ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 58 **2022** **Вып. 2**

УДК 621.391.1 : 519.713 : 517.977.5

© 2022 г. **А.В. Колногоров**

ПУАССОНОВСКИЙ ДВУРУКИЙ БАНДИТ: НОВЫЙ ПОДХОД[1]

Рассматривается новый подход к задаче о двуруком бандите с непрерывным временем, в которой доходы описываются пуассоновскими процессами. Для этого, во-первых, горизонт управления разбивается на равные последовательные полуинтервалы, на которых стратегия остается постоянной, а доходы поступают пакетами, соответствующими этим полуинтервалам. Для нахождения оптимальной кусочно-постоянной байесовской стратегии и соответствующего ей байесовского риска получено рекуррентное разностное уравнение. Установлено существование предельной величины байесовского риска, если количество полуинтервалов неограниченно растет, и получено дифференциальное уравнение в частных производных для его нахождения. Во-вторых, в отличие от рассмотренных ранее постановок этой задачи мы исследуем зависимость стратегии от текущей предыстории управляемого процесса, а не от эволюции апостериорного распределения. Это позволяет снять требование конечности множества допустимых параметров, которое накладывалось в прежних постановках. Численные эксперименты показывают, что для практического нахождения байесовских и минимаксных стратегий и рисков достаточно разбить поступающие доходы на 30 пакетов. В случае минимаксной постановки показано, что оптимальная обработка поступающих доходов по одному не является более эффективной, чем оптимальная пакетная обработка, если горизонт управления неограниченно растет.

Ключевые слова: пуассоновский двурукий бандит, байесовский и минимаксный подходы, асимптотическая минимаксная теорема, пакетная обработка.

DOI: 10.31857/S0555292322020065, **EDN:** DZKOCQ

§ 1. Введение

Рассматривается задача о двуруком бандите [1,2], известная также как задача об адаптивном управлении [3, 4] и целесообразном поведении в случайной среде [5, 6], имеющая приложения в медицине, интернет-технологиях, обработке данных. Двурукий бандит – это устройство с двумя рукоятками, называемыми также действиями. Каждый выбор одного из действий сопровождается случайным доходом, распределение которого зависит только от выбранного действия, фиксировано, но неизвестно игроку. Количество игр против двурукого бандита определено заранее и известно. Требуется, наблюдая статистику игры, определить лучшее действие и обеспечить его преимущественное применение с целью максимизации математического ожидания полного дохода, т.е. это задача оптимального управления.

Особенностью рассматриваемой постановки являются непрерывное время и описание доходов пуассоновскими процессами. Такая постановка естественно дополняет

[1] Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (номер проекта 20-01-00062).

классическую формулировку задачи о бернуллиевском двуруком бандите, в которой доходы могут принимать значения 0 и 1. Наиболее известными исследованиями по пуассоновскому двурукому бандиту являются работы [2, 7], в которых установлен ряд важных результатов, в частности, существование оптимальной стратегии и ее пороговый характер, а также описана процедура синтеза оптимального управления. Отметим, что постановка задачи в [2] является даже более общей, чем задача о двуруком бандите. Однако существенным недостатком работ [2,7] является то, что в них рассматриваются только конечные множества допустимых значений параметра управляемого процесса. Это вызвано тем, что в этих работах стратегии управления зависят от эволюции апостериорного распределения, описываемой системой обыкновенных дифференциальных уравнений, размерность которой как раз равна количеству параметров. В качестве других постановок задачи о двуруком бандите с непрерывным временем отметим [1, 8], где рассмотрено управление винеровскими процессами. В [8] исходная задача о бернуллиевском одноруком бандите ставится в байесовской постановке и дискретном времени, а непрерывное время возникает в результате предельного перехода, когда горизонт управления неограниченно растет. В [1] также рассмотрена задача об одноруком бандите в байесовской постановке, однако здесь время сразу предполагается непрерывным. Наконец, в [9] рассмотрена постановка с непрерывным временем для многоруких бандитов, у которых априорные распределения для различных действий являются независимыми, а доходы дисконтируются на бесконечном горизонте управления.

Формально, пуассоновский двурукий бандит – это непрерывный справа скачкообразный управляемый случайный процесс $\{X(t),\ 0 \leqslant t \leqslant T\}$, значения которого интерпретируются как текущие доходы, увеличивающиеся на единицу в моменты скачков. Управление осуществляется с использованием двух действий. Будем использовать обозначение $y((t, t+\varepsilon]) = \ell$, если на полуинтервале $t' \in (t, t+\varepsilon]$, $\varepsilon > 0$, постоянно выбиралось действие $y(t') = \ell$ $(\ell = 1, 2)$. При использовании такого постоянного управления приращения процесса $X(t)$ зависят от выбираемых действий следующим образом:

$$\Pr\Big(X(t+\varepsilon) - X(t) = i \,|\, y((t, t+\varepsilon]) = \ell\Big) = p(i, \varepsilon; \lambda_\ell) = \frac{(\lambda_\ell \varepsilon)^i}{i!} e^{-\lambda_\ell \varepsilon}, \tag{1.1}$$

$i = 0, 1, 2, \ldots$, $\ell = 1, 2$. Величину $X(t+\varepsilon) - X(t)$ будем интерпретировать как пакет доходов, полученных на полуинтервале $(t, t+\varepsilon]$. Как известно, математическое ожидание и дисперсия $X(t+\varepsilon) - X(t)$ в этом случае равны

$$\begin{aligned} &\mathbf{E}\big(X(t+\varepsilon) - X(t) \,|\, y((t, t+\varepsilon]) = \ell\big) = \\ &= \mathbf{D}\big(X(t+\varepsilon) - X(t) \,|\, y((t, t+\varepsilon]) = \ell\big) = \lambda_\ell \varepsilon, \end{aligned} \tag{1.2}$$

$\ell = 1, 2$. Кроме того, при малых ε справедливы приближенные формулы

$$p(0, \varepsilon; \lambda_\ell) = 1 - \lambda_\ell \varepsilon + o(\varepsilon), \quad p(1, \varepsilon; \lambda_\ell) = \lambda_\ell \varepsilon + o(\varepsilon), \quad p(i, \varepsilon; \lambda_\ell) = o(\varepsilon), \tag{1.3}$$

$i = 2, 3, \ldots$, $\ell = 1, 2$. Таким образом, векторный параметр $\theta = (\lambda_1, \lambda_2)$, где λ_1, λ_2 характеризуют интенсивности поступления единичных доходов при выборе первого и второго действий, полностью описывает пуассоновский двурукий бандит. Множество Θ допустимых значений параметра предполагается известным, измеримым относительно меры Лебега и ограниченным, т.е. $\lambda_\ell \leqslant C < \infty$, $\ell = 1, 2$.

Для управления используются кусочно-постоянные стратегии. Для этого горизонт управления разбивается на равные временны́е полуинтервалы длины ε, на которых выбранные действия не меняются, т.е. рассматривается дискретное приближение задачи с непрерывным временем. Стратегия управления σ в момент времени $t = n\varepsilon$, соответствующий началу очередного временно́го полуинтервала, определяет выбор (вообще говоря, рандомизированный) действия $y((t, t+\varepsilon])$ в зависимости от

известной текущей предыстории. Такая предыстория имеет достаточно общий вид, состоящий из последовательности примененных действий и полученных в ответ доходов

$$y((0,\varepsilon]),\quad X(\varepsilon)-X(0),\quad y((\varepsilon,2\varepsilon]),\quad X(2\varepsilon)-X(\varepsilon),\quad \ldots,$$
$$y(((n-1)\varepsilon,n\varepsilon]),\quad X(n\varepsilon)-X((n-1)\varepsilon).$$

Однако можно показать, что в качестве предыстории можно ограничиться достаточной статистикой вида (X_1,t_1,X_2,t_2), где t_1,t_2 – текущие полные времена применения обоих действий ($t_1+t_2=t$), а X_1,X_2 – соответствующие полные доходы. В частности, для рассматриваемых в данной статье байесовских стратегий это следует из представленных в § 3 уравнений, описывающих оптимальное управление. Более подробно кусочно-постоянные стратегии обсуждаются в § 2.

Обозначим текущие значения доходов X_1 и X_2 в момент времени t через $X_1(t)$ и $X_2(t)$. Если бы значения интенсивностей λ_1,λ_2 были известны, то следовало бы всегда применять действие, соответствующее большей из них; при этом полный ожидаемый доход на всем горизонте управления T был бы равен $T\max(\lambda_1,\lambda_2)$. Но поскольку для управления используется стратегия σ, то полный ожидаемый доход меньше максимального на величину

$$L_T(\sigma,\theta)=T\max(\lambda_1,\lambda_2)-\mathbf{E}_{\sigma,\theta}\left(X_1(T)+X_2(T)\right), \tag{1.4}$$

которая называется функцией потерь и вызвана неполнотой информации об управляемом процессе. Здесь $\mathbf{E}_{\sigma,\theta}$ обозначает математическое ожидание, вычисленное по мере, порожденной стратегией σ и параметром θ. Отметим, что из ограниченности множества Θ следует ограниченность функции потерь (1.4).

Зададим априорную плотность распределения $\mu(\theta)=\mu(\lambda_1,\lambda_2)$ на множестве параметров Θ. Тогда математическое ожидание потерь, вычисленных относительно априорной плотности распределения $\mu(\theta)$, равно

$$L_T(\sigma,\mu)=\int\limits_{\Theta} L_T(\sigma,\theta)\mu(\theta)\,d\theta. \tag{1.5}$$

Байесовский риск, вычисленный относительно плотности $\mu(\theta)$, равен

$$R_T^B(\mu)=\inf_{\{\sigma\}} L_T(\sigma,\mu), \tag{1.6}$$

и соответствующая оптимальная стратегия σ^B называется байесовской. Минимаксный риск на множестве Θ равен

$$R_T^M(\Theta)=\inf_{\{\sigma\}}\sup_{\Theta} L_T(\sigma,\theta), \tag{1.7}$$

и соответствующая оптимальная стратегия σ^M называется минимаксной.

Прямого метода для нахождения минимаксных стратегии и риска не существует. Однако их можно найти с использованием основной теоремы теории игр, согласно которой имеет место равенство

$$R_T^M(\Theta)=R_T^B(\mu_0)=\sup_{\{\mu\}} R_T^B(\mu), \tag{1.8}$$

т.е. минимаксный риск равен байесовскому риску, вычисленному относительно наихудшего априорного распределения, на котором байесовский риск достигает максимума, а минимаксная стратегия совпадает с соответствующей байесовской. Отметим, что в случае конечного множества Θ численное нахождение минимаксного

риска в соответствии с равенством (1.8) не представляет труда, поскольку байесовский риск является вогнутой функцией априорного распределения.

Известны общие асимптотические оценки для функции потерь, байесовского и минимаксного рисков при $T \to \infty$, которые справедливы для рассматриваемого дискретного приближения задачи. Асимптотические оценки для функции потерь (1.4) в случае фиксированного, но неизвестного параметра θ даны, например, в [10] и имеют порядок $\ln(T)$. Там же при некоторых ограничениях на априорную плотность распределения $\mu(\theta)$ даны асимптотические оценки для байесовского риска (1.6), которые имеют порядок $\ln^2(T)$. Отметим, что в [2, 7] в ряде случаев байесовский риск оказался асимптотически ограничен (см. замечание 2 в § 3), что противоречит приведенной оценке, однако во всех этих случаях не были выполнены ограничения на априорное распределение, указанные в [10]. Наконец, асимптотическая оценка для минимаксного риска (1.7) вытекает из результатов работы [11] и имеет порядок $T^{1/2}$.

Статья имеет следующую структуру. Параграф 2 посвящен более подробной характеристике рассматриваемых кусочно-постоянных стратегий, в том числе показано, что для этих стратегий выполнены условия основной теоремы теории игр. Стандартное рекуррентное уравнение типа уравнения Беллмана для нахождения функции потерь, а также байесовских стратегии и риска дано в § 3. Отметим, что рассматриваемый подход отличается от представленного в [2,7], поскольку пересчет байесовского риска основан на текущей известной предыстории процесса (X_1, t_1, X_2, t_2), в то время как в [2, 7] байесовский риск рассматривается как функция апостериорного распределения. В § 3 также представлена другая, более удобная для анализа версия рекуррентного уравнения для вычисления байесовских стратегии и риска. В § 4 установлен пороговый характер байесовской стратегии управления и рассмотрен предельный случай, когда число полуинтервалов, на которых определена кусочно-постоянная стратегия, неограниченно растет. Такое управление можно интерпретировать как обработку поступающих доходов по одному. В этом случае доказано существование предела байесовского риска и получено дифференциальное уравнение в частных производных для его нахождения. В § 5 для случая, когда горизонт управления T неограниченно растет, с помощью выбора подходящего априорного распределения получена асимптотическая оценка снизу для минимаксного риска. Согласно этой оценке оптимальная обработка поступающих доходов по одному не является более эффективной, чем оптимальная пакетная обработка, если горизонт управления и количество пакетов неограниченно растут. В § 6 приведены результаты численных экспериментов, которые показывают, что нахождение близкого к оптимальному управления не требует больших вычислительных ресурсов. Например, число полуинтервалов, на которых задана кусочно-постоянная стратегия и формируются поступающие пакеты доходов, достаточно выбрать равным 30, а наихудшее априорное распределение при рассмотрении умеренных горизонтов управления можно сконцентрировать на шести парах параметров. В § 7 содержится заключение.

§ 2. Кусочно-постоянные стратегии управления

В этом параграфе вводится класс непрерывных слева стратегий, кусочно-постоянных на множестве последовательных полуинтервалов. Разобьем горизонт управления T на N последовательных полуинтервалов одинаковой длины ε, так что $T = N\varepsilon$. На каждом из этих полуинтервалов выбранное действие не меняется. Именно, для любых целых n_1, n_2, таких что $n_1 \geqslant 0$, $n_2 \geqslant 0$ и $n_1 + n_2 < N$, положим

$$t_1 = n_1\varepsilon, \quad t_2 = n_2\varepsilon, \quad t = t_1 + t_2.$$

Тогда

$$\Pr\big(y((t,t+\varepsilon])=\ell\,|\,X_1,t_1,X_2,t_2\big)=\sigma_\ell(X_1,t_1,X_2,t_2),$$

где $\sigma_\ell(X_1,t_1,X_2,t_2)$ определяется текущей статистикой (X_1,t_1,X_2,t_2) и постоянна на временно́м полуинтервале $t'\in(n\varepsilon,(n+1)\varepsilon]$. В частности, если стратегия предписывает выбор одного из действий с вероятностью 1, то это действие и будет применяться на всем временно́м полуинтервале.

Если же на временно́м полуинтервале $(t,t+\varepsilon]$ требуется выбирать действия рандомизированно с вероятностями $\varkappa_1,\varkappa_2$, то решение о том, какое действие будет применяться, принимается в начале этого полуинтервала: либо с вероятностью $\varkappa_1$ на всем полуинтервале будет применяться первое действие, обеспечивающее поток событий (единичных доходов) с интенсивностью λ_1, либо с вероятностью $\varkappa_2$ будет применяться второе действие, обеспечивающее поток событий с интенсивностью λ_2, причем этот выбор не зависит от предыстории процесса. В [2] предложен другой способ решения этой проблемы: считать, что в этом случае применяются оба действия одновременно с распределением ресурса между ними, обеспечивая интенсивности потоков событий $\varkappa_1\lambda_1$ и $\varkappa_2\lambda_2$ соответственно, и следовательно, суммарную интенсивность $\varkappa_1\lambda_1+\varkappa_2\lambda_2$. Ясно, что эти способы не эквивалентны, хотя дают одинаковое математическое ожидание полного числа событий на полуинтервале длины ε, равное $(\varkappa_1\lambda_1+\varkappa_2\lambda_2)\varepsilon$. Отметим также, что определение стратегии в [2] не предполагает разбиения горизонта управления на полуинтервалы.

Покажем, что предлагаемый способ в рамках рассматриваемого в данной статье дискретного приближения задачи эквивалентен предложенному в [2], если количество полуинтервалов, на которые разбивается горизонт управления, неограниченно растет. Справедлива следующая

Лемма 1. *Пусть исходный полуинтервал длины ε, на котором оба действия применяются с вероятностями $\varkappa_1$ и $\varkappa_2$, разбит на K последовательных полуинтервалов длины ε/K, на каждом из которых независимо осуществляется данное смешанное управление, и пусть $K\to\infty$. Тогда распределение, характеризующее поток событий на указанном полуинтервале, слабо сходится к распределению пуассоновского процесса с интенсивностью $\varkappa_1\lambda_1+\varkappa_2\lambda_2$.*

Доказательство. В этом случае полное число событий, соответствующих применению ℓ-го действия, равно $\xi_{\ell,1}+\ldots+\xi_{\ell,K}$, где $\xi_{\ell,j}$ соответствует доходу за применение ℓ-го действия на j-м полуинтервале и с учетом (1.3) характеризуется распределением

$$\begin{aligned}
&\Pr(\xi_{\ell,j}=1)=\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K),\\
&\Pr(\xi_{\ell,j}=0)=1-\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K),\\
&\Pr(\xi_{\ell,j}=i)=o(\varepsilon/K),\quad i=2,3,\ldots
\end{aligned}$$

Поэтому соответствующая характеристическая функция равна

$$\varphi_{\xi_{\ell,j}}(t)=\mathbf{E}\,e^{it\xi_{\ell,j}}=1+(e^{it}-1)\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K).$$

Так как все $\{\xi_{\ell,j}\}$ независимы, то характеристическая функция полного дохода $\xi_{\ell,1}+\ldots+\xi_{\ell,K}$ за применение ℓ-го действия на полуинтервале длины ε равна

$$\varphi_{\xi_{\ell,1}+\ldots+\xi_{\ell,K}}(t)=\big(1+(e^{it}-1)\varkappa_\ell\lambda_\ell\varepsilon/K+o(\varepsilon/K)\big)^K,$$

поэтому

$$\varphi_{\xi_{\ell,1}+\ldots+\xi_{\ell,K}}(t)\to e^{\varkappa_\ell\lambda_\ell\varepsilon(e^{it}-1)}\quad\text{при } K\to\infty,$$

т.е. сходится к характеристической функции пуассоновского процесса с интенсивностью $\varkappa_\ell\lambda_\ell$ на полуинтервале длины ε. Поскольку события в обоих потоках независимы, то результирующий поток также имеет распределение Пуассона с суммарной интенсивностью $\varkappa_1\lambda_1 + \varkappa_2\lambda_2$. ▲

Еще один аргумент в пользу кусочно-постоянных стратегий следует из результатов [9], где так же, как и в [2], допускается одновременное применение нескольких действий. В [9] установлено, что для многоруких бандитов со случайными доходами диффузионного типа одновременное применение нескольких действий возможно только с нулевой вероятностью. По-видимому, пуассоновский двурукий бандит обладает сходным свойством. В § 6 приведена типичная траектория отклонений текущих доходов X_1 от их пороговых стратегий, которая ведет себя так же, как соответствующая траектория для диффузионных многоруких бандитов.

Покажем теперь, что для рассматриваемой задачи при использовании кусочно-постоянных стратегий справедлива основная теорема теории игр.

Лемма 2. *Пусть* Θ *– компактное множество, а в качестве стратегий используются определенные выше кусочно-постоянные стратегии. Тогда минимаксный риск* (1.7) *и байесовские риски* (1.6) *связаны равенством*

$$R_T^M(\Theta) = \sup_{\{\mu\}} R_T^B(\mu). \tag{2.1}$$

Доказательство. Рассмотрим следующее сужение множества стратегий. Для достаточно большого M зафиксируем вероятности выбора действий для предысторий (X_1, t_1, X_2, t_2), на которых выполнено условие

$$\max(X_1, X_2) > M$$

(например, в этом случае они всегда выбирают только первое действие). Отметим, что за счет выбора M вероятность появления таких предысторий может быть сделана сколь угодно малой. Так как интенсивности λ_1, λ_2 ограничены, то из малости вероятности наступления события $\max(X_1, X_2) > M$ следует близость функций потерь, вычисленных по всем стратегиям и по рассматриваемому сужению. Поэтому функцию потерь, вычисляемую на исходном множестве стратегий, можно сколь угодно точно приблизить с помощью стратегий из данного более узкого класса.

Покажем, что для этого класса стратегий выполнено равенство (1.8). В соответствии с первой фундаментальной теоремой теории игр (см., например, [12]) для этого достаточно показать, что множество таких стратегий $\{\sigma\}$ является компактным, а функция потерь (1.4) непрерывна по совокупности переменных θ, σ. Эти свойства следуют из замечания 1 (см. § 3). Поэтому равенство (1.8) выполнено для указанного сужения множества стратегий. Устремляя M к бесконечности, получаем, что для кусочно-постоянных стратегий без введенного ограничения выполнено (2.1). ▲

Отметим, что хотя доказательство леммы 2 не гарантирует существования наихудшей априорной плотности распределения в соответствии с равенством (1.8), это можно установить, если ввести подходящее расстояние на исходном классе стратегий, превращающее его в компактное множество. Такой подход использован, например, в [13], где справедливость основной теоремы теории игр установлена для двурукого бандита с нормально распределенными доходами. Отметим также следующее свойство функции потерь, связанное с использованием смешанных стратегий: для любых двух стратегий σ_1, σ_2 и любого $0 < \varkappa < 1$ существует такая стратегия σ, что равенство

$$L_{\varepsilon,T}(\sigma, \theta) = \varkappa L_{\varepsilon,T}(\sigma_1, \theta) + (1 - \varkappa) L_{\varepsilon,T}(\sigma_2, \theta)$$

выполнено при всех θ. Это равенство может быть проверено непосредственно, если выписать полное выражение для $L_{\varepsilon,T}(\sigma,\theta)$ с использованием (3.3)–(3.6). Оно означает, что для выбора смешанной стратегии не требуется выполнять рандомизацию в начале управления. Данное свойство имеет место для двуруких бандитов с любыми распределениями одношаговых доходов. По-видимому, впервые оно отмечено в [14] для бернуллиевского двурукого бандита.

§ 3. Рекуррентные уравнения для нахождения байесовских потерь и риска

В этом параграфе сперва дается стандартное рекуррентное уравнение для вычисления потерь, соответствующих применению некоторой кусочно-постоянной стратегии. Это уравнение позволяет определить оптимальную стратегию, минимизирующую эти потери, т.е. вычислить байесовский риск, для нахождения которого также приведено стандартное рекуррентное уравнение. Затем эти уравнения преобразованы в формы, более удобные для дальнейшего анализа.

Пусть к моменту времени $t = t_1 + t_2$ первое и второе действия применялись на промежутках времени общей длины t_1 и t_2 соответственно, при этом полные доходы за применение первого и второго действий оказались равны X_1 и X_2. Тогда апостериорная плотность распределения в момент времени t равна

$$\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)=\frac{p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)}{\mu(X_1,t_1,X_2,t_2)}, \tag{3.1}$$

где $p(X_\ell,t_\ell;\lambda_\ell)$, $\ell=1,2$, определены в (1.1),

$$\mu(X_1,t_1,X_2,t_2)=\iint\limits_{\Theta} p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2. \tag{3.2}$$

Так как $p(0,0;\lambda)=1$, то формула (3.1) сохраняется при $t_1=0$ и/или $t_2=0$.

Пусть $\sigma_\ell(X_1,t_1,X_2,t_2)$ – определенная в § 2 кусочно-постоянная стратегия, причем дополнительно предположим, что на начальном промежутке времени длины $2t_0$ действия применяются по очереди – каждое на отрезке времени длины t_0. Данное условие в предположении, что t_0 достаточно мало, удобно использовать при рассмотрении предельного и асимптотического описаний байесовского риска, представленных в §§ 4, 5. Обозначим через

$$L_{T-t}(\sigma,(\lambda_1,\lambda_2))=(T-t)\max(\lambda_1,\lambda_2)-\mathbf{E}_{\sigma,\theta}\left(X_1(T)-X_1(t)+X_2(T)-X_2(t)\right)$$

функцию потерь на горизонте управления $(T-t,T]$, а через

$$L^B_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=\iint\limits_{\Theta} L_{T-t}(\sigma,(\lambda_1,\lambda_2))\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)\,d\lambda_1\,d\lambda_2$$

– ее математическое ожидание, вычисленное относительно апостериорной плотности распределения $\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)$. Запишем стандартное рекуррентное уравнение для нахождения функции потерь (1.5) относительно апостериорного распределения (3.1). Обозначим $x^+=\max(x,0)$. Тогда

$$L^B_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=\sum_{\ell=1}^{2}\sigma_\ell(X_1,t_1,X_2,t_2)\times L^{B,\ell}_\varepsilon(\sigma;X_1,t_1,X_2,t_2), \tag{3.3}$$

где

$$L^{B,1}_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=L^{B,2}_\varepsilon(\sigma;X_1,t_1,X_2,t_2)=0, \tag{3.4}$$

если $t_1 + t_2 = T$, и далее

$$
\begin{aligned}
L_\varepsilon^{B,1}(\sigma; X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_2 - \lambda_1)^+ \varepsilon + \sum_{j=0}^{\infty} L_\varepsilon^B(\sigma; X_1 + j, t_1 + \varepsilon, X_2, t_2) p(j, \varepsilon; \lambda_1) \right) d\lambda_1 \, d\lambda_2, \\
L_\varepsilon^{B,2}(\sigma; X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_1 - \lambda_2)^+ \varepsilon + \sum_{j=0}^{\infty} L_\varepsilon^B(\sigma; X_1, t_1, X_2 + j, t_2 + \varepsilon) p(j, \varepsilon; \lambda_2) \right) d\lambda_1 \, d\lambda_2
\end{aligned}
\tag{3.5}
$$

при $2t_0 \leqslant t < T$. Здесь $\{L_\varepsilon^{B,\ell}(\sigma; X_1, t_1, X_2, t_2)\}$ описывают ожидаемые потери на оставшемся горизонте управления, если сначала на горизонте длины ε применялось ℓ-е действие, а затем управление осуществлялось в соответствии со стратегией σ. В частности, $(\lambda_2 - \lambda_1)^+ \varepsilon$ и $(\lambda_1 - \lambda_2)^+ \varepsilon$ в силу (1.2) описывают ожидаемые потери дохода на полуинтервале длины ε за применение первого и второго действий соответственно. Нижний индекс ε указывает, что для управления используются кусочно-постоянные стратегии на промежутках времени длины ε. Потери (1.5) вычисляются по формуле

$$
\begin{aligned}
L_{\varepsilon,T}(\sigma, \mu) &= t_0 \iint\limits_\Theta |\lambda_1 - \lambda_2| \mu(\lambda_1, \lambda_2) \, d\lambda_1 \, d\lambda_2 + \\
&+ \sum_{X_1=0}^{\infty} \sum_{X_2=0}^{\infty} L_\varepsilon^B(\sigma; X_1, t_0, X_2, t_0) \mu(X_1, t_0, X_2, t_0),
\end{aligned}
\tag{3.6}
$$

где первое слагаемое описывает потери на начальном горизонте длины $2t_0$, когда действия применяются по очереди, а второе слагаемое – потери на заключительном горизонте длины $T - 2t_0$. Ясно, что если $t_0 = 0$, то

$$L_{\varepsilon,T}(\sigma, \mu) = L_\varepsilon^B(\sigma; 0, 0, 0, 0).$$

Отметим, что для нахождения функции потерь (1.4) надо взять вырожденную априорную плотность распределения, сосредоточенную на параметре (λ_1, λ_2), при этом все апостериорные плотности также останутся вырожденными.

Замечание 1. Рассмотрим стратегии, у которых фиксированы вероятности выбора действий для предысторий, характеризуемых условием $\max(X_1, X_2) > M$, а остальные вероятности $\{\sigma_\ell(X_1, t_1, X_2, t_2)\}$ могут произвольно меняться от 0 до 1 при условии, что при всех предысториях выполнено равенство

$$\sigma_1(X_1, t_1, X_2, t_2) + \sigma_2(X_1, t_1, X_2, t_2) = 1.$$

Определим расстояние между двумя такими стратегиями $\sigma^{(1)}$ и $\sigma^{(2)}$ как

$$\max \left| \sigma_1^{(1)}(X_1, t_1, X_2, t_2) - \sigma_1^{(2)}(X_1, t_1, X_2, t_2) \right|,$$

где максимум берется по всем предысториям (X_1, t_1, X_2, t_2), удовлетворяющим условию $\max(X_1, X_2) \leqslant M$. Расстояние между параметрами $\theta^{(1)} = \left(\lambda_1^{(1)}, \lambda_2^{(1)}\right)$ и $\theta^{(2)} = \left(\lambda_1^{(2)}, \lambda_2^{(2)}\right)$ определим как

$$\max \left(\left| \lambda_1^{(1)} - \lambda_1^{(2)} \right|, \left| \lambda_2^{(1)} - \lambda_2^{(2)} \right| \right).$$

Так как при $\max(\lambda_1, \lambda_2) \leqslant C$ все функции потерь ограничены величиной TC, то все бесконечные суммы в (3.5), (3.6) сходятся равномерно. Поэтому из (3.3)–(3.6) следует, что функция потерь (1.4), вычисляемая при вырожденной априорной плотности распределения, непрерывна относительно введенных расстояний. При этом стратегии полностью определяются вероятностями $\{\sigma_1(X_1, t_1, X_2, t_2)\}$, заданными для предысторий, удовлетворяющим условию $\max(X_1, X_2) \leqslant M$, а их множество эквивалентно единичному кубу соответствующей размерности, который является компактным.

Уравнения (3.3)–(3.5) позволяют найти стратегию, минимизирующую полные потери. Для этого надо, начиная с момента $t_1 + t_2 = T - \varepsilon$ и заканчивая моментом $t_1 + t_2 = 2t_0$, при каждой предыстории (X_1, t_1, X_2, t_2) выбирать то действие, которому соответствует меньшее из значений $L^{B,\ell}(\sigma; X_1, t_1, X_2, t_2)$. Полученные полные потери характеризуют байесовский риск и могут быть вычислены с помощью стандартного рекуррентного уравнения

$$R_\varepsilon^B(X_1, t_1, X_2, t_2) = \min\bigl(R_\varepsilon^{B,1}(X_1, t_1, X_2, t_2), R_\varepsilon^{B,2}(X_1, t_1, X_2, t_2)\bigr), \tag{3.7}$$

где

$$R_\varepsilon^{B,1}(X_1, t_1, X_2, t_2) = R_\varepsilon^{B,2}(X_1, t_1, X_2, t_2) = 0, \tag{3.8}$$

если $t_1 + t_2 = T$, и далее

$$\begin{aligned}
R_\varepsilon^{B,1}(X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_2 - \lambda_1)^+ \varepsilon + \sum_{j=0}^{\infty} R_\varepsilon^B(X_1 + j, t_1 + \varepsilon, X_2, t_2) p(j, \varepsilon; \lambda_1) \right) d\lambda_1\, d\lambda_2, \\
R_\varepsilon^{B,2}(X_1, t_1, X_2, t_2) &= \iint\limits_\Theta \mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2) \times \\
&\times \left((\lambda_1 - \lambda_2)^+ \varepsilon + \sum_{j=0}^{\infty} R_\varepsilon^B(X_1, t_1, X_2 + j, t_2 + \varepsilon) p(j, \varepsilon; \lambda_2) \right) d\lambda_1\, d\lambda_2
\end{aligned} \tag{3.9}$$

при $2t_0 \leqslant t < T$. Здесь $\{R_\varepsilon^{B,\ell}(X_1, t_1, X_2, t_2)\}$ описывают ожидаемые потери на горизонте управления $(t, T]$, вычисленные относительно апостериорной плотности распределения $\mu(\lambda_1, \lambda_2 \,|\, X_1, t_1, X_2, t_2)$, если сначала на горизонте длины ε применялось ℓ-е действие, а затем управление осуществлялось оптимально, а $\{R_\varepsilon^B(X_1, t_1, X_2, t_2)\}$ описывают соответствующие байесовские риски. Байесовский риск (1.6) вычисляется по формуле

$$\begin{aligned}
R_{\varepsilon,T}^B(\mu) &= t_0 \iint\limits_\Theta |\lambda_1 - \lambda_2| \mu(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2 + \\
&+ \sum_{X_1=0}^{\infty} \sum_{X_2=0}^{\infty} R_\varepsilon^B(X_1, t_0, X_2, t_0) \mu(X_1, t_0, X_2, t_0).
\end{aligned} \tag{3.10}$$

Если $t_0 = 0$, то

$$R_{\varepsilon,T}(\mu) = R_\varepsilon^B(0, 0, 0, 0).$$

Наряду с байесовским риском уравнения (3.7)–(3.9) позволяют найти байесовскую стратегию. Байесовская стратегия предписывает выбирать ℓ-е действие, т.е.

$$\sigma_\ell(X_1,t_1,X_2,t_2)=1 \quad \text{при} \quad t'\in(t,t+\varepsilon],$$

если меньшую величину имеет $R_\varepsilon^{B,\ell}(X_1,t_1,X_2,t_2)$ $(\ell=1,2)$. В случае равенства $R_\varepsilon^{B,1}(X_1,t_1,X_2,t_2)=R_\varepsilon^{B,2}(X_1,t_1,X_2,t_2)$ выбор действия может быть произвольным.

Замечание 2. В случае конечного множества параметров

$$\{\theta_i=(\lambda_{1,i},\lambda_{2,i}),\ i=1,\dots,K\}$$

из результатов [2] вытекает следующее интересное свойство. Если в каждом из множеств $\{\lambda_{1,1},\dots,\lambda_{1,K}\}$ и $\{\lambda_{2,1},\dots,\lambda_{2,K}\}$ нет совпадающих элементов, то байесовский риск $R_{\varepsilon,T}^B(\mu)$ асимптотически конечен при $T\to\infty$.

Замечание 3. Для некоторого $k>0$ рассмотрим следующую замену переменных:

$$\begin{aligned}
&\lambda'_\ell=k\lambda_\ell, \quad \mu'(\lambda'_1,\lambda'_2)=k^{-2}\mu(\lambda_1,\lambda_2), \quad \varepsilon'=k^{-1}\varepsilon,\\
&t'_0=k^{-1}t_0, \quad T'=k^{-1}T, \quad t'_\ell=k^{-1}t_\ell, \quad X'_\ell=X_\ell,\\
&\sigma'_\ell(X'_1,t'_1,X'_2,t'_2)=\sigma_\ell(X_1,t_1,X_2,t_2), \quad \ell=1,2.
\end{aligned} \tag{3.11}$$

Тогда при всех X_1,t_1,X_2,t_2 справедливы равенства

$$\begin{aligned}
&R'^{B,\ell}_{\varepsilon'}(X'_1,t'_1,X'_2,t'_2)=R_\varepsilon^{B,\ell}(X_1,t_1,X_2,t_2),\\
&L'^{B,\ell}_{\varepsilon'}(\sigma';X'_1,t'_1,X'_2,t'_2)=L_\varepsilon^{B,\ell}(\sigma;X_1,t_1,X_2,t_2), \quad \ell=1,2,
\end{aligned} \tag{3.12}$$

где через $R'_{\varepsilon'}(\cdot)$ и $L'_{\varepsilon'}(\cdot)$ обозначены байесовские риски и потери, вычисленные относительно априорной плотности распределения μ'. Из (3.12) следует также выполнение равенств

$$R'^{B}_{\varepsilon'}(X'_1,t'_1,X'_2,t'_2)=R_\varepsilon^{B}(X_1,t_1,X_2,t_2)$$

и

$$L'^{B}_{\varepsilon'}(\sigma';X'_1,t'_1,X'_2,t'_2)=L_\varepsilon^{B}(\sigma;X_1,t_1,X_2,t_2)$$

при всех X_1,t_1,X_2,t_2. В том числе справедливы равенства

$$R^B_{\varepsilon',T'}(\mu')=R^B_{\varepsilon,T}(\mu), \quad L^B_{\varepsilon',T'}(\sigma',\mu')=L^B_{\varepsilon,T}(\sigma,\mu). \tag{3.13}$$

Равенства (3.12), (3.13) устанавливаются выполнением замены переменных (3.11) в формулах (3.3)–(3.6) и (3.7)–(3.10). Равенства (3.12), (3.13) означают, что задачу о пуассоновском двуруком бандите всегда можно рассматривать на единичном горизонте управления $T=1$.

Получим более удобную для вычислений и анализа форму рекуррентного уравнения (3.7)–(3.9). Положим

$$\begin{aligned}
&\widetilde{p}(X_\ell,t_\ell;\lambda_\ell)=\lambda_\ell^{X_\ell}e^{-\lambda_\ell t_\ell}=\frac{p(X_\ell,t_\ell;\lambda_\ell)X_\ell!}{t_\ell^{X_\ell}},\\
&\widetilde{\mu}(X_1,t_1,X_2,t_2)=\iint\limits_\Theta \widetilde{p}(X_1,t_1;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
&R_\varepsilon(X_1,t_1,X_2,t_2)=R_\varepsilon^B(X_1,t_1,X_2,t_2)\widetilde{\mu}(X_1,t_1,X_2,t_2).
\end{aligned} \tag{3.14}$$

Отметим, что

$$\mu(\lambda_1,\lambda_2\,|\,X_1,t_1,X_2,t_2)=\frac{\widetilde{p}(X_1,t_1;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)}{\widetilde{\mu}(X_1,t_1,X_2,t_2)}. \tag{3.15}$$

Справедлива следующая

Теорема 1. *Рассмотрим рекуррентное разностное уравнение*

$$R_\varepsilon(X_1,t_1,X_2,t_2)=\min\bigl(R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2),R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)\bigr), \tag{3.16}$$

где

$$R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2)=R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)=0, \tag{3.17}$$

если $t_1+t_2=T$, *и далее*

$$\begin{aligned}
R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2)&=\varepsilon g^{(1)}(X_1,t_1,X_2,t_2)+\mathbf{T}_\varepsilon^{(1)}R_\varepsilon(X_1,t_1+\varepsilon,X_2,t_2),\\
R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)&=\varepsilon g^{(2)}(X_1,t_1,X_2,t_2)+\mathbf{T}_\varepsilon^{(2)}R_\varepsilon(X_1,t_1,X_2,t_2+\varepsilon)
\end{aligned} \tag{3.18}$$

при $2t_0\leqslant t<T$. *Здесь функции* $\{g^{(\ell)}(X_1,t_1,X_2,t_2)\}$ *и операторы* $\{\mathbf{T}_\varepsilon^{(\ell)}\}$ *таковы*:

$$\begin{aligned}
g^{(1)}(X_1,t_1,X_2,t_2)&=\iint\limits_\Theta(\lambda_2-\lambda_1)^+\lambda_1^{X_1}e^{-\lambda_1t_1}\lambda_2^{X_2}e^{-\lambda_2t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
g^{(2)}(X_1,t_1,X_2,t_2)&=\iint\limits_\Theta(\lambda_1-\lambda_2)^+\lambda_1^{X_1}e^{-\lambda_1t_1}\lambda_2^{X_2}e^{-\lambda_2t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
\mathbf{T}_\varepsilon^{(1)}F(X_1,t_1,X_2,t_2)&=\sum_{j=0}^{\infty}F(X_1+j,t_1,X_2,t_2)\times\frac{\varepsilon^j}{j!},\\
\mathbf{T}_\varepsilon^{(2)}F(X_1,t_1,X_2,t_2)&=\sum_{j=0}^{\infty}F(X_1,t_1,X_2+j,t_2)\times\frac{\varepsilon^j}{j!}.
\end{aligned} \tag{3.19}$$

Байесовская стратегия предписывает выбирать ℓ*-е действие* (*иными словами*, $\sigma_\ell(X_1,t_1,X_2,t_2)=1$), *если* $R_\varepsilon^{(\ell)}(X_1,t_1,X_2,t_2)$ *имеет меньшую величину* ($\ell=1,2$). *В случае равенства* $R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2)=R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)$ *выбор действия может быть произвольным. Байесовский риск* (1.6) *вычисляется по формуле*

$$\begin{aligned}
R_{\varepsilon,T}(\mu)&=t_0\iint\limits_\Theta|\lambda_1-\lambda_2|\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2+\\
&+\sum_{X_1=0}^{\infty}\sum_{X_2=0}^{\infty}R_\varepsilon(X_1,t_0,X_2,t_0)\frac{t_0^{X_1}t_0^{X_2}}{X_1!\,X_2!},
\end{aligned} \tag{3.20}$$

в частности, $R_{\varepsilon,T}(\mu)=R_\varepsilon(0,0,0,0)$ *при* $t_0=0$.

Доказательство. Левую и правую части первого уравнения в (3.9) домножим на $\widetilde{\mu}(X_1,t_1,X_2,t_2)$. С учетом (3.15) получим первое уравнение из (3.18), где $g^{(1)}(X_1,t_1,X_2,t_2)$ имеет вид (3.19). Далее,

$$\mathbf{T}_\varepsilon^{(1)}R_\varepsilon(X_1,t_1+\varepsilon,X_2,t_2)=\sum_{j=0}^{\infty}R_\varepsilon(X_1+j,t_1+\varepsilon,X_2,t_2)\times h_\varepsilon(j),$$

где $h_\varepsilon(j)$ с учетом (3.9), (3.14) при $t_1 > 0$ равна

$$\frac{\iint\limits_\Theta \widetilde{p}(X_1,t_1;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)p(j,\varepsilon;\lambda_1)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2}{\iint\limits_\Theta \widetilde{p}(X_1+j,t_1+\varepsilon;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2} =$$
$$= \frac{\widetilde{p}(X_1,t_1;\lambda_1)p(j,\varepsilon;\lambda_1)}{\widetilde{p}(X_1+j,t_1+\varepsilon;\lambda_1)} = \frac{\varepsilon^j}{j!},$$

что соответствует (3.19). При $t_1 = 0$ также $X_1 = 0$, поэтому $h_\varepsilon(j)$ равна

$$\frac{\iint\limits_\Theta \widetilde{p}(X_2,t_2;\lambda_2)p(j,\varepsilon;\lambda_1)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2}{\iint\limits_\Theta \widetilde{p}(j,\varepsilon;\lambda_1)\widetilde{p}(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2} = \frac{p(j,\varepsilon;\lambda_1)}{\widetilde{p}(j,\varepsilon;\lambda_1)} = \frac{\varepsilon^j}{j!},$$

что также соответствует (3.19). Проверка второго равенства (3.18) выполняется аналогично. Равенство (3.20) следует из (3.10) с учетом (3.14). ▲

Запишем также новую форму рекуррентного уравнения для нахождения потерь. Справедлива следующая

Теорема 2. *Для заданной стратегии $\sigma(X_1,t_1,X_2,t_2)$ рассмотрим рекуррентное уравнение*

$$L_\varepsilon(\sigma;X_1,t_1,X_2,t_2) = \sum_{\ell=1}^{2} \sigma_\ell(X_1,t_1,X_2,t_2) \times L_\varepsilon^{(\ell)}(\sigma;X_1,t_1,X_2,t_2), \tag{3.21}$$

где

$$L_\varepsilon^{(1)}(\sigma;X_1,t_1,X_2,t_2) = L_\varepsilon^{(2)}(\sigma;X_1,t_1,X_2,t_2) = 0, \tag{3.22}$$

если $t_1 + t_2 = T$, и далее

$$\begin{aligned} L_\varepsilon^{(1)}(\sigma;X_1,t_1,X_2,t_2) &= \varepsilon g^{(1)}(X_1,t_1,X_2,t_2) + \mathbf{T}_\varepsilon^{(1)} L_\varepsilon(\sigma;X_1,t_1+\varepsilon,X_2,t_2),\\ L_\varepsilon^{(2)}(\sigma;X_1,t_1,X_2,t_2) &= \varepsilon g^{(2)}(X_1,t_1,X_2,t_2) + \mathbf{T}_\varepsilon^{(2)} L_\varepsilon(\sigma;X_1,t_1,X_2,t_2+\varepsilon) \end{aligned} \tag{3.23}$$

при $2t_0 \leqslant t < T$, где $\{g^{(\ell)}(X_1,t_1,X_2,t_2)\}$ и $\{\mathbf{T}_\varepsilon^{(\ell)}\}$ определены в (3.19). *Тогда полные потери* (1.5) *вычисляются по формуле*

$$\begin{aligned} L_{\varepsilon,T}(\sigma,\mu) &= t_0 \iint\limits_\Theta |\lambda_1-\lambda_2|\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 +\\ &+ \sum_{X_1=0}^{\infty}\sum_{X_2=0}^{\infty} L_\varepsilon(\sigma;X_1,t_0,X_2,t_0)\frac{t_0^{X_1}t_0^{X_2}}{X_1!\,X_2!}, \end{aligned} \tag{3.24}$$

в частности, $L_{\varepsilon,T}(\sigma,\mu) = L_\varepsilon(0,0,0,0)$ при $t_0 = 0$.

Доказательство проводится аналогично доказательству теоремы 1, если подставить $L(\sigma;X_1,t_1,X_2,t_2) = L^B(\sigma;X_1,t_1,X_2,t_2)\widetilde{\mu}(X_1,t_1,X_2,t_2)$ в формулы (3.3)–(3.6).

Замечание 4. Нетрудно проверить, что при использовании замены переменных (3.11) для любых X_1,t_1,X_2,t_2 будут выполнены равенства

$$\begin{aligned} R_{\varepsilon'}^{\prime(\ell)}(X_1',t_1',X_2',t_2') &= k^{X_1+X_2} R_\varepsilon^{(\ell)}(X_1,t_1,X_2,t_2),\\ L_{\varepsilon'}^{\prime(\ell)}(\sigma';X_1',t_1',X_2',t_2') &= k^{X_1+X_2} L_\varepsilon^{(\ell)}(\sigma;X_1,t_1,X_2,t_2), \end{aligned}$$

$\ell = 1, 2$, где через $R'_{\varepsilon'}(\cdot)$ и $L'_{\varepsilon'}(\cdot)$ обозначены риски и потери, вычисленные относительно априорной плотности μ'. При этом равенства (3.13) сохранятся.

§ 4. Предельное описание

В этом параграфе сначала устанавливается пороговый характер стратегии управления. Далее устанавливается существование непрерывного по t_1, t_2 предела риска $R_\varepsilon(X_1, t_1, X_2, t_2)$ при $\varepsilon \to +0$. Наконец, для предельного риска получено дифференциальное уравнение в частных производных.

Установим пороговый характер байесовской стратегии управления. Он выражается в том, что байесовскую стратегию всегда можно выбрать так, что на множествах значений статистики $(X_1, t_1 - t, X_2, t_2 + t)$, где X_1, t_1, X_2, t_2 фиксированы, а t возрастает, смена оптимального действия со второго на первое произойдет не более чем при одном t. Аналогичным свойством обладают статистики $(X_1 + x, t_1, X_2 - x, t_2)$, где X_1, t_1, X_2, t_2 фиксированы, а x возрастает. Обозначим

$$\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) - R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2).$$

Ясно, что критериями выбора первого и второго действий являются неравенства $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) < 0$ и $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) > 0$ соответственно, а в случае, когда $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = 0$, действия можно выбирать произвольно. Справедлива следующая

Теорема 3. *При любой априорной плотности распределения $\mu(\lambda_1, \lambda_2)$ функции*

$$\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t) \quad \textit{и} \quad \Delta R_\varepsilon(X_1 + x, t_1, X_2 - x, t_2) \tag{4.1}$$

являются монотонно невозрастающими функциями t и x соответственно.

Доказательство. Обозначим через $R_\varepsilon^{(12)}(X_1, t_1, X_2, t_2)$ и $R_\varepsilon^{(21)}(X_1, t_1, X_2, t_2)$ потери, если сначала по очереди применяются первое и второе (соответственно, второе и первое) действия, а затем управление осуществляется оптимально. Положим

$$\Delta R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) = R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) - R_\varepsilon^{(12)}(X_1, t_1, X_2, t_2),$$
$$\Delta R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2) = R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2) - R_\varepsilon^{(21)}(X_1, t_1, X_2, t_2).$$

Ясно, что

$$\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = \Delta R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) - \Delta R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2).$$

Введем также обозначения $x^+ = \max(x, 0)$ и $x^- = \max(-x, 0)$. Из первого уравнения в (3.18) следует, что

$$R_\varepsilon^{(12)}(X_1, t_1, X_2, t_2) = \varepsilon g^{(1)}(X_1, t_1, X_2, t_2) + \mathbf{T}_\varepsilon^{(1)} R_\varepsilon^{(2)}(X_1, t_1 + \varepsilon, X_2, t_2),$$

поэтому, вычитая это уравнение из первого уравнения (3.18), получаем

$$\Delta R_\varepsilon^{(1)}(X_1, t_1, X_2, t_2) = -\mathbf{T}_\varepsilon^{(1)} \Delta R_\varepsilon^{-}(X_1, t_1 + \varepsilon, X_2, t_2). \tag{4.2}$$

Аналогично,

$$-\Delta R_\varepsilon^{(2)}(X_1, t_1, X_2, t_2) = \mathbf{T}_\varepsilon^{(2)} \Delta R_\varepsilon^{+}(X_1, t_1, X_2, t_2 + \varepsilon). \tag{4.3}$$

Проверим монотонность функции $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ по t по индукции. При $t_1 + t_2 = T - \varepsilon$ имеем

$$\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t) = \varepsilon \iint\limits_\Theta \Big((\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{-(\lambda_2 - \lambda_1)t} -$$

$$- (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{(\lambda_1 - \lambda_2)t} \Big) \mu(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2.$$

Так как

$$(\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{-(\lambda_2 - \lambda_1)t} - (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} e^{(\lambda_1 - \lambda_2)t}$$

является монотонно невозрастающей функцией t, то $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ – также монотонно невозрастающая функция t при $t_1 + t_2 = T - \varepsilon$. Далее, если $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ – монотонно невозрастающая функция t, то такими же являются функции $\Delta R_\varepsilon^+(X_1, t_1 - t, X_2, t_2 + t)$ и $-\Delta R_\varepsilon^-(X_1, t_1 - t, X_2, t_2 + t)$. Поэтому из (4.2), (4.3) и равенства

$$\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t) = \Delta R_\varepsilon^{(1)}(X_1, t_1 - t, X_2, t_2 + t) -$$

$$- \Delta R_\varepsilon^{(2)}(X_1, t_1 - t, X_2, t_2 + t)$$

следует, что если $\Delta R_\varepsilon(X_1, t_1 - t, X_2, t_2 + t)$ является монотонно невозрастающей функцией t при некоторых $t_1 + t_2 = \tau + \varepsilon$, то это свойство сохранится и при $t_1 + t_2 = \tau$.

Проверка того, что $\Delta R_\varepsilon(X_1 + x, t_1, X_2 - x, t_2)$ является монотонно невозрастающей функцией x, выполняется аналогично по индукции, причем при $t_1 + t_2 = T - \varepsilon$ следует рассмотреть выражение

$$\Delta R_\varepsilon(X_1 + x, t_1, X_2 - x, t_2) = \varepsilon \iint\limits_\Theta \Big((\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} (\lambda_1/\lambda_2)^x -$$

$$- (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} (\lambda_1/\lambda_2)^x \Big) \mu(\lambda_1, \lambda_2)\, d\lambda_1\, d\lambda_2. \quad \blacktriangle$$

Таким образом, для обеспечения порогового характера стратегии достаточно обеспечить его при $\Delta R_\varepsilon(X_1, t_1, X_2, t_2) = 0$. Например, можно в этом случае всегда выбирать первое действие. При этом возможно, что одно из действий не будет выбрано ни при какой предыстории (X_1, t_1, X_2, t_2); например, так будет, если $\mu(\lambda_1, \lambda_2) > 0$ только при $\lambda_1 > \lambda_2$. Отметим также, что теорема 3 является аналогом теоремы 5 из [14], в которой установлен пороговый характер стратегии для бернуллиевского двурукого бандита.

Перейдем к оценкам рисков $R_\varepsilon(X_1, t_1, X_2, t_2)$.

Лемма 3. *При всех допустимых X_1, t_1, X_2, t_2 справедливы оценки*

$$R_\varepsilon(X_1, t_1, X_2, t_2) \leqslant (T - t) g(X_1, t_1, X_2, t_2), \tag{4.4}$$

где $t = t_1 + t_2$, $g(X_1, t_1, X_2, t_2) = \min\big(g^{(1)}(X_1, t_1, X_2, t_2), g^{(2)}(X_1, t_1, X_2, t_2)\big)$, а функции $g^{(1)}(X_1, t_1, X_2, t_2)$ и $g^{(2)}(X_1, t_1, X_2, t_2)$ определены в (3.19).

Доказательство. Выражение в (4.4) справа характеризует потери, обеспечиваемые стратегией, которая для текущей статистики (X_1, t_1, X_2, t_2) на всем оставшемся горизонте управления длины $T - t$ выбирает действие, которому соответствует меньшее из значений $g^{(1)}(X_1, t_1, X_2, t_2)$, $g^{(2)}(X_1, t_1, X_2, t_2)$. Ясно, что для оптимальной стратегии потери будут не больше указанных. $\blacktriangle$

Следующая лемма является вспомогательной.

Лемма 4. *При $\ell = 1, 2$ справедливы равенства*

$$\mathbf{T}^{(\ell)}_{\varepsilon_1}\mathbf{T}^{(\ell)}_{\varepsilon_2}F(X_1,t_1,X_2,t_2) = \mathbf{T}^{(\ell)}_{\varepsilon_1+\varepsilon_2}F(X_1,t_1,X_2,t_2), \tag{4.5}$$

где $F(X_1,t_1,X_2,t_2)$ – произвольная функция,

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon}g^{(\ell)}(X_1,t_1,X_2,t_2) = g^{(\ell)}(X_1,t_1-\varepsilon,X_2,t_2) \quad \textit{при } t_1 > 0,\\
&\mathbf{T}^{(2)}_{\varepsilon}g^{(\ell)}(X_1,t_1,X_2,t_2) = g^{(\ell)}(X_1,t_1,X_2,t_2-\varepsilon) \quad \textit{при } t_2 > 0,
\end{aligned} \tag{4.6}$$

а также оценки

$$\begin{aligned}
&\mathbf{T}^{(\ell)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) \leqslant (e^{C\varepsilon}-1)(T-t)g(X_1,t_1,X_2,t_2),\\
&\mathbf{T}^{(1)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) - \varepsilon R_\varepsilon(X_1+1,t_1,X_2,t_2) \leqslant\\
&\leqslant (e^{C\varepsilon}-1-\varepsilon C)(T-t)g(X_1,t_1,X_2,t_2),\\
&\mathbf{T}^{(2)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) - \varepsilon R_\varepsilon(X_1,t_1,X_2+1,t_2) \leqslant\\
&\leqslant (e^{C\varepsilon}-1-\varepsilon C)(T-t)g(X_1,t_1,X_2,t_2),
\end{aligned} \tag{4.7}$$

где $C = \max\limits_{\Theta}\max\limits_{\ell=1,2}\lambda_\ell$.

Доказательство. Проверим равенство (4.5). Достаточно рассмотреть случай $\ell = 1$. Тогда

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon_1}\mathbf{T}^{(1)}_{\varepsilon_2}F(X_1,t_1,\cdot) = \sum_{i=0}^{\infty}\left(\sum_{j=0}^{\infty}F(X_1+j+i,t_1,\cdot)\times\frac{\varepsilon_2^j}{j!}\right)\times\frac{\varepsilon_1^i}{i!} =\\
&= \sum_{k=0}^{\infty}F(X_1+k,t_1,\cdot)\times\left(\sum_{j=0}^{k}\frac{\varepsilon_2^j}{j!}\times\frac{\varepsilon_1^{k-j}}{(k-j)!}\right) = \sum_{k=0}^{\infty}F(X_1+k,t_1,\cdot)\times\frac{(\varepsilon_1+\varepsilon_2)^k}{k!},
\end{aligned}$$

что соответствует (4.5). Проверим равенство (4.6). Достаточно рассмотреть $\mathbf{T}^{(1)}_{\varepsilon}$ при $\ell = 1$. В этом случае

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon}g^{(1)}(X_1,t_1,X_2,t_2) =\\
&= \sum_{j=0}^{\infty}\frac{\varepsilon^j}{j!}\iint\limits_{\Theta}(\lambda_2-\lambda_1)^+\lambda_1^{X_1+j}e^{-\lambda_1 t_1}\lambda_2^{X_2}e^{-\lambda_2 t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 =\\
&= \iint\limits_{\Theta}(\lambda_2-\lambda_1)^+\left(\sum_{j=0}^{\infty}\frac{(\lambda_1\varepsilon)^j}{j!}\right)\lambda_1^{X_1}e^{-\lambda_1 t_1}\lambda_2^{X_2}e^{-\lambda_2 t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 =\\
&= g^{(1)}(X_1,t_1-\varepsilon,X_2,t_2),
\end{aligned}$$

что соответствует (4.6).

Проверим первую оценку в (4.7). Достаточно рассмотреть случай $\ell = 1$. С учетом (3.19), (4.4) и неравенства $g(X_1,t_1,X_2,t_2) \leqslant g^{(1)}(X_1,t_1,X_2,t_2)$ получаем

$$\begin{aligned}
&\mathbf{T}^{(1)}_{\varepsilon}R_\varepsilon(X_1,t_1,X_2,t_2) - R_\varepsilon(X_1,t_1,X_2,t_2) \leqslant\\
&\leqslant (T-t)\sum_{j=1}^{\infty}\frac{\varepsilon^j}{j!}\times g^{(1)}(X_1+j,t_1,X_2,t_2) =
\end{aligned}$$

$$= (T-t)\iint\limits_{\Theta}(\lambda_2-\lambda_1)^+\left(\sum_{j=1}^{\infty}\frac{(\lambda_1\varepsilon)^j}{j!}\right)\lambda_1^{X_1}e^{-\lambda_1 t_1}\lambda_2^{X_2}e^{-\lambda_2 t_2}\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2 \leqslant$$
$$\leqslant (T-t)\left(e^{C\varepsilon}-1\right)g^{(1)}(X_1,t_1,X_2,t_2).$$

Поскольку также выполнено неравенство

$$\mathbf{T}_{\varepsilon}^{(1)}R_{\varepsilon}(X_1,t_1,X_2,t_2)-R_{\varepsilon}(X_1,t_1,X_2,t_2)\leqslant (T-t)\left(e^{C\varepsilon}-1\right)g^{(2)}(X_1,t_1,X_2,t_2),$$

то отсюда следует первая оценка в (4.7). Вторая и третья оценки в (4.7) проверяются аналогично. ▲

Установим условия Липшица для $R_{\varepsilon}(X_1,t_1,X_2,t_2)$ по t_1,t_2. Справедлива следующая

Лемма 5. *Пусть $\delta=K\varepsilon$, где K – целое число. Тогда имеют место оценки*

$$\begin{aligned}
&|R_{\varepsilon}(X_1,t_1,X_2,t_2)-R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2)|\leqslant\\
&\leqslant \delta g^{(1)}(X_1,t_1,X_2,t_2)+(e^{C\delta}-1)(T-t-\delta)g(X_1,t_1+\delta,X_2,t_2),\\
&|R_{\varepsilon}(X_1,t_1,X_2,t_2)-R_{\varepsilon}(X_1,t_1,X_2,t_2+\delta)|\leqslant\\
&\leqslant \delta g^{(2)}(X_1,t_1,X_2,t_2)+(e^{C\delta}-1)(T-t-\delta)g(X_1,t_1,X_2,t_2+\delta).
\end{aligned}\tag{4.8}$$

Доказательство. Достаточно установить первую оценку. Обозначим через $R_{\varepsilon}^{(1,K)}(X_1,t_1,X_2,t_2)$ потери, если сначала K раз применялось первое действие, а затем управление осуществлялось оптимально. Справедливо уравнение

$$\begin{aligned}
&R_{\varepsilon}^{(1,i)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)=\varepsilon g^{(1)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)+\\
&+\mathbf{T}_{\varepsilon}^{(1)}R_{\varepsilon}^{(1,i-1)}(X_1,t_1+(K-i+1)\varepsilon,X_2,t_2),
\end{aligned}\tag{4.9}$$

где $i=1,2,\ldots,K$, причем $R_{\varepsilon}^{(1,0)}(X_1,t_1+K\varepsilon,X_2,t_2)=R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2)$. Из (4.9) следует, что

$$\begin{aligned}
&R_{\varepsilon}^{(1,1)}(X_1,t_1+(K-1)\varepsilon,X_2,t_2)=\varepsilon g^{(1)}(X_1,t_1+(K-1)\varepsilon,X_2,t_2)+\\
&+\mathbf{T}_{\varepsilon}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2).
\end{aligned}$$

Далее, с учетом (4.5), (4.6) получаем, что

$$\begin{aligned}
&R_{\varepsilon}^{(1,i)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)=i\varepsilon g^{(1)}(X_1,t_1+(K-i)\varepsilon,X_2,t_2)+\\
&+\mathbf{T}_{i\varepsilon}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2)\quad\text{при } i=2,\ldots,K.
\end{aligned}$$

При $i=K$ получаем

$$R_{\varepsilon}^{(1,K)}(X_1,t_1,X_2,t_2)=\delta g^{(1)}(X_1,t_1,X_2,t_2)+\mathbf{T}_{\delta}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2).\tag{4.10}$$

С другой стороны, справедлива оценка

$$R_{\varepsilon}(X_1,t_1,X_2,t_2)\geqslant\mathbf{T}_{\delta}^{(1)}R_{\varepsilon}(X_1,t_1+\delta,X_2,t_2),\tag{4.11}$$

которая следует из того, что байесовский риск на меньшем горизонте управления и при наличии дополнительной информации, обусловленной K-кратным применением первого действия, не превосходит исходного байесовского риска. Из (4.10), (4.11) и неравенства

$$R_{\varepsilon}(X_1,t_1,X_2,t_2)\leqslant R_{\varepsilon}^{(1,K)}(X_1,t_1,X_2,t_2)$$

следует, что

$$\left|R_\varepsilon(X_1,t_1,X_2,t_2)-\mathbf{T}_\delta^{(1)}R_\varepsilon(X_1,t_1+\delta,X_2,t_2)\right|\leqslant\delta g^{(1)}(X_1,t_1,X_2,t_2).$$

С учетом первой оценки (4.7) отсюда следует (4.8). ▲

Далее считаем, что $\varepsilon\to 0$. Отметим, что при малых ε управление соответствует обработке доходов по одному. Справедлива следующая

Теорема 4. *При $t_1\geqslant t_0$, $t_2\geqslant t_0$ существует предел*

$$R(X_1,t_1,X_2,t_2)=\lim_{\varepsilon\to+0}R_\varepsilon(X_1,t_1,X_2,t_2). \tag{4.12}$$

Этот предел ограничен в соответствии с оценкой (4.4) *и удовлетворяет условиям Липшица по t_1,t_2 в соответствии с оценками* (4.8). *Байесовский риск* (1.6) *равен*

$$R_T(\mu)=\lim_{t_0\to+0}R(0,t_0,0,t_0). \tag{4.13}$$

Доказательство. Для некоторого ε рассмотрим последовательность $\varepsilon_i=2^{-i}\varepsilon$, $i=1,2,\ldots$ Так как уменьшение величины ε_i означает, что действия можно менять чаще, то $R_{\varepsilon_i}(X_1,t_1,X_2,t_2)$ при фиксированных X_1,t_1,X_2,t_2 является неубывающей функцией ε_i. Поскольку $R_{\varepsilon_i}(X_1,t_1,X_2,t_2)\geqslant 0$, то предел (4.12) существует при всех $\{t_\ell\}$ вида $t_\ell=t_0+k\varepsilon_i$, $\ell=1,2$, $i=1,2,\ldots$ Выполнение оценок (4.4) и (4.8) для $R(X_1,t_1,X_2,t_2)$ устанавливается предельным переходом по $\varepsilon_i\to 0$. Поэтому полученный предел можно по непрерывности доопределить на все $t_1\geqslant t_0$, $t_2\geqslant t_0$. Формула (4.13) следует из (3.20) и (4.4), так как первое слагаемое в правой части (3.20) и все

$$\frac{R(X_1,t_0,X_2,t_0)t_0^{X_1+X_2}}{X_1!\,X_2!}$$

при $X_1+X_2>0$ стремятся к нулю при $t_0\to 0$. ▲

Покажем, что $R(X_1,t_1,X_2,t_2)$ удовлетворяет дифференциальному уравнению в частных производных

$$\begin{aligned}&\min\biggl(\frac{\partial R}{\partial t_1}+R(X_1+1,t_1,X_2,t_2)+g^{(1)}(X_1,t_1,X_2,t_2),\\&\frac{\partial R}{\partial t_2}+R(X_1,t_1,X_2+1,t_2)+g^{(2)}(X_1,t_1,X_2,t_2)\biggr)=0\end{aligned} \tag{4.14}$$

с начальным условием

$$R(X_1,t_1,X_2,t_2)=0\quad\text{при } t_1+t_2=T, \tag{4.15}$$

при этом байесовский риск (1.6) вычисляется по формуле (4.13). Дифференциальное уравнение (4.14) одновременно описывает не только эволюцию байесовского риска $R(X_1,t_1,X_2,t_2)$, но и байесовскую стратегию, которая предписывает выбирать ℓ-е действие, если ℓ-й член в левой части (4.14) имеет меньшее значение; в случае их равенства выбор действия может быть произвольным. Существования такой предельной стратегии, в свою очередь, достаточно для строгого вывода уравнения (4.14). Однако пока этого сделать не удалось, хотя можно, как это сделано в [15], доказать, что этот предел существует в некоторых областях.

Зафиксируем некоторое $\varepsilon>0$, и пусть $\varepsilon_i=2^{-i}\varepsilon$, $i=1,2,\ldots$ Из (4.10) и второй оценки в (4.7) следует уравнение

$$\begin{aligned}&R_{\varepsilon_i}^{(1,2^i)}(X_1,t_1,X_2,t_2)=\varepsilon g^{(1)}(X_1,t_1,X_2,t_2)+\\&+R_{\varepsilon_i}(X_1,t_1+\varepsilon,X_2,t_2)+\varepsilon R_{\varepsilon_i}(X_1+1,t_1+\varepsilon,X_2,t_2)+\alpha(\varepsilon),\end{aligned} \tag{4.16}$$

где $|\alpha(\varepsilon)| \leqslant (e^{C\varepsilon} - 1 - \varepsilon C)(T-t)g(X_1,t_1,X_2,t_2) = O(\varepsilon^2)$. Аналогично,

$$R_{\varepsilon_i}^{(2,2^i)}(X_1,t_1,X_2,t_2) = \varepsilon g^{(2)}(X_1,t_1,X_2,t_2) + \\ + R_{\varepsilon_i}(X_1,t_1,X_2,t_2+\varepsilon) + \varepsilon R_{\varepsilon_i}(X_1,t_1,X_2+1,t_2+\varepsilon) + \alpha(\varepsilon), \tag{4.17}$$

Если при всех t_1', t_2', таких что $t_1' \geqslant t_1$, $t_2' \geqslant t_2$, $t_1' + t_2' < t_1 + t_2 + \varepsilon$, оптимальным является одно и то же действие, то уравнение (3.16), которым следует дополнить уравнения (4.16), (4.17), можно записать в виде

$$\min\Big(R_{\varepsilon_i}^{(1,2^i)}(X_1,t_1,X_2,t_2) - R_{\varepsilon_i}(X_1,t_1,X_2,t_2), \\ R_{\varepsilon_i}^{(2,2^i)}(X_1,t_1,X_2,t_2) - R_{\varepsilon_i}(X_1,t_1,X_2,t_2)\Big) = 0. \tag{4.18}$$

Выполняя предельные переходы сначала по $i \to \infty$, а затем по $\varepsilon \to 0$, из (4.16)–(4.18) получаем уравнение (4.14).

Отметим, что для численного решения уравнения (4.14) с начальным условием (4.15) следует использовать уравнения (3.16)–(3.18), в которых для вычисления операторов $\{\mathbf{T}_\varepsilon^{(\ell)}\}$ надо ограничиться слагаемыми порядка не выше ε.

§ 5. Асимптотическая оценка минимаксного риска снизу

Рассмотрим теперь асимптотическое описание байесовского риска при $T \to \infty$, которое в значительной степени аналогично приведенному в [16] описанию для бернуллиевского двурукого бандита. Будет показано, что при подходящем выборе априорного распределения он описывается тем же дифференциальным уравнением в частных производных второго порядка, что и байесовский риск для гауссовского двурукого бандита. Поскольку минимаксный риск не меньше любого байесовского, а гауссовский двурукий бандит описывает пакетную обработку, эти результаты означают, что минимаксный риск для пуассоновского двурукого бандита при обработке доходов по одному не может быть сделан меньше минимаксного риска, соответствующего оптимальной пакетной обработке, если $T \to \infty$.

Отметим, что в [16] асимптотическое описание было получено для риска

$$\widehat{R}(X_1,t_1,X_2,t_2) = R^B(X_1,t_1,X_2,t_2)\mu(X_1,t_1,X_2,t_2). \tag{5.1}$$

В этом случае при подходящей нормировке переменных X_1,t_1,X_2,t_2 и самого риска можно получить дифференциальное уравнение в частных производных второго порядка. Однако получить соответствующее дифференциальное уравнение для $R_\varepsilon(X_1,t_1,X_2,t_2)$, описываемого разностными уравнениями (3.16)–(3.19), а в предельном случае – дифференциальным уравнением (4.14), не удается по той причине, что этот риск отличается от риска $\widehat{R}_\varepsilon(X_1,t_1,X_2,t_2)$ множителем $X_1!\,X_2!\,t_1^{-X_1}t_2^{-X_2}$, который не является медленно меняющимся при любых изменениях X_1 и X_2.

Получим дифференциальное уравнение для $\widehat{R}(X_1,t_1,X_2,t_2)$. Снова предполагаем, что на начальном этапе управления оба действия применяются по t_0 раз, а затем при $t_1 > t_0$, $t_2 > t_0$ управление осуществляется оптимально в соответствии с уравнением (4.14). Если $t_0 \ll T$, то такая стратегия практически не приводит к увеличению байесовского риска. Справедлива следующая

Теорема 5. *Риск* (5.1) *удовлетворяет дифференциальному уравнению*

$$\min_{\ell=1,2}\left(\frac{\partial\widehat{R}}{\partial t_\ell} + D^{(\ell)}\widehat{R}(X_1,t_1,X_2,t_2) + \widehat{g}^{(\ell)}(X_1,t_1,X_2,t_2)\right) = 0 \tag{5.2}$$

с начальным условием

$$\widehat{R}(X_1,t_1,X_2,t_2)=0 \quad \textit{при } t_1+t_2=T, \tag{5.3}$$

где

$$\begin{aligned}
&D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)=\big(\widehat{R}(X_1+1,t_1,X_2,t_2)(X_1+1)-\widehat{R}(X_1,t_1,X_2,t_2)X_1\big)t_1^{-1},\\
&D^{(2)}\widehat{R}(X_1,t_1,X_2,t_2)=\big(\widehat{R}(X_1,t_1,X_2+1,t_2)(X_2+1)-\widehat{R}(X_1,t_1,X_2,t_2)X_2\big)t_2^{-1},\\
&\widehat{g}^{(1)}(X_1,t_1,X_2,t_2)=\iint\limits_{\Theta}(\lambda_2-\lambda_1)^+p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,\\
&\widehat{g}^{(2)}(X_1,t_1,X_2,t_2)=\iint\limits_{\Theta}(\lambda_1-\lambda_2)^+p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2.
\end{aligned} \tag{5.4}$$

Байесовский риск (1.6) *вычисляется по формуле*

$$R_T(\mu)=t_0\iint\limits_{\Theta}|\lambda_1-\lambda_2|\mu(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2+\sum_{X_1=0}^{\infty}\sum_{X_2=0}^{\infty}\widehat{R}(X_1,t_0,X_2,t_0). \tag{5.5}$$

Доказательство. Из (5.1) и (3.14) следует, что

$$R(X_1,t_1,X_2,t_2)=\widehat{R}(X_1,t_1,X_2,t_2)X_1!\,X_2!\,t_1^{-X_1}t_2^{-X_2}.$$

Подставляя это выражение в первый член в левой части (4.14), получаем при $X_1>0$

$$\begin{aligned}
&\left(\widehat{R}(X_1,t_1,X_2,t_2)\frac{X_1!\,X_2!}{t_1^{X_1}t_2^{X_2}}\right)'_{t_1}+\widehat{R}(X_1+1,t_1,X_2,t_2)\frac{(X_1+1)!\,X_2!}{t_1^{X_1+1}t_2^{X_2}}+\\
&+g(X_1,t_1,X_2,t_2)=\Big(\widehat{R}'_{t_1}(X_1,t_1,X_2,t_2)+D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)+\\
&+\widehat{g}(X_1,t_1,X_2,t_2)\Big)\frac{X_1!\,X_2!}{t_1^{X_1}t_2^{X_2}},
\end{aligned}$$

что с точностью до множителя $X_1!\,X_2!\,t_1^{-X_1}t_2^{-X_2}$ равно члену в левой части (5.2) при $\ell=1$. При $X_2>0$ так же связаны второй член в левой части (4.14) и член в левой части (5.2) при $\ell=2$. Проверка показывает, что при $X_1=0$ и/или $X_2=0$ эти выражения сохраняются. Поэтому из (4.14), (4.15) следуют (5.2), (5.3). Формула (5.5) следует из (5.1) и (3.10). ▲

Для некоторого $\lambda>0$ положим

$$\begin{aligned}
&\lambda_1=\lambda+(m+w)T^{-1/2},\quad \lambda_2=\lambda+(m-w)T^{-1/2},\\
&X_\ell=\lambda t_\ell+x_\ell T^{1/2},\quad \tau_\ell=t_\ell/T,\quad \ell=1,2.
\end{aligned}$$

В качестве множества параметров выберем

$$\Theta=\big\{\theta=(\lambda+(m+w)T^{-1/2},\lambda+(m-w)T^{-1/2}):\ |w|\leqslant c,\ |m|\leqslant a_T\big\},$$

где $c>0$ – достаточно большая фиксированная константа, $a_T=T^\alpha$, $0<\alpha<1/2$ и T достаточно велико. В дальнейшем удобно изменить параметризацию и от $\theta=(\lambda_1,\lambda_2)$ перейти к $\theta'=(m,w)$. Априорную плотность распределения выберем в виде $T\varkappa_a(m)\varrho(w)$, где $\varkappa_a(m)=(2a_T)^{-1}$ – плотность равномерного распределения на отрезке $|m|\leqslant a_T$, а $\varrho(w)$ – произвольная плотность на отрезке $|w|\leqslant c$.

Далее положим

$$D_1 = D_2 = \lambda, \quad t_\ell^* = t_\ell D_\ell, \quad t'_\ell = t_\ell / D_\ell, \quad \tau_\ell^* = \tau_\ell D_\ell, \quad \tau'_\ell = \tau_\ell / D_\ell, \quad \tau' = \tau'_1 + \tau'_2,$$
$$\varepsilon_0 = t_0/T, \quad \varepsilon = T^{-1}, \quad \delta = T^{-1/2}, \quad \widehat{R}(X_1, t_1, X_2, t_2) = T^{-1/2}\widehat{r}(x_1, \tau_1, x_2, \tau_2).$$

Отметим, что в этом параграфе ε и δ не характеризуют длины полуинтервалов. Будем писать

$$x_T \sim y_T, \quad \text{если } \lim_{T\to\infty} \frac{x_T}{y_T} = 1,$$
$$x_T \lesssim y_T, \quad \text{если } \lim_{T\to\infty} \frac{x_T}{y_T} \leqslant 1.$$

Если t_0 достаточно велико, то при $t_1 \geqslant t_0$, $t_2 \geqslant t_0$ справедливы оценки

$$\begin{aligned} &p(X_1, t_1; \lambda_1) \sim T^{-1/2} f_{\tau_1^*}(x_1 \,|\, (m+w)\tau_1), \\ &p(X_2, t_2; \lambda_2) \sim T^{-1/2} f_{\tau_2^*}(x_2 \,|\, (m-w)\tau_2). \end{aligned} \tag{5.6}$$

Действительно, в силу центральной предельной теоремы для плотностей имеем

$$p(X_\ell, t_\ell; \lambda_\ell) \sim f_{t_\ell^*}(X_\ell \,|\, \lambda_\ell t_\ell) = f_{t_\ell^*}(X_\ell - \lambda t_\ell \,|\, \lambda_\ell t_\ell - \lambda t_\ell).$$

Отсюда с учетом сделанной замены переменных следует (5.6).

Выберем $b_T > 0$ из условий $b_T \to +\infty$, $b_T/a_T \to 0$ при $T \to \infty$. Сделаем замену переменных $y = (\bar{x}_1\tau'_1 + \bar{x}_2\tau'_2)/\tau'$, $z = x_1\tau_2 - x_2\tau_1$, где $\bar{x}_1 = x_1/\tau_1$, $\bar{x}_2 = x_2/\tau_2$. Так же, как в [16], с учетом (5.6) устанавливаются оценки

$$\begin{aligned} &\widehat{g}^{(\ell)}(x_1, t_1, X_2, t_2) \sim T^{-3/2}(2a_T)^{-1}\widehat{g}^{(\ell)}(z, \tau_1, \tau_2) \quad \text{при } |y| \leqslant a_T - b_T, \\ &\widehat{g}^{(\ell)}(x_1, t_1, X_2, t_2) \lesssim T^{-3/2}(2a_T)^{-1}\widehat{g}^{(\ell)}(z, \tau_1, \tau_2) \quad \text{при } a_T - b_T < |y| \leqslant a_T + b_T, \\ &\widehat{g}^{(\ell)}(x_1, t_1, X_2, t_2) = T^{-3/2}(2a_T)^{-1} o(e^{-\gamma(y-a_T)^2}) \quad \text{при } |y| > a_T + b_T, \ \gamma > 0, \end{aligned} \tag{5.7}$$

где

$$\begin{aligned} \widehat{g}^{(1)}(z, \tau_1, \tau_2) &= \int_{-c}^{0} 2|w| f_{\tau_1^*\tau_2^*\tau'}(z - 2w\tau_1\tau_2)\varrho(w)\,dw, \\ \widehat{g}^{(2)}(z, \tau_1, \tau_2) &= \int_{0}^{c} 2|w| f_{\tau_1^*\tau_2^*\tau'}(z - 2w\tau_1\tau_2)\varrho(w)\,dw. \end{aligned} \tag{5.8}$$

Кроме того, так же, как в [16], устанавливаются оценки

$$\begin{aligned} &\widehat{r}(x_1, \tau_1, x_2, \tau_2) = (2a_T)^{-1} O(1) \quad \text{при } |y| \leqslant a_T + b_T, \\ &\widehat{r}(x_1, \tau_1, x_2, \tau_2) = (2a_T)^{-1} o(e^{-\gamma(y-a_T)^2}) \quad \text{при } |y| > a_T + b_T, \ \gamma > 0. \end{aligned} \tag{5.9}$$

Покажем, что для

$$\mathrm{r}(z, \tau_1, \tau_2) = (2a_T)\widehat{r}(x_1, \tau_1, x_2, \tau_2)$$

при $|y| \leqslant a_T - b_T$ и $T \to \infty$ справедливо дифференциальное уравнение в частных производных второго порядка

$$\min_{\ell=1,2}\Big(\mathrm{r}'_{\tau_\ell} + \tau_\ell^{-1}\mathrm{r} + z\tau_\ell^{-1}\mathrm{r}'_z + 0{,}5D_\ell(\tau_{3-\ell})^2\mathrm{r}''_{zz} + \widehat{g}^{(\ell)}(z, \tau_1, \tau_2)\Big) = 0 \tag{5.10}$$

с начальным условием

$$\mathrm{r}(z,\tau_1,\tau_2)=0 \quad \text{при } \tau_1+\tau_2=1. \tag{5.11}$$

При этом байесовский риск равен

$$R_N^B(\lambda)\sim T^{1/2}\biggl(\varepsilon_0\int\limits_{-c}^{c}2|w|\varrho(w)\,dw+\int\limits_{-\infty}^{\infty}\mathrm{r}(z,\varepsilon_0,\varepsilon_0)\,dz\biggr). \tag{5.12}$$

Запишем уравнение (5.2) с использованием переменных x_1,τ_1,x_2,τ_2. Чтобы выразить в новых переменных выражение $D^{(\ell)}\widehat{R}(X_1,t_1,X_2,t_2)$, заметим, что паре (X_ℓ,t_ℓ) соответствует (x_ℓ,τ_ℓ) по определению, а паре $(X_\ell+1,t_\ell)$ соответствует $(x_\ell+\delta,\tau_\ell)$. Действительно,

$$X_\ell+1=\lambda t_\ell+x_\ell T^{1/2}+1=\lambda t_\ell+x'_\ell T^{1/2},$$

откуда $x'_\ell=x_\ell+\delta$. Обозначим $\widetilde{x}_\ell=x_\ell/\tau_\ell$, $\ell=1,2$. Пусть $\ell=1$. Так как

$$(X_1+1)/t_1=(\lambda\tau_1T+x_1T^{1/2}+1)/(\tau_1T)=\lambda+\widetilde{x}_1\delta+\varepsilon/\tau_1,\quad X_1/t_1=\lambda+\widetilde{x}_1\delta,$$

то

$$\begin{aligned}&\widehat{R}(X_1+1,t_1,\cdot\,)(X_1+1)/t_1=\delta\times\widehat{r}(x_1+\delta,\tau_1,\cdot\,)(\lambda+\widetilde{x}_1\delta+\varepsilon/\tau_1),\\&\widehat{R}(X_1,t_1,\cdot\,)(X_1/t_1)=\delta\times\widehat{r}(x_1,\tau_1,\cdot\,)(\lambda+\widetilde{x}_1\delta).\end{aligned} \tag{5.13}$$

В этих обозначениях опущена зависимость $\widehat{R}(X_1,t_1,X_2,t_2)$ от X_2,t_2 и зависимость $\widehat{r}(x_1,\tau_1,x_2,\tau_2)$ от x_2,τ_2, поскольку при $\ell=1$ эти переменные не меняются. Если $r(x_1,\tau_1,\cdot\,)$ – достаточно гладкая функция x_1, то разлагая ее в ряд Тейлора до членов порядка ε и учитывая (5.13), получаем

$$\begin{aligned}&D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)=\widehat{R}(X_1+1,t_1,\cdot\,)(X_1+1)/t_1-\widehat{R}(X_1,t_1,\cdot\,)(X_1/t_1)=\\&=\delta(\widehat{r}+\widehat{r}'_{x_1}\delta+0{,}5\varepsilon\widehat{r}''_{x_1x_1}+O(\varepsilon^{3/2}))(\lambda+\widetilde{x}_1\delta+\varepsilon/\tau_1)-\delta\times\widehat{r}\times(\lambda+\widetilde{x}_1\delta)=\\&=\delta\Bigl(\widehat{r}\varepsilon/\tau_1+\widehat{r}'_{x_1}(\lambda\delta+\widetilde{x}_1\varepsilon)+0{,}5\varepsilon\lambda\widehat{r}''_{x_1x_1}+O(\varepsilon^{3/2})\Bigr),\end{aligned} \tag{5.14}$$

где $\widehat{r}=\widehat{r}(x_1,\tau_1,x_2,\tau_2)$. Так как $x_1=T^{-1/2}(X_1-\lambda t_1)$, $\tau_1=t_1/T$, то

$$\begin{aligned}&\widehat{R}'_{t_1}(X_1,t_1,X_2,t_2)=T^{-1/2}\widehat{r}'_{t_1}(T^{-1/2}(X_1-\lambda t_1),t_1/T,x_2,\tau_2)=\\&=T^{-1/2}\Bigl(\widehat{r}'_{x_1}(x_1,\tau_1,x_2,\tau_2)(-\lambda T^{-1/2})+\widehat{r}'_{\tau_1}(x_1,\tau_1,x_2,\tau_2)T^{-1}\Bigr)=-\lambda\varepsilon\widehat{r}'_{x_1}+\delta\varepsilon\widehat{r}'_{\tau_1}.\end{aligned}$$

Поэтому с учетом (5.14), (5.7) левая часть (5.2) при $\ell=1$ принимает вид

$$\begin{aligned}&\widehat{R}'_{t_1}(X_1,t_1,X_2,t_2)+D^{(1)}\widehat{R}(X_1,t_1,X_2,t_2)+g^{(1)}(X_1,t_1,X_2,t_2)=\\&=-\lambda\varepsilon r'_{x_1}+\delta\varepsilon r'_{\tau_1}+\delta\Bigl(r\varepsilon/\tau_1+r'_{x_1}(\lambda\delta+\widetilde{x}_1\varepsilon)+0{,}5\lambda r''_{x_1x_1}\varepsilon+O(\varepsilon^{3/2})\Bigr)+\\&+\varepsilon\delta(2a_T)^{-1}\widehat{g}^{(1)}(z,\tau_1,\tau_2)=\\&=\varepsilon\delta\Bigl(r'_{\tau_1}+r/\tau_1+r'_{x_1}\widetilde{x}_1+0{,}5\lambda r''_{x_1x_1}+(2a_T)^{-1}\widehat{g}^{(1)}(z,\tau_1,\tau_2)+O(\varepsilon^{1/2})\Bigr).\end{aligned} \tag{5.15}$$

Положим

$$\widehat{r}(x_1,\tau_1,x_2,\tau_2)=(2a_T)^{-1}\mathrm{r}(z,\tau_1,\tau_2),\quad \text{где}\quad z=x_1\tau_2-x_2\tau_1.$$

Таблица 1

Байесовский риск в зависимости от ε и $\Delta\lambda$

t_0	$\varepsilon \diagdown \Delta\lambda$	0,08	0,16	0,20	0,24	0,28	0,32	0,36	0,40
0	1	0,3478	0,5310	0,5731	0,5904	0,5891	0,5748	0,5519	0,5246
0	0,5	0,3469	0,5276	0,5678	0,5830	0,5794	0,5627	0,5377	0,5082
0	0,25	0,3467	0,5268	0,5666	0,5813	0,5772	0,5599	0,5343	0,5042
0	0,125	0,3467	0,5266	0,5663	0,5809	0,5766	0,5592	0,5334	0,5033
1	0,125	0,3472	0,5297	0,5716	0,5890	0,5881	0,5748	0,5540	0,5295

Тогда

$$\widehat{r}'_{\tau_1} = (2a_T)^{-1}(-x_2\mathrm{r}'_z + \mathrm{r}'_{\tau_1}), \quad \widehat{r}'_{x_1} = (2a_T)^{-1}\mathrm{r}'_z\tau_2, \quad \widehat{r}''_{x_1x_1} = (2a_T)^{-1}\mathrm{r}''_{zz}\tau_2^2.$$

Так как $a_T = T^\alpha$, $0 < \alpha < 1/2$, то (5.15) принимает вид

$$\varepsilon\delta(2a_T)^{-1}\Big(\mathrm{r}'_{\tau_1} + \mathrm{r}/\tau_1 + (z/\tau_1)\mathrm{r}'_z + 0{,}5\lambda\tau_2^2\mathrm{r}''_{zz} + \widehat{g}^{(1)}(z,\tau_1,\tau_2) + O(\varepsilon^{1/2-\alpha})\Big). \tag{5.16}$$

Аналогично,

$$\widehat{R}'_{t_2}(X_1,t_1,X_2,t_2) + D^{(2)}\widehat{R}(X_1,t_1,X_2,t_2) + g^{(2)}(X_1,t_1,X_2,t_2) = \varepsilon\delta(2a_T)^{-1} \times$$
$$\times \Big(\mathrm{r}'_{\tau_2} + \mathrm{r}/\tau_2 + (z/\tau_2)\mathrm{r}'_z + 0{,}5\lambda\tau_1^2\mathrm{r}''_{zz} + \widehat{g}^{(2)}(z,\tau_1,\tau_2) + O(\varepsilon^{1/2-\alpha})\Big). \tag{5.17}$$

Подставляя (5.16), (5.17) в (5.2), учитывая, что $\lambda = D_1 = D_2$, и переходя к пределу при $T \to \infty$, что соответствует $\varepsilon \to 0$, получаем уравнение (5.10). Начальные условия (5.11) следуют из (5.3). Получение формулы (5.12) из (5.5) выполняется с учетом (5.9) так же, как формулы (6.18) в [16]. При этом в [16] показано, что (5.10)–(5.12) описывают байесовский риск, вычисленный для гауссовского двурукого бандита относительно наихудшего априорного распределения (см. [16, теорема 8]).

§ 6. Численные эксперименты

О качестве управления в зависимости от значения ε можно судить по соответствующей величине байесовского риска. Результаты представлены в табл. 1. Априорное распределение выбрано равномерным на множестве из 18 параметров

$$\theta_{1,i} = (\lambda_i + \Delta\lambda, \lambda_i - \Delta\lambda), \quad \theta_{2,i} = (\lambda_i - \Delta\lambda, \lambda_i + \Delta\lambda),$$

где $\lambda_i = 0{,}8 + 0{,}05\,(i-1)$, $i = 1,\ldots,9$. Горизонт управления выбран равным $T = 30$, а возможные значения $\{\Delta\lambda\}$ и $\{\varepsilon\}$ представлены в верхней строке и втором слева столбце таблицы. В первых четырех строках при задании стратегии принято $t_0 = 0$, в последней строке $t_0 = 1$ (для сравнения результатов). Отметим, что значение $\varepsilon = 1$ соответствует 30 полуинтервалам переключения стратегии, что уже обеспечивает высокое качество управления.

О поведении стратегии на конкретной траектории управления можно судить по рис. 1. В этом случае байесовская стратегия была вычислена для указанного выше априорного распределения при $\Delta\lambda = 0{,}1$. Кроме того, выбраны $T = 120$, $t_0 = \varepsilon = 2$. Затем моделировалось управление с использованием найденной стратегии при $\lambda_1 = 1 + \Delta\lambda$, $\lambda_2 = 1 - \Delta\lambda$. Для текущей статистики (X_1, t_1, X_2, t_2), наблюдаемой в момент времени $t = t_1 + t_2$, величина x определялась из условия $x = X_1 - \widetilde{X}_1$ (или, эквивалентно, $x = \widetilde{X}_2 - X_2$), где $(\widetilde{X}_1, t_1, \widetilde{X}_2, t_2)$ – статистика, при которой происходит переключение оптимального действия с первого на второе (см. теорему 3). При этом при $x > 0$ выбирается первое действие, а при $x < 0$ – второе. На рис. 1 сплошной линией представлена типичная траектория отклонений X_1 от пороговых

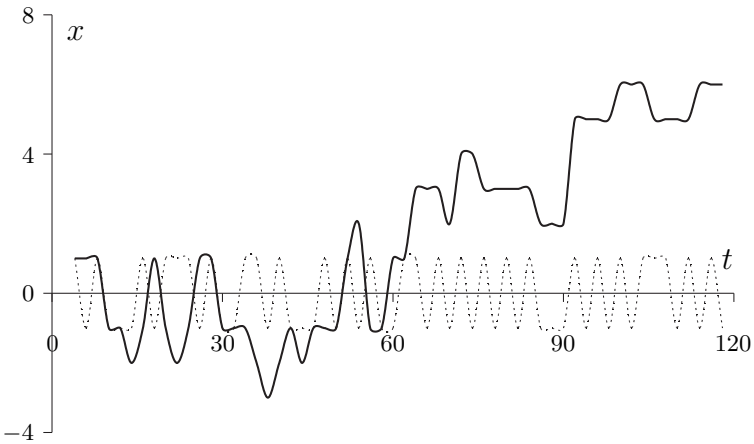

Рис. 1. Динамика отклонений X_1 от пороговых значений

значений $\widetilde{X}_1$. Такое поведение отклонений с учетом аргументации, представленной в [9, раздел 6], позволяет предположить, что в рассматриваемой постановке одновременное применение действий (с распределением ресурса между ними) практически не требуется. Такая необходимость в распределении ресурса возникает, когда отклонения минимальны при выборе сколь угодно малых значений ε, как это было бы в случае траектории, представленной пунктирной линией.

В качестве примера приближенного определения минимаксных стратегии и риска рассмотрим множество параметров

$$\Theta = \big\{(\lambda_1, \lambda_2) :\ 0{,}8 \leqslant (\lambda_1 + \lambda_2)/2 \leqslant 1{,}2;\ |\lambda_1 - \lambda_2|/2 \leqslant 0{,}4\big\}.$$

Приблизительно наихудшее априорное распределение было выбрано симметричным на множестве из шести пар параметров

$$\theta_{1,i} = (\lambda_i + \Delta\lambda_i, \lambda_i - \Delta\lambda_i), \quad \theta_{2,i} = (\lambda_i - \Delta\lambda_i, \lambda_i + \Delta\lambda_i), \quad i = 1, \ldots, 6,$$

т.е. $\Pr(\theta_{1,i}) = \Pr(\theta_{2,i}) = \mu_i$. Значения $\lambda_1 = 0{,}8$, $\lambda_2 = 0{,}85$, $\lambda_3 = 0{,}95$, $\lambda_4 = 1{,}05$, $\lambda_5 = 1{,}15$, $\lambda_6 = 1{,}2$ были фиксированы, а $\{\Delta\lambda_i\}$, $\{\mu_i\}$ требовалось найти в соответствии с условием (1.8).

В рассматриваемом случае байесовский риск становится функцией конечного числа переменных $\{\Delta\lambda_i\}$, $\{\mu_i\}$, поэтому для поиска максимума этой функции был использован градиентный метод. Начальные значения переменных были выбраны из условий $\mu_i = 1/12$, $\Delta\lambda_i = 1{,}6\,(\lambda_i/T)^{1/2}$, $i = 1, \ldots, 6$ (такие $\{\Delta\lambda_i\}$ соответствуют наихудшему распределению при больших T). При $T = 150$, $t_0 = 0$, $\varepsilon = 5$ параметры приблизительно наихудшего априорного распределения оказались следующими: $\Delta\lambda_1 \approx 0{,}138$, $\Delta\lambda_2 \approx 0{,}134$, $\Delta\lambda_3 \approx 0{,}127$, $\Delta\lambda_4 \approx 0{,}122$, $\Delta\lambda_5 \approx 0{,}123$, $\Delta\lambda_6 \approx 0{,}129$, $\mu_1 \approx 0{,}043$, $\mu_2 \approx 0$, $\mu_3 \approx 0{,}140$, $\mu_4 \approx 0{,}036$, $\mu_5 \approx 0$, $\mu_6 \approx 0{,}281$. Соответствующее значение нормированного байесовского риска равно $T^{-1/2} R_T^B(\mu) \approx 0{,}627$. Затем для найденной байесовской стратегии σ^B были вычислены нормированные потери

$$l_i(\Delta\lambda) = T^{-1/2} L_T^B(\sigma^B, (\lambda_i + \Delta\lambda, \lambda_i - \Delta\lambda)),$$

которые представлены на рис. 2. Номера линий $i = 1, 2, 3, 4, 5, 6$ соответствуют индексам $l_i(\Delta\lambda)$, при этом линии *4–6* почти совпадают. Максимальные потери, как и байесовский риск, приблизительно равны 0,627, поэтому найденные стратегию и риск можно считать приблизительно минимаксными. Отметим, что нормированный байесовский риск для начального распределения приблизительно равен 0,613,

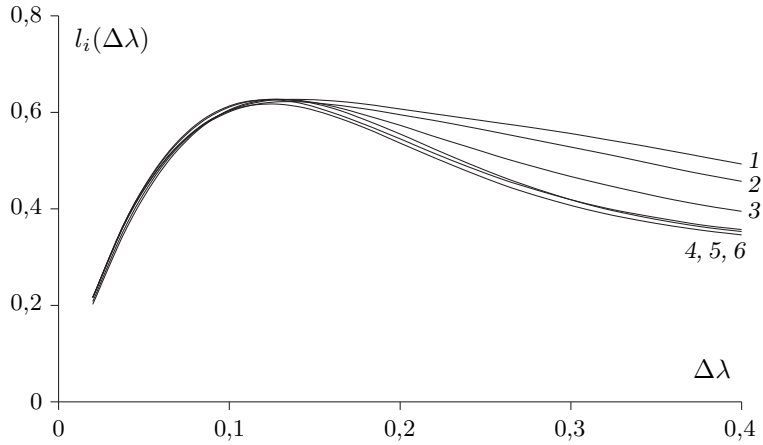

Рис. 2. Потери, обеспечиваемые приблизительно минимаксной стратегией

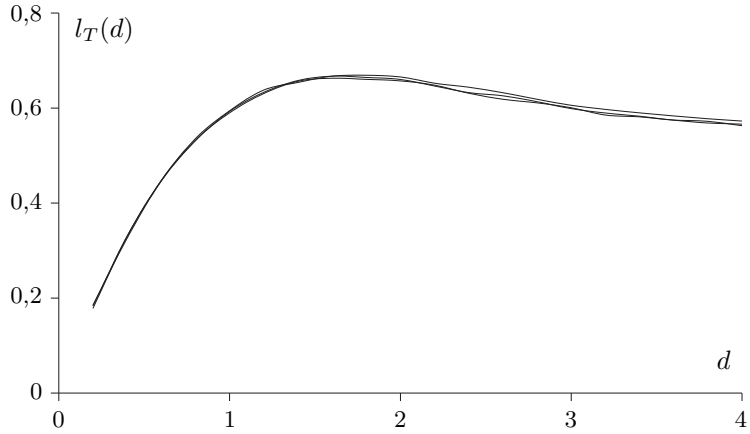

Рис. 3. Потери, обеспечиваемые пакетной обработкой

а максимальные нормированные потери $\{l_i(\Delta\lambda)\}$ приблизительно равны 0,659, т.е. эти параметры уже обеспечивают неплохое приближение минимаксного управления.

Минимаксное управление на больших горизонтах управления T выполнялось с использованием стратегий, полученных для гауссовского двурукого бандита, описывающего пакетную обработку. В [16] показано, что в этом случае максимальные потери достигаются на множестве "близких" распределений, для которых математические ожидания одношаговых доходов различаются на величину порядка $T^{-1/2}$, а нормированный минимаксный риск не меньше величины 0,637, к которой стремится, если количество пакетов неограниченно растет. При моделировании пакетной обработки нормированные потери определялись как

$$l_T(d) = (\lambda T)^{-1/2} L_T(\sigma, \theta_d),$$

где

$$\theta_d = (\lambda + d(\lambda/T)^{1/2}, \lambda - d(\lambda/T)^{1/2}),$$

а σ – минимаксная стратегия для гауссовского двурукого бандита. Интенсивность λ всегда выбиралась равной единице, а T выбиралось равным 600, 3000 и 15000.

Количество пакетов (полуинтервалов переключения стратегии) было равно 30, т.е. t_0 и ε выбирались равными 20, 100 и 500 соответственно. Так как стратегия пакетной обработки требует знания дисперсий одношаговых доходов, то на первых двух этапах, когда действия применяются поровну, делались оценки интенсивности как

$$\widehat{\lambda} = \frac{X_1 + X_2}{2t_0},$$

где X_1, X_2 – начальные доходы за применение первого и второго действий. На оставшихся 28 этапах $\widehat{\lambda}$ использовалась в качестве оценки обеих дисперсий. На рис. 3 представлены полученные результаты моделирования нормированных потерь методом Монте-Карло. Видно, что все кривые близки, хотя кривая, соответствующая $T = 600$, проходит чуть выше. Если оценку интенсивности не делать, а сразу принять $\lambda = 1$, то все три кривые практически совпадают с нижними двумя кривыми на рис. 3. Отметим, что максимум нормированных потерь при этом приблизительно равен 0,667, т.е. превышает минимально возможный менее чем на 5%.

§ 7. Заключение

Рассмотрено использование кусочно-постоянных стратегий в задаче о пуассоновском двуруком бандите в байесовской и минимаксной постановках. Такое управление соответствует пакетной обработке поступающих доходов. Результаты численных экспериментов показывают, что разбиение горизонта управления на 30 равных полуинтервалов, на которых стратегия остается постоянной, обеспечивает высокое качество управления. Для умеренных горизонтов управления минимаксные стратегию и риск можно искать как байесовские, вычисленные относительно наихудшего априорного распределения, которое с высокой степенью точности можно считать сосредоточенным на конечном множестве параметров. Для больших горизонтов можно использовать пакетные стратегии управления, ранее полученные для гауссовского двурукого бандита. При этом оптимальное управление, найденное для пуассоновского двурукого бандита, не позволяет уменьшить минимаксный риск, обеспечиваемый оптимальными пакетными стратегиями, если горизонт управления и количество пакетов неограниченно растут.

Автор выражает глубокую признательность рецензенту за внимание к статье и полезные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. *Berry D.A., Fristedt B.* Bandit Problems: Sequential Allocation of Experiments. London, New York: Chapman & Hall, 1985.
2. *Пресман Э.Л., Сонин И.М.* Последовательное управление по неполным данным. Байесовский подход. М.: Наука, 1982.
3. *Срагович В.Г.* Адаптивное управление. М.: Наука, 1981.
4. *Назин А.В., Позняк А.С.* Адаптивный выбор вариантов: рекуррентные алгоритмы. М.: Наука, 1986.
5. *Цетлин М.Л.* Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969.
6. *Варшавский В.И.* Коллективное поведение автоматов. М.: Наука, 1973.
7. *Пресман Э.Л.* Пуассоновский вариант задачи о «двуруком бандите» с дисконтированием // Теория вероятн. и ее примен. 1990. Т. 35. № 2. С. 318–328. `http://mi.mathnet.ru/tvp999`
8. *Chernoff H., Ray S.N.* A Bayes Sequential Sampling Inspection Plan // Ann. Math. Statist. 1965. V. 36. № 5. P. 1387–1407. `https://doi.org/10.1214/aoms/1177699898`

9. *Mandelbaum A.* Continuous Multi-Armed Bandits and Multiparameter Processes // Ann. Probab. 1987. V. 15. № 4. P. 1527–1556. https://doi.org/10.1214/aop/1176991992
10. *Lai T.L.* Adaptive Treatment Allocation and the Multi-Armed Bandit Problem // Ann. Statist. 1987. V. 15. № 3. P. 1091–1114. https://doi.org/10.1214/aos/1176350495
11. *Vogel W.* An Asymptotic Minimax Theorem for the Two Armed Bandit Problem // Ann. Math. Statist. 1960. V. 31. P. 444–451. https://doi.org/10.1214/aoms/1177705907
12. *Боровков А.А.* Математическая статистика. Дополнительные главы. М.: Наука, 1984.
13. *Колногоров А.В.* Нахождение минимаксных стратегии и риска в случайной среде (задаче о двуруком бандите) // АиТ. 2011. № 5. С. 127–138. http://mi.mathnet.ru/at1708
14. *Fabius J., van Zwet W.R.* Some Remarks on the Two-Armed Bandit // Ann. Math. Statist. 1970. V. 41. № 6. P. 1906–1916. https://doi.org/10.1214/aoms/1177696692
15. *Колногоров А.В.* К предельному описанию робастного параллельного управления в случайной среде // АиТ. 2015. № 7. С. 111–126. http://mi.mathnet.ru/at14258
16. *Колногоров А.В.* Гауссовский двурукий бандит: предельное описание // Пробл. передачи информ. 2020. Т. 56. № 3. С. 86–111. https://doi.org/10.31857/S0555292320030055

Колногоров Александр Валерианович
Новгородский государственный университет
им. Ярослава Мудрого, кафедра
прикладной математики и информатики
kolnogorov53@mail.ru

Поступила в редакцию
31.05.2021
После доработки
09.04.2022
Принята к публикации
18.04.2022