal values. Taking into account (5) one can choose $\beta_i \asymp N_i^{-1/2}$, $i = 1, \dots, I$. If in addition the following condition holds

$$\sum_{i=1}^{I} \beta_i = 1$$

then one can say that the prior distribution $\Pr(N = N_i) = \beta_i$, $i = 1, \dots, I$ is assigned on the set of control horizons.

Given n satisfying condition $N_{i-1} < n \leq N_i$, let's put

$$\gamma_n = \sum_{j=i}^{I} \beta_j.$$

Using (1) and (6) one obtains

$$L(\sigma, \theta, \{N_i\}) = E_{\sigma,\theta}\left(\sum_{n=1}^{N} \gamma_n \left((m_1 \vee m_2) - \xi_n\right)\right) \tag{7}$$

In the sequel we'll denote $\hat{L}_N(\sigma, \theta) = L(\sigma, \theta, \{N_i\})$. Corresponding minimax and Bayesian risks are defined as

follows

$$\hat{R}_N^M(\Theta) = \inf_{\Sigma} \sup_{\Theta} \hat{L}_N(\sigma, \theta), \tag{8}$$

$$\hat{R}_N^B(\lambda) = \inf_{\Sigma} \int_{\Theta} \hat{L}_N(\sigma, \theta)\lambda(\theta)d\theta. \tag{9}$$

Taking into account (7) an using reasonings similar to those in [10], one can prove that the main theorem of the theory of games holds in considered setting as well as in the case of definite horizon. It means that minimax risk (8) can be determined as Bayesian risk (9) calculated with respect to the worst-case prior distribution, i.e.

$$\hat{R}_N^M(\Theta) = \hat{R}_N^B(\lambda_0) = \sup_{\lambda} \hat{R}_N^B(\lambda) \tag{10}$$

and minimax strategy is equal to corresponding Bayesian strategy as well.

Bayesian risk can be calculated recursively. Denote by

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

the Normal distribution density with mathematical expectation M and variance D. Denote by $\lambda(m_1, m_2)$ the prior distribution density on the set of parameters Θ. Let history of control up to instant of time n be described by (X_1, n_1, X_2, n_2). Here n_1, n_2 are total numbers of applications of both actions ($n_1 + n_2 = n$) and X_1, X_2 are corresponding total incomes. Let $X_\ell = 0$ if $n_\ell = 0$. The posterior distribution density is thus equal to

$$\lambda(m_1, m_2|X_1, n_1, X_2, n_2) = \frac{f_{n_1}(X_1|n_1m_1)f_{n_2}(X_2|n_2m_2)\lambda(m_1, m_2)}{\iint\limits_{\Theta} f_{n_1}(X_1|n_1m_1)f_{n_2}(X_2|n_2m_2)\lambda(m_1, m_2)dm_1, dm_2}.$$

If additionally it is assumed that $f_n(X|nm) = 1$ at $n = 0$ then this expression holds true if $n_1 = 0$ and/or $n_2 = 0$ as well.

Denote by $\hat{R}^B_{N-n}(\lambda; X_1, n_1, X_2, n_2)$ Bayesian risk at the latter $(N-n)$ steps calculated with respect to the posterior distribution density $\lambda(m_1, m_2|X_1, n_1, X_2, n_2)$. Let $x^+ = \max(x, 0)$. Then

$$\hat{R}^B_{N-n}(\cdot) = \min(\hat{R}^{(1)}_{N-n}(\cdot), \hat{R}^{(2)}_{N-n}(\cdot)), \tag{11}$$

where $\hat{R}^{(1)}_0(\cdot) = \hat{R}^{(2)}_0(\cdot) = 0$,

$$\begin{aligned}\hat{R}^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint\limits_{\Theta} (\gamma_{n+1}(m_2 - m_1)^+ \\ +E^{(1)}_x \hat{R}^B_{N-(n+1)}(\lambda; X_1 + x, n_1 + 1, X_2, n_2)) \\ \times\lambda(m_1, m_2|X_1, n_1, X_2, n_2)dm_1dm_2,\end{aligned} \tag{12}$$

$$\begin{aligned}\hat{R}^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint\limits_{\Theta} (\gamma_{n+1}(m_1 - m_2)^+ \\ +E^{(2)}_x \hat{R}^B_{N-(n+1)}(\lambda; X_1, n_1, X_2 + x, n_2 + 1)) \\ \times\lambda(m_1, m_2|X_1, n_1, X_2, n_2)dm_1dm_2\end{aligned} \tag{13}$$

and

$$E^{(\ell)}_x \hat{R}(x) = \int\limits_{-\infty}^{+\infty} \hat{R}(x)f(x|m_\ell)dx, \quad \ell = 1, 2.$$

Bayesian strategy prescribes currently to choose action corresponding to the smaller value of $\hat{R}^{(1)}_{N-n}(\cdot)$, $\hat{R}^{(2)}_{N-n}(\cdot)$, the choice may be arbitrary if these values are equal.

3 Recursive Bellman-type Equation for Calculation of the Bayesian Risk and Expected Losses

Let's recall that Bayesian risk is a concave function of the prior distribution density, i.e.

$$\hat{R}^B_N(\alpha\lambda + \tilde{\alpha}\tilde{\lambda}) \geq \alpha\hat{R}^B_N(\lambda) + \tilde{\alpha}\hat{R}^B_N(\tilde{\lambda}),$$

if $\alpha + \tilde{\alpha} = 1$; $\alpha, \tilde{\alpha} > 0$. This property allows to specify the worst-case prior distribution. Like in [10], one can prove that the following transformations $\tilde{\lambda}$ of the prior distribution density λ do not change the Bayesian risk, i.e. $\hat{R}^B_N(\tilde{\lambda}) = \hat{R}^B_N(\lambda)$:

1. $\tilde{\lambda}^{(1)}(m_1, m_2) = \lambda(m_2, m_1)$ (for all m_1, m_2). This property means that expected losses do not change if one swaps the arms of the bandit.

2. $\tilde{\lambda}^{(2)}(m_1, m_2) = \lambda(m_1 + m, m_2 + m)$ (for all m_1, m_2 and any fixed m). This property means that expected losses do not change if one equally shifts both mathematical expectations.

So, if λ is the worst-case prior distribution then $\alpha\lambda+\tilde{\alpha}\tilde{\lambda}$ is the worst-case prior as well. It means that the worst-case prior distribution does not change if the above transformations are implemented.

In the sequel it is convenient to modify parameterization. Let $m_1 = u + v$, $m_2 = u - v$, then $\theta = (u + v, u - v)$ and $\Theta = \{\theta : |v| \leq C\}$. Taking into account the Jacobian $|\partial(m_1, m_2)/\partial(u, v)| = 2$, a prior distribution density is equal to $\nu(u, v) = 2\lambda(u+v, u-v)$. Then the following transformations of the prior distribution densities $\tilde{\nu}^{(1)}(u, v) = \nu(u, -v)$, $\tilde{\nu}^{(2)}(u, v) = \nu(u + m, v)$ (for any fixed m) do not change the value of Bayesian risk. These properties allow to describe the worst-case prior. Namely, asymptotically the worst-case prior distribution density can be chosen the following one:

$$\nu_a(u, v) = \kappa_a(u)\rho(v), \tag{14}$$

where $\kappa_a(u)$ is the uniform density on the interval $|u| \leq a$, $\rho(v)$ is a symmetric density (i.e. $\rho(-v) = \rho(v)$) on the interval $|v| \leq C$ and $a \to \infty$. This prior does not change under the first transformation and asymptotically (as $a \to \infty$) does not change under the second transformation.

Now let's write the dynamic programming equation for calculation the Bayesian risk with respect to (14). This equation follows from (11)–(13) if the prior distribution density is formally assumed to be constant with respect to u and this gives true expressions for the posterior densities if $n_1 \geq 1$, $n_2 \geq 1$. At the former two steps actions should be chosen turn-by-turn. Note that equation is more simple for risks

$$\hat{R}_{n_1,n_2}(X_1, X_2) = \hat{R}^B_{N-n}(X_1, n_1, X_2, n_2)p_{n_1,n_2}(X_1, X_2)$$

with

$$p_{n_1,n_2}(X_1, X_2) = \iint\limits_{\Theta} f_{n_1}(X_1|n_1m_1)f_{n_2}(X_2|n_2m_2)\lambda(m_1, m_2)dm_1dm_2.$$

Denote by $\hat{R}_{n_1,n_2}(Z) = \hat{R}_{n_1,n_2}(X_1, X_2)$ with $Z = X_1 n_2 - X_2 n_1$. Then

$$\hat{R}_{n_1,n_2}(\cdot) = \min(\hat{R}^{(1)}_{n_1,n_2}(\cdot), \hat{R}^{(2)}_{n_1,n_2}(\cdot)), \tag{15}$$

with $\hat{R}^{(1)}_{n_1,n_2}(Z) = \hat{R}^{(2)}_{n_1,n_2}(Z) = 0$ at $n_1 + n_2 = N$,

$$\begin{aligned}\hat{R}^{(1)}_{n_1,n_2}(Z) &= \int_0^C 2\gamma_{n+1} v g_{n_1,n_2}(Z, v)\rho(v)dv \\ &+n_2^{-1}\int_{-\infty}^{+\infty} \hat{R}_{n_1+1,n_2}(Z + z)h_{n_1}\left(\tfrac{Z-n_1 z}{n_2}\right)dz,\end{aligned} \tag{16}$$

$$\begin{aligned}\hat{R}^{(2)}_{n_1,n_2}(Z) &= \int_0^C 2\gamma_{n+1} v g_{n_1,n_2}(Z, -v)\rho(v)dv \\ &+n_1^{-1}\int_{-\infty}^{+\infty} \hat{R}_{n_1,n_2+1}(Z + z)h_{n_2}\left(\tfrac{Z-n_2 z}{n_1}\right)dz\end{aligned} \tag{17}$$

at $n = n_1 + n_2 < N$, $n_1 \geq 1$, $n_2 \geq 1$. Here

$$\begin{aligned}g_{n_1,n_2}(Z, v) &= \frac{1}{(2\pi n_1 n_2(n_1 + n_2))^{1/2}} \\ &\times \exp\left(-\frac{(Z + 2vn_1 n_2)^2}{2n_1 n_2(n_1 + n_2)}\right),\end{aligned} \tag{18}$$

$$h_n(z) = \left(\frac{n+1}{2\pi n}\right)^{1/2} \times \exp\left(-\frac{z^2}{2n(n+1)}\right). \tag{19}$$

Bayesian risk (9) is calculated according to the formula

$$\lim_{a\to\infty} \hat{R}^B_N(\nu_a(u, v)) = \hat{L}(\rho(v)) + \int_{-\infty}^{\infty} \hat{R}_{1,1}(z)dz \tag{20}$$

where

$$\hat{L}(\rho(v)) = \int_0^C (\gamma_1 + \gamma_2)2v\rho(v)dv.$$

Now let's present a recursive equation for expected losses corresponding to strategy $\sigma_\ell(Z, n_1, n_2) = \Pr(y_n = \ell|Z, n_1, n_2)$:

$$\begin{aligned}\hat{L}_{n_1,n_2}(Z) &= \sigma_1(Z, n_1, n_2)\hat{L}^{(1)}_{n_1,n_2}(Z) \\ &+\sigma_2(Z, n_1, n_2)\hat{L}^{(2)}_{n_1,n_2}(Z),\end{aligned} \tag{21}$$

with $\hat{L}^{(1)}_{n_1,n_2}(Z) = \hat{L}^{(2)}_{n_1,n_2}(Z) = 0$ at $n_1 + n_2 = N$,

$$\begin{aligned}\hat{L}^{(1)}_{n_1,n_2}(Z) &= \int_0^C 2\gamma_{n+1} v g_{n_1,n_2}(Z, v)\rho(v)dv \\ &+n_2^{-1}\int_{-\infty}^{+\infty} \hat{L}_{n_1+1,n_2}(Z + z)h_{n_1}\left(\tfrac{Z-n_1 z}{n_2}\right)dz,\end{aligned} \tag{22}$$

$$\begin{aligned}\hat{L}^{(2)}_{n_1,n_2}(Z) &= \int_0^C 2\gamma_{n+1} v g_{n_1,n_2}(Z, -v)\rho(v)dv \\ &+n_1^{-1}\int_{-\infty}^{+\infty} \hat{L}_{n_1,n_2+1}(Z + z)h_{n_2}\left(\tfrac{Z-n_2 z}{n_1}\right)dz\end{aligned} \tag{23}$$

if $n = n_1 + n_2 < N$, $n_1 \geq 1$, $n_2 \geq 1$. Expected losses are calculated according to the formula

$$\lim_{a\to\infty} \hat{L}^B_N(\nu_a(u, v)) = \hat{L}(\rho(v)) + \int_{-\infty}^{\infty} \hat{L}_{1,1}(z)dz. \tag{24}$$

4 Numerical Results

Minimax risk and minimax strategy were sought for the case $N_1 = 20$, $N_2 = N = 30$. Factors were chosen as follows: $\beta_1 = \beta^{-1}N_1^{-1/2}$, $\beta_2 = \beta^{-1}N_2^{-1/2}$ with $\beta = N_1^{-1/2} + N_2^{-1/2}$, so that $\beta_1 + \beta_2 = 1$. Bayesian risk was calculated by (15)–(20). The worst-case prior distribution corresponding to the maximum of the Bayesian risk is concentrated at $d \approx \pm 1.9$ and maximal value of the Bayesian risk is approximately equal to 0.594. Then expected losses were calculated by (21)–(24) using determined strategy. Results are presented on Figure 1. One can see that maximal value of expected losses does not exceed the value 0.594. Hence, the determined strategy is minimax strategy for considered problem.

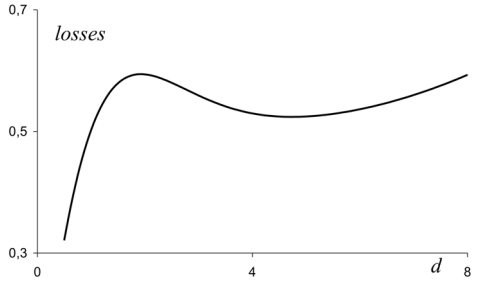

Figure 1. Expected losses corresponding to minimax strategy, $N_1 = 20$, $N_2 = 30$.

5 Conclusion

Minimax approach to the two-armed bandit problem is considered with a priori assigned distribution on the control horizon. The problem can be reduced to the classical two-armed bandit problem with discounted incomes. The algorithm of determination of the minimax strategy and minimax risk as Bayesian ones corresponding to the worst-case prior distribution is obtained and numerical results are presented.

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.

References

[1] D.A. Berry, B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments* (Chapman and Hall, London, New York, 1985)

[2] E.L. Presman, I.M. Sonin, *Sequential Control with Incomplete Information* (Academic Press, New York, 1990)

[3] M.L. Tsetlin, *Automation Theory and Modeling of Biological Systems* (Academic Press, New York, 1973)

[4] V.I. Varshavsky, *Collective Behavior of Automata* (Nauka, Moscow, 1973) (In Russian)

[5] V.G. Sragovich, *Mathematical Theory of Adaptive Control* (World Scientific. Interdisciplinary Mathematical Sciences, New Jersey, London, **4**, 2006)

[6] A.V. Nazin, A.S. Poznyak, *Adaptive Choice of Alternatives* (Nauka, Moscow, 1986) (In Russian)

[7] H. Robbins, Bulletin AMS. **58**, 527–535 (1952)

[8] J. Fabius, W.R. van Zwet, Ann. Math. Statist. **41**, 1906–1916 (1970)

[9] W. Vogel, Ann. Math. Stat. **31**, 444–451 (1960)

[10] A.V. Kolnogorov, Automation and Remote Control **72**, 1017–1027 (2011)

[11] A.V. Kolnogorov, Automation and Remote Control **73**, 689–701 (2012)

[12] A.V. Kolnogorov, Automation and Remote Control **76**, 1229–1241 (2015)

[13] A.V. Kolnogorov, International Journal of Mathematics and Computers in Simulation **10**, 129–132 (2016)

[14] G. Lugosi, N. Cesa-Bianchi, *Prediction, Learning and Games* (Cambridge University Press, New York, 2006)

[15] A. Juditsky, A.V. Nazin, A.B. Tsybakov, N. Vayatis, Proc. 17th World Congress IFAC, 11560–11563 (2008)

[16] A.V. Gasnikov, Yu.E. Nesterov, V.G. Spokoiny, Computational Mathematics and Mathematical Physics **55**, 580–596 (2015)

[17] T.L. Lai, B. Levin, H. Robbins, D. Siegmund, Proc. Nati. Acad. Sci. USA **77**, 3135–3138 (1980)