SOS-2010

Minimax Strategy for Bernoulli Two-Armed Bandit with One Known Probability of Income

Alexander Kolnogorov

Novgorod State University
B.St-Petersburgskaya Str. 41, Velikiy Novgorod,
Russian Federaration

Alexander.Kolnogorov@novsu.ru

Novgorod

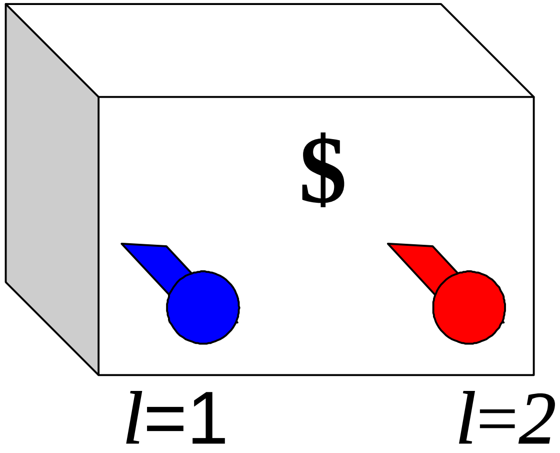
Monument to
Millennium of Russia

Sophia Cathedral

Novgorod State
University

Outline

- What is the Two-Armed Bandit?
- Object of Control, Strategy and Loss Function
- Relation of Minimax and Bayesian Approaches
- Determination of the Minimax Risk
- Examples
- Calculation of the Minimax Risk for Considered Set of Parameters
- An Asymptotic Approximation
- Conclusion



What is the Two-Armed Bandit?

Two-Armed Bandit is a slot machine with 2 arms. If the l-th arm has been chosen then the gambler gets random income which is equal to 1 with probability p_l and to 0 with probability q_l $(p_l + q_l = 1)$.

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

A Dilemma "Information vs Control"

For the gambler it would be optimal always to choose the arm corresponding to the maximal value of probabilities p_1, p_2.
However, to determine this arm he should test both of them and this diminishes his total expected income.

Object of Control, Strategy and Loss Function

Formally, incomes are considered as a controlled process $\xi_1, \xi_2,..,\xi_N$, which values depend on currently chosen alternatives $y_1, y_2,\dots,y_N$ only, i.e

$$P\{x_n = 1 \mid y_n = l\} = p_l, \quad P\{x_n = 0 \mid y_n = l\} = q_l, \qquad l = 1,2.$$

Controlled process can be described by a vector parameter $\theta=(p_1,p_2)$.

Control strategy σ prescribes the choice of alternatives y_n, $n=1,\dots,N$ and can use all history of the process $y_1,\xi_1,\dots,y_{n\text{-}1},\xi_{n\text{-}1}$. Instead of detailed history it is sufficient to know 4 current numbers:

n_1, n_2 - total numbers of choices of all alternatives,

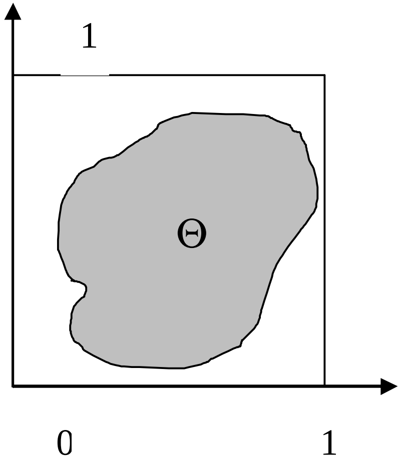
m_1, m_2 - total numbers of nonzero incomes for their applications.

The total number of probabilities describing the strategy is

$$S(N) \sim N^4/4!$$

The loss function is defined as

$$L_N(s\ ,q) = N \max_{l=1,2} p_l - E_{s\ ,q}\left(\sum_{n=1}^{N} x_n\right).$$



Minimax Approach

Given a set of parameters *Θ*, the minimax risk is defined as

$$R_N^M(\Theta) = \min_s \max_\Theta L_N(s\ ,q\ ).$$

Corresponding strategy σ^M is called a minimax strategy.

Robustness of Minimax Approach

If the minimax strategy σ^M is applied then the values of loss function do not exceed the value of the minimax risk on the whole set *Θ*, i.e.

$$L_N(s^{\ M}, q\ ) \leq R_N^M(\Theta) \quad \text{for all} \quad q \in \Theta.$$

However, a direct determination of the minimax strategy is very difficult. As Fabius and van Zwet (1970) write:

> “the algebra involved becomes progressively more complicated with increasing N and seems to remain prohibitive already for N as small as 5".

Bayesian Approach

According to Bayesian approach the Bayes loss function should be minimized

$$R_N^B(l) = \min_{s} \int_{\Theta} L_N(s,q)\, l(dq),$$

where $\lambda(d\theta)$ is a prior distribution of the parameter. The minimum $R^B{}_N(\lambda)$ is called a Bayes risk and corresponding strategy σ^B is called a Bayes strategy.

A simple recurrent algorithm of determination of the Bayes risk and Bayes strategy is well known. As Berry and Fristedt (1985) 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".

Adaptive Learning and Bayesian Formalism

Bayes risk can be calculated by dynamic programming equation

$$R_n^B(l) = \min_{\mathbf{l}=1,2} E_l \left(x + R_{n-1}^B \left(l\,(y_1 = \mathbf{l}, x_1 = x) \right) \mid y_1 = \mathbf{l} \right)$$

and Bayes strategy σ^B prescribes to choose at the first step the alternative corresponding to the minimal term in the right-hand side of the equation.

Comments to the equation

recalculation of the posterior distribution
provides ***identification*** of the parameter

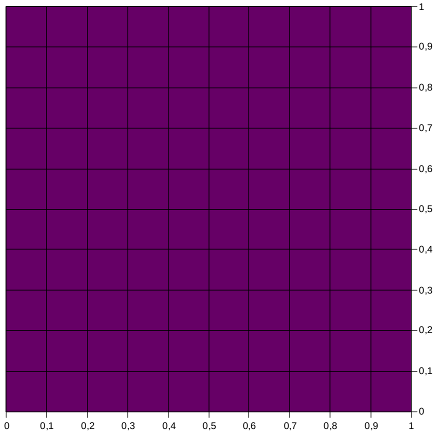
$$R_n^B(l) = \underline{\min_{\mathbf{l}=1,2}}\, \underline{E_l \left(x + R_{n-1}^B \left(l\,\overline{(y_1 = \mathbf{l}, x_1 = x)} \right) \mid y_1 = \mathbf{l} \right)}$$

minimization of the
income provides the goal
of the ***control***

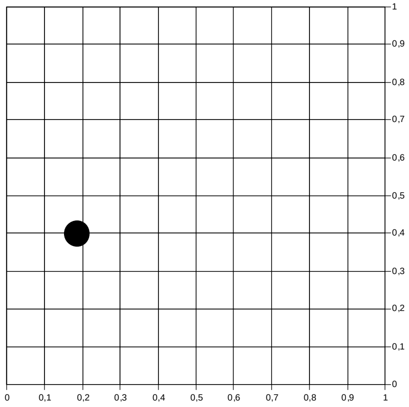
minimal total expected income
provided that at the first step the ℓ-th
alternative has been chosen

Prior and Posterior Distributions -1

Uniform prior distribution

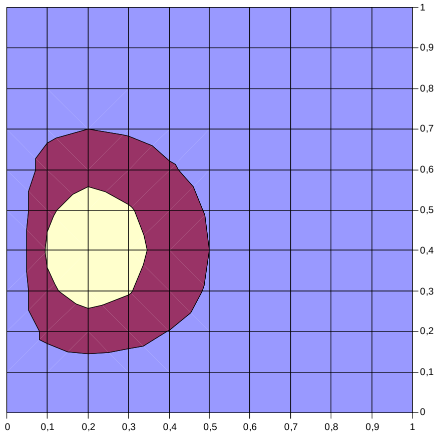


Actual value of the parameter

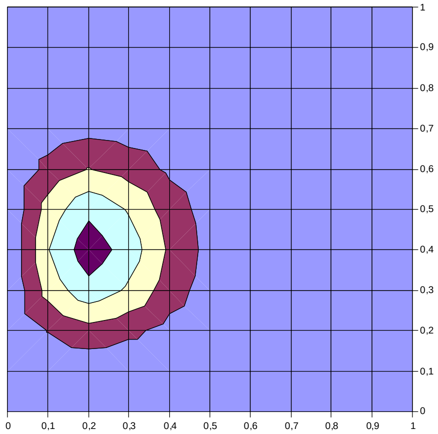


Posterior distributions

m_1=1, n_1=5,
m_2=2, n_2=5



m_1=2, n_1=10,
m_2=4, n_2=10



Main Theorem of the Theory of Games

Under mild conditions the minimax risk is equal to Bayes risk calculated over the worst prior distribution, corresponding to the maximum of Bayes risk, i.e.

$$R_N^M(\Theta) = \max_l R_N^B(l),$$

and minimax strategy σ^M is equal to some Bayes strategy σ^B (the latter may be ambiguously determined).

Some References

Sragovich, V.G. (1981). Adaptive Control. Nauka, Moscow. (In Russian)

Nazin, A.V., and Poznyak, A.S. (1986). Adaptive Choice of Alternatives. Nauka, Moscow. (In Russian)

Presman E. L., Sonin I.M. Sequential control with incomplete information данным. – Nauka, Moscow. 1982 (In Russian).

Kolnogorov, A.V. (1989). A minimax approach to optimal expedient behavior in stationary environments over finite time. Sov. J. Comput. Syst. Sci., volume 27, No.4, 33 – 35. (Translation from Russian)

Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin AMS., volume 58(5), 527-535.

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

Fabius, J., and van Zwet, W.R. (1970). Some remarks on the two-armed bandit. Ann. Math. Statist., volume 41, 1906 -1916.

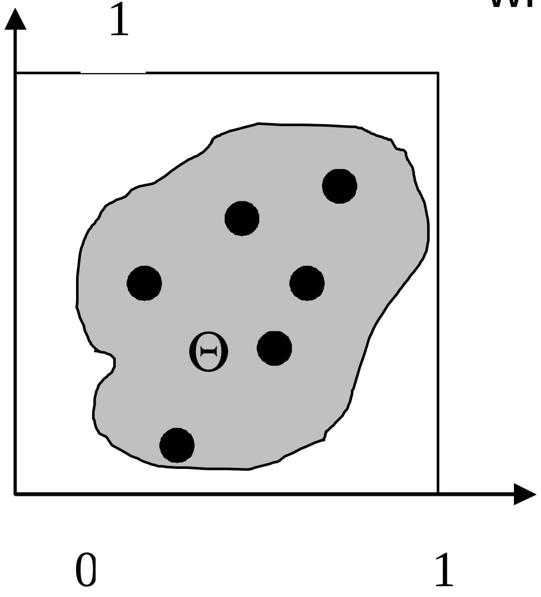
Bradt R.N., Johnson S.M., Karlin S. On sequential designs for maximizing the sum of n observations. -- Ann. Math. Statist, 1956. V.27. P.1060-1074.

Determination of the Minimax Risk -1

Minimax risk on the whole set of parameters is equal to the minimax risk on some its finite subset and minimax strategy on the whole set of parameters is equal to the minimax strategy on that finite subset. The finite subset of parameters satisfies a condition

$$R_N^M(\Theta) = R_N^M(q_1^0,\ldots,q_s^0) = \max_{q_1,\ldots,q_r} R_N^M(q_1,\ldots,q_r)$$

where



$$R_N^M(q_1,\ldots,q_r) = \min_s \max_{i=1,\ldots,r} L_T(s\ ,q_i)$$

For *s* the estimate holds

$$s \le 2S(N) \sim \frac{N^4}{12}$$

Determination of the Minimax Risk -2

Minimax strategy on the finite set of parameters is equal to some Bayes strategy calculated over the worst prior distribution on this set, i.e. the following equality holds

$$R_N^M(q_1,...,q_r) = R_N^B(q_1,...,q_r;\, l_1^0,...,\, l_r^0) =$$

$$= \max_{l_1,...,l_r} R_N^B(q_1,...,q_r;\, l_1,...,\, l_r)$$

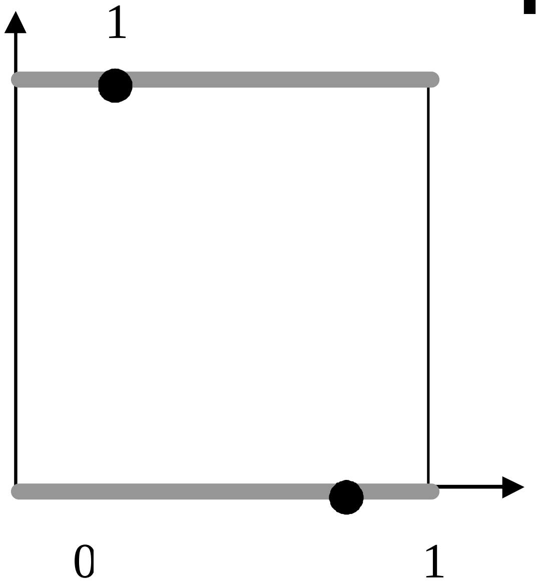
where

$$R_N^B(q_1,...,q_r;\, l_1,...,\, l_r) = \min_s \sum_{i=1}^r l_i \cdot L_N(s\,, q_i)$$

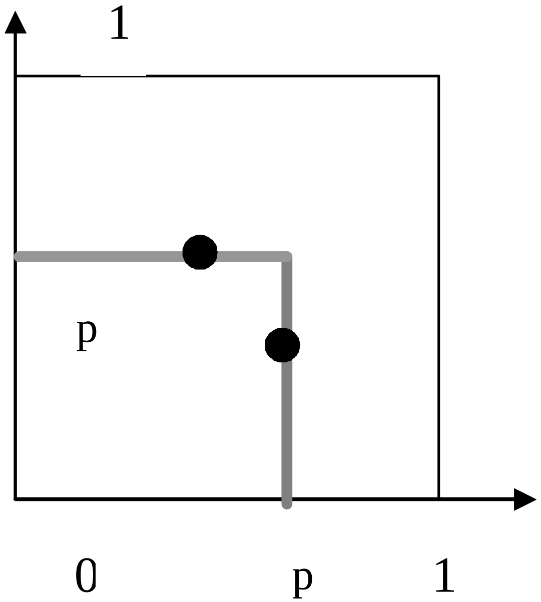
Some probabilities of the Bayes strategy can be arbitrary ones. They should be chosen to satisfy the system of equations

$$L_N(s^{\,M}, q_i^{\,0}) = R_N^B(q_1^{\,0},...,q_s^{\,0};\, l_1^0,...,\, l_s^0), \qquad i = 1,...,n,$$

Examples -1

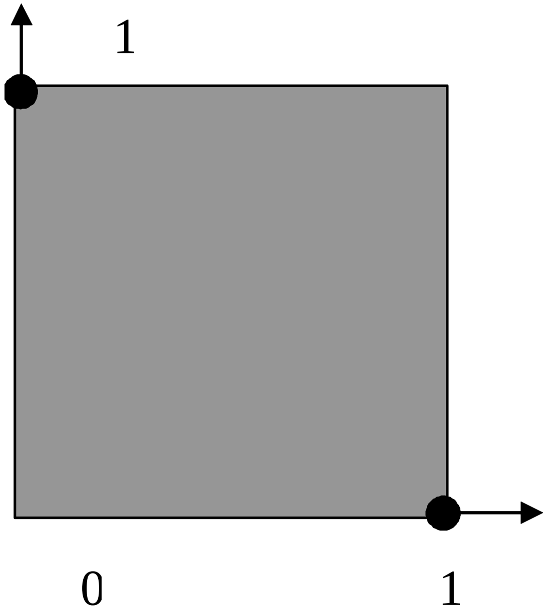


1. From Berry, D.A. and Fristedt, B.
   N=3,
   $l(0.881,0)\approx 0.591$,
   $l(0.218,1)\approx 0.409$

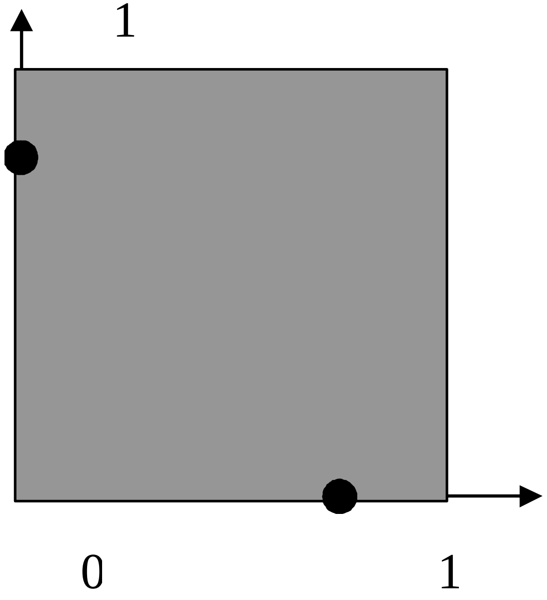


2. $l(p,p\text{-}1.25(D/N)^{1/2})=$
   $l(p\text{-}1.25(D/N)^{1/2},p)=0.5$,
   $D=p(1\text{-}p)$

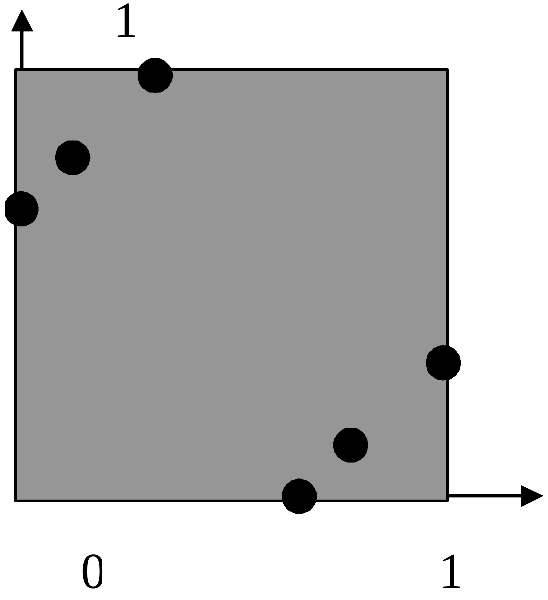
Examples – 2



3. N=1,2.
$l(0,1)=$
$l(1,0)=0.5$

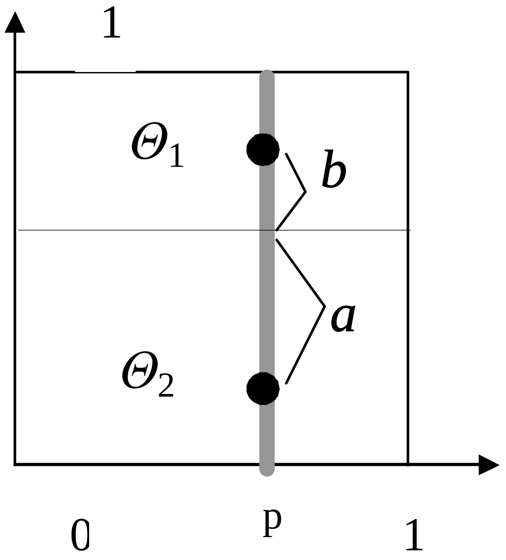


4. N=3.
$l(0,0.75)=$
$l(0.75,0)=0.5$



5. N=4.
$l(0,a)=l(a,0)\approx 0.29$
$l(1,1\text{-}a)=\ l(1\text{-}a,1)\approx 0.16$
$l(0.7,0.7\text{-}a)=$
$=l(0.7\text{-}a,0.7)\approx 0.05$
$a\approx 0.654$

Calculation of the Bayes Risk for Considered Set of Parameters - 1

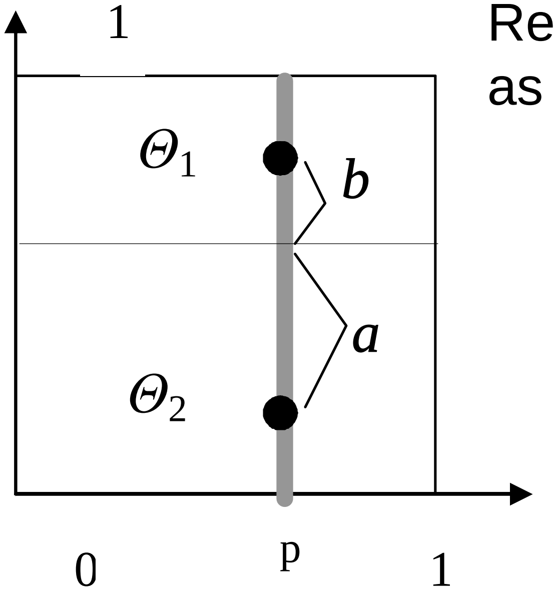


Consider the set $\Theta = \{\theta_1=(p,p+b), \theta_2=(p,p-a)\}$ with a prior distribution $\mathrm{P}(\theta=\theta_1)=\lambda$, $\mathrm{P}(\theta=\theta_2)=1-\lambda$, где p, a, b are known, $0<p<1$, $0<a<p$, $0<b<1-p$. The Bayes risk is defined as

$$R_N^B(l\,,a,b)=\min_s\left(l\cdot L_N(s\,,q_1)+(1-l\,)\cdot L_N(s\,,q_2)\right)$$

The choice l=1 does not give any additional information about parameter. Hence, if l=1 is optimal at some step it will remain to be optimal at all further steps. It means that optimal strategy prescribes to choose l=2 until some stopping time and then to choose l=1 till the end of control.

Calculation of the Bayes Risk for Considered Set of Parameters - 2



Recurrent equation for calculation of the Bayes risk is as follows

$$R_n^B(l,a,b) = \min(lbn,(1-l)a +$$
$$p(l)R_{n-1}^B(l_1,a,b) + q(l)R_{n-1}^B(l_0,a,b))$$

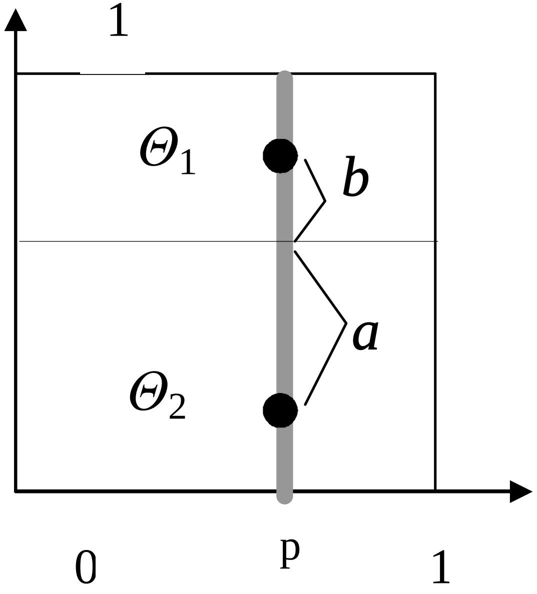
with initial conditions $R_0^B(l,a,b) = 0.$

Here $p(\lambda)=\lambda(p+b)+(1–\lambda)(p–a)$, $q(\lambda)=\lambda(q–b)+(1–\lambda)(q+a)$ are probabilities to get 1 and 0 incomes if the choice *l*=2 was made and for current posterior distribution $\mathrm{P}(\theta=\theta_1)=\lambda$, $\mathrm{P}(\theta=\theta_2)=1–\lambda$, $q=1–p$. And $\lambda_1=\lambda(p+b)/p(\lambda)$, $\lambda_0=\lambda(q–b)/q(\lambda)$ are posterior probabilities of $\mathrm{P}(\theta=\theta_1)$ provided that 1 and 0 incomes were got respectively.

Calculation of the Minimax Risk for Considered Set of Parameters

It can be determined as

$$R_N^M(\Theta) = \max_{a,b,l} R_N^M(l, a, b)$$



N= 10, p=0.6,

$l\,(p,p+b) \approx 0.276$,

$l\,(p,p\text{-}a) \approx 0.724$,

$a \approx 0.451$, $b \approx 0.209$

An Asymptotic Approximation -1

Let a=$\alpha(pq/N)^{1/2}$, b=$\beta(pq/N)^{1/2}$, τ=n/N, $r(t, l, a, b) = R_n^B(l, a, b)\cdot(pqN)^{-1/2}$.
If $N \to \infty$ then the following differential equation holds

$$r'_t(t, l, a, b) = a(1-l) + 0.5\,l^2(1-l)^2(a+b)^2\, r''_{ll}(t, l, a, b)$$

with initial condition $r(t, l, a, b) = 0.$

Boundary conditions if 0 ≤ τ ≤ 1 are as follows:

$r(t, 1, a, b) = 0,$

and at the point of tangency $(s.t.\ r(t, l, a, b) = tbl)$

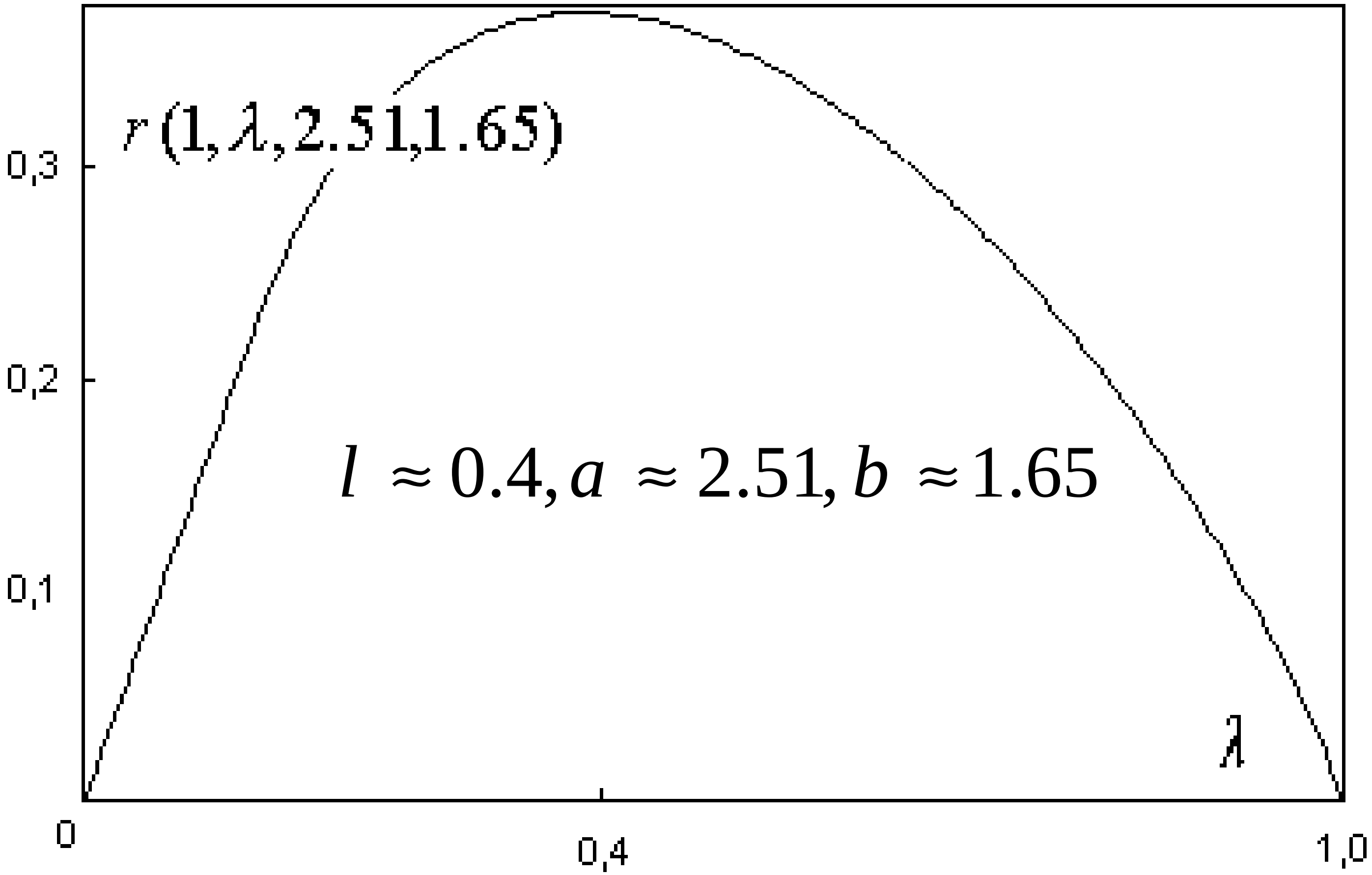
$r'_l(t, l, a, b) = tb.$

To the left of the point of tangency $r(t, l, a, b) = tbl.$

An Asymptotic Approximation -2

Minimax risk can be estimated as

$$(pqN)^{1/2}\max_{l,a,b} r(1,l,a,b) \approx 0.38(pqN)^{1/2}$$



$$l \approx 0.4, a \approx 2.51, b \approx 1.65$$

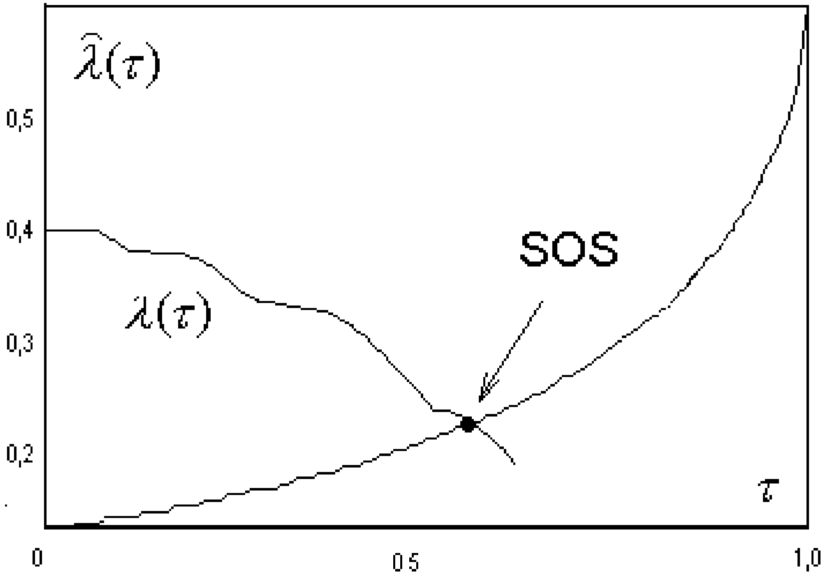
Asymptotically Minimax Strategy

The gambler should take parameters mentioned above

$$l \approx 0.4, a \approx 2.51, b \approx 1.65,$$

apply the alternative ℓ=2 and recalculate the posterior probability until its current value becomes less than the current value at the point of tangency $\hat{l}(t)$ where

$$\hat{l}(1-t) = \inf(l : r(t, l, a, b) < tbl)$$



It is a SOS point of time. Then the alternative ℓ=1 should be applied till the end of control.

Conclusion

- Bernoulli two-armed bandit problem with one known probability of income is considered in minimax setting which provides the robustness of the control. However, a direct solution of the problem is impossible for practical magnitudes of horizon;
- According to the main theorem of the theory of games the problem is reduced to the solution in Bayes setting. It provides a convenient mathematical formalism for adaptive learning;
- In Bayes setting the problem is solved for the worst prior distribution. This worst prior distribution is concentrated on two parameters.
- An asymptotic approximation is considered and asymptotically minimax strategy is described.

Thank you for attention