Математическая Теория Игр и её Приложения, т.14, в.2, с. 3–30

УДК 519.832, 519.245
ББК 22.18

АДАПТАЦИЯ СТРАТЕГИИ UCB ДЖ. БАСЕРА ДЛЯ ГАУССОВСКОГО МНОГОРУКОГО БАНДИТА*

Сергей В. Гарбарь
Александр В. Колногоров
Новгородский государственный университет
им. Ярослава Мудрого
173003, Великий Новгород, ул. Б.С.-Петербургская, 41
e-mail: Sergey.Garbar@novsu.ru, Alexander.Kolnogorov@novsu.ru

Рассмотрена адаптация стратегии UCB, впервые предложенной Дж. Басером для бернуллиевского двурукого бандита, на случай гауссовского многорукого бандита, описывающего пакетную обработку данных. Эта задача оптимального управления имеет классическую интерпретацию как игра с природой, в которой платежной функцией игрока является математическое ожидание потерь полного дохода, вызванное неполнотой информации. Цель управления сформулирована в минимаксной постановке. Для рассмотренной игры с природой построено инвариантное описание управления с горизонтом равным единице, позволяющее выполнять расчеты двумя способами: с использованием моделирования Монте-Карло и аналитически методом динамического программирования. Для различных конфигураций рассматриваемой игры с природой численными методами найдены седловые точки, характеризующие оптимальное управление и наихудшее распределение параметров многорукого бандита.

* Исследование выполнено при финансовой поддержке РФФИ, научный проект номер 20-01-00062

Ключевые слова: задача о многоруком бандите, гауссовский многорукий бандит, минимаксный подход, правило UCB, инвариантное описание, моделирование Монте-Карло, динамическое программирование.

Поступила в редакцию: 10.10.21 *После доработки*: 03.03.22 *Принята к публикации*: 16.05.22

1. Введение

Рассматривается задача о многоруком бандите (см., например, [8, 17]). Название происходит от игрального автомата с двумя или более рукоятками, которые далее будем называть действиями. Каждый выбор одного из действий приносит случайный доход, распределение которого зависит только от текущего выбранного действия, фиксировано в процессе игры, однако неизвестно игроку. Целью игрока является максимизация математического ожидания полного полученного дохода. Для достижения цели следует в процессе игры определить действие, математическое ожидание дохода за выбор которого будет максимальным, и обеспечить его преимущественное применение. Этот процесс всегда так или иначе сопровождается решением дилеммы «информация или управление», которая состоит в том, что для максимизации полного ожидаемого дохода хотелось бы применять только то действие, которому соответствует наибольший ожидаемый одношаговый доход, однако для определения этого действия требуется сравнивать его с остальными, применение которых ведет к уменьшению полного дохода.

Задача о многоруком бандите также известна как задача о целесообразном поведении [2, 13] и об адаптивном управлении в случайной среде [7, 12]. Это классическая задача обучения с подкреплением, широко известная в машинном обучении [10, 14, 15, 23]. Многорукие бандиты используются для оптимизации процессов в медицине, индустрии, исследовательских проектах, IT-технологиях и т.д. [19].

Далее будет рассматриваться гауссовский многорукий бандит с $J \geq 2$ действиями. Формально это управляемый случайный процесс $\xi(n)$, $n = 1, 2, \ldots, N$, где N называется горизонтом управления. Значение $\xi(n)$ в момент времени n зависит только от выбранного действия y_n и является нормально распределенной случайной величиной

с плотностью

$$f_{D_\ell}(x|m_l) = (2\pi D_\ell)^{-1/2} \exp\left(-\frac{(x-m_\ell)^2}{2D_\ell}\right) \tag{1.1}$$

при $y_n = \ell,\ \ell = 1, \ldots, J$. Дисперсии $D_1, \ldots, D_J$ предполагаются известными, причем без ограничения общности считаем, что их значения упорядочены, т.е. $D_1 \geq D_2 \geq \ldots \geq D_J$ (в противном случае действия можно перенумеровать). В дальнейшем предположение об априорной известности дисперсий будет снято. Математические ожидания $m_1, \ldots, m_J$ считаются неизвестными и расположены в произвольном порядке. Такой многорукий бандит полностью описывается параметром $\theta = (m_1, \ldots, m_J)$, множество допустимых значений которого обозначим Θ. Это множество предполагается известным и будет определено ниже.

Отметим, что классическим объектом управления в задаче о многоруком бандите является бернуллиевский управляемый случайный процесс, принимающий два возможных значения 1, 0 с вероятностями p_ℓ, q_ℓ соответственно ($p_\ell + q_\ell = 1$), если текущее выбранное действие есть $y_n = \ell$ (см., например, [2, 7, 13, 17]). В качестве приложения задачи о многоруком бандите в данной статье рассматривается обработка данных. В этом случае значения процесса 1 и 0 соответствуют успешно и неуспешно обработанному элементу данных, действиями являются альтернативные методы обработки, а цель состоит в максимизации математического ожидания полного количества успешно обработанных данных. Гауссовский многорукий бандит возникает при рассмотрении пакетной обработки, при которой одинаковые действия (методы) применяются к пакетам поступающих данных, а затем суммарные доходы в пакетах (количества успешно обработанных данных) используются для управления [3, 5, 6]. Если размеры пакетов достаточно велики, то при широких предположениях в силу центральной предельной теоремы распределения суммарных доходов в них будут близки к гауссовским.

Основное свойство пакетной обработки состоит в том, что если данные разбиты на достаточно большое число пакетов (на практике достаточно разбить их на 50 пакетов), то максимальные потери будут практически такими же, как и при оптимальной обработке данных по

одному [3, 5]. При этом пакетные стратегии управления в любом случае являются более удобными, так как позволяют менять действия значительно реже. Если же при этом допускается параллельная обработка данных, то они позволяют значительно уменьшить полное время обработки, так как оно теперь определяется количеством обрабатываемых пакетов, а не полным числом данных. Отметим, что ранее пакетная обработка рассматривалась в приложении к лечению пациентов альтернативными лекарствами (см., например, [22, 24]). В этом случае используется подход, при котором пациенты разбиваются на малое число групп, размеры которых возрастают с ростом номера этапа управления (обычно бывает два, редко – три или четыре этапа), так как пациенты не могут долго ждать.

Стратегия σ определяет выбор действия y_{n+1} с учетом имеющейся на текущий момент времени n информации об истории процесса управления. В этой статье мы изучаем стратегии, построенные на основе правила UCB, предложенного Дж. Басером в [16]. Пусть к моменту времени n действие ℓ было выбрано n_ℓ раз, обозначим через $X_\ell(n)$ текущий полный доход за использование этого действия ($\ell = 1, \ldots, J$). В таком случае $X_\ell(n)/n_\ell$ является точечной оценкой математического ожидания m_ℓ. Так как целью является максимизация математического ожидания полного дохода, то может показаться разумным в момент времени $n+1$ использовать действие, которому соответствует наибольшее текущее значение точечной оценки $X_\ell(n)/n_\ell$. Это известное правило play-the-leader. То, что это правило может привести к большим потерям, обсуждалось рядом авторов (см., например, [21]), поскольку в случае, когда начальная оценка $X_\ell(n)/n_\ell$, соответствующая максимальному значению m_ℓ, случайным образом будет иметь небольшое значение, лучшее действие может совсем не быть использовано в дальнейшем.

Дж. Басер предложил вместо точечных оценок математических ожиданий использовать для выбора действия верхние границы их интервальных оценок. В начале управления предложенная им стратегия предписывает использовать каждое из действий единожды, а затем в каждый момент времени $n+1$ следует выбирать действие, которому соответствует наибольшее значение величины

$$Q_\ell(n) = \frac{X_\ell(n)}{n_\ell} + \frac{4 + n_\ell^{1/2}}{15 n_\ell}\ (2 + \zeta_\ell(n)), \tag{1.2}$$

где $\zeta_l(n)$ – независимые одинаково распределенные случайные величины с экспоненциальной плотностью распределения равной e^{-x} при $x \geq 0$ и 0 при $x < 0$; $\ell = 1, 2, \ldots, J$; $n = J, \ldots, N-1$. Нетрудно видеть, что при использовании данной стратегии на бесконечном горизонте управления каждое действие с вероятностью 1 будет применено бесконечное число раз и, следовательно, точечные оценки $X_\ell(n)/n_\ell$ будут стремиться к математическим ожиданиям m_ℓ. Это приведет к тому, что доля применения действий с наибольшим значением m_ℓ будет стремиться к 1. Подобные стратегии называются правилами UCB (upper confidence bound – верхней границы доверительного интервала), они рассматривались, например, в [11, 14, 15, 20, 21, 23, 25].

Перейдем к формулировке цели управления. Если бы параметр $\theta = (m_1, \ldots, m_J)$ был известен, то выбором действия, соответсвующего максимальному одношаговому доходу, можно было бы обеспечить математическое ожидание полного дохода равное $N \max(m_1, \ldots, m_J)$. В случае применения стратегии σ полный доход будет меньше максимально возможного на величину

$$L_N(\sigma, \theta) = N \max(m_1, \ldots, m_J) - \mathbf{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi(n) \right), \tag{1.3}$$

называемую функцией потерь и обусловленную неполнотой информации. Здесь $\mathbf{E}_{\sigma,\theta}$ обозначает математическое ожидание, вычисленное по мере, порожденной стратегией σ и параметром θ. Через

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta) \tag{1.4}$$

обозначим минимаксный риск, вычисленный по функции потерь (1.3) на множестве параметров Θ для рассматриваемого множества стратегий. Соответствующая оптимальная стратегия σ^M, если она существует, называется минимаксной.

Задача о многоруком бандите является примером игры с природой [1]. В этом случае $L_N(\sigma, \theta)$ является платежной функцией, записанной в нормальной форме и характеризующей потери лица, осуществляющего управление; его стратегиями являются $\{\sigma\}$. Природа

свои стратегии выбирает из множества параметров Θ. Данная игра является антагонистической, но из двух игроков только лицо, осуществляющее управление, заинтересовано в минимизации платежной функции; природа к результатам игры равнодушна.

Опишем полученные результаты. В параграфе 2 предложена пакетная версия стратегии (1.2) и получено инвариантное описание управления на единичном горизонте. Эти результаты позволяют вычислять функцию потерь с помощью моделирования Монте-Карло. В параграфе 3 представлен другой подход к инвариантному описанию, основанный на использовании рекуррентного интегро-разностного уравнения динамического программирования в инвариантной форме с горизонтом управления равным единице. В пределе, когда количество пакетов неограниченно растет, данное рекуррентное уравнение переходит в дифференциальное уравнение в частных производных второго порядка. Отметим, что инвариантные описания справедливы в области так называемых «близких» распределений, которые характеризуются тем, что разности математических ожиданий имеют порядок $N^{-1/2}$. Именно в этой области функция потерь (1.3) достигает максимальных значений, имеющих порядок $N^{1/2}$ [26].

В параграфе 4 приведены результаты численных экспериментов с использованием моделирования Монте-Карло и решения рекуррентного и дифференциального уравнений. В результате для ряда конфигураций многоруких бандитов численными методами найдены седловые точки функции потерь (1.3), каждая из которых характеризует соответствующее минимаксное правило UCB, наихудшее априорное распределение параметра θ и значение минимаксного риска. В частности, для двурукого бандита с равными дисперсиями доходов $D_1 = D_2 = D$ оптимальное управление с использованием правила UCB при разбиении данных на 50 пакетов позволяет ограничить нормализованные (к величине $(DN)^{1/2}$) потери значением 0.73, характеризующим нормализованный минимаксный риск. Для сравнения, оптимальное управление, использующее всю предысторию процесса, в качестве такой границы дает 0.637, при этом стратегия UCB намного проще в реализации. Отметим, что в [16] в качестве оценки нормализованного минимаксного риска для правила UCB указано 0.72, однако без обоснования этого результата.

Кроме того, результаты параграфа 4 показывают, что функция потерь мало меняется при значительном изменении дисперсий, например, при изменении на 5–10%. Это означает, что дисперсии можно оценить во время обработки первых пакетов, когда все действия применяются по очереди, и в дальнейшем использовать эти оценки для управления, поэтому предположение об априорном знании дисперсий может быть опущено.

Параграф 5 содержит заключение. Отметим, что в случае равных дисперсий инвариантное описание управления с использованием правила UCB и результаты моделирования Монте-Карло были ранее представлены в работах авторов [4, 18].

2. Инвариантное описание управляющего алгоритма

Стратегия (1.2) была предложена в [16] для бернуллиевского двурукого бандита. Рассмотрим ее обобщение на случай гауссовского многорукого бандита с различными дисперсиями доходов. Обозначим $D = D_1 = \max(D_1, \ldots, D_J)$ и $\gamma_\ell = D_\ell / D$, $\ell = 1, \ldots, J$. Модифицированные границы (1.2) будем выбирать следующими:

$$Q_\ell(n) = \frac{X_\ell(n)}{n_\ell} + \frac{a_\ell(\gamma_1, \ldots, \gamma_J) D_\ell^{1/2}}{n_\ell^{1/2}} \, (2 + \zeta_\ell(n)), \qquad (2.1)$$

где $a_\ell(\gamma_1, \ldots, \gamma_J) > 0$ при $1 = \gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_J > 0$, а последовательность $\{\zeta_\ell(n)\}$ определена выше. Множители $(D_\ell / n_\ell)^{1/2}$ характеризует порядок ширины доверительных интервалов.

Функции $\{a_\ell(\gamma_1, \ldots, \gamma_J)\}$ являются параметрами стратегии UCB. Оптимальный выбор этих функций позволяет минимизировать максимальные потери (1.3) на множестве Θ, т.е. обеспечить достижение минимаксной цели управления (1.4). Ясно, что $a_l(\gamma_1, \ldots, \gamma_J) = a_k(\gamma_1, \ldots, \gamma_J)$, если $\gamma_l = \gamma_k$, так как равные дисперсии можно располагать в произвольном порядке. В частности, $a_\ell(\gamma_1, \ldots, \gamma_J) = a$ при всех $\ell = 1, \ldots, J$, если $D_1 = D_2 = \ldots = D_J$, где a зависит только от количества действий J. Отметим, что стратегии (1.2) и (2.1) эквивалентны с точностью до слагаемого порядка n_ℓ^{-1} (это слагаемое исключено для того, чтобы сделать возможным инвариантное описание) и множителя перед вторым слагаемым.

В параграфе 1 упоминалось, что максимальные потери достигаются в области «близких» распределений, которые характеризуются

тем, что разности математических ожиданий имеют порядок $N^{-1/2}$. Поясним этот факт, ограничившись случаем $J = 2$. Действительно, если $|m_1 - m_2| = cN^{-1/2}$, где c фиксировано, то на горизонте управления N невозможно надежно (т.е. с вероятностью сколь угодно близкой к 1) различить большее и меньшее математические ожидания. Поэтому одношаговые потери всегда будут не меньше, чем $cN^{-1/2}p_e$, где p_e – вероятность ошибки в определении большего из m_1, m_2, а потери на всем горизонте управления N будут не меньше, чем $cp_eN^{1/2}$, т.е. имеют порядок $N^{1/2}$. Также интуитивно ясно, что малые изменения дисперсий, например, на несколько процентов, скорее всего не приведут к значительному изменению полных потерь. Наконец, отметим, что для «далеких» распределений потери будут иметь меньшие значения. В частности, в [21, 22] установлено, что они асимптотически имеют порядок $\ln(N)$, если $|m_1 - m_2| \geq \delta > 0$.

Поскольку стратегия (1.2) была предложена для бернуллиевского двурукого бандита, для которого математические ожидания и дисперсии доходов равны p_ℓ и $p_\ell q_\ell$ соответственно ($\ell = 1, 2$), то в этом случае при больших N «близость» математических ожиданий означает и приблизительное равенство дисперсий. Так как максимальные потери достигаются при максимальных значениях дисперсий, то в этом случае в (2.1) следует положить $D_1 = D_2 = 0.25$, которые имеют место при $p_1 = p_2 = 0.5$. Тогда из (1.2) следует, что в этом случае в правиле (2.1) оптимально выбирать $a = 2/15$.

Перейдем к инвариантному описанию процесса управления. Рассмотрим следующее множество допустимых значений параметра Θ

$$m_\ell = m + d_\ell (D/N)^{1/2};\ m \in (-\infty, +\infty),\ |d_\ell| \leq c,\ \ell = 1, \ldots, J, \quad (2.2)$$

где $c > 0$ достаточно большое фиксированное число. Данное множество описывает «близкие» распределения, именно на нем потери достигают максимальной величины. Далее мы рассматриваем стратегии, которые могут менять выбор действия только после того, как оно было применено M раз подряд. Данные стратегии позволяют выполнять пакетную, а также параллельную обработку. Предположим, что $N = MK$, где K – количество пакетов. Для пакетной версии стратегии верхние границы доверительных интервалов имеют вид

$$\hat{Q}_\ell(k) = \frac{\hat{X}_\ell(k)}{k_\ell} + \frac{a_\ell(\gamma_1, \ldots, \gamma_J)(MD_\ell)^{1/2}}{k_\ell^{1/2}}\ (2 + \zeta_\ell(k)), \qquad (2.3)$$

где k – количество обработанных пакетов, k_ℓ – количество пакетов, для которых было применено ℓ-е действие, $\hat{X}_\ell(k)$ – соответствующий ℓ-му действию доход после обработки k пакетов ($k = 2, \ldots, K-1$). Обозначим через

$$\mathbf{I}_\ell(k) = \begin{cases} 1, & \text{если } \hat{Q}_\ell(k-1) = \max(\hat{Q}_1(k-1), \ldots, \hat{Q}_J(k-1)), \\ 0, & \text{в противном случае} \end{cases}$$

индикаторную функцию выбранного действия при обработке k-го пакета в соответствии с рассматриваемым правилом при $k \geq J+1$. При $k \leq J$ каждое из действий будет по очереди применено ровно к одному пакету и можно, например, положить $\mathbf{I}_\ell(k) = \delta_{\ell,k}$, $k = 1, \ldots, J$, где $\delta_{\ell k}$ –символ Кронекера. Обратим внимание, что при каждом k с вероятностью 1 ровно одно из значений $\{\mathbf{I}_l(k)\}$ будет равно 1. Тогда общий доход, полученный от использования каждого из действий, равен

$$\hat{X}_\ell(k) = k_\ell M(m + d_\ell(D/N)^{1/2}) + \sum_{i=1}^{k} \mathbf{I}_\ell(i)\eta_\ell(MD_\ell; i), \qquad (2.4)$$

где $\eta_\ell(MD_\ell; i) \sim \mathcal{N}(0, \sqrt{MD_\ell})$ – независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией MD_ℓ. Введем обозначения $t = kK^{-1}$, $t_\ell = k_\ell K^{-1}$, $\varepsilon = K^{-1}$ и с учетом (2.4) запишем верхние границы (2.3) как

$$\hat{Q}_\ell(k) = Mm + \left(\frac{MD}{K}\right)^{1/2} \times \qquad (2.5)$$

$$\times \left(d_\ell + \frac{\sum_{i=1}^{k} \mathbf{I}_\ell(i)\eta_\ell(\gamma_\ell\varepsilon; i)}{t_\ell} + \frac{a_\ell(\gamma_1, \ldots, \gamma_J)\gamma_\ell^{1/2}}{t_\ell^{1/2}}\ (2 + \zeta_\ell(k)) \right),$$

$\ell = 1, 2, \ldots, J$. Далее приме́ним к границам (2.5) следующее линейное преобразование, которое не влияет на их порядок расстановки:

$$\hat{q}_\ell(t) = (\hat{Q}_l(k) - Mm)\left(\frac{K}{MD}\right)^{1/2}.$$

В результате получаем выражение для верхних границы доверительных интервалов для стратегии UCB с горизонтом управления, равным 1, т.е. в инвариантной форме, зависящей не от размера горизонта N, а от относительного размера пакета ε:

$$\hat{q}_\ell(t) = d_\ell + \frac{\sum_{i=1}^{k} \mathbf{I}_\ell(i)\eta_\ell(\gamma_\ell\varepsilon; i)}{t_\ell} + \frac{a_\ell(\gamma_1, \ldots, \gamma_J)\gamma_\ell^{1/2}}{t_\ell^{1/2}}(2 + \zeta_\ell(t)), \quad (2.6)$$

$\ell = 1, 2, \ldots, J$. Следующей задачей является нахождение в инвариантной форме выражения для функции потерь (1.3). Положим $d_{l_0} = \max(d_1, \ldots, d_J)$. Тогда с учетом (2.2)

$$L_N(\sigma, \theta) = \sum_{\ell=1}^{J}(m_{l_0} - m_\ell)\mathbf{E}_{\sigma,\theta}\left(\sum_{i=1}^{K} M\mathbf{I}_\ell(i)\right) =$$

$$= (D/N)^{1/2}\sum_{\ell=1}^{J} M(d_{l_0} - d_\ell)\mathbf{E}_{\sigma,\theta}\,(k_\ell) = (DN)^{1/2}\sum_{\ell=1}^{J}(d_{l_0} - d_\ell)\mathbf{E}_{\sigma,\theta}(t_\ell).$$

В нормализованном виде (к величине $(DN)^{1/2}$) получаем следующее выражение для функции потерь:

$$(DN)^{-1/2}L_N(\sigma, \theta) = \sum_{\ell=1}^{J}(d_{l_0} - d_\ell)\mathbf{E}_{\sigma,\theta}(t_\ell). \quad (2.7)$$

Полученные результаты можно сформулировать в виде теоремы.

Теорема 2.1. *Для гауссовского многорукого бандита с J действиями, фиксированными известными дисперсиями $D = D_1 \geq \ldots \geq D_J$ и неизвестными математическими ожиданиями $m_1, \ldots, m_J$, принадлежащими множеству* (2.2)*, пакетная версия стратегии UCB с границами, задаваемыми соотношениями* (2.3)*, имеет инвариантное описание с единичным горизонтом управления и границами* (2.6)*. Нормализованная (к величине $(DN)^{1/2}$) функция потерь* (1.3) *описывается выражением* (2.7)*.*

3. Другой подход к инвариантному описанию

В этом параграфе мы рассматриваем другой подход к инвариантному описанию стратегии и функции потерь. Этот подход использует

результаты [3, 5]. Ограничимся случаем $J = 2$, причем считаем, что $D_1 \geq D_2$. Отметим, что перенос результатов, представленных ниже, на случай $J > 2$ не может быть выполнен автоматически.

При $J = 2$ параметр имеет вид $\theta = (m_1, m_2)$, а множество параметров сначала выберем в виде $\Theta = \{|m_1 - m_2| \leq 2C\}$. Зададим априорную плотность распределения $\lambda(\theta)$ и будем рассматривать усредненную функцию потерь

$$L_N(\sigma, \lambda) = \int_\Theta L_N(\sigma, \theta)\lambda(\theta)d\theta, \tag{3.1}$$

где $L_N(\sigma, \theta)$ определяется аналогично (1.3), но при $J = 2$. Верхние границы доверительных интервалов ниже удобно обозначить так:

$$Q_\ell(X_\ell, n_\ell, \zeta_\ell) = \frac{X_\ell}{n_\ell} + \frac{a_\ell(\gamma_1, \gamma_2)D_\ell^{1/2}}{n_\ell^{1/2}}\ (2 + \zeta_\ell),$$

где n_ℓ, X_ℓ – текущие значения полного числа выборов ℓ-го действия и соответствующего полного дохода, ζ_ℓ – случайные величины с плотностью стандартного экспоненциального распределения равной e^{-x} при $x \geq 0$ и 0 при $x < 0$, независимые между собой, $\gamma_\ell = D_\ell/D_1$, $\ell = 1, 2$. Обозначим $n_\ell^* = n_\ell D_\ell$. Для текущей предыстории процесса X_1, n_1, X_2, n_2 апостериорная плотность распределения имеет вид

$$\lambda(m_1, m_2|X_1, n_1, X_2, n_2) = \tag{3.2}$$
$$= \frac{f_{n_1^*}(X_1|n_1 m_1) f_{n_2^*}(X_2|n_2 m_2)\lambda(m_1, m_2)}{p(\lambda; X_1, n_1, X_2, n_2)},$$

где

$$p(\lambda; X_1, n_1, X_2, n_2) = \tag{3.3}$$
$$= \iint_\Theta f_{n_1^*}(X_1|n_1 m_1) f_{n_2^*}(X_2|n_2 m_2)\lambda(m_1, m_2)dm_1 dm_2.$$

Снова предполагаем, что обработка ведется пакетами по M данных, причем первые два пакета имеют размер M_0 и обрабатываются обоими действиями по очереди (считаем, что $N - 2M_0$ кратно M). Стратегия управления $\sigma_\ell(X_1, n_1, X_2, n_2)$ характеризует вероятности выбора действий для обработки следующего пакета при текущей предыстории X_1, n_1, X_2, n_2. Обозначим через $L_{N-n}(\lambda; X_1, n_1, X_2, n_2)$

функцию потерь на оставшемся горизонте управления длины $N-n$, где $n = n_1 + n_2$, вычисленную относительно апостериорной плотности распределения (3.2). Тогда для нахождения функции потерь (3.1) нужно решать следующее стандартное рекуррентное уравнение динамического программирования

$$L_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \sigma_1(X_1, n_1, X_2, n_2) L^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) + \\ + \sigma_2(X_1, n_1, X_2, n_2) L^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2), \tag{3.4}$$

где $L^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = L^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = 0$ при $n = N$ и далее

$$L^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint_{\Theta} \lambda(m_1, m_2 | X_1, n_1, X_2, n_2) \times \\ \times \left(M(m_2 - m_1)^+ + \mathbf{E}^{(1)}_x L_{N-n-M}(\lambda; X_1 + x, n_1 + M, X_2, n_2) \right) dm_1 dm_2, \\ L^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint_{\Theta} \lambda(m_1, m_2 | X_1, n_1, X_2, n_2) \times \tag{3.5} \\ \times \left(M(m_1 - m_2)^+ + \mathbf{E}^{(2)}_x L_{N-n-M}(\lambda; X_1, n_1, X_2 + x, n_2 + M) \right) dm_1 dm_2$$

при $2M_0 \le n < N$. Здесь $x^+ = \max(x, 0)$, а

$$\mathbf{E}^{(\ell)}_x L(x) = \int_{-\infty}^{\infty} L(x) f_{MD_\ell}(x | M m_\ell) dx$$

обозначает математическое ожидание относительно плотности распределения, соответствующей применению ℓ-го действия к одному пакету данных ($\ell = 1, 2$). Легко видеть, что $L^{(\ell)}_{N-n}(\lambda; X_1, n_1, X_2, n_2)$ описывает математическое ожидание потерь на оставшемся горизонте $N - n$, если к первому пакету применено ℓ-е действие, а затем управление осуществляется в соответствии со стратегией σ. Стратегия UCB Дж. Басера определяется следующим образом:

$$\sigma_\ell(X_1, n_1, X_2, n_2) = \iint_{\mathcal{A}_\ell} e^{-(s_1+s_2)} ds_1 ds_2, \quad \ell = 1, 2, \tag{3.6}$$

где $\mathcal{A}_1 = \{s_1 \ge 0,\ s_2 \ge 0,\ Q_1(X_1, n_1, s_1) \ge Q_2(X_2, n_2, s_2)\}$, $\mathcal{A}_2 = \{s_1 \ge 0,\ s_2 \ge 0,\ Q_2(X_1, n_1, s_1) \ge Q_1(X_2, n_2, s_2)\}$. Таким образом, это рандомизированная стратегия, т.е. при каждой предыстории X_1, n_1, X_2, n_2 оба действия выбираются с некоторыми вероятностями отличными

от 0 и 1; при этом вероятность события $Q_1(X_1, n_1, s_1) = Q_2(X_2, n_2, s_2)$ равна нулю. Функция потерь (3.1) находится по формуле

$$L_N(\sigma, \lambda) = \iint_{\Theta} M_0|m_1 - m_2|\lambda(m_1, m_2)dm_1dm_2 + \tag{3.7}$$
$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L_{N-2M_0}(\lambda; X_1, M_0, X_2, M_0)p(\lambda; X_1, M_0, X_2, M_0)dX_1dX_2,$$

где $p(\lambda; X_1, M_0, X_2, M_0)$ определена в (3.3). Первое слагаемое в (3.7) соответствует начальному этапу, когда к первым двум пакетам действия применяются по очереди. Второе слагаемое относится к управлению на оставшемся горизонте. Отметим, что если требуется найти функцию потерь (1.3), то нужно выбрать вырожденную априорную плотность распределения, сосредоточенную на единственном параметре θ; в этом случае все апостериорные плотности также останутся вырожденными.

Формулы (3.4)–(3.7) в принципе позволяют находить функцию потерь, однако при выполнении очень большого объема вычислений. Кардинально уменьшить объем вычислений можно, если установить ряд свойств стратегии Дж. Басера и функции потерь.

Лемма 3.1. *Пусть стратегия σ такова, что*

$$\sigma_\ell(X_1, n_1, X_2, n_2) = \sigma_\ell(X_1 + n_1z, n_1, X_2 + n_2z, n_2) \tag{3.8}$$

при всех предысториях X_1, n_1, X_2, n_2, $\ell = 1, 2$ и для некоторого фиксированного z. Тогда справедливо равенство

$$L_N(\sigma, \lambda(m_1, m_2)) = L_N(\sigma, \lambda(m_1 + z, m_2 + z)). \tag{3.9}$$

Доказательство. Свойство устанавливается методом математической индукции. Действительно, выполняя замены $m_\ell \leftarrow m_\ell + z$, $X_\ell \leftarrow X_\ell + n_\ell z$ для обоих m_1, m_2 и всех возможных X_ℓ, n_ℓ, видим, что для плотности (3.2) справедливы равенства

$$\lambda(m_1, m_2|X_1, n_1, X_2, n_2) =$$
$$= \lambda(m_1 + z, m_2 + z|X_1 + n_1z, n_1, X_2 + n_2z, n_2).$$

При этом не меняются выражения вида $(m_1 - m_2)^+$, $(m_2 - m_1)^+$, $|m_1 - m_2|$. Поэтому из уравнений (3.4)–(3.5) и свойства (3.8) после-

довательно вытекают равенства

$$L_N^{(\ell)}(\lambda(m_1,m_2),X_1,n_1,X_2,n_2)=$$
$$=L_N^{(\ell)}(\lambda(m_1+z,m_2+z),X_1+n_1z,n_1,X_2+n_2z,n_2),\ \ell=1,2,$$
$$L_N(\lambda(m_1,m_2),X_1,n_1,X_2,n_2)=$$
$$=L_N(\lambda(m_1+z,m_2+z),X_1+n_1z,n_1,X_2+n_2z,n_2).$$

Поскольку в (3.7) $p(\lambda(m_1,m_2);X_1,n_1,X_2,n_2)=p(\lambda(m_1+z,m_2+z);X_1+n_1z,n_1,X_2+n_2z,n_2)$, а пределы интегрирования по X_1,X_2 бесконечны, то отсюда следует (3.9). □

Лемма 3.2. *Положим $\overline{\ell}=3-\ell$. Пусть стратегия σ такова, что*

$$\sigma_\ell(X_1,n_1,X_2,n_2)=\sigma_{\overline{\ell}}(X_2,n_2,X_1,n_1), \tag{3.10}$$

при всех предысториях X_1,n_1,X_2,n_2 и $\ell=1,2$. Если $D_1=D_2$, то справедливо равенство

$$L_N(\sigma,\lambda(m_1,m_2))=L_N(\sigma,\lambda(m_2,m_1)). \tag{3.11}$$

Доказательство. Свойство устанавливается методом математической индукции. Действительно, выполняя замены $m_\ell\leftarrow m_{\overline{\ell}}$, $X_\ell\leftarrow X_{\overline{\ell}}$, $n_\ell\leftarrow n_{\overline{\ell}}$ для обоих m_1,m_2 и всех возможных X_ℓ,n_ℓ, видим, что для плотности (3.2) справедливы равенства

$$\lambda(m_1,m_2|X_1,n_1,X_2,n_2)=\lambda(m_2,m_1|X_2,n_2,X_1,n_1).$$

Поэтому из уравнений (3.4)–(3.5) и свойства (3.10) последовательно вытекают равенства

$$L_N^{(\ell)}(\lambda(m_1,m_2),X_1,n_1,X_2,n_2)=L_N^{(\overline{\ell})}(\lambda(m_2,m_1),X_2,n_2,X_1,n_1),\ \ell=1,2,$$
$$L_N(\lambda(m_1,m_2),X_1,n_1,X_2,n_2)=L_N(\lambda(m_2,m_1),X_2,n_2,X_1,n_1).$$

Так как $p(\lambda(m_1,m_2);X_1,n_1,X_2,n_2)=p(\lambda(m_2,m_1);X_2,n_2,X_1,n_1)$, то отсюда следует (3.11). □

Для дальнейшего удобно изменить параметризацию следующим образом: $m_1=m+v,m_2=m-v$. В новых переменных априорную плотность распределения обозначим $\mu(m,v)$. Тогда для стратегий, удовлетворяющих условию леммы 3.1, выполнено свойство

$L_N(\sigma, \mu(m, v)) = L_N(\sigma, \mu(m + z, v))$ для любого фиксированного z, а для стратегий, удовлетворяющих условию леммы 3.2, при $D_1 = D_2$ выполнено свойство $L_N(\sigma, \mu(m, v)) = L_N(\sigma, \mu(m, -v))$.

Рассмотрим плотность распределения $\mu(m, v) = \rho(v)\beta(m)$, где $\beta(m)$ – произвольная плотность. Нетрудно видеть, что справедливо равенство

$$L_N(\sigma, \rho(v)\beta(m)) = \int_{-\infty}^{\infty} L_N(\sigma, \rho(v)\delta(m - z))\beta(z)dz \tag{3.12}$$

где $\delta(m - z)$ – дельта-функция Дирака. Поскольку для стратегий, удовлетворяющих условию леммы 3.1, все $L_N(\sigma, \rho(v)\delta(m - z))$ равны между собой, то из (3.12) следует, что в этом случае справедливо равенство $L_N(\sigma, \rho(v)\beta_1(m)) = L_N(\sigma, \rho(v)\beta_2(m)) = L_N(\sigma, \rho(v))$ для любых плотностей $\beta_1(m), \beta_2(m)$, а обозначение $L_N(\sigma, \rho(v))$ подчеркивает, что функция потерь определяется только плотностью $\rho(v)$.

Пусть теперь $\rho_1(v)$, $\rho_2(v)$ таковы, что $\rho_1(v) = 0$ при $v \leq 0$, а $\rho_2(v) = 0$ при $v \geq 0$ и $L_N(\sigma, \rho_1(v)) = L_N(\sigma, \rho_2(v))$. Тогда

$$L_N(\sigma, \alpha_1\rho_1(v)\beta_1(m) + \alpha_2\rho_2(v)\beta_2(m)) = L_N(\sigma, \rho_1(v)) = L_N(\sigma, \rho_2(v))$$

для любых плотностей $\beta_1(m)$, $\beta_1(m)$ и неотрицательных чисел α_1, α_2 таких, что $\alpha_1 + \alpha_2 = 1$.

Если же выполнено условие леммы 3.2 и $D_1 = D_2$, то выбирая в качестве плотности распределения $\tilde{\rho}(v) = 0.5(\rho(v) + \rho(-v))\beta(m)$, получаем, что плотность $\tilde{\rho}(v)$ можно выбрать симметричной, т.е. $\tilde{\rho}(v) = \tilde{\rho}(-v)$. Выбирая $\rho_1(v)$, $\rho_2(v)$ из условий $\rho_1(v) = 2\tilde{\rho}(v)$ при $v \geq 0$, $\rho_1(v) = 0$ при $v < 0$ и $\rho_2(v) = 2\tilde{\rho}(v)$ при $v \leq 0$, $\rho_2(v) = 0$ при $v > 0$, получаем, что в этом случае

$$L_N(\sigma, \alpha_1\rho_1(v)\beta_1(m) + \alpha_2\rho_2(v)\beta_2(m)) = L_N(\sigma, \tilde{\rho}(v)) = L_N(\sigma, \rho(v)).$$

для любых плотностей $\beta_1(m)$, $\beta_1(m)$ и неотрицательных чисел α_1, α_2 таких, что $\alpha_1 + \alpha_2 = 1$.

Для дальнейшего выберем априорную плотность в виде

$$\mu(m, v) = \rho(v)\kappa_a(m), \tag{3.13}$$

где $\kappa_a(m)$ – плотность равномерного распределения на отрезке $m \in [-a, a]$ и a достаточно велико.

Лемма 3.3. *Стратегия σ, определяемая правилом UCB Дж. Басера, удовлетворяет условиям лемм* 3.1, 3.2 *и при всех n_1, n_2 зависит только от текущей статистики (U, n_1, n_2), где $U = (X_1 n_2 - X_2 n_1)/n'$, $n' = n'_1 + n'_2$, $n'_1 = n_1/D_1$, $n'_2 = n_2/D_2$. При фиксированных n_1, n_2 функция $\sigma_1(U, n_1, n_2)$ монотонно возрастает по U.*

Доказательство. Выполнение условий леммы 3.2 следует из формулы (3.6) и определения областей $\mathcal{A}_1$, $\mathcal{A}_2$. Для проверки выполнения условий леммы 3.1 найдем явное выражение для $\sigma_1(X_1, n_1, X_2.n_2)$; при этом $\sigma_2(X_1, n_1, X_2.n_2) = 1 - \sigma_1(X_1, n_1, X_2.n_2)$. Обозначим $\alpha_\ell = a_\ell(\gamma_1, \gamma_2)(D_\ell/n_\ell)^{1/2}$, $\ell = 1, 2$, $Y = X_1/n_1 - X_2/n_2$. Тогда область $\mathcal{A}_1$ определяется условиями $s_1 \geq 0$, $s_2 \geq 0$, $s_1 \geq k s_2 + B$, где $k = \alpha_1^{-1}\alpha_2 > 0$, $B = \alpha_1^{-1}(2\alpha_2 - 2\alpha_1 - Y)$. Так как при возрастании Y область $\mathcal{A}_1$ увеличивается, то функция $\sigma_1(X_1, n_1, X_2.n_2)$ является возрастающей функцией Y. Значение интеграла в (3.6) по области $\mathcal{A}_1$ получим при $B \geq 0$:

$$\int_0^\infty e^{-s_2} ds_2 \int_{ks_2+B}^\infty e^{-s_1} ds_1 = \int_0^\infty e^{-(1+k)s_2 - B} ds_2 = e^{-B}(1+k)^{-1}.$$

Обозначим $C = -B/k$. При $B < 0$ рассмотрим область $\mathcal{A}_2$, которая определяется условиями $s_1 \geq 0,\ s_2 \geq 0,\ s_2 \geq k^{-1}s_1 + C$. Ясно, что соответствующее значение интеграла в (3.6) по области $\mathcal{A}_2$ равно $e^{-C}(1 + k^{-1})^{-1}$. Поэтому

$$\sigma_1(X_1, n_1, X_2, n_2) = \begin{cases} e^{-B}(1+k)^{-1} & \text{при } B \geq 0, \\ 1 - e^{-C}(1+k^{-1})^{-1} & \text{при } B < 0, \end{cases} \tag{3.14}$$

где k, B, C, α_1, α_2 определены выше. Поэтому $\sigma_1(X_1, n_1, X_2, n_2)$ зависит только от текущей статистики (Y, n_1, n_2), причем возрастает по Y. Следовательно, условия леммы 3.1 выполнены для статистики (Y, n_1, n_2). Так как $U = Y n_1 n_2 (n')^{-1}$, то условия леммы выполнены и для статистики (U, n_1, n_2). □

Для стратегий, которые выбирают действия в зависимости от текущей статистики (U, n_1, n_2), и для априорной плотности распределения (3.13) при условии, что $a \to +\infty$, в [3], Теорема 3, показано, что из уравнения (3.4)–(3.5) можно получить уравнение динамического программирования, которое последовательно, «с конца» вычисляет

функцию потерь, зависящую от текущей статистики (U, n_1, n_2). В [5], Теорема 5, это уравнение представлено в инвариантной форме с горизонтом управления равным единице в области «близких» распределений, которые характеризуются тем, что величина $|m_1 - m_2|$ имеет порядок $N^{-1/2}$. Как отмечено в параграфе 2, наибольшие потери достигаются именно в этой области. Более детальный анализ «близких» распределений приведен в [5].

Приведем рекуррентное уравнение динамического программирования в инвариантной форме с горизонтом управления равным единице, справедливое в области «близких» распределений. Сделаем следующую замену переменных:

$$C = cN^{-1/2}, w = N^{1/2}v, \varrho(w) = N^{-1/2}\rho(v),$$
$$x_\ell = X_\ell N^{-1/2}, u = UN^{-1/2},$$
$$t_\ell = n_\ell N^{-1}, t_\ell^* = n_\ell^* N^{-1}, t'_\ell = n'_\ell N^{-1}, t = t_1 + t_2, t' = t'_1 + t'_2,$$
$$\varepsilon_0 = M_0 N^{-1}, \varepsilon = MN^{-1}, \varepsilon_\ell^* = \varepsilon D_\ell,\ \varepsilon'_\ell = \varepsilon / D_\ell,\ \ell = 1, 2.$$

Из (3.6) следует, что $\sigma_\ell(X_1, n_1, X_2, n_2) = \sigma_\ell(x_1, t_1, x_2, t_2)$ для всех предысторий (X_1, n_1, X_2, n_2), поэтому $\sigma_\ell(U, n_1, n_2) = \sigma_\ell(u, t_1, t_2)$ для всех предысторий (U, n_1, n_2) $(\ell = 1, 2)$. Положим $D_g^2 = D_1 D_2$, $D_h^{-1} = 0.5(D_1^{-1} + D_2^{-1})$, $f_D(x) = f_D(x|0)$. Используем стандартное обозначение для свертки функций $F(x) * G(x) = \int_{-\infty}^{\infty} F(x-y)G(y)dy$. Требуемое рекуррентное уравнение имеет вид

$$l_\varepsilon(u, t_1, t_2) = \sigma_1(u, t_1, t_2) l_\varepsilon^{(1)}(u, t_1, t_2) + \sigma_2(u, t_1, t_2) l_\varepsilon^{(2)}(u, t_1, t_2), \quad (3.15)$$

где $l_\varepsilon^{(1)}(u, t_1, t_2) = l_\varepsilon^{(2)}(u, t_1, t_2) = 0$ при $t_1 + t_2 = 1$ и далее

$$\begin{aligned} &l_\varepsilon^{(1)}(u, t_1, t_2) = \varepsilon g^{(1)}(u, t_1, t_2) + \\ &+ l_\varepsilon(u, t_1 + \varepsilon, t_2) * f_{\varepsilon_1^* t_2^2 (t')^{-1} (t' + \varepsilon'_1)^{-1}}(u), \\ &l_\varepsilon^{(2)}(u, t_1, t_2) = \varepsilon g^{(2)}(u, t_1, t_2) + \\ &+ l_\varepsilon(u, t_1, t_2 + \varepsilon) * f_{\varepsilon_2^* t_1^2 (t')^{-1} (t' + \varepsilon'_2)^{-1}}(u), \end{aligned} \quad (3.16)$$

если $t_1 + t_2 < 1$, $t_1 \geq \varepsilon_0$ и $t_2 \geq \varepsilon_0$. Функция потерь (3.1) вычисляется в соответствии с формулой

$$L_N(\sigma; \rho(v)) = N^{1/2} l_\varepsilon(\sigma; \varrho(w)), \quad (3.17)$$
$$l_\varepsilon(\sigma; \varrho(w)) = \varepsilon_0 \int_{-c}^{c} 2|w| \varrho(w) dw + \int_{-\infty}^{\infty} f_{0.5\varepsilon_0 D_g^2 D_h}(u) l_\varepsilon(u, \varepsilon_0, \varepsilon_0) du.$$

Здесь

$$g^{(1)}(u,t_1,t_2)=\int_{-c}^{0}2|w|g(w;u,t_1,t_2)\varrho(w)dw,$$
$$g^{(2)}(u,t_1,t_2)=\int_{0}^{c}2wg(w;u,t_1,t_2)\varrho(w)dw,$$
$$g(w;u,t_1,t_2)=\exp\left(2D_g^{-2}(uw-w^2t_1t_2(t')^{-1})\right).$$

Отметим, что здесь ε_0 и ε характеризуют относительные размеры начальных и последующих пакетов.

Лемма 3.4. *Пусть при некотором $D>0$ выполнены следующие преобразования: $\hat{D}_1=D^{-1}D_1$, $\hat{D}_2=D^{-1}D_2$, $\hat{w}=D^{-1/2}w$, $\hat{c}=D^{-1/2}c$, $\hat{\varrho}(\hat{w})=D^{1/2}\varrho(w)$, $\hat{u}=D^{-3/2}u$, $\hat{\varepsilon}'_\ell=D\varepsilon'_\ell$, $\hat{\varepsilon}^*_\ell=D^{-1}\varepsilon^*_\ell$, $\hat{t}'_\ell=Dt'_\ell$, $\hat{t}^*_\ell=D^{-1}t^*_\ell$ (при этом t_1, t_2 не меняются), $\hat{\sigma}_\ell(\hat{u},t_1,t_2)=\sigma_\ell(u,t_1,t_2)$, $\ell=1,2$. Тогда соответствующие потери удовлетворяют равенству*

$$l_\varepsilon(\sigma,\varrho(w))=D^{1/2}\hat{l}_\varepsilon(\hat{\sigma},\hat{\varrho}(\hat{w})). \tag{3.18}$$

Доказательство. Формула (3.18) устанавливается по индукции путем выполнения указанных замен в (3.15)–(3.17). □

Пусть $D=D_1\geq D_2$. В этом случае лемма 3.4 позволяет рассматривать только пары дисперсий вида (γ_1,γ_2), где $1=\gamma_1\geq\gamma_2>0$. При этом формулы (3.17) и (3.18) в совокупности являются аналогом формулы (2.7).

Предположим теперь, что при всех u,t_1,t_2, при которых определено решение уравнения (3.15)–(3.16), существует предел $l(u,t_1,t_2)=\lim\limits_{\varepsilon\to+0}l_\varepsilon(u,t_1,t_2)$, имеющий производные требуемых порядков по u,t_1,t_2, и получим для него дифференциальное уравнение в частных производных второго порядка. Используя разложение в ряд Тэйлора $l_\varepsilon(u-x,t_1+\varepsilon,t_2)=l_\varepsilon(u,t_1+\varepsilon,t_2)-x(l_\varepsilon)'_u(u,t_1+\varepsilon,t_2)+0.5x^2(l_\varepsilon)''_{uu}(u,t_1+\varepsilon,t_2)+o(\varepsilon)$ и с учетом равенств

$$\int_{-\infty}^{\infty}f_\varepsilon(x)dx=1,\ \int_{-\infty}^{\infty}xf_\varepsilon(x)dx=0,\ \int_{-\infty}^{\infty}x^2f_\varepsilon(x)dx=\varepsilon,$$

получаем, что уравнения (3.16) принимают вид

$$\begin{gathered}
l_\varepsilon^{(1)}(u,t_1,t_2) = \varepsilon g^{(1)}(u,t_1,t_2) + l_\varepsilon(u,t_1+\varepsilon,t_2)+ \\
+0.5\varepsilon_1^* t_2^2 (t')^{-1}(t'+\varepsilon_1')^{-1}(l_\varepsilon)''_{uu}(u,t_1+\varepsilon,t_2) + o(\varepsilon), \\
l_\varepsilon^{(2)}(u,t_1,t_2) = \varepsilon g^{(2)}(u,t_1,t_2) + l_\varepsilon(u,t_1,t_2+\varepsilon)+ \\
+0.5\varepsilon_2^* t_1^2 (t')^{-1}(t'+\varepsilon_2')^{-1}(l_\varepsilon)''_{uu}(u,t_1,t_2+\varepsilon) + o(\varepsilon).
\end{gathered}$$

Дополняя их уравнением (3.15), которое сейчас запишем в виде

$$\sum_{\ell=1}^{2} \sigma_\ell(u,t_1,t_2)(l_\varepsilon^{(\ell)}(u,t_1,t_2) - l_\varepsilon(u,t_1,t_2)) = 0,$$

получим в пределе при $\varepsilon \to +0$ уравнение для $l = l(u,t_1,t_2)$:

$$\sum_{\ell=1}^{2} \sigma_\ell(u,t_1,t_2) \left(\frac{\partial l}{\partial t_\ell} + \frac{0.5 D_\ell t_\ell^2}{(t')^2} \times \frac{\partial^2 l}{\partial u^2} + g^{(\ell)}(u,t_1,t_2) \right) = 0 \qquad (3.19)$$

с начальным условием

$$l(u,t_1,t_2) = 0 \quad \text{при } t_1 + t_2 = 1. \qquad (3.20)$$

Для нахождения функции потерь (3.1) следует использовать формулу (3.17). в которой $l_\varepsilon(u,t_1,t_2)$ следует заменить на $l(u,t_1,t_2)$. Для численного решения уравнения (3.19) следует решать разностное уравнение

$$l(u,t_1,t_2) = \sigma_1(u,t_1,t_2) l^{(1)}(u,t_1,t_2) + \sigma_2(u,t_1,t_2) l^{(2)}(u,t_1,t_2), \qquad (3.21)$$

где

$$\begin{gathered}
l^{(1)}(u,t_1,t_2) = l(u,t_1+\Delta t,t_2)+ \\
+\Delta t \left(g^{(1)}(u,t_1,t_2) + 0.5 D_1 t_2^2 (t')^{-2} D^2 l(u,t_1+\Delta t,t_2) \right), \\
l^{(2)}(u,t_1,t_2) = l(u,t_1,t_2+\Delta t)+ \\
+\Delta t \left(g^{(2)}(u,t_1,t_2) + 0.5 D_2 t_1^2 (t')^{-2} D^2 l(u,t_1,t_2+\Delta t) \right)
\end{gathered} \qquad (3.22)$$

и

$$D^2 l(u,\cdot,\cdot) = \frac{l(u+\Delta u,\cdot,\cdot) - 2l(u,\cdot,\cdot) + l(u-\Delta u,\cdot,\cdot)}{\Delta u^2}$$

с начальным условием (3.20).

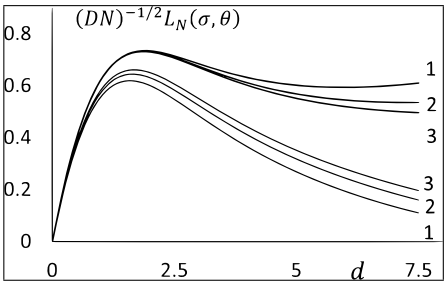

Рисунок 1. Нормализованные потери при $N = 30$, 40, 50, $a = 1/3$.

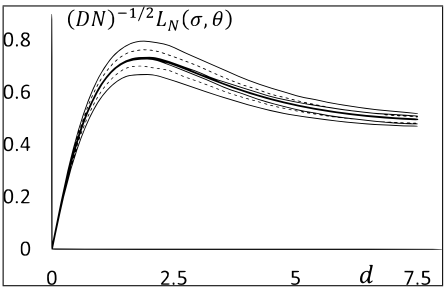

Рисунок 2. Нормализованные потери при $N = 50$ и неточно указанных дисперсиях.

4. Результаты моделирования

На рис. 1 линиями 1, 2, 3 даны результаты вычислений нормализованных потерь при горизонтах управления $N = 30$, 40, 50 и параметре стратегии $a = 1/3$. Верхние линии характеризуют полные потери, а нижние – потери без начального этапа, когда действия применяются поровну. Параметр управляемого процесса выбирался равным $\theta = (d(D/N)^{1/2}, -d(D/N)^{1/2})$, а общая дисперсия равной $D = 1$. Везде при вычислениях d менялось с шагом 0.3 от 0 до 7.5. Потери вычислялись двумя способами: методом Монте-Карло и аналитически по формулам (3.14)–(3.17) в предположении, что плотность $\varrho(w)$ вырождена и сосредоточена в точке d. Отметим, что при $d < 0$ в силу равенства дисперсий нормализованные потери такие же, как представленные на графике при $|d|$. Методом Монте Карло выполнялось 400000 усреднений, а при аналитическом вычислении полагались $\varepsilon_0 = \varepsilon = N^{-1}$, шаг численного интегрирования $\Delta u = 0.005$, область интегрирования $|u| \leq 2.5$. В результате графики потерь, вычисленных обоими способами, оказались визуально идентичными и на рис. 1 отображаются совпадающими линиями. Однако с точки зрения времени вычисления метод Монте-Карло оказался намного быстрее. При $N = 30$ он был быстрее примерно в 11 раз, а при $N = 50$ – примерно в 17 раз (время работы метода Монте-Карло приблизительно пропорционально N, а аналитического метода – приблизительно пропорционально N^2). Поэтому в дальнейшем вычисления нормали-

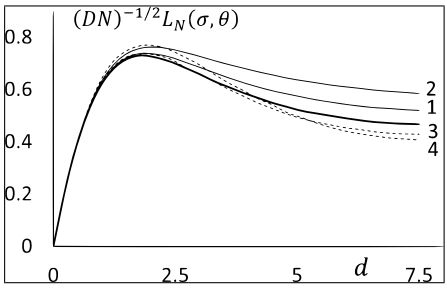

Рисунок 3. Определение оптимального a при $N = 50$.

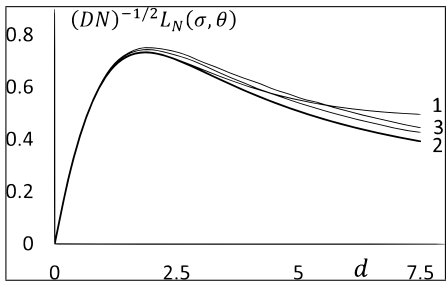

Рисунок 4. Предельное поведение, $a = 1/3$.

зованных потерь, как правило, выполнялись методом Монте-Карло с 400000 усреднениями.

На рис. 2 даны графики нормализованных потерь, использующих правило UCB с параметром $a = 1/3$ на горизонте $N = 50$, в случае неточного указания дисперсий. Фактические значения параметра управляемого процесса и дисперсий здесь такие же, как и выше, однако в правиле UCB дисперсии указывались с отклонениями от действительных значений парой определенных ранее чисел γ_1, γ_2. Жирная центральная линия соответствует точному указанию дисперсий, т.е. $\gamma_1 = \gamma_2 = 1$. Верхняя и нижняя тонкие сплошные линии соответствуют случаям $\gamma_1 = 0.9$, $\gamma_2 = 1.1$ и $\gamma_1 = 1.1$, $\gamma_2 = 0.9$, верхняя и нижняя тонкие пунктирные – случаям $\gamma_1 = 0.95$, $\gamma_2 = 1.05$ и $\gamma_1 = 1.05$, $\gamma_2 = 0.95$ соответственно. Тонкие сплошные линии, соответствующие значениям $\gamma_1 = \gamma_2 = 0.9$ и $\gamma_1 = \gamma_2 = 1.1$ также отображены на рис. 1, они в области максимальных потерь практически совпадают с основной линией, соответствующей $\gamma_1 = \gamma_2 = 1$. Линии, соответствующие случаям $\gamma_1 = \gamma_2 = 0.95$ и $\gamma_1 = \gamma_2 = 1.05$ на рисунке не отображены, но они еще ближе к основной линии. Эти результаты говорят о том, что потери мало меняются при малом изменении значений дисперсий в правиле UCB, поэтому в случае априори неизвестных дисперсий можно выполнять их оценки на начальном этапе, когда действия применяются поровну, а затем использовать эти оценки для управления.

На рис. 3 при $N = 50$ показано, как можно определить опти-

мальное значение параметра a. Центральная жирная линия соответствует $a = 0.30$, для нее значение максимальных нормализованных потерь приблизительно равно 0.73 при $d \approx 1.8$. Линии 1 и 2 соответствуют значениям параметра 0.36 и 0.42 и приближенным значениям максимумов 0.74 и 0.76, линии 3 и 4 соответствуют значениям параметра 0.24 и 0.18 и приближенным значениям максимумов 0.74 и 0.77; асе максимумы достигаются при $d \approx 1.8$. Таким образом, $a = 0.30$ минимизирует максимальные потери. Так как значение максимума при $d > 0$ единственно, то с учетом результатов параграфа 3 соответствующая априорная плотность распределения имеет вид $\mu(m, v) = \alpha_1\beta_1(m)\delta(v - dN^{-1/2}) + \alpha_2\beta_2(m)\delta(v + dN^{-1/2})$, где $N = 50$, $d \approx 1.8$, $\beta_1(m)$, $\beta_2(m)$ – произвольные плотности, $\delta(\cdot)$ – дельта-функция Дирака, α_1, α_2 – неотрицательные числа, такие что $\alpha_1 + \alpha_2 = 1$. Пара $(\sigma, \mu(m, v))$ при найденных значениях a, d характеризует седловую точку функции потерь (3.1). Отметим также, что значению параметра $a = 1/3$ соответствуют максимальные потери приблизительно равные 0.73 при $d \approx 1.8$. А значению $a = 2/15$, которое в соответствии с высказанным в параграфе 2 предположением должно быть оптимальным, соответствуют потери, приблизительно равные 0.82 при $d \approx 2.1$, т.е. существенно бóльшие минимальных. Объяснение такому расхождению в оценках оптимального a может быть в том, что в [16] было рассмотрено управление бернуллиевскими процессами на небольших горизонтах N (например, $N = 50$). В этой ситуации дополнительные слагаемые порядка n_ℓ^{-1}, присутствующие в правиле (1.2), могут существенно влиять на значения $\{Q_\ell(n)\}$.

На рис. 4 анализируется предельное поведение нормализованных потерь при $a = 1/3$ и $N \to \infty$. Тонкие линии 1, 2, 3 соответствуют значениям горизонта $N = 50$, 250, 1000. Жирная линия характеризует предельные нормализованные потери, вычисленные по формуле (3.17) в соответствии с разностной схемой (3.21)–(3.22) и начальным условием (3.20). При этом $\varepsilon_0 = 0.005$, $\Delta t = 2000^{-1}$, $\Delta u = 0.025$, $|u| \leq 2.5$. Отметим, что при выполнении вычислений по разностной схеме (3.21)–(3.22) для обеспечения устойчивости требуется выполнение условия $\Delta t/\Delta u^2 < 1$ (см., например, [9]).

На рис. 5 представлены нормализованные потери при $N = 50$ и различных дисперсиях. Везде $D_1 = D = 1$, параметр θ определен как

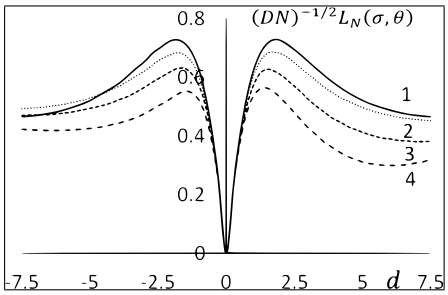

Рисунок 5. Потери при различных дисперсиях.

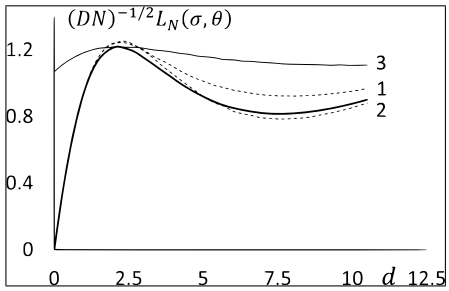

Рисунок 6. Потери для трехрукого бандита.

и выше, a_1 и a_2 характеризуют параметры правила UCB, d_1 и d_2 – положительное и отрицательное значения d, при которых достигаются максимальные нормализованные потери. Приближенные значения параметров и максимальных нормализованных потерь, соответствующих линиям 1, 2, 3 и 4, представлены в таблице 1. Отметим, что данные значения параметров a_1, a_2, и d_1, d_2 характеризуют седловые точки функции потерь.

Таблица 1.

линия	γ_2	a_1	a_2	d_1	d_2	$\max_\theta (DN)^{-1/2}L_N(\sigma,\theta)$
1	1	0.30	0.30	1.8	-1.8	0.73
2	0.75	0.34	0.36	1.8	-1.8	0.69
3	0.5	0.32	0.36	1.5	-1.5	0.63
4	0.25	0.27	0.32	1.5	-1.5	0.55

На рис. 6 при $N = 50$ представлены нормализованные потери для трехрукого бандита. Здесь $\theta = \left(d_1(D/N)^{1/2}, -d_1(D/N)^{1/2}, -d(D/N)^{1/2}\right)$ и $D = 1$, $d_1 > 0$, $d > 0$. Оказалось, что минимум максимальных нормализованных потерь, представленных на рисунке жирной линией и вычисленных при равных d_1, d, обеспечивается значением параметра $a \approx 0.21$, приблизительно равен 1.22 и достигается при $d_1 = d \approx 2.1$. Пунктирные линии 1 и 2 получены при равных d_1, d и значениях параметра $a = 0.30$ и $a = 0.15$ соответственно, т.е. подтверждают оптимальность параметра $a \approx 0.21$ при равных d_1, d. Наконец, тонкая сплошная линия 3 получена при $a = 0.21$, фиксированном $d_1 = 2.1$

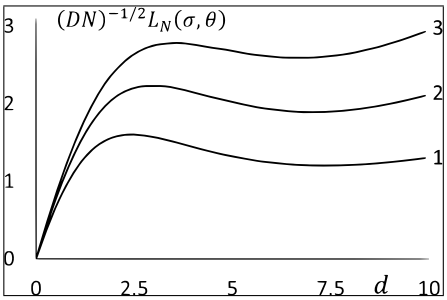

Рисунок 7. Потери для J-рукого бандита.

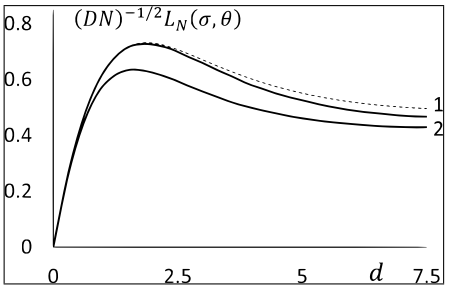

Рисунок 8. Обработка бернуллиевских доходов.

и переменном d, что подтверждает оптимальность значений $d_1 = d$ при $a = 0.21$.

Далее рассмотрим J-рукий бандит, характеризуемый параметром $\theta = (m_1, m_2, \ldots, m_J)$, где $m_1 = d(D/N)^{1/2}$, $m_k = -d(D/N)^{1/2}$, $k = 2, \ldots, J$, $D = 1$, $d > 0$, т.е. ищем максимальные потери в предположении, что есть одно максимальное математическое ожидание и остальные равные минимальные. На рис. 7 приведены результаты расчетов при $N = 50$. Линии 1, 2, 3 характеризуют нормализованные потери при $J = 4$, 6, 8, соответствующие максимумы приблизительно равны 1.61, 2.23, 2.79 и достигаются при $d \approx 2.4$, 3, 3.6. Отметим, что рост нормализованных потерь при больших d вызван равным применением действий на начальных этапах (см. также рис. 1). Чтобы его уменьшить, надо увеличить число обрабатываемых пакетов N.

Наконец, на рис. 8 представлены полученные в результате моделирования Монте-Карло нормализованные потери при пакетной обработке бернуллиевских доходов $\xi(n)$, $n = 1, 2, \ldots, N$, характеризуемых распределением

$$\Pr(\xi(n) = 1 | y_n = \ell) = p_\ell, \quad \Pr(\xi(n) = 0 | y_n = \ell) = 1 - p_\ell, \quad \ell = 1, 2.$$

Полное число данных равно $N = 5000$, размер пакета равен $M = 100$, поэтому полное число пакетов равно $K = 50$. Параметр процесса имеет вид $\theta = (p_1, p_2) = \big(p + d(D/N)^{1/2}, p - d(D/N)^{1/2}\big)$, где $D = 0.25$ – максимальная дисперсия одношагового дохода, которая достигается при $p = 0.5$. На рис. 8 жирная линия 1 соответствует $p = 0.5$, при

этом пунктирная линия дает значения потерь для гауссовских доходов. Отклонения линии 1 от пунктирной вызваны тем, что с ростом d дисперсии бернуллиевских доходов все больше отклоняются от 0.25. Жирная линия 2 характеризует одинаковые нормализованные потери при $p = 0.25$ и $p = 0.75$, которые имеют одинаковые дисперсии. При других p, отличных от 0.5 максимальные потери также не превосходят представленных на рис. 8.

5. Заключение

Мы рассмотрели адаптацию стратегии UCB, предложенной в [16], использовав ее для управления гауссовским многоруким бандитом. Получено инвариантное описание стратегии и нормализованных потерь при ее использовании. Данное описание позволяет определить оптимальный параметр стратегии и наихудшее априорное распределение с помощью моделирования Монте-Карло и методом динамического программирования. Полученное значение нормализованных потерь равное 0.73, хорошо согласуется с результатом, представленным в [16]. Подход к инвариантному описанию на основе использования рекуррентного уравнения динамического программирования позволяет понять структуру наихудшего априорного распределения.

СПИСОК ЛИТЕРАТУРЫ

1. Боровков А.А. *Математическая статистика. Дополнительные главы: Учебное пособие для вузов*. М.: Наука. Главная редакция физико-математической литературы, 1984.

2. Варшавский В.И. *Коллективное поведение автоматов*. М.: Наука, 1973.

3. Колногоров А.В. *Гауссовский двурукий бандит и оптимизация групповой обработки данных* // Пробл. передачи информ. 2018. Т. 54. № 1. С. 93–111.

4. Колногоров А.В., Гарбарь С. В. *Машинное обучение, задача о многоруком бандите и пакетное правило UCB* //XIII Всероссийское совещание по проблемам управления. ВСПУ-2019: Тру-

ды [Электронный ресурс] 17–20 июня 2019 г., Москва / Под общ. ред. Д.А. Новикова. – М.: ИПУ РАН, 2019. С. 2120–2124.

5. Колногоров А.В. *Гауссовский двурукий бандит: предельное описание* // Пробл. передачи информ. 2020. Т. 56. № 3. С. 86–111.

6. Колногоров А.В, Назин А.В., Шиян Д.Н. *Задача о двуруком бандите и пакетная версия алгоритма зеркального спуска* // МТИП. 2021. Т. 13. № 2. С. 9—39.

7. Назин А.В., Позняк А.С. *Адаптивный выбор вариантов.* М.: Наука, 1986.

8. Пресман Э.Л., Сонин И.М. *Последовательное управление по неполным данным.* М.: Наука, 1982.

9. Самарский А.А. *Теория разностных схем.* М.: Наука. 1989.

10. Саттон Р. С., Барто Э. Дж. *Обучение с подкреплением: Введение.* 2-е изд. М.: ДМК Пресс, 2020.

11. Смирнов Д.С., Громова Е.В. *Модель принятия решений при наличии экспертов как модифицированная задача о многоруком бандите* // МТИП. 2017. Т. 9. № 4. 69–87.

12. Срагович В.Г. *Адаптивное управление.* М.: Наука, 1981.

13. Цетлин М.Л. *Исследования по теории автоматов и моделированию биологических систем.* М.: Наука, 1969.

14. Auer P. *Using Confidence Bounds for Exploitation-Exploration Tradeoffs* // Journal of Machine Learning Research. 2002. V. 3. P. 397–422.

15. Auer P., Cesa-Bianchi N., Fischer P. *Finite-time analysis of the multi-armed bandit problem* // Machine learning. 2002. V. 47. No 2–3. P. 235–256.

16. Bather J.A. *The Minimax Risk for the Two-Armed Bandit Problem* // Mathematical Learning Models – Theory and Algorithms. Lecture Notes in Statistics. New York Inc.: Springer-Verlag. V. 20. P. 1–11, 1983.

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

18. Garbar S.V., Kolnogorov A.V. *Invariant description for regret of UCB strategy for Gaussian multi-arm bandit* //Proceedings of the 6th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop (SMTDA2020), 2–5 June 2020, P. 251–260.

19. Gittins J.C. *Multi-Armed Bandit Allocation Indices* // Wiley-Interscience Series in Systems and Optimization, Chichester: John Wiley & Sons, Ltd., 1989.

20. Kaufmann E. *On Bayesian Index Policies for Sequential Resource Allocation* // Annals of Statistics. 2018. V. 46, No 2. P. 842–865.

21. Lai T. L. *Adaptive Treatment Allocation and the Multi-Armed Bandit Problem* // Annals of Statist. 1987. V. 25. P. 1091–1114.

22. Lai T.L., Levin B., Robbins H., Siegmund D. *Sequential medical trials (stopping rules/asymptotic optimality)* // Proc. Nati. Acad. Sci. USA. 1980. V. 77. No 6. P. 3135–3138.

23. Lugosi G., Cesa-Bianchi N. *Prediction, Learning and Games*. Cambridge: Cambridge University Press, 2006.

24. Perchet V., Rigollet P., Chassang S. and Snowberg E. *Batched Bandit Problems.* // Annals of Statistics. 2016. V. 44, No. 2, P. 660–681.

25. Lattimore T., Szepesvari C. *Bandit Algorithms*. Cambridge: Cambridge University Press, 2020.

26. Vogel W. *An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem* // Ann. Math. Stat. 1960. V. 31. P. 444–451.

CUSTOMIZATION OF J. BATHER UCB STRATEGY FOR A GAUSSIAN MULTI-ARMED BANDIT

Sergey V. Garbar, Yaroslav-the-Wise Novgorod State University, senior lecturer (Sergey.Garbar@novsu.ru),
Alexander V. Kolnogorov, Yaroslav-the-Wise Novgorod State University, Dr.Sc., professor (Alexander.Kolnogorov@novsu.ru).

Abstract: We consider the customization of the UCB strategy, which was first proposed by J. Bather for Bernoulli two-armed bandit, to the case of a Gaussian multi-armed bandit describing batch data processing. This optimal control problem has classical interpretation as a game with nature, in which the payment function of the player is the expected loss of total income caused by incomplete information. The goal is stated in minimax setting. For the considered game with nature, we present an invariant description of the control with a horizon equal to one, which allows to perform computations in two ways: using Monte-Carlo simulations and analytically by dynamic programming technique. For various configurations of the considered game with nature, we have found saddle points, which characterize the optimal control and the worst-case distribution of the parameters of the multi-armed bandit.

Keywords: multi-armed bandit problem, Gaussian multi-armed bandit, minimax approach, UCB rule, invariant description, Monte-Carlo simulations, dynamic programming.