ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 31 № 3 2025

УДК 519.83 + 519.244

МИНИМАКСНЫЙ ПОДХОД К ЗАДАЧЕ О ГАУССОВСКОМ МНОГОРУКОМ БАНДИТЕ[1,2]

А. В. Колногоров

Рассматривается задача о многоруком бандите в приложении к пакетной обработке больших данных, если имеется более двух альтернативных методов обработки с различными, априори неизвестными, эффективностями. В процессе обработки требуется определить наиболее эффективный метод и обеспечить его преимущественное использование. Управление осуществляется на основе суммарных доходов в пакетах, которые в силу центральной предельной теоремы имеют приблизительно гауссовское распределение. Важной особенностью пакетной обработки является то, что она почти не приводит к увеличению максимальных потерь полного ожидаемого дохода, т. е. к увеличению минимаксного риска, если количество обрабатываемых данных и пакетов, на которые они разбиты, достаточно велико. Это означает, что гауссовский многорукий бандит обеспечивает универсальный подход к оптимальному управлению обработкой больших данных, если одношаговые доходы удовлетворяют центральной предельной теореме. Согласно этому подходу минимаксные стратегия и риск ищутся с использованием основной теоремы теории игр как байесовские, вычисленные относительно наихудшего априорного распределения, на котором байесовский риск максимален. Для этого дана характеризация наихудшего априорного распределения и получены рекуррентные уравнения в обычной и инвариантной форме с горизонтом управления, равным единице. В предельном случае, когда количество обрабатываемых пакетов бесконечно растет, получено дифференциальное уравнение в частных производных второго порядка. Приведен численный пример нахождения минимаксного риска и стратегии для трехрукого бандита.

Ключевые слова: гауссовский многорукий бандит, игра с природой, байесовский и минимаксный подходы, основная теорема теории игр, пакетная обработка.

A. V. Kolnogorov. Minimax approach to the Gaussian multi-armed bandit.

We consider the multi-armed bandit problem in an application to batch processing of big data if there are more than two alternative processing methods with different a priori unknown efficiencies. During processing, it is necessary to determine a more effective method and ensure its preferential use. Control is performed on the basis of cumulative incomes in batches, which, by virtue of the central limit theorem, have approximately Gaussian distributions. An important feature of batch processing is that it almost does not lead to an increase in maximum losses of the total expected income, i.e., to an increase in minimax risk if the numbers of processed data and batches into which the data is divided are large enough. This means that Gaussian multi-armed bandit provides universal approach to optimal control of big data when one-step incomes satisfy the central limit theorem. According to this approach, minimax strategy and risk are sought using the main theorem of game theory as Bayesian ones calculated relative to the worst-case prior distribution at which the Bayesian risk is maximal. For this purpose, the characterization of the worst-case prior distribution is given and recursive equations in a usual and invariant form with a control horizon equal to one are obtained. In the limiting case when the number of processed batches grows infinitely, a second-order partial differential equation is obtained. A numerical example of finding minimax risk and strategy for a 3-armed bandit is given.

Keywords: Gaussian multi-armed bandit, Game with nature, Bayesian and minimax approaches, Main theorem of game theory, Batch processing.

MSC: 91A35, 62L05, 62C10

DOI: 10.21538/0134-4889-2025-31-3-fon-06

1. Введение

Рассматривается задача о многоруком бандите [1; 2], которая является частью последовательного управления с неполной информацией [3]. Название происходит от игрового автомата

[1]Исследование выполнено в ходе реализации НИР “Математическое моделирование природных процессов”, выполняемой в рамках государственного задания в сфере научной деятельности.

[2]Исследование выполнено с использованием инфраструктуры Центра коллективного пользования “Высокопроизводительные вычисления и большие данные” (ЦКП “Информатика”) ФИЦ ИУ РАН (г. Москва).

с несколькими рукоятками (далее — действиями), выбор каждой из которых сопровождается случайным доходом игрока. Распределение одношагового дохода зависит только от выбранного действия, фиксировано во время игры, но неизвестно игроку. В процессе игры необходимо определить более эффективное (т. е. соответствующее более высокому математическому ожиданию одношагового дохода) действие и обеспечить его преимущественное использование. Задача имеет приложения в моделировании биологических систем [4], оптимальном управлении случайными процессами [5], медицине [6], интернет-технологиях [2;7], оптимизации обработки данных [8] и других областях.

Ниже задача рассматривается в приложении к пакетной обработке больших данных, если имеются два или более альтернативных метода обработки, которые в данном случае являются действиями, с различными, априори неизвестными, эффективностями. Для этого данные разбиваются на пакеты равного объема, одинаковый метод обработки применяется ко всем данным одного и того же пакета, а для управления используются количества успешно обработанных данных в пакетах, которые интерпретируются как доходы. Если размеры пакетов достаточно велики, то при широких предположениях доходы в пакетах являются приблизительно гауссовскими. Таким образом, рассматривается задача о гауссовском многоруком бандите. Отметим, что первоначально пакетная обработка была предложена для лечения групп пациентов альтернативными лекарствами. Для этого сначала различные лекарства даются сравнительно небольшим тестовым группам пациентов, а затем лучшее по результатам лечения тестовых групп — всем оставшимся. Хороший обзор и литература по этому подходу имеются в [6].

Формально гауссовский многорукий бандит — это управляемый случайный процесс, значения которого ξ_k зависят только от действий y_k, выбранных в моменты времени k, и имеют плотности распределения $(2\pi)^{-1/2}\exp(-(x-m_\ell)^2/2)$, если $y_k = \ell$ $(k = 1, 2, \ldots, K)$, где $\ell = 1, 2, \ldots, J$; $J \geq 2$. Таким образом, рассматриваемый гауссовский многорукий бандит характеризуется параметром $\theta = (m_1, m_2, \ldots, m_J)$, компоненты которого равны математическим ожиданиям одношаговых доходов, а все дисперсии доходов равны единице.

Особенность задачи о многоруком бандите заключается в том, что многие результаты, которые справедливы для случая двух действий, перестают быть таковыми, если количество действий становится равным трем или более. Отметим, например, хорошо известный результат [9], где задача о двуруком бандите рассматривается в байесовской постановке на паре параметров $\theta_1 = (m_1, m_2)$, $\theta_2 = (m_2, m_1)$ и где доказано, что оптимальная стратегия управления зависит только от текущих апостериорных вероятностей этих параметров. В случае $J > 2$, как установлено в [10;11], эта стратегия может быть обобщена только на наборы из J параметров вида

$$\theta_1 = (m_1, m_2, \ldots, m_2), \quad \theta_2 = (m_2, m_1, \ldots, m_2), \quad \ldots, \quad \theta_J = (m_2, m_2, \ldots, m_1),$$

причем $m_1 > m_2$, но не на перестановки компонент параметра $(m_1, m_2, \ldots, m_J)$. В качестве еще одного примера укажем, что в случае двурукого бандита байесовская стратегия часто имеет пороговый характер для произвольных априорных распределений (см., например, [12]), но при $J > 2$ это свойство, вообще говоря, больше не выполняется.

В данной статье исследуется обобщение ряда результатов, полученных ранее в [8;13] для случая двурукого бандита ($J = 2$), на случай $J = 3$, т. е. рассматривается трехрукий бандит. Это сделано для простоты. При этом нетрудно понять, как перенести результаты на случай $J > 3$, а также какую (более простую) форму они примут при $J = 2$.

Чтобы определить цель управления, рассмотрим стратегии σ, использующие всю известную историю управляемого процесса, которая характеризуется полными количествами применения всех действий и соответствующими полными доходами, т. е.

$$\sigma_\ell(X_1, X_2, X_3; k_1, k_2, k_3) = \Pr(y_{k+1} = \ell | X_1, X_2, X_3; k_1, k_2, k_3), \quad \ell = 1, 2, 3.$$

Обозначим $m^* = \max(m_1, m_2, m_3)$. Если бы был известен параметр $\theta = (m_1, m_2, m_3)$, то оптимальная стратегия позволяла бы получить максимально возможный ожидаемый доход Km^*.

Если же применяется стратегия σ, то ожидаемый доход будет меньше максимального на величину

$$L_K(\sigma,\theta) = Km^* - \mathbf{E}_{\sigma,\theta}\Big(\sum_{k=1}^{K}\xi_k\Big), \tag{1.1}$$

которая называется *функцией потерь*. Здесь $\mathbf{E}_{\sigma,\theta}$ — знак математического ожидания по мере, порождаемой стратегией σ и параметром θ. Отметим, что многие исследования были посвящены изучению асимптотического (при $K\to\infty$) поведения $L_K(\sigma,\theta)$, если параметр θ фиксирован, но априори неизвестен. В результате были найдены точная асимптотическая оценка функции потерь порядка $\ln K$ и соответствующая оптимальная стратегия [14]. Впоследствии такой же порядок роста функции потерь был установлен для ряда известных стратегий, таких как EXP3, UCB, сэмплирование Томпсона и так далее (см., например, [2; 7]).

Рассмотрим множество параметров

$$\Theta = \{(m_1,m_2,m_3) : m_\ell = m + v_\ell,\ |v_\ell|\le C,\ \ell = 1,2,3;\ v_1+v_2+v_3 = 0,\ |m|\le C_1\},$$

где $C > 0$, $C_1 > 0$ и константа C_1 достаточно велика. Предполагается, что множество Θ известно. Для априорной плотности распределения $\lambda(\theta) = \lambda(m_1,m_2,m_3)$ на Θ байесовский риск определяется как

$$R_K^B(\lambda) = \inf_\sigma \int_\Theta L_K(\sigma,\theta)d\theta; \tag{1.2}$$

соответствующая оптимальная стратегия σ^B называется байесовской стратегией. Байесовский подход очень популярен, поскольку позволяет найти байесовские стратегию и риск, решая в обратном времени уравнение динамического программирования Беллмана для любого априорного распределения. Минимаксный риск вычисляется по формуле

$$R_K^M(\Theta) = \inf_\sigma \sup_\Theta L_K(\sigma,\theta); \tag{1.3}$$

соответствующая оптимальная стратегия σ^M называется минимаксной стратегией. Минимаксный подход является робастным, поскольку обеспечивает выполнение неравенства $L_K(\sigma^M,\theta)\le R_K^M(\Theta)$ для всех $\theta\in\Theta$. К сожалению, прямого метода нахождения минимаксных стратегии и риска не существует. Например, в [12] минимаксные стратегия и риск были точно найдены для бернуллиевского двурукого бандита только горизонтов управления $K\le 4$. Однако в [15] была получена асимптотическая оценка минимаксного риска, которая имеет вид $r(DK)^{1/2}$ при $K\to\infty$, где $D = 0.25$ — максимальная дисперсия одношагового бернуллиевского дохода, а множитель r ограничен значениями $0.53 < r < 0.74$. Насколько нам известно, пока не получены хорошие верхние и нижние оценки минимаксного риска в случае $J > 2$. Взамен были получены верхние границы минимаксного риска порядка $K^{1/2}$ для некоторых известных стратегий, например алгоритма зеркального спуска [16], UCB [17], сэмплирования Томпсона [18] и так далее (см. также [2; 7]).

Чтобы найти минимаксный риск (1.3), можно использовать основную теорему теории игр, согласно которой выполнено равенство

$$R_K^M(\Theta) = R_K^B(\lambda^0) = \sup_\lambda R_K^B(\lambda), \tag{1.4}$$

т. е. минимаксный риск равен байесовскому, вычисленному относительно наихудшего априорного распределения, на котором байесовский риск максимален, а минимаксная стратегия совпадает с соответствующей байесовской. Поэтому данную задачу оптимального управления часто рассматривают как игру с природой. В этой игре $\{\sigma\}$ — набор стратегий игрока, а Θ — набор стратегий природы. Игра является антагонистической, хотя в ее результатах заинтересован только игрок; природа к ним равнодушна.

Прямое использование равенства (1.4) практически невозможно из-за его высокой вычислительной сложности. В [8;13] были проанализированы свойства наихудшего априорного распределения. Оказалось, что это распределение асимптотически равномерно вследствие вогнутости байесовского риска. Более того, оно является вырожденным, поскольку должно концентрироваться в точках максимума функции потерь (1.1) и, следовательно, в рассматриваемом случае зависит от единственного параметра. Чтобы найти этот параметр, необходимо максимизировать соответствующий байесовский риск. В случае $J = 2$ указанный подход позволил оценить минимаксный риск как $r(DK)^{1/2}$ при $K \to \infty$, где $r \approx 0.637$, а D — дисперсия одношагового дохода (см. [8;13]). В настоящей статье мы развиваем данный подход для случая многорукого бандита с $J \geq 3$ действиями с конечной (будущей) целью получения точных оценок минимаксного риска.

Обсудим теперь преимущества пакетной обработки. Во-первых, если возможно дополнительно использовать параллельную обработку данных, то она позволяет значительно сократить полное время обработки. Во-вторых, поскольку доходы в пакетах приблизительно гауссовские, найденные стратегии становятся универсальными и могут быть применены к процессам с произвольными распределениями одношаговых доходов, если последние удовлетворяют центральной предельной теореме. В-третьих, *когда задача рассматривается в минимаксной постановке, минимаксный риск пакетной обработки практически не увеличивается по сравнению со значением, соответствующим обработке данных по одному, если количества данных и пакетов, на которые они разбиты, достаточно велики* (см., например, [6;8;13]).

Поясним последнее свойство на примере, показывающем, как рассматриваемая постановка задачи может возникать на практике. Пусть есть бернуллиевский многорукий бандит ζ_n, $n = 1, 2, \ldots, N$ $(N = MK)$, который характеризуется распределением $\Pr(\zeta_n = 1|y_n = \ell) = p_\ell$, $\Pr(\zeta_n = 0|y_n = \ell) = 1 - p_\ell$, $\ell = 1, \ldots, J$. Значения процесса $\zeta_n = 1$ и $\zeta_n = 0$ соответствуют успешно и неуспешно обработанной единице данных с номером n. Известно, что максимальные значения функции потерь достигаются в области "близких распределений", которая характеризуется условием, что $|p_i - p_j| \leq cN^{-1/2}$ для всех p_i, p_j и $c > 0$ — достаточно большой фиксированной константы, причем все p_i, p_j близки к некоторому $p \in (0, 1)$ $(\ell = 1, \ldots, J)$ (см., например, [15] и [13]). Это является следствием того, что в области "близких распределений" максимальная из неизвестных вероятностей $p_1, \ldots, p_J$ определяется с ошибкой, вероятность которой будет не меньше некоторого $q_e > 0$. Следовательно, одношаговая величина потерь будет не меньше чем $cN^{-1/2}q_e$, а полная величина функции потерь — не меньше чем $cN^{1/2}q_e$, т. е. имеет порядок $N^{1/2}$.

В области "близких распределений" приме́ним одинаковые действия к пакетам из M данных и используем для управления значения процесса $\xi_k = (DM)^{-1/2} \sum_{n=(k-1)M+1}^{kM} \zeta_n$, $k = 1, \ldots, K$, где $D = p(1 - p)$. Если M достаточно велико, то распределения процесса ξ_k, $k = 1, 2, \ldots, K$, близки к гауссовским с дисперсиями, равными единице. Тогда в случае использования минимаксной стратегии максимальные значения функции потерь для процесса $\{\xi_k\}$ при $K \to \infty$ будут асимптотически равны $rK^{1/2}$ (где $r \approx 0.637$ при $J = 2$), а для процесса $\{\zeta_n\}$ с учетом сделанной нормировки — $(DM)^{1/2}rK^{1/2} = r(DN)^{1/2}$ независимо от того, на сколько пакетов были разбиты исходные данные.

Статья организована следующим образом. В разд. 2 получено рекуррентное уравнение для вычисления байесовских стратегии и риска относительно априорного распределения общего вида. В разд. 3 описываются свойства наихудшего априорного распределения, на котором байесовский риск достигает своих максимальных значений, т. е. асимптотическая равномерность и симметричность. В разд. 4 получено рекуррентное уравнение для вычисления байесовских стратегии и риска относительно наихудшего априорного распределения. В разд. 5 это уравнение представлено в инвариантной форме с горизонтом управления, равным единице, и уравнением в частных производных второго порядка в предельном случае, когда количество обрабатываемых пакетов неограниченно растет. Результаты численных экспериментов представлены в разд. 6.

2. Рекуррентное уравнение для нахождения байесовского риска

Получим рекуррентное уравнение для нахождения байесовских стратегии и риска относительно априорной плотности распределения общего вида. Обозначим через $\{X,k\} = (X_1,k_1,X_2,k_2,X_3,k_3)$ текущую предысторию процесса, где k_1, k_2, k_3 — это текущие полные применения всех действий, а X_1, X_2, X_3 — соответствующие полные доходы. Если на множестве Θ задана априорная плотность распределения $\lambda(\theta) = \lambda(m_1,m_2,m_3)$, то апостериорная плотность определяется как

$$\lambda(\theta|\{X,k\}) = \frac{\Big(\prod_{\ell=1}^{3} f_{k_\ell}(X_\ell|m_\ell k_\ell)\Big)\lambda(\theta)}{P(\{X,k\})}, \tag{2.1}$$

где

$$P(\{X,k\}) = \int\limits_{\Theta} \Big(\prod_{\ell=1}^{3} f_{k_\ell}(X_\ell|m_\ell k_\ell)\Big)\lambda(\theta)d\theta, \tag{2.2}$$

$d\theta = dm_1dm_2dm_3$ и

$$f_D(x|m) = \frac{1}{(2\pi D)^{1/2}}\exp\Big(-\frac{(x-m)^2}{2D}\Big). \tag{2.3}$$

В частности, обозначим $f_D(x) = f_D(x|0)$ и $f(x) = f_1(x)$. *В дальнейшем ограничимся рассмотрением стратегий управления, которые сначала применяют каждое из трех действий k_0 раз по очереди, а затем используют всю известную предысторию процесса.* Заметим, что если $k_0 \ll K$, то такое ограничение практически не приводит к снижению общего ожидаемого дохода. Обозначим $m^* = \max(m_1,m_2,m_3)$. Тогда стандартное рекуррентное уравнение динамического программирования для нахождения байесовских стратегии и риска имеет вид

$$R^B(\{X,k\}) = \min\left(R_1^B(\{X,k\}), R_2^B(\{X,k\}), R_3^B(\{X,k\})\right), \tag{2.4}$$

где $R_1^B(\{X,k\}) = R_2^B(\{X,k\}) = R_3^B(\{X,k\}) = 0$ if $k = K$, а затем

$$\begin{aligned}
R_1^B(\{X,k\}) &= \int\limits_{\Theta} \lambda(\theta|\{X,k\})\left((m^*-m_1) + \mathbf{E}_1 R^B(X_1+x_1,k_1+1,X_2,k_2,X_3,k_3)\right)d\theta,\\
R_2^B(\{X,k\}) &= \int\limits_{\Theta} \lambda(\theta|\{X,k\})\left((m^*-m_2) + \mathbf{E}_2 R^B(X_1,k_1,X_2+x_2,k_2+1,X_3,k_3)\right)d\theta,\\
R_3^B(\{X,k\}) &= \int\limits_{\Theta} \lambda(\theta|\{X,k\})\left((m^*-m_3) + \mathbf{E}_3 R^B(X_1,k_1,X_2,k_2,X_3+x_3,k_3+1)\right)d\theta
\end{aligned} \tag{2.5}$$

при $k = K-1,\ldots,3k_0$ и $k_1 \ge k_0$, $k_2 \ge k_0$, $k_3 \ge k_0$, где $\mathbf{E}_\ell R(x) = \displaystyle\int_{-\infty}^{\infty} R(x)f_1(x|m_\ell)dx$. Значение $R_\ell^B(\{X,k\})$ равно математическому ожиданию потерь, если в момент времени $k+1$, где $k = k_1+k_2+k_3$, применяется ℓ-е действие, а затем управление выполняется оптимально. Байесовская стратегия предписывает в момент времени $k+1$ (при $k \ge 3k_0$) выбирать действие, соответствующее наименьшему значению $R_\ell^B(\{X,k\})$; если таких значений несколько, то можно выбрать любое из них. Байесовский риск (1.2) определяется как

$$R_K^B(\lambda) = k_0\int\limits_{\Theta}(3m^* - m_1 - m_2 - m_3)\,\lambda(\theta)d\theta + \mathbf{E}R^B(\{X,k_0\}), \tag{2.6}$$

где $\{X,k_0\} = (X_1,k_0,X_2,k_0,X_3,k_0)$ и

$$\begin{gathered}
\mathbf{E}R^B(\{X,k_0\}) = \iiint\limits_{A_1} R^B(\{X,k_0\})P(\{X,k_0\})dX_1dX_2dX_3,\\
A_1 = \{-\infty < X_\ell < \infty;\ \ell = 1,2,3\}.
\end{gathered} \tag{2.7}$$

Получим эквивалентную форму формул (2.4)–(2.7), которая не требует нахождения функций (2.2) и, следовательно, более удобна для вычислений. Обозначим

$$\tilde{R}(\{X,k\}) = R^B(\{X,k\})P(\{X,k\}).$$

Теорема 1. *Чтобы найти байесовские стратегию и риск, следует решить рекуррентное уравнение*

$$\tilde{R}(\{X,k\}) = \min\Big(\tilde{R}_1(\{X,k\}), \tilde{R}_2(\{X,k\}), \tilde{R}_3(\{X,k\})\Big), \tag{2.8}$$

где $\tilde{R}_1(\{X,k\}) = \tilde{R}_2(\{X,k\}) = \tilde{R}_3(\{X,k\}) = 0$, *если* $k = K$, *а затем*

$$\begin{aligned}
\tilde{R}_1(\{X,k\}) &= \tilde{G}_1(\{X,k\}) + \int\limits_{-\infty}^{\infty} \tilde{R}(X_1+x_1, k_1+1, X_2, k_2, X_3, k_3) H_1(X_1,k_1,x_1)dx_1,\\
\tilde{R}_2(\{X,k\}) &= \tilde{G}_2(\{X,k\}) + \int\limits_{-\infty}^{\infty} \tilde{R}(X_1, k_1, X_2+x_2, k_2+1, X_3, k_3) H_2(X_2,k_2,x_2)dx_2,\\
\tilde{R}_3(\{X,k\}) &= \tilde{G}_3(\{X,k\}) + \int\limits_{-\infty}^{\infty} \tilde{R}(X_1, k_1, X_2, k_2, X_3+x_3, k_3+1) H_3(X_3,k_3,x_3)dx_3
\end{aligned} \tag{2.9}$$

при $k = K-1,\ldots,3k_0$ *и* $k_1 \ge k_0$, $k_2 \ge k_0$, $k_3 \ge k_0$, *где*

$$\tilde{G}_\ell(\{X,k\}) = \int\limits_{\Theta} (m^* - m_\ell)\Big(\prod_{j=1}^{3} f_{k_j}(X_j|m_jk_j)\Big)\lambda(\theta)d\theta, \tag{2.10}$$

$$H_\ell(X_\ell,k_\ell,x_\ell) = (k_\ell+1)\, f_{k_\ell(k_\ell+1)}\left(k_\ell x_\ell - X_\ell\right). \tag{2.11}$$

Байесовская стратегия предписывает в момент времени $k+1$ (*при* $k \ge 3k_0$) *выбирать действие, соответствующее наименьшему значению* $\tilde{R}_\ell(\{X,k\})$; *если таких значений несколько, можно выбрать любое из них. Байесовский риск* (1.2) *определяется как*

$$R_K^B(\lambda) = k_0 \int\limits_{\Theta} (3m^* - m_1 - m_2 - m_3)\,\lambda(\theta)d\theta + \iiint\limits_{A_1} \tilde{R}(\{X,k_0\})dX_1dX_2dX_3. \tag{2.12}$$

Доказательство. Следует умножить левую и правую части (2.4), (2.5) на $P(\{(X,k\})$. В результате получим (2.8), (2.9), где $\tilde{G}_\ell(\{X,k\})$ вычисляется по формуле (2.10), а выражение для $H_\ell(X_\ell,k_\ell,x_\ell)$ в случае $H_1(X_1,k_1,x_1)$ имеет вид

$$\begin{aligned}
H_1(X_1,k_1,x_1) &= \frac{\displaystyle\int_\Theta \Big(\prod_{\ell=1}^{3} f_{k_\ell}(X_\ell|m_\ell k_\ell)\Big) f_1(x_1|m_1)\lambda(\theta)d\theta}{P(X_1+x_1,k_1+1,X_2,k_2,X_3,k_3)}\\
&= \frac{f_{k_1}(X_1|m_1k_1)f_1(x_1|m_1)}{f_{(k_1+1)}(X_1+x_1|m_1(k_1+1))} = \Big(\frac{k_1+1}{2\pi k_1}\Big)^{1/2}\exp\Big(-\frac{(k_1x_1-X_1)^2}{2k_1(k_1+1)}\Big)\\
&= (k_1+1)\, f_{k_1(k_1+1)}\left(k_1x_1 - X_1\right).
\end{aligned}$$

Это соответствует (2.11). Формула (2.12) следует из (2.6), (2.7). □

3. Свойства наихудшего априорного распределения

Дадим краткую характеризацию класса априорных распределений, к которому принадлежит наихудшее, т. е. априорных распределений, на которых байесовский риск максимален. Это можно сделать более подробно, чем представлено ниже, аналогично результатам в [8;19]. Прежде всего отметим, что байесовский риск является вогнутой функцией априорного распределения, т. е. справедлива

Лемма 1. *Пусть $\lambda_1(\theta)$, $\lambda_2(\theta)$ являются априорными плотностями распределения. Тогда*

$$R_K^B\left(\alpha_1\lambda_1+\alpha_2\lambda_2\right)\geq\alpha_1R_K^B\left(\lambda_1\right)+\alpha_2R_K^B\left(\lambda_2\right).$$

Доказательство следует из цепочки (не)равенств

$$R_K^B\left(\alpha_1\lambda_1+\alpha_2\lambda_2\right)=\inf_\sigma\int_\Theta L_K(\sigma,\theta)\left(\alpha_1\lambda_1(\theta)+\alpha_2\lambda_2(\theta)\right)d\theta$$

$$\geq\alpha_1\inf_\sigma\int_\Theta L_K(\sigma,\theta)\lambda_1(\theta)d\theta+\alpha_2\inf_\sigma\int_\Theta L_K(\sigma,\theta)\lambda_2(\theta)d\theta=\alpha_1R_K^B\left(\lambda_1\right)+\alpha_2R_K^B\left(\lambda_2\right).$$

В дальнейшем удобно изменить параметризацию и положить $m_\ell=m+v_\ell$, $\ell=1,2,3$; при этом $v_1+v_2+v_3=0$. В новых переменных рассмотрим априорную плотность распределения $\nu(m,v_1,v_2,v_3)$. Для некоторой плотности $\nu(\theta)$ определим плотность $\nu_c(\theta)$ условием $\nu_c(m,v_1,v_2,v_3)=\nu(m+c,v_1,v_2,v_3)$. Рассмотрим также плотности вида $\nu_{12}(m,v_1,v_2,v_3)=\nu(m,v_2,v_1,v_3)$, $\nu_{13}(m,v_1,v_2,v_3)=\nu(m,v_3,v_2,v_1)$, $\nu_{23}(m,v_1,v_2,v_3)=\nu(m,v_1,v_3,v_2)$.

Лемма 2. *Справедливы равенства*

$$R_K^B(\nu_c)=R_K^B(\nu),R_K^B(\nu_{ij})=R_K^B(\nu)$$

для любого c и всех ij. Байесовская стратегия, соответствующая плотности ν_c, на первом шаге выбирает то же действие, что и стратегия, соответствующая плотности ν.

Доказательство выполняется прямым использованием уравнения (2.4), (2.5). Однако эти свойства интуитивно понятны. В первом случае все доходы изменяются на одну и ту же величину c, что не приведет к изменению потерь относительно максимально возможного дохода. Во втором случае некоторые два действия меняются местами одновременно с соответствующими им распределениями. □

Лемма 3. *Априорные плотности распределения, к которым относится наихудшая, асимптотически равномерны по m и симметричны относительно v_1,v_2,v_3, т. е.*

$$\nu_a(m,v_1,v_2,v_3)=\kappa_a(m)\rho(v_1,v_2,v_3),\tag{3.1}$$

где $\kappa_a(m)$ — равномерная плотность распределения на отрезке $m\in[-a,a]$, причем a достаточно велико, а $\rho_{ij}(v_1,v_2,v_3)=\rho(v_1,v_2,v_3)$ для всех ij.

Доказательство. Напомним, что множество Θ таково, что $|m|\leq C_1$. Для $a\gg C_1$ и некоторой плотности $\nu(m,v_1,v_2,v_3)$ рассмотрим $\tilde\nu(m,v_1,v_2,v_3)=(2a)^{-1}\int_{-a}^{a}\nu(m+x,v_1,v_2,v_3)dx$. Легко видеть, что эта плотность постоянна по m при $|m|\leq a-C_1$, уменьшается до 0, если $|m|\in[a-C_1,a+C_1]$, и равна 0 при $|m|>a+C_1$. Поскольку $C_1\ll a$, эта плотность может быть сколь угодно точно приближена плотностью (3.1). Согласно лемме 1 байесовский риск для нее не меньше исходного. Аналогично плотность $(1/6)\kappa_a(m)(\rho(v_1,v_2,v_3)+$

$\rho(v_1,v_3,v_2)+\rho(v_2,v_1,v_3)+\rho(v_2,v_3,v_1)+\rho(v_3,v_1,v_2)+\rho(v_3,v_2,v_1))$ симметрична относительно v_1,v_2,v_3, и в силу леммы 1 байесовский риск на ней не меньше исходного. □

З а м е ч а н и е 1. Из леммы 3 следует, что для любого априорного распределения $\nu(m,v_1,v_2,v_3)$ на множестве Θ, можно определить априорное распределение $\nu_a(m,v_1,v_2,v_3)$ вида (3.1), на котором байесовский риск не меньше исходного. Однако его носитель, т. е. множество параметров, на которым задано $\nu_a(m,v_1,v_2,v_3)$, отличается от Θ. Как и в [8;19], можно доказать, что байесовский риск $R_K^B(\nu_a(m,v_1,v_2,v_3))$ имеет предел при $a\to\infty$, и, следосвательно, близкое к наихудшему априорное распределение может быть определено на исходном множестве Θ, если C_1 достаточно велико.

Следующая лемма является вспомогательной.

Лемма 4. *Если $k_\ell>0$, $\ell=1,2,3$, то справедливо равенство*

$$\prod_{\ell=1}^{3} f_{k_\ell}(X_\ell|m_\ell k_\ell)=f_{k^{-1}}(m+\hat v-Y)\tilde f(\{Z,k\},\{v\}) \tag{3.2}$$

с функциями $f_{k_\ell}(X_\ell|m_\ell k_\ell)$, $f_{k^{-1}}(m+\hat v-Y)$, определенными в соответствии с (2.3),

$$\begin{gathered}\tilde f(\{Z,k\},\{v\})=\frac{1}{2\pi(k_1k_2k_3k)^{1/2}}\\ \times\exp\left(-\frac{k_1k_2(Z_1-v_{12})^2+k_2k_3(Z_2-v_{23})^2+k_3k_1(Z_3-v_{31})^2}{2k}\right),\end{gathered} \tag{3.3}$$

где $\{Z,k\}=(Z_1,k_1,Z_2,k_2,Z_3,k_3)$, $\{v\}=(v_1,v_2,v_3)$, $Z_1=\overline{X}_1-\overline{X}_2$, $Z_2=\overline{X}_2-\overline{X}_3$, $Z_3=\overline{X}_3-\overline{X}_1$, $Y=(X_1+X_2+X_3)/k$, $\hat v=(v_1k_1+v_2k_2+v_3k_3)/k$, $v_{ij}=v_i-v_j$.

Д о к а з а т е л ь с т в о. Непосредственно проверяется равенство

$$\sum_i k_i(m-a_i)^2=k\Big(m-\frac{\sum_i a_ik_i}{k}\Big)^2+\frac{\sum_{i>j}k_ik_ja_{ij}^2}{k}, \tag{3.4}$$

где $k=\sum_i k_i$, $a_{ij}=a_i-a_j$. Справедливость леммы 4 проверяется с использованием (3.4) и представления $f_{k_i}(X_i|(m+v_i)k_i)=(2\pi k_i)^{-1/2}\exp\left(-k_i(m-a_i)^2/2\right)$, где $a_i=\overline{X}_i-v_i$, $\overline{X}_i=X_i/k_i$, в левой части (3.2). □

Для получения рекуррентного уравнения относительно наихудшей априорной плотности распределения (3.1) сделаем несколько предварительных замечаний. Во-первых, отметим, что из (2.1) следует, что плотность апостериорного распределения не изменится, если априорную плотность $\lambda(\theta)$ (или, что эквивалентно, $\nu(m,v_1,v_2,v_3)$) умножить на произвольную постоянную. Значит, можно формально рассмотреть априорную плотность вида (3.1) с $\kappa(m)=1$ для всех $m\in(-\infty,\infty)$, т. е. рассмотреть множество параметров Θ с $C_1=\infty$.

Во-вторых, вспомним, что проанализированные стратегии применяют все действия k_0 раз по очереди в начале управления. Учитывая сделанные выше комментарии и равенство (3.2), получим, что апостериорная плотность (2.1), соответствующая априорной плотности распределения (3.1) с $\kappa(m)=1$, имеет вид

$$\nu(m,\{v\}|\{Z,k\},Y)=\frac{f_{k^{-1}}(m+\hat v-Y)\tilde f(\{Z,k\},\{v\})\rho(\{v\})}{P(\{Z,k\})} \tag{3.5}$$

с $\tilde f(\{Z,k\},\{v\})$, представленной в (3.3), и

$$\begin{aligned}P(\{Z,k\})&=\int_{-\infty}^{\infty}\iint_{\Theta_v} f_{k^{-1}}(m+\hat v-Y)\tilde f(\{Z,k\},\{v\})\rho(\{v\})dv_1\,dv_2dm\\ &=\iint_{\Theta_v}\tilde f(\{Z,k\},\{v\})\rho(\{v\})dv_1\,dv_2,\end{aligned}$$

где $\{Z,k\}$, $\{v\}$, Z_1, Z_2, Z_3, Y, $\hat{v}$, v_{ij} определены в лемме 4, а $\Theta_v = \{|v_\ell| \le C,\ \ell = 1,2,3\}$. Отметим, что $Z_1+Z_2+Z_3 = 0$, поэтому только две переменные, например Z_1 и Z_2, независимы.

В-третьих, из (3.5) видно, что плотности $\nu(m,\{v\}|\{Z,k\},Y)$ для различных Y могут быть получены друг из друга путем сдвига по m, т. е. имеет место равенство $\nu(m,\{v\}|\{Z,k\},Y) = \nu_{-Y}(m,\{v\}|\{Z,k\},0)$. В силу леммы 2 байесовские стратегии и риски, соответствующие этим плотностям, не зависят от Y, т. е. $R^B(\{X,k\}) = R^B(\{Z,k\},Y) = R^B(\{Z,k\})$. Выбирая Z_1, Z_2, Y в качестве независимых переменных, согласно (2.7) получаем равенство

$$\begin{aligned}\mathbf{E}R^B(\{X,k_0\}) &= k_0^3 \iint\limits_{A_2} \mathbf{E}_Y R^B\left(\{Z,k_0\},Y\right) P(\{Z,k_0\}) dZ_1 dZ_2 \\ &= k_0^3 \iint\limits_{A_2} \tilde{R}(\{Z,k_0\}) dZ_1 dZ_2,\end{aligned} \tag{3.6}$$

где $A_2 = \{-\infty < Z_\ell < \infty;\ \ell = 1,2\}$, а $k_0^3 = \partial(X_1,X_2,X_3)/\partial(Z_1,Z_2,Y)$ — якобиан преобразования переменных при $k_1 = k_2 = k_3 = k_0$.

4. Рекуррентное уравнение относительно наихудшего априорного распределения

Получим рекуррентное уравнение для нахождения байесовских стратегии и риска относительно плотности распределения вида (3.1).

Теорема 2. *Чтобы найти байесовские стратегию и риск относительно плотности распределения вида* (3.1)*, следует решить рекуррентное уравнение*

$$\tilde{R}(\{Z,k\}) = \min\left(\tilde{R}_1(\{Z,k\}), \tilde{R}_2(\{Z,k\}), \tilde{R}_3(\{Z,k\})\right), \tag{4.1}$$

где $\tilde{R}_1(\{Z,k\}) = \tilde{R}_2(\{Z,k\}) = \tilde{R}_3(\{Z,k\}) = 0$*, если* $k = K$*, а затем*

$$\begin{aligned}\tilde{R}_1(\{Z,k\}) &= \tilde{G}_1(\{Z,k\}) + \int\limits_{-\infty}^{\infty} \tilde{R}(Z_1+z_1,k_1+1,Z_2,k_2,Z_3-z_1,k_3) H_1(k_1,z_1)dz_1, \\ \tilde{R}_2(\{Z,k\}) &= \tilde{G}_2(\{Z,k\}) + \int\limits_{-\infty}^{\infty} \tilde{R}(Z_1-z_2,k_1,Z_2+z_2,k_2+1,Z_3,k_3) H_2(k_2,z_2)dz_2, \\ \tilde{R}_3(\{Z,k\}) &= \tilde{G}_3(\{Z,k\}) + \int\limits_{-\infty}^{\infty} \tilde{R}(Z_1,k_1,Z_2-z_3,k_2,Z_3+z_3,k_3+1) H_3(k_3,z_3)dz_3\end{aligned} \tag{4.2}$$

при $k = K-1,\dots,3k_0$ *и* $k_1 \ge k_0$, $k_2 \ge k_0$, $k_3 \ge k_0$, *где*

$$\tilde{G}_\ell(\{Z,k\}) = \iint\limits_{\Theta_v} (v^* - v_\ell)\tilde{f}(\{Z,k\},\{v\})\rho(\{v\})dv_1 dv_2, \tag{4.3}$$

$$H_\ell(k_\ell,z_\ell) = (k_\ell+1)k_\ell^{-1} f_{1/(k_\ell(k_\ell+1))}(z_\ell), \tag{4.4}$$

$v^* = \max(v_1,v_2,v_3)$, *а* $\tilde{f}(\{Z,k\},\{v\})$ *представлена в* (3.3)*. Байесовская стратегия в момент времени* $k+1$ (*при* $k \ge 3k_0$) *предписывает выбирать действие, соответствующее наименьшему значению* $\tilde{R}_\ell(\{Z,k\})$*; если таких значений несколько, то можно выбрать любое из них. Байесовский риск* (1.2) *определяется по формуле*

$$R_K^B(\nu) = k_0 \iint\limits_{\Theta_v} 3v^*\rho(\{v\})dv_1 dv_2 + k_0^3 \iint\limits_{A_2} \tilde{R}(\{Z,k_0\})dZ_1 dZ_2. \tag{4.5}$$

Д о к а з а т е л ь с т в о. Рассмотрим первое уравнение (2.9). Так как $(X_1+x_1)/(k_1+1)-X_1/k_1=(k_1x_1-X_1)/(k_1(k_1+1))=z_1$, то $\tilde R(X_1+x_1,k_1+1,X_2,k_2,X_3,k_3)=\tilde R(Z_1+z_1,k_1+1,Z_2,k_2,Z_3-z_1,k_3)$, $dx_1=(k_1+1)dz_1$, $H_1(X_1,k_1,x_1)dx_1=\left[k_1^{-1}f_{1/(k_1(k_1+1))}(z_1)\right][(k_1+1)dz_1]=H_1(k_1,z_1)dz_1$. Далее, с учетом (3.2) и предполагая $C_1=\infty$ в Θ, имеем

$$\tilde G_\ell(\{Z,k\})=\iiint\limits_{\Theta}(v^*-v_\ell)f_{k^{-1}}(m+\hat v-Y)\tilde f(\{Z,k\},\{v\})\rho(\{v\})dv_1\,dv_2dm$$
$$=\iint\limits_{\Theta_v}(v^*-v_\ell)\tilde f(\{Z,k\},\{v\})\rho(\{v\})dv_1\,dv_2,$$

откуда следует справедливость (4.2)–(4.4) для первого уравнения. Остальные два уравнения рассматриваются аналогично. Формула (4.5) следует из (2.6), (3.6), согласно равенству $v_1+v_2+v_3=0$.

Следствие 1. *Байесовские стратегия и риск, вычисленные относительно плотности распределения вида* (3.1)*, могут быть найдены в результате решения рекуррентного уравнения*

$$R(\{Z,k\})=\min\left(R_1(\{Z,k\}),R_2(\{Z,k\}),R_3(\{Z,k\})\right),\tag{4.6}$$

где $R_1(\{Z,k\})=R_2(\{Z,k\})=R_3(\{Z,k\})=0$*, если* $k=K$*, а затем*

$$R_1(\{Z,k\})=G_1(\{Z,k\})+\int\limits_{-\infty}^{\infty}R(Z_1+z_1,k_1+1,Z_2,k_2,Z_3-z_1,k_3)f_{1/(k_1(k_1+1))}(z_1)dz_1,$$
$$R_2(\{Z,k\})=G_2(\{Z,k\})+\int\limits_{-\infty}^{\infty}R(Z_1-z_2,k_1,Z_2+z_2,k_2+1,Z_3,k_3)f_{1/(k_2(k_2+1))}(z_2)dz_2,\tag{4.7}$$
$$R_3(\{Z,k\})=G_3(\{Z,k\})+\int\limits_{-\infty}^{\infty}R(Z_1,k_1,Z_2-z_3,k_2,Z_3+z_3,k_3+1)f_{1/(k_3(k_3+1))}(z_3)dz_3$$

при $k=K-1,\ldots,3k_0$ *и* $k_1\ge k_0$, $k_2\ge k_0$, $k_3\ge k_0$, *где* $G_\ell(\{Z,k\})=k_1k_2k_3\tilde G_\ell(\{Z,k\})$ *и* $\tilde G_\ell(\{Z,k\})$ *определены в* (4.3)*. Байесовская стратегия в момент времени* $k+1$ *(при* $k\ge3k_0$*) предписывает выбирать действие, соответствующее наименьшему значению* $R_\ell(\{Z,k\})$*; если таких значений несколько, то можно выбрать любое из них. Байесовский риск* (1.2) *вычисляется как*

$$R_K^B(\nu)=k_0\iint\limits_{\Theta_v}3v^*\rho(\{v\})dv_1\,dv_2+\iint\limits_{A_2}R(\{Z,k_0\})dZ_1\,dZ_2.\tag{4.8}$$

Формулы (4.6)–(4.8) *более удобны при переходе к предельному описанию интегро-разностного уравнения.*

Д о к а з а т е л ь с т в о. Следует положить

$$\tilde R(\{Z,k\})=(k_1k_2k_3)^{-1}R(\{Z,k\}),\quad \tilde R_\ell(\{Z,k\})=(k_1k_2k_3)^{-1}R_\ell(\{Z,k\}),$$
$$\tilde G_\ell(\{Z,k\})=(k_1k_2k_3)^{-1}G_\ell(\{Z,k\})$$

в (4.1)–(4.2) и (4.5).

Нетрудно понять, как следует вычислять функцию потерь. □

Следствие 2. *Рассмотрим вырожденную плотность распределения* $\rho(\{v\})$, *сосредоточенную на параметрах* $\theta = (m+v_1, m+v_2, m+v_3)$, $m \in (-\infty, \infty)$, *и решим рекуррентное уравнение*

$$L(\{Z,k\}) = \sum_{\ell=1}^{3} \sigma_\ell(\{Z,k\}) L_\ell(\{Z,k\}), \tag{4.9}$$

где $L_1(\{Z,k\}) = L_2(\{Z,k\}) = L_3(\{Z,k\}) = 0$ *при* $k = K$, *а затем*

$$\begin{aligned}
L_1(\{Z,k\}) &= G_1(\{Z,k\}) + \int_{-\infty}^{\infty} L(Z_1+z_1, k_1+1, Z_2, k_2, Z_3-z_1, k_3) f_{1/(k_1(k_1+1))}(z_1)dz_1, \\
L_2(\{Z,k\}) &= G_2(\{Z,k\}) + \int_{-\infty}^{\infty} L(Z_1-z_2, k_1, Z_2+z_2, k_2+1, Z_3, k_3) f_{1/(k_2(k_2+1))}(z_2)dz_2, \\
L_3(\{Z,k\}) &= G_3(\{Z,k\}) + \int_{-\infty}^{\infty} L(Z_1, k_1, Z_2-z_3, k_2, Z_3+z_3, k_3+1) f_{1/(k_3(k_3+1))}(z_3)dz_3
\end{aligned} \tag{4.10}$$

при $k = K-1, \ldots, 3k_0$ *и* $k_1 \geq k_0$, $k_2 \geq k_0$, $k_3 \geq k_0$. *Функция потерь* (1.1) *определяется как*

$$L_K(\sigma, \theta) = k_0 \iint_{\Theta_v} 3v^* \rho(\{v\}) dv_1 dv_2 + \iint_{A_2} L(\{Z, k_0\}) dZ_1 dZ_2. \tag{4.11}$$

З а м е ч а н и е 2. Дадим оценку временно́й сложности численного решения уравнения (4.6)–(4.7). Отметим, что функции $\{G_\ell(\{Z,k\})\}$ фактически не требуют интегрирования, поскольку наихудшая априорная плотность $\rho(\{v\})$ является вырожденной. Поэтому численное интегрирование требуется только для второго слагаемого в правой части (4.7). Предположим, что для одного численного интегрирования необходимо N операций, соответствующих общему количеству используемых дискретных значений каждой переменной Z_ℓ. Следовательно, для некоторого фиксированного набора целых чисел $k_1, \ldots, k_J$, количество операций для вычисления $\{R_\ell(\{Z,k\})\}$ для всех $Z_1, \ldots, Z_{J-1}$ и ℓ может быть оценено как JN^J.

Обозначим через $n(J,k)$ количество наборов неотрицательных целых чисел $k_1, \ldots, k_J$, удовлетворяющих условию $k_1 + \cdots + k_J = k$. Тогда для достаточно большого k имеем

$$n(2,k) = k+1, \quad n(3,k) = \sum_{j=0}^{k} n(2,j) \sim k^2/2, \ldots, \quad n(J,k) = \sum_{j=0}^{k} n(J-1,j) \sim k^{J-1}/(J-1)!.$$

Поскольку в начале управления все действия выполняются k_0 раз по очереди, общее количество наборов $k_1, \ldots, k_J$, для которых необходимо решить уравнение, имеет порядок $(K-k_0J)^J/J!$. Значит, общее количество вычислений можно оценить как $N^J(K-k_0J)^J/(J-1)!$.

С другой стороны, эти вычисления допускают распараллеливание. Для заданного момента времени k все $\{R_\ell(\{Z,k\})\}$, такие что $k_1 + \cdots + k_J = k$, могут быть вычислены параллельно. В этом случае общее время, необходимое для вычислений, пропорционально $N(K-k_0J)$.

5. Предельное описание. Дифференциальное уравнение в частных производных

Получим сначала инвариантную форму формул (4.6)–(4.8) с горизонтом управления равным единице. Соответствующее уравнение справедливо в области “близких распределений”, которая обсуждалась в разд. 1 для бернуллиевского многорукого бандита. Для гауссовского трехрукого бандита она описывается множеством параметров $\Theta_v = \{|v_\ell| \leq cK^{-1/2}, \ell = 1,2,3\}$,

где $c > 0$ — достаточно большая фиксированная константа. Сделаем замены: $v_\ell = K^{-1/2}w_\ell$, $\rho(v_1, v_2, v_3) = K\varrho(w_1, w_2, w_3)$, $Z_\ell = S_\ell K^{-1/2}$, $z_\ell = s_\ell K^{-1/2}$, $k_\ell = t_\ell K$, $\varepsilon = K^{-1}$, $k_0 = t_0 K$, $R(\{Z, k\}) = K^{3/2} r(\{S, t\})$, $G_\ell(\{Z, k\}) = K^{1/2} g_\ell(\{S, t\})$. Справедлива следующая теорема.

Теорема 3. *Байесовские стратегия и риск, вычисленные относительно распределения вида (3.1), могут быть найдены в результате решения рекуррентного уравнения*

$$r(\{S,t\}) = \min\big(r_1(\{S,t\}), r_2(\{S,t\}), r_3(\{S,t\})\big), \tag{5.1}$$

где $r_1(\{S,t\}) = r_2(\{S,t\}) = r_3(\{S,t\}) = 0$, *если* $t = t_1 + t_2 + t_3 = 1$, *а затем*

$$\begin{aligned}
r_1(\{S,t\}) &= \varepsilon g_1(\{S,t\}) + \int\limits_{-\infty}^{\infty} r(S_1 + s_1, t_1 + \varepsilon, S_2, t_2, S_3 - s_1, t_3) f_{\varepsilon/(t_1(t_1+\varepsilon))}(s_1) ds_1, \\
r_2(\{S,t\}) &= \varepsilon g_2(\{S,t\}) + \int\limits_{-\infty}^{\infty} r(S_1 - s_2, t_1, S_2 + s_2, t_2 + \varepsilon, S_3, t_3) f_{\varepsilon/(t_2(t_2+\varepsilon))}(s_2) ds_2, \\
r_3(\{S,t\}) &= \varepsilon g_3(\{S,t\}) + \int\limits_{-\infty}^{\infty} r(S_1, t_1, S_2 - s_3, t_2, S_3 + s_3, t_3 + \varepsilon) f_{\varepsilon/(t_3(t_3+\varepsilon))}(s_3) ds_3
\end{aligned} \tag{5.2}$$

при $t = 1 - \varepsilon, \ldots, 3t_0$ *и* $t_1 \geq t_0$, $t_2 \geq t_0$, $t_3 \geq t_0$, *где*

$$g_\ell(\{S,t\}) = t_1 t_2 t_3 \iint\limits_{\Theta_w} (w^* - w_\ell) \tilde{f}(\{S,t\}, \{w\}) \varrho(\{w\}) dw_1 dw_2, \tag{5.3}$$

с $\Theta_w = \{|w_\ell| \leq c,\ \ell = 1,2,3\}$ *и* $w_{ij} = w_i - w_j$. *Байесовская стратегия в момент времени* $t + \varepsilon$ (*при* $t \geq 3t_0$) *предписывает выбирать действие, соответствующее наименьшему значению* $r_\ell(\{S,t\})$; *если таких значений несколько, то можно выбрать любое из них. Байесовский риск (1.2) определяется по формуле*

$$R_K^B(\nu) = K^{1/2}\left(t_0 \iint\limits_{\Theta_w} 3w^* \varrho(\{w\}) dw_1 dw_2 + \iint\limits_{A_3} r(\{S, t_0\}) dS_1 dS_2\right), \tag{5.4}$$

где $A_3 = \{-\infty < S_\ell < \infty;\ \ell = 1,2\}$.

Доказательство. Следует выполнить указанные выше замены в (4.6)–(4.8).

Получим дифференциальное уравнение в предельном случае при $\varepsilon \to 0$. Рассмотрим первое уравнение (5.2). Предположим, что $r(S_1, t_1, S_2, t_2, S_3, t_3)$ имеет частные производные требуемых порядков по всем переменным. Представляя $r(S_1 + s_1, t_1 + \varepsilon, S_2, t_2, S_3 - s_1, t_3)$ в виде ряда Тейлора в окрестности точки $(S_1, t_1 + \varepsilon, S_2, t_2, S_3, t_3)$ и используя равенства

$$\int\limits_{-\infty}^{\infty} f_\varepsilon(s) ds = 1, \quad \int\limits_{-\infty}^{\infty} s f_\varepsilon(s) ds = 0, \quad \int\limits_{-\infty}^{\infty} s^2 f_\varepsilon(s) ds = \varepsilon,$$

имеем, что первое уравнение принимает вид

$$r_1(\{S,t\}) = \varepsilon g_1(\{S,t\}) + r(S_1, t_1 + \varepsilon, S_2, t_2, S_3, t_3) + \frac{\varepsilon}{2t_1(t_1+\varepsilon)}\left(r''_{S_1S_1} - 2r''_{S_1S_3} + r''_{S_3S_3}\right) + o(\varepsilon).$$

Выписывая аналогичные уравнения для второго и третьего уравнений (5.2) и дополняя их уравнением (5.1), представленным в виде $\min_{\ell=1,2,3}\left(r_\ell(\{S,t\}) - r(\{S,t\})\right) = 0$, получаем в пределе при $\varepsilon \to 0$ дифференциальное уравнение в частных производных второго порядка

$$\min_{\ell=1,2,3}\left\{\frac{\partial r}{\partial t_\ell} + \frac{1}{2t_\ell^2}\left(\frac{\partial^2 r}{\partial S_\ell^2} - 2\frac{\partial^2 r}{\partial S_\ell \partial S_{\overline{\ell}}} + \frac{\partial^2 r}{\partial S_{\overline{\ell}}^2}\right) + g_\ell\left(\{S,t\}\right)\right\} = 0, \tag{5.5}$$

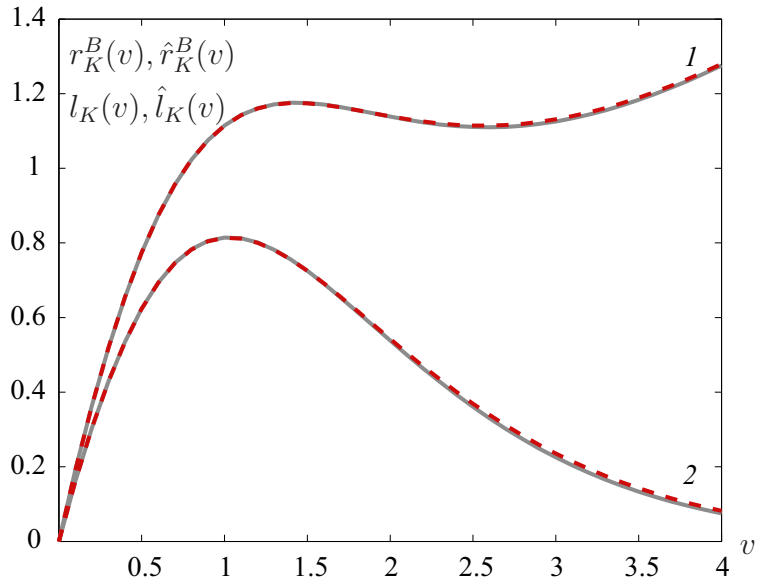

Рис. 1. Нормализованные байесовские риски и потери как функции v.

где $\overline{\ell}(1) = 3$, $\overline{\ell}(2) = 1$, $\overline{\ell}(3) = 2$ (или, эквивалентно, $\overline{\ell} - 1 \equiv \ell + 1 \,(\mathrm{mod}\ 3)$) и $g_\ell(\{S,t\})$ даны в (5.3). Начальное условие имеет вид $r(S_1,t_1,S_2,t_2,S_3,t_3) = 0$ при $t = t_1 + t_2 + t_3 = 1$. Дифференциальное уравнение в частных производных (5.5) следует решать в обратном времени при $3t_0 \le t \le 1$, $t_\ell \ge t_0$, $\ell = 1,2,3$. Байесовская стратегия предписывает выбирать действие, которое соответствует наименьшему текущему значению выражения в левой части (5.5); если таких значений несколько, то можно выбрать любое из них. Байесовский риск (1.2) дается выражением (5.4). □

З а м е ч а н и е 3. В случае $J > 3$ байесовские стратегию и риск, вычисленные относительно распределения типа (3.1), можно найти, решив рекуррентное уравнение

$$r(\{S,t\}) = \min_{\ell=1,\dots,J} r_\ell(\{S,t\}),$$

где $r_\ell(\{S,t\}) = 0$, $\ell = 1,\dots,J$, если $t = t_1 + \dots + t_J = 1$, а затем

$$r_\ell(\{S,t\}) = \varepsilon g_\ell(\{S,t\}) + \int_{-\infty}^{\infty} r(\dots, S_\ell + s_\ell, t_\ell + \varepsilon, \dots, S_{\overline{\ell}} - s_\ell, t_{\overline{\ell}}, \dots) f_{\varepsilon/(t_\ell(t_\ell+\varepsilon))}(s_\ell) ds_\ell \qquad (5.6)$$

при $t = 1-\varepsilon,\dots,Jt_0$ и $t_\ell \ge t_0$, $\ell = 1,\dots,J$, где $\overline{\ell} - 1 \equiv \ell + J - 2 \,(\mathrm{mod}\ J)$ и

$$g_\ell(\{S,t\}) = (t_1 \dots t_J) \int\cdots\int_{\Theta_w} (w^* - w_\ell) \tilde{f}(\{S,t\},\{w\}) \varrho(\{w\}) dw_1 \dots dw_{J-1} \qquad (5.7)$$

с $\Theta_w = \{|w_\ell| \le c, \ell = 1,\dots,J\}$ и $w^* = \max(w_1,\dots,w_J)$. Байесовская стратегия в момент времени $t+1$ (при $t \ge Jt_0$) предписывает выбирать действие, соответствующее наименьшему значению $r_\ell(\{S,t\})$; если таких значений несколько, то можно выбрать любое из них. Байесовский риск (1.2) может быть найден по формуле

$$R_K^B(\nu) = K^{1/2}\left(t_0 \int\cdots\int_{\Theta_w} Jw^* \varrho(\{w\}) dw_1 \dots dw_{J-1} + \int\cdots\int_{A_J} r(\{S,t_0\}) dS_1 \dots dS_{J-1}\right),$$

где $A_J = \{-\infty < S_\ell < \infty;\ \ell = 1, \ldots, J-1\}$. В пределе при $\varepsilon \to 0$ рекуррентное уравнение (5.6) превращается в дифференциальное уравнение в частных производных

$$\min_{\ell=1,\ldots,J} \left\{ \frac{\partial r}{\partial t_\ell} + \frac{1}{2t_\ell^2} \left(\frac{\partial^2 r}{\partial S_\ell^2} - 2\frac{\partial^2 r}{\partial S_\ell \partial S_{\overline{\ell}}} + \frac{\partial^2 r}{\partial S_{\overline{\ell}}^2} \right) + g_\ell\left(\{S,t\}\right) \right\} = 0$$

с начальным условием $r(\{S,t\}) = 0$ при $t_1 + \cdots + t_J = 1$.

Здесь

$$\{S,t\} = (S_1, t_1, \ldots, S_J, t_J), \quad \{w\} = (w_1, \ldots, w_J), \quad S_j = \overline{x}_j - \overline{x}_{j+1},$$
$$j = 1, \ldots, J-1, \quad S_J = \overline{x}_J - \overline{x}_1, \quad \text{где} \ \ \overline{x}_j = K^{1/2}(X_j/k_j), \quad j = 1, \ldots, J.$$

Таким образом, $S_1 + \cdots + S_J = 0$, $w_1 + \cdots + w_J = 0$. Функция $\tilde{f}(\{S,t\},\{w\})$ в (5.7) может быть определена аналогично тому, как это сделано в лемме 4:

$$\tilde{f}(\{S,t\},\{w\}) = \frac{1}{(2\pi)^{(J-1)/2}(t_1 t_2 \ldots t_J t)^{1/2}} \exp\Big(- \frac{\sum_{i>j} t_i t_j (\overline{x}_i - \overline{x}_j - w_{ij})^2}{2t} \Big),$$

где переменные $\{\overline{x}_i\}$ должны быть выражены через $\{S_i\}$.

6. Численные результаты

При проведении численных экспериментов использовалось следующее важное свойство: функция потерь, соответствующая минимаксной стратегии, принимает максимальные значения для параметров, на которых сосредоточено наихудшее априорное распределение. Поэтому с учетом симметрии предполагалось, что плотность наихудшего априорного распределения $\rho(v_1, v_2, v_3)$ является вырожденной и сосредоточена в трех точках $(2d, -d, -d)$, $(-d, 2d, -d)$, $(-d, -d, 2d)$ с вероятностями $1/3$, где $d = vK^{-1/2}$. В системе координат (m_1, m_2, m_3) это распределение сосредоточено на трех прямых, параллельных главной диагонали и пересекающих оси в точках $(3d, 0, 0)$, $(0, 3d, 0)$, $(0, 0, 3d)$. Неизвестное значение v было найдено из условия, что наихудшее априорное распределение соответствует максимуму байесовского риска, который, согласно основной теореме теории игр совпадает с минимаксным.

Поскольку массивы данных в случае трехрукого бандита велики и, следовательно, объем вычислений также велик, расчеты были выполнены при $K = 20$. На рис. 1 сплошными линиями *1* и *2* представлены нормализованные байесовские риски

$$r_K^B(v) = K^{-1/2} R_K^B(\nu) \quad \text{и} \quad \hat{r}_K^B(v) = K^{-1/2} \iint\limits_{-\infty}^{\infty} R(\{Z, k_0\}) dZ_1 dZ_2$$

соответственно, где второе выражение не учитывает потери на начальном этапе управления, когда действия применяются однократно по очереди, и, значит, имеет меньшее значение. Эти риски вычислялись с использованием формул (4.6)–(4.8) в диапазоне $0.1 \le v \le 4.0$ с шагом 0.1, а в окрестностях максимумов — с шагом 0.01. В первом случае внутренний максимум равен приблизительно 1.18 и достигается в точке $v \approx 1.43$, во втором случае максимум равен приблизительно 0.81 и достигается в точке $v \approx 1.03$. Затем для найденных байесовских стратегий с использованием формул (4.9)–(4.11) были рассчитаны значения функций потерь

$$l_K(v) = K^{-1/2} L_K(\sigma, \theta) \quad \text{и} \quad \hat{l}_K(v) = K^{-1/2} \iint\limits_{-\infty}^{\infty} L(\{Z, k_0\}) dZ_1 dZ_2$$

в диапазоне $0.1 \le v \le 4.0$ с шагом 0.1, на рис. 1 они представлены пунктирными линиями *1* и *2* соответственно. Видно, что максимальные значения потерь не превышают соответствующих

максимальных значений байесовских рисков, т. е. найденные стратегии являются минимаксными (вторая — на горизонте управления $k = 4, \ldots, 20$, т. е. без учета потерь на начальном этапе, когда действия применяются по очереди).

Для выполнения расчетов использовалась программа, разработанная в среде C++. Риски $R(\{Z, k\})$ и потери $L(\{Z, k\})$ представлялись массивами, заданными по Z_1, Z_2 в диапазоне от -5.6 до $+5.6$ с шагом 0.08, численное интегрирование выполнялось методом прямоугольников. Для вычисления массива функций потерь модифицировалась программа для вычисления байесовского риска. Поскольку эта программа также находит и байесовскую стратегию, то при вычислении байесовского риска, соответствующего наихудшему априорному распределению, для соответствующей байесовской стратегии одновременно вычислялся массив функций потерь для всего массива значений v.

Отметим, что значения функции потерь почти совпали с рисками (разница не превышает величины 0.01), что, по-видимому, можно объяснить тем фактом, что оптимальная стратегия при $K = 20$ мало зависит от v в указанном диапазоне. В частности, можно проверить, что на последнем шаге она является следующей: необходимо выбирать первое действие, если $Z_1 > 0$, $Z_3 < 0$; второе действие, если $Z_2 > 0$, $Z_1 < 0$; третье действие, если $Z_3 > 0$, $Z_2 < 0$; на границах можно выбирать любое действие из смежных областей. Таким образом, в данном случае оптимальная стратегия вообще не зависит от v. Наконец, отметим, что при проверке оптимальности стратегии точки вида $(2v, -v+x, -v+y)$, $(-v+x, 2v, -v+y)$, $(-v+x, -v+y, 2v)$ также рассматривались при различных x, y в окрестности максимума. Значения функции потерь в этих точках не превышали максимума.

СПИСОК ЛИТЕРАТУРЫ

1. **Berry D.A., Fristedt B.** Bandit problems: sequential allocation of experiments. London, NY: Chapman and Hall, 1985. 275 p. https://doi.org/10.1007/978-94-015-3711-7
2. **Lattimore T., Szepesvari C.** Bandit algorithms. Cambridge: Cambridge Univ. Press, 2020. 586 p. https://doi.org/10.1017/9781108571401
3. **Пресман Э.Л., Сонин И.М.** Последовательное управление по неполным данным. М.: Наука, 1982. 256 с.
4. **Цетлин М.Л.** Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969. 316 с.
5. **Срагович В.Г.** Адаптивное управление. М.: Наука, 1981. 384 с.
6. **Perchet V., Rigollet P., Chassang S., Snowberg E.** Batched bandit problems // Ann. Statist. 2016. Vol. 44, no. 2. P. 660–681. https://doi.org/10.1214/15-aos1381
7. **Slivkins A.** Introduction to multi-armed bandits. 184 p. https://arxiv.org/pdf/1904.07272
8. **Kolnogorov A.V.** Determination of the minimax risk for the normal two-armed bandit // IFAC Proceedings Volumes. 2010. Vol. 43, no. 10. P. 231–236. https://doi.org/10.3182/20100826-3-TR-4015.00044
9. **Feldman D.** Contributions to the 'two-armed bandit' problem // Ann. Math. Stat. 1962. Vol. 33. P. 847–856. https://doi.org/https://doi.org/10.1214/aoms/1177704454
10. **Заборскис А.А.** Последовательный байесовый план выбора лучшего метода лечения // Автоматика и телемеханика. 1976. № 11. С. 144–153.
11. **Rodman L.** On the many-armed bandit problem // Ann. Probabil. 1978. Vol. 6, no. 3. P. 491–498. https://doi.org/10.1214/aop/1176995533
12. **Fabius J., van Zwet W.R.** Some remarks on the two-armed bandit // Ann. Math. Statist. 1970. Vol. 41. P. 1906–1916. https://doi.org/10.1214/aoms/1177696692
13. **Колногоров А.В.** К предельному описанию робастного параллельного управления в случайной среде // Автоматика и телемеханика. 2015. № 7. С. 111–126.
14. **Lai T.L., Robbins H.** Asymptotically efficient adaptive allocation rules // Adv. Appl. Math. 1985. Vol. 6. P. 4–22. https://doi.org/10.1016/0196-8858(85)90002-8
15. **Vogel W.** An asymptotic minimax theorem for the two-armed bandit problem // Ann. Math. Stat. 1960. Vol. 31, P. 444–451. https://doi.org/10.1214/aoms/1177705907

16. **Juditsky A., Nazin A.V., Tsybakov A.B., Vayatis N.** Gap-free bounds for stochastic multi-armed bandit // Proc. 17th World Congress IFAC. 2008. P. 11560–11563. https://doi.org/10.3182/20080706-5-KR-1001.01959
17. **Ershov M.A., Voroshilov A.S.** UCB strategy for Gaussian and Bernoulli multi-armed bandits // Commun. in Comp. and Inform. Sci. 2023. Vol. 1881. P. 67–78. https://doi.org/10.1007/978-3-031-43257-6_6
18. **Russo D., Van Roy B.** Learning to optimize via posterior sampling // Math. Oper. Research. 2014. Vol. 39, no. 4. P. 1221–1243. https://doi.org/10.1287/moor.2014.0650
19. **Колногоров А.В.** Гауссовский двурукий бандит и оптимизация групповой обработки данных // Пробл. передачи информации. 2018. Т. 54, № 1. С. 93–111.

Поступила 12.05.2025
После доработки 16.06.2025
Принята к публикации 23.06.2025
Опубликована онлайн 26.06.2025

Колногоров Александр Валерианович
д-р физ.-мат. наук, профессор
Новгородский государственный университет им. Ярослава Мудрого
г. Великий Новгород
e-mail: kolnogorov53@mail.ru

REFERENCES

1. Berry D.A., Fristedt B. *Bandit problems: sequential allocation of experiments*. London, NY, Springer Dordrecht, 1985, 275 p. https://doi.org/10.1007/978-94-015-3711-7
2. Lattimore T., Szepesvari C. *Bandit algorithms*. Cambridge, Cambridge Univ. Press, 2020, 518 p. https://doi.org/10.1017/9781108571401
3. Presman E.L., Sonin I.M. *Sequential control with incomplete information*. NY, Acad. Press, 1990, 290 p. ISBN-13: 9780125644358. Original Russian text published in Presman E. L., Sonin I. M. *Posledovatel'noye upravleniye po nepolnym dannym*, Moscow, Nauka Publ., 1982, 256 p.
4. Tsetlin M.L. *Automation theory and modeling of biological systems*, vol. 102. NY, Acad. Press, 1973. doi: 10.1016/s0076-5392(08)x6049-7. Original Russian text published in Tsetlin M. L. *Issledovaniya po teorii avtomatov i modelirovaniyu biologicheskikh sistem*, Moscow, Nauka Publ., 1969, 316 p.
5. Sragovich V.G. *Mathematical theory of adaptive control*, vol. 4. NJ, Interdisc. Math. Sci., London, World Sci., 2006, 473 p.https://doi.org/10.1142/5857. Original Russian text published in Sragovich V. G. *Adaptivnoye upravleniye*, Moscow, Nauka Publ., 1981, 381 p.
6. Perchet V., Rigollet P., Chassang S., Snowberg E. Batched bandit problems. *Ann. Stat.*, 2016, vol. 44, no. 2, pp. 660–681. https://doi.org/10.1214/15-aos1381
7. Slivkins A. *Introduction to multi-armed bandits*. 2019, 188 p. https://arxiv.org/pdf/1904.07272
8. Kolnogorov A.V. Determination of the minimax risk for the normal two-armed bandit. *IFAC Proc. Vol.*, 2010, vol. 43, iss. 10, pp. 231–236. https://doi.org/10.3182/20100826-3-TR-4015.00044
9. Feldman D. Contributions to the "Two-armed bandit" problem. *Ann. Math. Statist.*, 1962, vol. 33, pp. 847–856. https://doi.org/10.1214/aoms/1177704454
10. Zaborskis A.A. A sequential Bayesian plan for choosing the best method of curing. *Autom. Remote Control*, 1976, vol. 37, no. 11, pp. 1750–1757.
11. Rodman L. On the many-armed bandit problem. *Ann. Probab.*, 1978, vol. 6, no. 3, pp. 491–498.
12. Fabius J., van Zwet W.R. Some remarks on the two-armed bandit. *Ann. Math. Statist.*, 1970, vol. 41, no. 6, pp. 1906–1916. https://doi.org/10.1214/aoms/1177696692
13. Kolnogorov A.V. On a limiting description of robust parallel control in a random environment. *Autom. Remote Control*, 2015, vol. 76, no. 7, pp. 1229–1241. https://doi.org/10.1134/S0005117915070085
14. Lai T.L., Robbins H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 1985, vol. 6, iss. 1, pp. 4–22. https://doi.org/10.1016/0196-8858(85)90002-8
15. Vogel W. An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Statist.*, 1960, vol. 31, no. 2, pp. 444–451. https://doi.org/10.1214/aoms/1177705907

16. Juditsky A., Nazin A.V., Tsybakov A.B., Vayatis N. Gap-free bounds for stochastic multi-armed bandit. In: *Proc. 17th World Congress IFAC*, Seoul, 2008, vol. 41, iss. 2, pp. 11560–11563. https://doi.org/10.3182/20080706-5-KR-1001.01959
17. Ershov M.A., Voroshilov A.S. UCB strategy for Gaussian and Bernoulli multi-armed bandits. In: Khachay M., Kochetov Yu., Eremeev A., Khamisov O., Mazalov V., Pardalos P. (eds.) *Proc. 22nd Inter. Conf. "Mathematical optimization theory and operations research: recent trends"* (*MOTOR* 2023), vol. 1881. Cham, Springer, 2023, pp. 67–78. https://doi.org/10.1007/978-3-031-43257-6_6
18. Russo D., Van Roy B. Learning to optimize via posterior sampling. *Math. Operat. Res.*, 2014, vol. 39, no. 4, pp. 1221–1243. https://doi.org/10.1287/moor.2014.0650
19. Kolnogorov A.V. Gaussian two-armed bandit and optimization of batch data processing. *Probl. Inf. Transm.*, 2018, vol. 54, no. 1, pp. 84–100. https://doi.org/10.1134/S0032946018010076

Received May 12, 2025
Revised June 16, 2025
Accepted June 23, 2025
Published online June 26, 2025

Funding Agency. 1. The research was conducted as part of the scientific research project "Mathematical Modeling of Natural Processes", implemented under the state assignment in the field of scientific activity. 2. The study was carried out using the infrastructure of the Shared Use Center "High-Performance Computing and Big Data" (SUC "Informatics") of the Federal Research Center for Information Technology and Control of the Russian Academy of Sciences (Moscow).

Alexander Valerianovich Kolnogorov, Dr. Phys.-Math. Sci., Prof., Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, 173003 Russia, e-mail: kolnogorov53@mail.ru .

Cite this article as: A. V. Kolnogorov. Minimax approach to the Gaussian multi-armed bandit. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, 2025, vol. 31, no. 3, pp. 150–166 .