An Asymptotic Minimax Theorem for Gaussian Two-Armed Bandit

A.V.Kolnogorov[1]

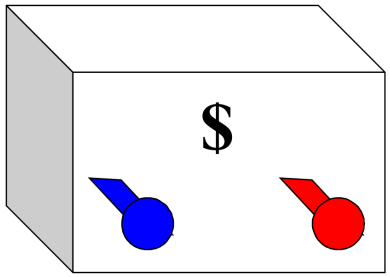
[1]Yaroslav-the-Wise Novgorod State University
Alexander.Kolnogorov@novsu.ru

ACMPT

Moscow

2017, October 26

Bernoulli Two-Armed Bandit



It is a slot machine with two arms. If the ℓ-th arm is chosen then the gambler gets unit income (+1) with probability p_ℓ and nothing (0) with probability q_ℓ ($p_\ell + q_\ell = 1$).

The gambler can chose arms N times totally. His goal is to maximize (in some sense) the total expected income. Probabilities p_1, p_2 are fixed during control process but unknown to the gambler.

A Dilemma "Information vs Control"

For the gambler it would be better always to chose the arm corresponding to the largest value of probabilities p_1, p_2. However, if he wants to determine this arm he should try them both and this diminishes his total expected income.

Formal Setup

Formally, let's consider a Bernoulli random controlled process ξ_n, $n = 1, \ldots, N$, s.t.

$$\Pr(\xi_n = 1 | y_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0 | y_n = \ell) = q_\ell, \quad \ell = 1, 2.$$

A strategy σ can use all currently available information of the process: n_1, n_2 – total numbers of both arms choices, X_1, X_2 – corresponding total incomes. The loss function is as follows

$$L_N(\sigma, \theta) = N(p_1 \vee p_2) - \mathbb{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi_n \right),$$

where $\theta = (p_1, p_2)$ is a parameter of the process.

Bayesian Approach

Let $\lambda(\theta)$ be a prior distribution density on $\Theta = \{(p_1, p_2) : 0 \le p_\ell \le 1, \ell = 1, 2\}$. The Bayesian risk is equal to

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

corresponding optimal strategy σ^B is called Bayesian strategy.

A Simple Recursive Algorithm of Determination

As Berry and Fristedt write: "...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".

Minimax Approach

The minimax risk is equal to

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

corresponding optimal strategy σ^M minimax strategy.

Robustness of Minimax Approach

If σ^M is applied then the following inequality holds

$$L_N(\sigma^M, \theta) \leq R_N^M(\Theta), \quad \forall \theta \in \Theta.$$

Impossibility of Direct Determination

As Fabius and van Zwet write about Bernoulli two-armed bandit: “the algebra involved becomes progressively more complicated with increasing N and seems to remain prohibitive already for N as small as 5”.

An Asymptotic Minimax Theorem

Theorem

The following inequality holds as $N \to \infty$ for Bernoulli two-armed bandit

$$0.612 \leq (DN)^{-1/2} R_N^M(\Theta) \leq 0.752.$$

The proof of the theorem uses some indirect estimates and techniques.

Vogel, W. An asymptotic minimax theorem for the two-armed bandit problem. Ann. Math. Stat., 1961, V. 31, P. 444–451

Specification of the Asymptotic Minimax Theorem

We propose the following specification of the AMT for *Gaussian* two-armed bandit

Theorem

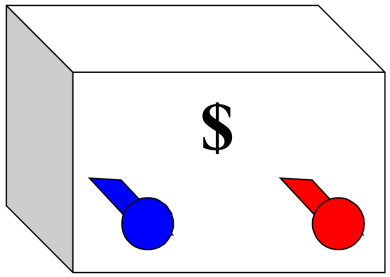
The following equality holds

$$\lim_{N\to\infty}(DN)^{-1/2}R_N^M(\Theta)\approx 0.637.$$

The following issues are to be discussed:

1. What is Gaussian two-armed bandit?
2. Why Gaussian two-armed bandit?
3. How are Gaussian and Bernoulli two-armed bandits related with?
4. How the theorem is proved?

What is Gaussian Two-Armed Bandit



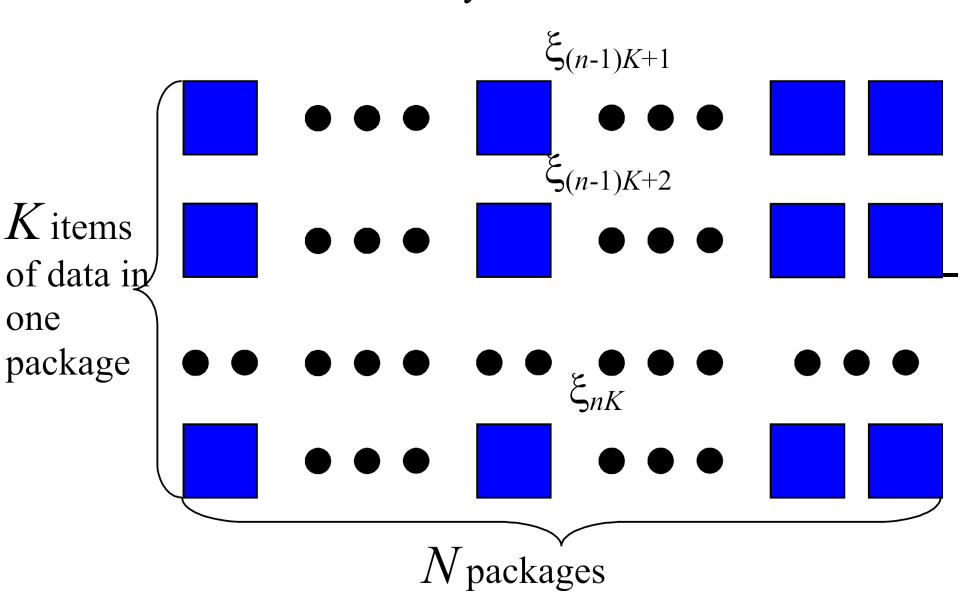
It is a slot machine with two arms. If the ℓ-th arm is chosen then the gambler gets random income. This income is normally distributed with unit variance and mathematical expectation m_ℓ.

The gambler can chose arms N times totally. His goal is to maximize (in some sense) the total expected income. Expectations m_1, m_2 are fixed during control process but unknown to the gambler.

A Dilemma "Information vs Control"

For the gambler it would be better always to chose the arm corresponding to the largest value of expectations m_1, m_2. However, if he wants to determine this arm he should try them both and this diminishes his total expected income.

Why Gaussian Two-Armed Bandit? Parallel processing



Assume that a large number $T = NK$ items of data are given, which can be processed by two alternative methods. Processing may be successful ($\xi_t = 1$) or unsuccessful ($\xi_t = 0$).

Probabilities of successful and unsuccessful processing depend on chosen methods (arms) and are equal to p_ℓ and q_ℓ respectively ($\ell = 1, 2$).
Assume that p_1, p_2 are close to p. Let's define the process

$$\xi'_n = (DK)^{-1/2} \sum_{t=(n-1)K+1}^{nK} \xi_t, \quad n = 1, \ldots, N, \quad D = p(1-p).$$

Distributions of ξ'_n are close to normal and their variances are close to 1.

Relation between Gaussian and Bernoulli Two-armed Bandits

1. In application to data processing, Gaussian two-armed bandit is a particular case of Bernoulli two-armed bandit which allows to process data in parallel. The data should be partitioned in a number of groups. Data in the same group are then processed in parallel by the same method.
2. For example, given 30 000 items of data, the data can be partitioned into 30 groups each containing 1000 items of data. Then they can be processed in 30 stages by 1000 items of data at each stage.
3. If the number of stages is large enough (e.g. 30 stages or more) then maximal losses for parallel data processing are almost the same as if the data were processed optimally one-by-one!
4. So, Bernoulli and Gaussian two-armed bandits are equivalent as $N \to \infty$.

Formal Setup of the Problem

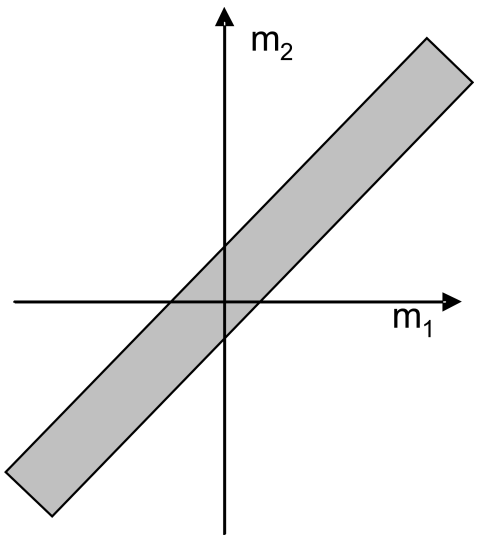
Formally incomes are considered as a controlled random process $\xi_1, \xi_2, \ldots, \xi_N$, which values depend only on currently chosen arms (in the sequel called actions) $y_1, y_2, \ldots, y_N$ and are normally distributed with unit variance and mathematical expectation m_ℓ if the ℓ-th action is chosen

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

This process is completely described by a vector parameter $\theta = (m_1, m_2)$. A control strategy σ prescribes the choice of actions y_n, $n = 1, \ldots, N$ and depends on complete prehistory of the process. In fact, it is sufficient to know four current values: n_1, n_2 – numbers of choices of both actions and X_1, X_2 -– total incomes for both actions. Loss function is defined as follows

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

Minimax and Bayesian Settings



Assume that the set of parameters is the following $\Theta = \{(m_1, m_2) : |m_1 - m_2| \leq 2c_1, |m_1 + m_2| \leq 2c_2\}$, with $0 < c_1 < \infty$, $0 < c_2 < \infty$ and c_2 is large enough. Minimax risk is defined as

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

corresponding strategy σ^M is called minimax strategy.

Consider a prior distribution $\lambda(m_1, m_2)$ on Θ. Bayesian risk is defined as

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

corresponding strategy σ^B is called Bayesian strategy.

Some Properties of Approaches

Robustness of Minimax Approach

If σ^M is applied then the following inequality holds

$$L_N(\sigma^M, \theta) \leq R_N^M(\Theta), \quad \forall \theta \in \Theta.$$

A Simple Recursive Algorithm for Bayesian Approach

There is a well-known recursive algorithm for Bayesian risk and Bayesian strategy numerical determination for any prior distribution.

Main Theorem of the Theory of Games

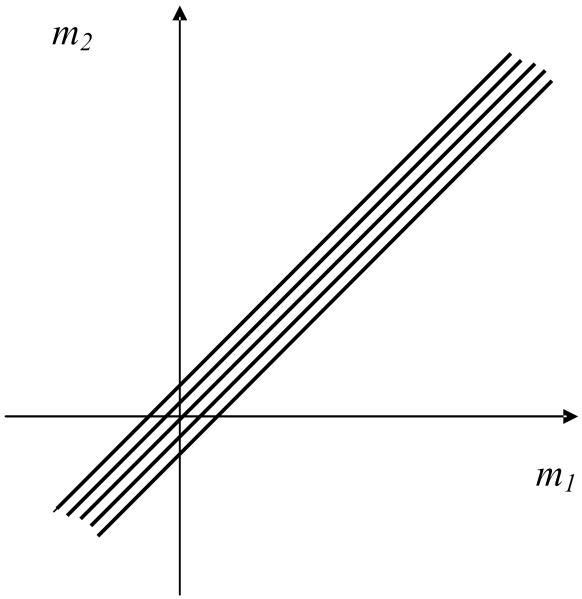
Under mild conditions the minimax risk is equal to Bayesian one corresponding to the worst-case prior distribution i.e.

$$R_N^M(\Theta) = \sup_{\{\Lambda\}} R_N^B(\Lambda) = R_N^B(\Lambda^0).$$

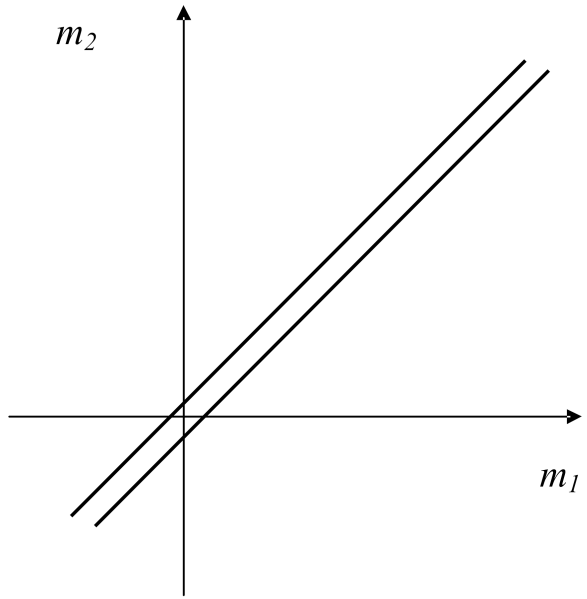
and minimax strategy is equal to corresponding Bayesian strategy as well. **In the sequel, minimax risk is searched as Bayesian one calculated with respect to the worst-case prior.**

Asymptotically the Worst-Case Prior Distribution

Asymptotically the worst-case prior is uniform along and symmetric crosswise the main diagonal



Calculations allow to expect that it concentrates on two parallel lines. Distance between them is the only unknown parameter



Change of Variables

Recall that n_1, n_2 denote total numbers of choices of both actions, X_1, X_2 are corresponding total incomes and m_1, m_2 are mathematical expectations. We assume that actions are applied to groups of the size M. This allows parallel processing.
Let's denote

$$u = \frac{X_1 n_2 - X_2 n_1}{n N^{1/2}},$$

$$t_1 = \frac{n_1}{N}, \quad t_2 = \frac{n_2}{N}, \quad \varepsilon = \frac{M}{N},$$

$$w = (m_1 - m_2) N^{1/2}$$

and let $\varrho(w)$ characterize a symmetric uniform prior distribution.

Invariant Recursive Equation with Unit Time Horizon

To determine the Bayesian risk one should solve Bellman type recursive equation

$$r_\varepsilon(u,t_1,t_2)=\min_{\ell=1,2} r_\varepsilon^{(\ell)}(u,t_1,t_2),$$

where $r_\varepsilon^{(1)}(u,t_1,t_2)=r_\varepsilon^{(2)}(u,t_1,t_2)=0$ for $t_1+t_2=1$,

$$r_\varepsilon^{(1)}(u,t_1,t_2)=\varepsilon g^{(1)}(u,t_1,t_2)+r_\varepsilon(u,t_1+\varepsilon,t_2)*f_{\varepsilon t_2^2 t^{-1}(t+\varepsilon)^{-1}}(u),$$
$$r_\varepsilon^{(2)}(u,t_1,t_2)=\varepsilon g^{(2)}(u,t_1,t_2)+r_\varepsilon(u,t_1,t_2+\varepsilon)*f_{\varepsilon t_1^2 t^{-1}(t+\varepsilon)^{-1}}(u),$$

for $t_1+t_2<1$, $t_1\geq\varepsilon$ and $t_2\geq\varepsilon$.

$$g^{(\ell)}(u,t_1,t_2)=\int_0^\infty 2wg(u,(-1)^{\ell+1}w,t_1,t_2)\varrho(w)dw,\quad \ell=1,2,$$
$$g(u,w,t_1,t_2)=\exp\left(-2uw-2w^2t_1t_2t^{-1}\right),$$
$$f_\varepsilon(x)=(2\pi\varepsilon)^{-1/2}\exp(x^2/(2\varepsilon)).$$

Passage to the Limit

Let $\varepsilon \to 0$. Then for all u and for all t_1, t_2 for which solution of the equation is well defined there exists a limit

$$r(u, t_1, t_2) = \lim_{\varepsilon \to +0} r_\varepsilon(u, t_1, t_2), \quad \ell = 1, 2.$$

Under some additional conditions $r(u, t_1, t_2)$ satisfies the second order partial differential equation

$$\min_{\ell=1,2} \left(\frac{\partial r}{\partial t_\ell} + \frac{t_{\overline{\ell}}^2}{2t^2} \times \frac{\partial^2 r}{\partial u^2} + g^{(\ell)}(u, t_1, t_2) \right) = 0,$$

$\overline{\ell} = 3 - \ell$, $\ell = 1, 2$, with initial conditions

$$\lim_{t_1 + t_2 \to 1} r(u, t_1, t_2) = 0 \quad \text{if} \quad t_1 > \varepsilon, \ t_2 > \varepsilon,$$

and boundary conditions

$$\lim_{u \to +\infty} r(u, t_1, t_2) = \lim_{u \to -\infty} r(u, t_1, t_2) = 0$$

Risk and Strategy

Minimax Risk

Limiting minimax risk is equal to the limiting Bayesian risk corresponding to the worst prior distribution and is calculated as

$$\lim_{N\to\infty} N^{-1/2} R_N^M(\Theta) = \sup_{\varrho} r(\varrho; u, t_1, t_2)\big|_{u=0,t_1=0,t_2=0}.$$

Optimal Strategy

Currently optimal is the ℓ-th action if

$$\frac{\partial r}{\partial t_\ell} + \frac{t_\ell^2}{2t^2} \times \frac{\partial^2 r}{\partial u^2} + g^{(\ell)}(u, t_1, t_2)$$

has the smaller value ($\ell = 1, 2$).

Numerical Experiments

Calculations were done under assumption that the worst prior distribution $\varrho(w)$ is concentrated at two points $w = \pm d$ with $0.5 \leq d \leq 2.5$. The maximal value of $r(\varrho; 0, 0, 0)$ was determined as

$$\max_d r(\varrho; 0, 0, 0) \approx 0.637$$

corresponding to $d \approx 1.57$.

Russian References

1. Tsetlin, M.L., Issledovaniya po teorii avtomatov i modelirovaniyu biologicheskikh sistem (Studies in Automata Theory and Modeling Biological Systems), Moscow: Nauka, 1969.
2. Varshavskii, V.I., Kollektivnoe povedenie avtomatov (Collective Behavior of Automata), Moscow: Nauka, 1973.
3. Sragovich, V.G., Adaptivnoe upravlenie (Adaptive Control), Moscow: Nauka, 1981.
4. Nazin, A.V. and Poznyak, A.S., Adaptivnyi vybor variantov (Adaptive Choice between Alternatives), Moscow: Nauka, 1986.
5. Presman, E.L. and Sonin, I.M., Posledovatel'noe upravlenie po nepolnym dannym (Sequential Control with Incomplete Data), Moscow: Nauka, 1982.

English References

1. Berry, D.A. and Fristedt, B. Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall. London, New York.,1985.
2. Lai, T.L. and Robbins, H. Asymptotically Efficient Adaptive Allocation Rules. Advances in Applied Mathematics, 1985, V. 6, P. 4-22.
3. Robbins, H. Some aspects of the sequential design of experiments. Bulletin of Amer. Math. Soc., 1952, V. 58, P.527-535.
4. Vogel, W. An asymptotic minimax theorem for the two-armed bandit problem. Ann. Math. Stat., 1961, V. 31, P.444-451
5. Fabius, J., and van Zwet, W.R. Some remarks on the two-armed bandit. Ann. Math. Stat., 1970, V. 41, 1906 -1916.

Previous Publications

1. Kolnogorov A. V. Finding Minimax Strategy and Minimax Risk in a Random Environment (the Two-Armed Bandit Problem) // Automation and Remote Control, 2011, Vol. 72, No. 5, pp. 1017-1027.
2. Kolnogorov A.V. On a Limiting Description of Robust Parallel Control in a Random Environment // Automation and Remote Control, Vol. 76, No. 7, pp. 1229 - 1241, 2015.

Thank you for attention

Thank you for attention