Математическая Теория Игр и её Приложения, т.13, в.2, с. 9–39

УДК 519.832, 519.245
ББК 22.18

ЗАДАЧА О ДВУРУКОМ БАНДИТЕ И ПАКЕТНАЯ ВЕРСИЯ АЛГОРИТМА ЗЕРКАЛЬНОГО СПУСКА

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

Александр В. Назин**
Институт проблем управления
им. В.А. Трапезникова РАН
117997, Москва, ул. Профсоюзная, 65
e-mail: nazine@ipu.ru

Дмитрий Н. Шиян***
Новгородский государственный университет
им. Ярослава Мудрого
173003, Великий Новгород, ул. Б.С.-Петербургская, 41
e-mail: dsqq-tm@ya.ru

Рассматривается минимаксная постановка задачи о двуруком бандите в приложении к обработке данных, если для обработки имеются два альтернативных метода с различными

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

** Исследование выполнено при финансовой поддержке РНФ, научный проект номер 16-11-10015

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

априори неизвестными эффективностями. Требуется определить более эффективный метод и обеспечить его преимущественное применение. Для этой цели используется алгоритм зеркального спуска (АЗС). Известно, что минимаксный риск, обеспечиваемый этим алгоритмом, имеет порядок $N^{1/2}$, где N характеризует количество обрабатываемых данных, причем этот порядок неулучшаем. Нами предложена версия АЗС, позволяющая обрабатывать данные пакетами, что особенно важно, если можно обеспечить параллельную обработку данных. В этом случае полное время обработки определяется количеством обрабатываемых пакетов, а не полным числом данных. Неожиданным оказался результат, что пакетный алгоритм ведет себя не так, как обычный, даже если количество пакетов, на которые разбиты данные, велико. Более того, пакетная версия позволила значительно уменьшить величину минимаксного риска, т.е. повысить качество управления. Для объяснения этого результата мы рассмотрели еще одну пакетную версию АЗС, демонстрирующую поведение, близкое к поведению обычного алгоритма и обеспечивающую близкое значение минимаксного риска. Наши оценки используют инвариантное описание алгоритмов, основанное на гауссовских аппроксимациях доходов в пакетах в области «близких» распределений и получены с помощью моделирования методом Монте-Карло.

Ключевые слова: задача о двуруком бандите, минимаксный подход, алгоритм зеркального спуска, EXP3, пакетная обработка.

Поступила в редакцию: 20.11.20 *После доработки*: 08.12.20 *Принята к публикации*: 01.03.21

1. Введение

Рассматривается задача о двуруком бандите (см., например, [8], [15]), известная также как задача о целесообразном поведении в случайной среде ([2], [11]) и как одна из проблем адаптивного управления ([6], [10]) в следующей постановке. Пусть ξ_n, $n = 1, \ldots, N$ – это управляемый случайный процесс, значения которого интерпретируются как доходы, зависят только от текущих выбранных действий y_n ($y_n \in \{1, 2\}$) и описываются распределениями

$$\Pr(\xi_n = 1 | y_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0 | y_n = \ell) = q_\ell, \tag{1.1}$$

где $p_\ell+q_\ell = 1$, $\ell = 1, 2$, т.е. рассматривается бернуллиевский двурукий бандит. Он характеризуется параметром $\theta = (p_1, p_2)$ с множеством допустимых значений $\Theta = \{\theta : 0 \le p_\ell \le 1;\ \ell = 1, 2\}$. Отметим, что математическое ожидание и дисперсия одношаговых доходов бернуллиевского двурукого бандита равны

$$m_\ell = \mathbf{E}(\xi_n | y_n = \ell) = p_\ell,\ D_\ell = \mathbf{D}(\xi_n | y_n = \ell) = p_\ell q_\ell.$$

Величина N называется горизонтом управления, она предполагается известной и достаточно большой.

В данной статье задача рассматривается в приложении к обработке данных, если для обработки имеются два альтернативных метода с различными априори неизвестными эффективностями. В этом случае доходы $\xi_n = 1$ и $\xi_n = 0$ интерпретируются соответственно как успешная и неуспешная обработки данного с номером n, действия соответствуют методам обработки, N – это число данных, а цель управления состоит в максимизации математического ожидания количества успешно обработанных данных.

Стратегия управления σ в момент времени n обеспечивает выбор действия y_n в зависимости от известной текущей предыстории процесса, т.е. откликов $x^{n-1} = x_1, \ldots, x_{n-1}$ на ранее примененные действия $y^{n-1} = y_1, \ldots, y_{n-1}$:

$$\sigma_\ell(y^{n-1}, x^{n-1}) = \Pr(y_n = \ell | y^{n-1}, x^{n-1}),\ \ell = 1, 2.$$

Такое множество стратегий, которое является наиболее общим, обозначим Σ_0. На множество стратегий могут быть наложены некоторые дополнительные ограничения. Например, через Σ_1 обозначим множество стратегий, реализующих алгоритм зеркального спуска (далее, АЗС), рассматриваемый в этой статье. Более подробно это и другие множества стратегий будут определены ниже.

Опишем цель управления. Если бы параметр θ был известен, то оптимальная стратегия всегда применяла бы действие, соответствующее большей из величин m_1, m_2. Математическое ожидание полного дохода в этом случае равно $N(m_1 \vee m_2)$, где $m_1 \vee m_2$ – максимум из m_1, m_2. В случае неизвестных значений m_1, m_2 функция потерь

$$L_N(\sigma, \theta) = N(m_1 \vee m_2) - \mathbf{E}_{\sigma,\theta}\left(\sum_{n=1}^{N} \xi_n\right) \tag{1.2}$$

описывает математическое ожидание потерь полного дохода относительно максимально возможной величины вследствие неполноты информации. Здесь через $\mathbf{E}_{\sigma,\theta}$ обозначено математическое ожидание, вычисленное по мере, порожденной стратегией σ и параметром θ. Величина

$$R_N^{(0)}(\Theta) = \inf_{\Sigma_0} \sup_{\Theta} L_N(\sigma, \theta) \tag{1.3}$$

характеризует минимаксный риск, вычисленный на множестве стратегий Σ_0, а соответствующая оптимальная стратегия σ_0^M называется минимаксной стратегией. Отметим, что применение стратегии σ_0^M влечет выполнение неравенства

$$L_N(\sigma_0^M, \theta) \le R_N^{(0)}(\Theta)$$

для всех $\theta \in \Theta$, что означает робастность управления.

Данная задача может рассматриваться как игра с природой (см., например, [1]). В этом случае множество Σ_0 описывает стратегии лица, осуществляющего управление, а множество Θ – стратегии природы. Формула (1.2) описывает платежную функцию игроков в нормальной форме. В развернутой форме игра может быть представлена в виде дерева, в вершинах которого, формирующихся в соответствии с историей игры $\{y^{n-1}, x^{n-1}\}$, выполняется выбор текущих действий $\{y_n\}$ и получение ответных доходов $\{\xi_n\}$. Так как информация о стратегиях природы поступает в виде выборки случайных величин (доходов $\{\xi_n\}$ в ответ на выбранные действия $\{y_n\}$), то данная игра является статистической. Наконец, отметим, что хотя эта игра является игрой с нулевой суммой, в ней только лицо, осуществляющее управление, заинтересовано в максимизации собственного выигрыша; природа к результатам игры равнодушна.

Минимаксный подход к задаче о двуруком бандите был предложен в [23]. В [16] было показано, что точное нахождение минимаксных стратегии и риска практически невозможно уже при $N > 4$. В [24] была установлена асимптотическая минимаксная теорема, согласно которой для минимаксного риска (1.3) справедливы оценки

$$0.612 \le (DN)^{-1/2} R_N^{(0)}(\Theta) \le 0.752 \tag{1.4}$$

при $N \to \infty$, где $D = 0.25$ характеризует максимальную дисперсию одношагового дохода. Оценка сверху обеспечивается стратегией, предложенной в [24]. Оценка снизу была получена как точное

значение минимаксного риска при дополнительных ограничениях на множество параметров и стратегий и неоднократно в дальнейшем улучшалась. Представленная здесь оценка снизу получена в [14].

Исследование пакетной обработки данных в задаче о двуруком бандите рассмотрено в [4, 5]. В этом случае ко всем данным пакета применяется одно и то же действие, а для управления используются суммарные доходы в пакетах. Основное свойство пакетной обработки в случае больших данных состоит в том, что она практически не увеличивает минимаксный риск в сравнении с обработкой данных по одному, если количество пакетов достаточно велико. Это особенно важно, если можно обеспечить параллельную обработку данных пакета; в этом случае полное время обработки определяется количеством сформированных пакетов, а не количеством данных. В разделе 2 пакетная обработка обсуждается более детально.

Известен ряд других подходов к робастному управлению в задаче о двуруком бандите, см., например, [3, 6, 12, 17, 20, 22]. В этих публикациях для управления используются метод стохастической аппроксимации, правило верхней границы доверительного интервала (upper confidence bound, UCB), правила EXP2, EXP3 и АЗС (mirror descent algorithm, MDA). Вместо минимаксного риска часто рассматривается эквивалентная характеристика – гарантированная скорость сходимости. Порядок минимаксного риска для указанных алгоритмов равен $N^{1/2}$ или близок к $N^{1/2}$. Отметим, что пакетные версии для данных алгоритмов ранее не рассматривались.

Наконец, отметим публикации [9, 13, 18, 19, 21], в которых изучается поведение платежной функции (1.2) при фиксированном значении параметра θ и $N \to \infty$. В этом случае при широких предположениях $L_N(\sigma, \theta)$ имеет порядок $\ln(N)$, и этот порядок является неулучшаемым.

В данной статье рассматривается АЗС для задачи о двуруком бандите, предложенный в [17]. Для этого алгоритма в [17] получена оценка минимаксного риска сверху как $r^{(1)}N^{1/2}$ и оценка множителя $r^{(1)} \leq 4.710$. С помощью моделирования методом Монте-Карло, мы улучшили оценку множителя как $r^{(1)} \leq 2.0$. Затем предложена пакетная версия алгоритма, которая делит применение обоих действий в каждом пакете пропорционально вероятностям выбора этих дей-

ствий. Для этой версии АЗС с использованием гауссовской аппроксимации доходов в пакетах получено инвариантное описание управления с горизонтом, равным единице. Это описание справедливо в области «близких» распределений, характеризуемых тем, что разность $|p_1 - p_2|$ имеет порядок $N^{-1/2}$. Отметим, что максимальные значения функции потерь достигаются именно в области «близких» распределений (см., например, [5]). С использованием инвариантного описания показано, что минимаксный риск для пакетной версии АЗС имеет порядок $N^{1/2}$, а множитель оценен как $r^{(2)} \approx 1.1$. Весьма неожиданным оказалось, что пакетная версия позволила существенно уменьшить величину минимаксного риска, поскольку значение множителя для нее меньше. На первый взгляд это противоречит здравому смыслу, поскольку пакетная обработка накладывает дополнительные ограничения на стратегию.

Для объяснения этого результата была рассмотрена еще одна пакетная версия АЗС, которая распределяет применение действий к данным пакета последовательно с вероятностями, определенными в начале обработки пакета. Эта версия АЗС ведет себя так же, как и обычный алгоритм, если количество пакетов достаточно велико. Для нее также получено инвариантное описание в области «близких» распределений с использованием гауссовской аппроксимации доходов в пакетах. Таким образом, более высокая эффективность первой пакетной версии АЗС объясняется тем, что она использует другой (не вероятностный) метод распределения данных по пакетам.

Важно понимать, что пакетные версии АЗС хорошо функционируют только в средах с «близкими» вероятностями p_1, p_2. Если p_1, p_2 существенно различаются, то могут возникать значительные потери при обработке первого пакета, когда оба действия применяются поровну. Чтобы этого избежать, предложены комбинированные версии АЗС, которые на коротком начальном этапе управления используют обычную версию АЗС, а затем переключаются на пакетную. Комбинированные алгоритмы хорошо функционируют в средах с любыми значениями вероятностей p_1, p_2.

Наконец, отметим, что численную оптимизацию параметров пакетных версий АЗС можно в принципе выполнить не только с помощью моделирования Монте-Карло, но и непосредственно, так как со-

ответствующие формулы для вычисления функций потерь получены (см. теоремы 4.1, 5.1), однако размерность вычислений резко возрастает с ростом числа пакетов. При этом метод Монте-Карло привлекает простотой реализации. Отметим также, что полученные оценки минимаксного риска не являются аккуратными аналитическими оценками, как представленные, например, в [5, 17, 24]. Это, скорее, оценки, полученные численными методами с достаточной степенью точности и демонстрирующие высокое качество управления, обеспечиваемое рассматриваемыми алгоритмами, которые по этой причине, а также в силу своей простоты могут найти практическое применение.

Структура статьи следующая. В разделе 2 представлены предварительные сведения и обозначения. В разделе 3 рассмотрено описание АЗС, предложенного в [17], и улучшена оценка минимаксного риска с помощью моделирования Монте-Карло. В разделе 4 представлена пакетная версия этого алгоритма и дано ее описание в области «близких» распределений. В разделе 5 предложена еще одна пакетная версия АЗС, которая ведет себя аналогично обычному алгоритму, если число пакетов велико. Комбинированные алгоритмы представлены в разделе 6. Раздел 7 содержит заключение.

2. Предварительные сведения и обозначения

Групповая обработка для задачи о двуруком бандите впервые была предложена для лечения большой группы пациентов двумя альтернативными лекарствами. Поскольку для проявления результата лечения требуется значительное время, было предложено давать одинаковые лекарства группам пациентов. Обсуждение этого подхода и библиографию можно найти, например, в [19]. В приложении к обработке данных пакетный подход исследовался в [4, 5]. Опишем коротко результаты [4, 5] применительно к рассматриваемой задаче.

Пусть требуется обработать $N = T \times M$ данных, характеризуемых распределениями (1.1), где доходы $\xi_n = 1$ и $\xi_n = 0$ соответствуют успешной и неуспешной обработкам данного с номером n. Предположим, что p_1, p_2 в (1.1) близки к некоторому p $(0 < p < 1)$. Разобьем все данные на T пакетов по M данных и будем обрабатывать данные каждого пакета одним и тем же методом. Для управления используем значения процесса $\zeta_t = M^{-1/2} \sum_{n=(t-1)M+1}^{tM} \xi_n$, $t = 1, \ldots, T$. В

силу центральной предельной теоремы распределения ζ_t, $t = 1, \ldots, T$ близки к нормальным, а их дисперсии близки к $\hat{D} = p(1-p)$.

Пусть управление осуществляется с использованием стратегий из Σ_0 на основе наблюдений процесса $\{\zeta_t\}$ и пусть N достаточно велико, т.е. рассматривается обработка больших данных. В [4, 5] показано, что в этом случае максимальные значения функции потерь (1.2) и минимаксный риск (1.3) достигаются в области «близких» распределений, для которых $|p_1 - p_2| \le 2cN^{-1/2}$, где $c > 0$ достаточно большая фиксированная константа. Минимаксный риск мало зависит от количества пакетов T, на которые разбиты данные, если T достаточно велико, и определяется в этом случае полным числом данных N и дисперсией $\hat{D}$, причем

$$\lim_{\substack{M \to \infty \\ T \to \infty}} (\hat{D}N)^{1/2} R_N^{(0)}(\Theta) = \hat{r}^{(0)}, \tag{2.1}$$

где $\hat{r}^{(0)} \approx 0.637$, что согласуется с (1.4) при $\hat{D} = 0.25$. Дисперсию $\hat{D}$ можно считать известной, так как она может быть оценена на коротком начальном отрезке управления, а минимаксный риск мало меняется при малом изменении дисперсии.

При этом оптимальная обработка больших данных, характеризуемых распределением (1.1), по одному не является более эффективной, чем оптимальная пакетная обработка, так как не позволяет уменьшить предельное значение нормированного минимаксного риска $(\hat{D}N)^{-1/2} R_N^{(0)}(\Theta)$. Таким образом, оценка (2.1) характеризует принципиальную нижнюю границу для нормированного минимаксного риска, которая в дальнейшем будет использоваться для оценки качества рассматриваемых алгоритмов.

Опишем теперь общую схему пакетных алгоритмов, рассматриваемых в этой статье. В отличие от [4, 5], где одно и то же действие применяется ко всем данным пакета, ниже рассматриваются схемы, где сначала M данных пакета с номером t распределяются между двумя пакетами меньшего объема пропорционально вероятностям $p_{t-1}^{(1)}$, $p_{t-1}^{(2)}$, определенным по результатам обработки предыдущих $t-1$ пакетов ($t = 1, \ldots, T$), а затем данные малых пакетов обрабатываются соответствующими им методами. Мы предполагаем, что временем распределения данных по малым пакетам можно пренебречь в срав-

нении с временем обработки данных, поэтому в случае параллельной обработки полное время работы алгоритма определяется числом пакетов T, а не полным числом данных N. Анализ ведется в области «близких» распределений в окрестности некоторого параметра (p, p), определяемой либо как $\hat{\theta}_N = (p+d(\hat{D}/N)^{1/2}, p-d(\hat{D}/N)^{1/2})$, либо как $\theta_N = (p + d(D/N)^{1/2}, p - d(D/N)^{1/2})$, где $0 < p < 1$, $\hat{D} = p(1-p)$, $D = \max\limits_{0<p<1} \hat{D} = 0.25$. Множество стратегий, реализующих алгоритм **Ak**, обозначается Σ_k. Для алгоритма **Ak** могут исследоваться нормированные функции потерь

$$\hat{l}_N^{(k)}(\beta, T, p, d) = (\hat{D}N)^{-1/2} L_N(\sigma, \hat{\theta}_N), \tag{2.2}$$

$$l_N^{(k)}(\beta, T, p, d) = (DN)^{-1/2} L_N(\sigma, \theta_N), \tag{2.3}$$

где $\sigma \in \Sigma_k$ – стратегия, реализующая алгоритм **Ak** с параметром β и размером пакета T. Для $\hat{l}_N^{(k)}(\beta, T, p, d)$, $l_N^{(k)}(\beta, T, p, d)$ определяются нормированные минимаксные риски

$$\hat{r}_N^{(k)}(T) = \min_{\beta} \max_{d} \hat{l}_N^{(k)}(\beta, T, p, d), \tag{2.4}$$

$$r_N^{(k)}(T) = \min_{\beta} \max_{p,d} l_N^{(k)}(\beta, T, p, d). \tag{2.5}$$

При $T = 1$ зависимость функций потерь и рисков от T не указывается. Основными характеристиками алгоритмов являются (2.3), (2.5). При этом (2.2), (2.4) являются дополнительными характеристиками: для **A2** они, как и в случае (2.1), асимптотически не зависят от p, а для **A1** и **A3**, напротив, зависят.

3. Алгоритм зеркального спуска для бернуллиевского двурукого бандита

В этом разделе описывается АЗС, предложенный в [17] для управления бернуллиевским двуруким бандитом. Отметим, что первоначально алгоритм зеркального спуска был предложен в [7] для эффективного решения задач выпуклого программирования большой размерности. Отметим также, что близкую формулировку имеет алгоритм EXP3 (см., например, [20]).

Рассмотрим вектор вероятностей выбора действий $\overline{p}_n = (p_n^{(1)}, p_n^{(2)})$, координаты которого удовлетворяют условиям $p_n^{(1)} \geq 0$, $p_n^{(2)} \geq 0$, $p_n^{(1)} +$

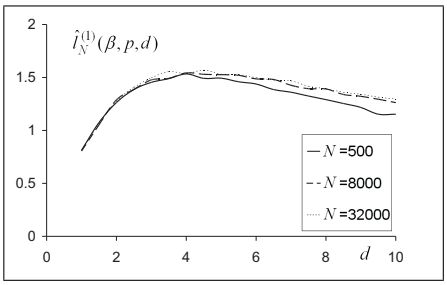

Рисунок 1. Алгоритм **A1**.
$\beta = 2.2$; $p = 0.5$;
$N = 500, 8000, 32000$.

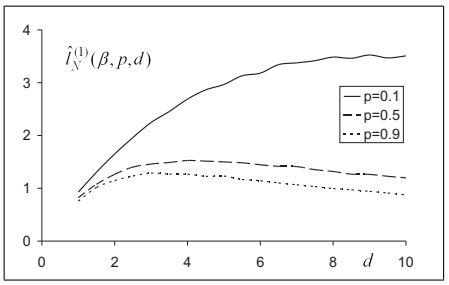

Рисунок 2. Алгоритм **A1**.
$\beta = 2.2$; $N = 2000$;
$p = 0.1, 0.5, 0.9$.

$p_n^{(2)} = 1$, дуальный вектор $\overline{\zeta}_n = (\zeta_n^{(1)}, \zeta_n^{(2)})$ и вектор стохастического градиента $\overline{u}_n = (u_n^{(1)}, u_n^{(2)})$. Определим распределение Гиббса

$$\overline{G}_\beta(\overline{\zeta}) = \{S_\beta(\overline{\zeta})\}^{-1} \left(e^{-\zeta^{(1)}/\beta}, e^{-\zeta^{(2)}/\beta}\right),$$

где $S_\beta(\overline{\zeta}) = e^{-\zeta^{(1)}/\beta} + e^{-\zeta^{(2)}/\beta}$.

АЗС для управления бернуллиевским двуруким бандитом определяется следующим образом.

Алгоритм A1.

---

1. Выбрать некоторое $\overline{\zeta}_0$ и положить $\overline{p}_0 = \overline{G}_{\beta_0}(\overline{\zeta}_0)$.

2. **for** $n = 1, 2, \ldots, N$ **do**

   (a) Выбрать y_n в соответствии с распределением:

   $$\Pr(y_n = \ell) = p_{n-1}^{(\ell)}, \quad \ell = 1, 2;$$

   (b) Применить действие y_n и получить случайный доход ξ_n в соответствии с распределением:

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

(c) Вычислить стохастический градиент $\overline{u}_n(\overline{p}_{n-1})$:

$$\overline{u}_n(\overline{p}_{n-1}) = \begin{cases} \left(\dfrac{1-\xi_n}{p^{(1)}_{n-1}}, 0\right), & \text{если} \quad y_n = 1, \\ \left(0, \dfrac{1-\xi_n}{p^{(2)}_{n-1}}\right), & \text{если} \quad y_n = 2; \end{cases}$$

(d) Обновить дуальный вектор и вектор вероятностей

$$\overline{\zeta}_n = \overline{\zeta}_{n-1} + \overline{u}_n(\overline{p}_{n-1}), \;\; \overline{p}_n = \overline{G}_{\beta_n}(\overline{\zeta}_n);$$

(e) **end for**

Обозначим через Σ_1 множество стратегий, порождаемых АЗС, и через

$$R^{(1)}_N(\Theta) = \inf_{\Sigma_1} \sup_{\Theta} L_N(\sigma, \theta) \tag{3.1}$$

соответствующий минимаксный риск. В [17] установлена

Теорема 3.1. *Рассмотрим алгоритм* **A1**. *Пусть* $\beta_n = \beta^* \times \{D(n+1)\}^{1/2}$, *где* $\beta^* = (8/\ln 2)^{1/2} \approx 3.397$, $D = 0.25$. *Тогда для любого горизонта управления* $N \geq 1$ *для минимаксного риска* (3.1) *справедлива оценка сверху*

$$R^{(1)}_N(\Theta) \leq r^* \{D(N+1)\}^{1/2}, \tag{3.2}$$

где $r^* = 4(2\ln 2)^{1/2} \approx 4.710$.

Замечание 3.1. Наше описание алгоритма отличается от представленного в [17] в следующих деталях. В [17] алгоритм предложен для задачи о минимизации полного дохода для многорукого бандита с произвольным конечным числом действий и сформулирован с использованием 2-го момента одношагового дохода.

Оценка (3.2) была получена аналитически. Отметим, что она приблизительно в 7.39 раза хуже оценки (2.1). Ее можно улучшить с помощью моделирования методом Монте-Карло. Для этого вычислим определенную в (2.2) величину $\hat{l}^{(1)}_N(\beta, p, d)$ с параметром алгоритма $\beta_n = \beta\{\hat{D}(n+1)\}^{1/2}$. Здесь и ниже мы выбираем $\zeta^{(1)}_0 = \zeta^{(2)}_0 = 0$ и,

следовательно, $p_0^{(1)} = p_0^{(2)} = 0.5$. Количество симуляций Монте-Карло всегда равно 10000.

На рис. 1 представлены значения $\hat{l}_N^{(1)}(\beta, p, d)$, вычисленные для различных горизонтов управления при $\beta = 2.2$, $p = 0.5$ и $1 \le d \le 10$. Результаты представлены для $N = 500; 8000; 32000$. Видно, что значения $\hat{l}_N^{(1)}(\beta, p, d)$ близки при больших N.

Можно предположить, что при больших N функция $\hat{l}_N^{(1)}(\beta, p, d)$ не зависит от $\hat{D}$, как это имеет место в (2.1) для стратегий из Σ_0. Однако, это не так. На рис. 2 представлены значения $\hat{l}_N^{(1)}(\beta, p, d)$, вычисленные при $\beta = 2.2$, $N = 2000$ и $0 \le d \le 10$ для $p = 0.1; 0.5; 0.9$. Видно, что соответствующие кривые не являются близкими и наибольшие потери соответствуют наименьшим p.

Поэтому далее мы сначала найдем параметр β, минимизирующий $\max\limits_d l_N^{(1)}(\beta, p, d)$ при $p = 0.1$, а затем проверим, что этот максимум не будет превышен и при других рассматриваемых d, p. Результаты вычисления $l_N^{(1)}(\beta, p, d)$ при $p = 0.1$, $N = 2000$ и $0 \le d \le 10$ представлены на рис. 3 при $\beta = 1.0; 2.0; 3.0$. Видно, что $\beta = 2.0$ – близкое к оптимальному значение параметра, потому что оно обеспечивает минимальное значение $\max\limits_d l_N^{(1)}(\beta, p, d)$. Более полное исследование дает, что $\beta = 2.2$ является приблизительно оптимальным значением параметра.

Результаты вычисления $l_N^{(1)}(\beta, p, d)$ при $\beta = 2.2$, $N = 2000$ представлены на рис. 4 при $p = 0.1; 0.5; 0.9$. Видно, что $l_N^{(1)}(\beta, p, d)$ достигает максимальных значений при $p = 0.1$. Следовательно, $\beta = 2.2$ – приблизительно оптимальное значение параметра и

$$r_N^{(1)} = \min_{\beta > 0} \max_{\substack{1 \le d \le 10, \\ 0.1 < p < 0.9}} l_N^{(1)}(\beta, p, d) \approx 2.0. \tag{3.3}$$

Эта оценка примерно в 2.37 раза лучше теоретической оценки (3.2).

Замечание 3.2. Можно усомниться в справедливости оценки (3.3), если N отлично от 2000, поскольку не было доказано существование предела $l_N^{(1)}(\beta, p, d)$ при $N \to \infty$. Действительно, предположение о существовании предела основывается только на анализе рис. 1. В разделе 5 предложена пакетная версия АЗС, поведение которой близко к поведению алгоритма **A1**, если количество пакетов достаточно

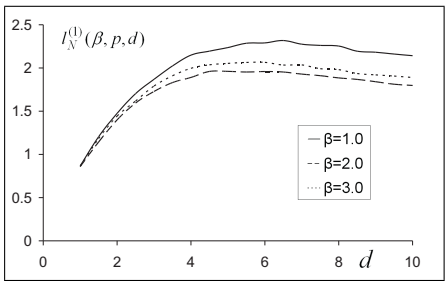

Рисунок 3. Алгоритм **A1**. $p = 0.1$, $N = 2000$, $\beta = 1.0; 2.0; 3.0$.

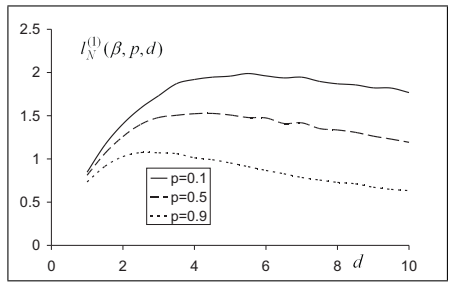

Рисунок 4. Алгоритм **A1**, $\beta = 2.2$; $N = 2000$, $p = 0.1; 0.5; 0.9$.

велико. Для этой пакетной версии АЗС доказано существование соответствующего предела.

4. Пакетная версия АЗС

Пусть требуется обработать $N = T \times M$ данных, где M – размер пакета, а T – количество обрабатываемых пакетов. Пусть $\overline{M}_t = (M_t^{(1)}, M_t^{(2)})$ – вектор, удовлетворяющий условиям $M_t^{(1)} \geq 0$, $M_t^{(2)} \geq 0$, $M_t^{(1)} + M_t^{(2)} = M$. Обозначим через $I_t^{(1)}$, $I_t^{(2)}$ ближайшие целые числа к $M_t^{(1)}$, $M_t^{(2)}$, причем $I_t^{(1)} + I_t^{(2)} = M$. Например, можно положить $I_t^{(1)} = [M_t^{(1)} + 0.5]$, $I_t^{(2)} = M - I_t^{(1)}$, где $[\cdot]$ – знак целой части числа.

При $0 < \varrho < 0.5$ определим оператор проекции вектора $\overline{p} = (p^{(1)}, p^{(2)})$ на отрезок $[\varrho, 1-\varrho]$ условиями

$$\mathcal{P}_\varrho(\overline{p}) = \begin{cases} \overline{p}, & \text{если } p^{(\ell)} \in [\varrho, 1-\varrho],\ \ell = 1,2, \\ (\varrho, 1-\varrho), & \text{если } p^{(1)} < \varrho, \\ (1-\varrho, \varrho), & \text{если } p^{(2)} < \varrho. \end{cases}$$

В отличие от обычной версии АЗС, пакетная на этапе с номером t применяет первое и второе действия к пакетам из $I_t^{(1)}$ и $I_t^{(2)}$ данных, где $M_t^{(\ell)} = M \times p_{t-1}^{(\ell)}$, $\ell = 1, 2$, причем оператор проекции гарантирует, что оба пакета не пусты, если M достаточно велико.

Алгоритм A2.

---

1. Выбрать некоторое $\overline{\zeta}_0$ и положить $\overline{p}'_0 = \overline{G}_{\beta_0}(\overline{\zeta}_0)$, $\overline{p}_0 = \mathcal{P}_\varrho(\overline{p}'_0)$.

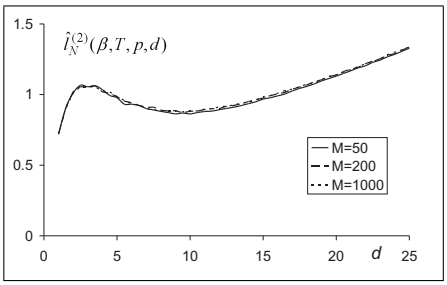

Рисунок 5. Алгоритм **A2**. $\beta = 1.0$; $\varrho = 0.02$; $p = 0.5$; $T = 100$; $M = 50; 200; 1000$.

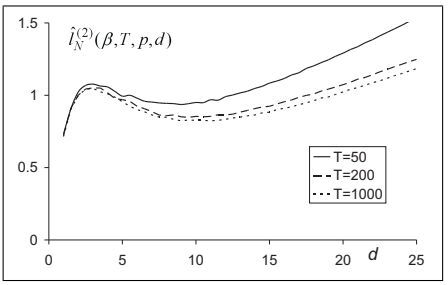

Рисунок 6. Алгоритм **A2**. $\beta = 1.0$; $\varrho = 0.02$; $p = 0.5$; $M = 100$; $T = 50; 200; 1000$.

2. **for** $t = 1, 2, \ldots, T$ **do**

(a) Положить $M_t^{(\ell)} = Mp_{t-1}^{(\ell)}$, $\ell = 1, 2$;

(b) i. Положить $\chi_t^{(1)} = 0$;

ii. **for** $n = (t-1)M + 1, \ldots, (t-1)M + I_t^{(1)}$ **do**

A. Применить действие $y_n = 1$, получить случайный доход ξ_n в соответствии с распределением

$$\Pr(\xi_n = 1 | y_n = 1) = p_1, \ \Pr(\xi_n = 0 | y_n = 1) = q_1;$$

B. Обновить $\chi_t^{(1)}$ в соответствии с правилом

$$\chi_t^{(1)} \leftarrow \chi_t^{(1)} + (1 - \xi_n);$$

C. **end for**

(c) i. Положить $\chi_t^{(2)} = 0$;

ii. **for** $n = (t-1)M + I_t^{(1)} + 1, \ldots, tM$ **do**

A. Применить действие $y_n = 2$, получить случайный доход ξ_n в соответствии с распределением

$$\Pr(\xi_n = 1 | y_n = 2) = p_2, \ \Pr(\xi_n = 0 | y_n = 2) = q_2;$$

B. Обновить $\chi_t^{(2)}$ в соответствии с правилом

$$\chi_t^{(2)} \leftarrow \chi_t^{(2)} + (1 - \xi_n);$$

C. end for

(d) Вычислить стохастический градиент $\overline{u}_t(\overline{p}_{t-1})$

$$\overline{u}_t(\overline{p}_{t-1}) = \left(\frac{\chi_t^{(1)}}{p_{t-1}^{(1)}}, \frac{\chi_t^{(2)}}{p_{t-1}^{(2)}}\right);$$

(e) Обновить дуальный вектор и вектор вероятностей

$$\overline{\zeta}_t = \overline{\zeta}_{t-1} + \overline{u}_t(\overline{p}_{t-1}),\ \ \overline{p}'_t = \overline{G}_{\beta_t}(\overline{\zeta}_t),\ \ \overline{p}_t = \mathcal{P}_\varrho(\overline{p}'_t);$$

(f) end for

Отметим, что доходы $\{\chi_t^{(\ell)}\}$ имеют биномиальное распределение и

$$\mathbf{E}(\chi_t^{(\ell)}) = q_\ell I_t^{(\ell)},\ \mathbf{D}(\chi_t^{(\ell)}) = D_\ell I_t^{(\ell)}, \tag{4.1}$$

где $D_\ell = p_\ell q_\ell \approx \hat{D} = pq$, $\ell = 1, 2$. Справедлива

Теорема 4.1. *Рассмотрим алгоритм* **A2** *с фиксированным числом пакетов T и параметром $\beta_t = \beta(\hat{D}M(t+0.5))^{1/2}$. Тогда при $N = M \times T$ существует предел*

$$\hat{l}^{(2)}(\beta, T, d) = \lim_{M\to\infty} \hat{l}_N^{(2)}(\beta, T, p, d), \tag{4.2}$$

который не зависит от p и $\hat{D}$, а только от β, T и d.

Доказательство. Положим

$$\eta_t^{(\ell)} = \left(\chi_t^{(\ell)} - \mathbf{E}(\chi_t^{(\ell)})\right)\left(\mathbf{D}(\chi_t^{(\ell)})\right)^{-1/2}, \tag{4.3}$$

через $f^{(\ell)}(x, I_t^{(\ell)})$ обозначим плотность распределения случайной величины $\eta_t^{(\ell)}$. Распределение $\eta_t^{(\ell)}$ зависит только от текущего выбранного действия ℓ и объема выборки $I_t^{(\ell)}$. Так как все $\{\eta_t^{(\ell)}\}$ сформированы на основе выборок объемом не меньше чем $M\varrho$, то в силу локальной предельной теоремы Муавра-Лапласа справедливы равенства

$$f^{(\ell)}(x, I_t^{(\ell)}) = \varphi(x)(1 + \delta_t^{(\ell)}(x, I_t^{(\ell)})), \quad t = 1, \ldots, T,\ \ell = 1, 2, \tag{4.4}$$

при $|x| \le A$ и любом $A > 0$, где $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ – плотность стандартного нормального распределения, а $\delta_t^{(\ell)}(x, I_t^{(\ell)}) \to 0$ равномерно по x, если $M \to \infty$. Пусть $q_\ell = q + w_\ell(\hat{D}/N)^{1/2}$, где $w_\ell = (-1)^\ell d,\ \ell = 1, 2;\ d > 0$. Из (4.3) следует, что

$$\chi_t^{(\ell)} = \left(q + w_\ell(\hat{D}/N)^{1/2}\right) I_t^{(\ell)} + \left(D_\ell I_t^{(\ell)}\right)^{1/2} \eta_t^{(\ell)}.$$

Отметим, что $|I_t^{(\ell)} - M_t^{(\ell)}| \le 0.5$ и $I_t^{(\ell)} \ge \varrho M$, поэтому $I_t^{(\ell)} = M_t^{(\ell)}(1 + \alpha_{t,M}^{(\ell)})$, где $\alpha_{t,M}^{(\ell)}$ – случайная величина, которая удовлетворяет условию $|\alpha_{t,M}^{(\ell)}| \le 0.5(\varrho M)^{-1}$. Следовательно,

$$\chi_t^{(\ell)} = \left(qMp_{t-1}^{(\ell)} + \varepsilon w_\ell(\hat{D}N)^{1/2} p_{t-1}^{(\ell)}\right)(1 + \alpha_{t,M}^{(\ell)}) + \\ + \left(D_\ell(1 + \alpha_{t,M}^{(\ell)})N\varepsilon p_{t-1}^{(\ell)}\right)^{1/2} \eta_t^{(\ell)}.$$

Здесь $\varepsilon = M/N = 1/T$. Далее

$$\frac{\chi_t^{(\ell)}}{p_{t-1}^{(\ell)}} = (qM + \varepsilon w_\ell(\hat{D}N)^{1/2})(1 + \alpha_{t,M}^{(\ell)}) + \left(\frac{D_\ell(1 + \alpha_{t,M}^{(\ell)})N\varepsilon}{p_{t-1}^{(\ell)}}\right)^{(1/2)} \eta_t^{(\ell)}.$$

Поэтому

$$\zeta_t^{(\ell)} = qMt + \tau w_\ell(\hat{D}N)^{1/2} + \\ + \sum_{i=1}^{t} \left(\frac{D_\ell(1 + \alpha_{i,M}^{(\ell)})N\varepsilon}{p_{i-1}^{(\ell)}}\right)^{1/2} \eta_i^{(\ell)} + Y_t^{(\ell)}, \tag{4.5}$$

где $\tau = t/T$ и

$$Y_t^{(\ell)} = \sum_{i=1}^{t} \left(qM + \varepsilon w_\ell(\hat{D}N)^{1/2}\right) \alpha_{i,M}^{(\ell)}.$$

С учетом оценок для $\{\alpha_{i,M}^{(\ell)}\}$ получаем, что

$$\limsup_{M\to\infty} |Y_t^{(\ell)}| \le qt(0.5\varrho)^{-1}. \tag{4.6}$$

Вспомним, что $\beta_t = \beta(\hat{D}M(t + 0.5))^{1/2}$. Из (4.5) и (4.6) получаем

$$\frac{\zeta_t^{(\ell)}}{\beta_t} = \frac{qMt}{\beta_t} + \frac{1}{\beta}\left(\frac{\tau w_\ell}{(\tau + 0.5\varepsilon)^{1/2}} + X_t^{(\ell)}\right) + Y_t^{(\ell)}/\beta_t,$$

где

$$X_t^{(\ell)} = \sum_{i=1}^{t} \left(\frac{\varepsilon(D_\ell/\hat{D})(1+\alpha_{i,M}^{(\ell)})}{(\tau+0.5\varepsilon)p_{i-1}^{(\ell)}} \right)^{1/2} \eta_i^{(\ell)},$$

причем $\limsup_{M\to\infty} |Y_t^{(\ell)}/\beta_t| = 0$, $\ell = 1, 2$. Поэтому

$$\frac{\zeta_t^{(1)} - \zeta_t^{(2)}}{\beta_t} = -\frac{2d\tau}{\beta(\tau+0.5\varepsilon)^{1/2}} + \frac{1}{\beta}\left(X_t^{(1)} - X_t^{(2)}\right) + \gamma_t, \tag{4.7}$$

где $\gamma_t = (Y_t^{(1)} - Y_t^{(2)})/\beta_t$. Из определения алгоритма **A2** следует, что

$$p_t'^{(2)} = \frac{1}{\exp\left(-\dfrac{\zeta_t^{(1)} - \zeta_t^{(2)}}{\beta_t}\right) + 1}, \tag{4.8}$$

$p_t'^{(1)} = 1 - p_t'^{(2)}$ и $\overline{p}_t = \mathcal{P}_\varrho(\overline{p}_t')$, при этом функция $\hat{l}_N^{(2)}(\beta, T, p, d)$ может быть представлена в виде

$$\begin{aligned} \hat{l}_N^{(2)}(\beta, T, p, d) &= (\hat{D}N)^{-1/2}(p_1 - p_2)\sum_{t=1}^{T} \mathbf{E}_{\mathbf{A2}}\left(Mp_{t-1}^{(2)}\right) = \\ &= 2d\sum_{t=1}^{T} \varepsilon\ \mathbf{E}_{\mathbf{A2}}\left(p_{t-1}^{(2)}\right), \end{aligned} \tag{4.9}$$

где через $\mathbf{E}_{\mathbf{A2}}$ обозначено математическое ожидание по мере, порожденной алгоритмом **A2**. Из (4.9) следует, что для доказательства (4.2) достаточно установить существования пределов

$$\lim_{M\to\infty} \mathbf{E}_{\mathbf{A2}}\left(p_t^{(2)}\right) \tag{4.10}$$

с аналогичными свойствами при всех $t = 0, \ldots, T-1$. Обозначим через $\overline{x}_t = (x_t^{(1)}, x_t^{(2)})$, $\overline{x}^t = (x_1^{(1)}, \ldots, x_t^{(1)}, x_1^{(2)}, \ldots, x_t^{(2)})$ текущие значения $(\eta_t^{(1)}, \eta_t^{(2)})$ в момент t и полную предысторию к моменту t, причем при $t = 0$ предполагаем, что $\overline{x}_t$ и $\overline{x}^t$ пусты. Тогда $\overline{p}_0 = \mathcal{P}_\varrho(\overline{G}_{\beta_0}(\overline{\zeta}_0))$; в частности, $p_0^{(1)} = p_0^{(2)} = 0.5$, если $\zeta_0^{(1)} = \zeta_0^{(1)} = 0$. При $t \geq 1$ имеем

$$\begin{aligned} \mathbf{E}_{\mathbf{A2}}\left(p_t^{(2)}\right) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p_t^{(2)}\left(\overline{x}^t\right) \times \\ \times \left(\prod_{i=1}^{t} f^{(1)}\left(x_i^{(1)}, I_i^{(1)}(\cdot)\right) f^{(2)}\left(x_i^{(2)}, I_i^{(2)}(\cdot)\right) \right) d\overline{x}^t, \end{aligned} \tag{4.11}$$

где использованы сокращения $d\overline{x}^t = dx_1^{(1)} \dots dx_{t-1}^{(1)} dx_1^{(2)} \dots dx_{t-1}^{(2)}$ и $I_i^{(\ell)}(\cdot) = I_i^{(\ell)}(\overline{p}_{i-1}(\overline{x}^{i-1}))$, $\ell = 1, 2$. Отметим, что для плотностей распределения $\left\{ f^{(\ell)}\left(x_i^{(\ell)}, I_i^{(1)}(\cdot)\right)\right\}$ справедливы оценки (4.4). Функции $p_t^{(\ell)}(\overline{x}^t)$ в соответствии с (4.7), (4.8) при $t = 1, \dots, T-1$ определяется рекуррентно:

$$\overline{p}_t(\overline{x}^t) = \mathcal{P}_\varrho(p'_t(\overline{x}^t)),$$

где

$$p_t'^{(2)}(\overline{x}^t) = \frac{1}{\exp\left(\dfrac{2d\tau}{\beta(\tau+0.5\varepsilon)^{1/2}} - \dfrac{X_t^{(1)}(\overline{x}^t) - X_t^{(2)}(\overline{x}^t)}{\beta} - \gamma_t(\overline{x}^t)\right) + 1},$$

$$X_t^{(\ell)}(\overline{x}^t) = \sum_{i=1}^{t} \left(\frac{\varepsilon(D_\ell/\hat{D})(1+\alpha_i^{(\ell)}(\overline{x}^i))}{(\tau+0.5\varepsilon)p_{i-1}^{(\ell)}(\overline{x}^{i-1})} \right)^{1/2} x_i^{(\ell)}, \quad \ell = 1, 2, \qquad (4.12)$$

$$p_t'^{(1)}(\overline{x}^t) = 1 - p_t'^{(2)}(\overline{x}^t),$$

$\tau = t/T$, $\varepsilon = 1/T$. Так как $D_\ell/\hat{D} \to 1$, $\alpha_i^{(\ell)}(\overline{x}^i) \to 0$, $\gamma_t(\overline{x}^t) \to 0$, $\delta_t^{(\ell)}(x, I_t^{(\ell)}) \to 0$ при $M \to \infty$ равномерно при всех $\ell = 1, 2$ и $t = 1, \dots, T-1$, то функции $p_t^{(\ell)}(\overline{x}^t)$, $f^{(\ell)}\left(x_i^{(\ell)}, I_i^{(\ell)}(\cdot)\right)$ имеют непрерывные пределы при $M \to \infty$ при всех $\ell = 1, 2$ и $t = 1, \dots, T-1$ как суперпозиции непрерывных функций. Для вычисления пределов следует положить $D_\ell/\hat{D} = 1$, $\alpha_i^{(\ell)}(\overline{x}^i) = \gamma_t(\overline{x}^t) = \delta_t^{(\ell)}(x, I_t^{(\ell)}) = 0$ в (4.4), (4.11), (4.12). Отсюда следует существование пределов (4.10) и (4.2), причем эти пределы зависят только от β, T и d. □

Замечание 4.1. Можно ожидать, что алгоритм **A2** сходится при $M \to \infty$, $T \to \infty$. Отметим, что при этом $\varepsilon = 1/T \to 0$. Положим $\varsigma_N^{(\ell)}(\tau) = (\hat{D}N)^{-1/2}\left(\zeta_t^{(\ell)} - qMt\right)$, $P_N^{(\ell)}(\tau) = p_{t-1}^{(\ell)}$, $\ell = 1, 2$, где $\tau = t/T$. Пусть $\varsigma_N^{(\ell)}(\tau)$, $P_N^{(\ell)}(\tau)$ слабо сходятся к $\varsigma^{(\ell)}(\tau)$, $P^{(\ell)}(\tau)$ при $M \to \infty$, $T \to \infty$; через $\hat{l}^{(2)}(\beta, d)$ обозначим соответствующий предел для $\hat{l}^{(2)}(\beta, T, d)$. Обозначим через $W_\ell(\tau)$, $\ell = 1, 2$ независимые винеровские процессы. С использованием (4.5), (4.8) и (4.9) предельное описание имеет вид

$$d\varsigma^{(\ell)}(\tau) = w_\ell d\tau + (P^{(\ell)}(\tau))^{-1/2} dW_\ell(\tau), \ \ell = 1, 2,$$

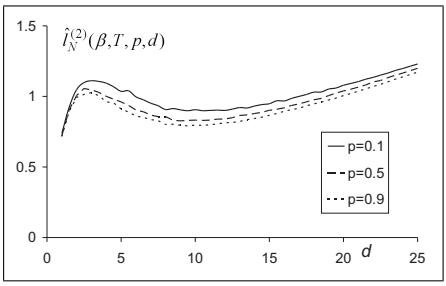

Рисунок 7. Алгоритм **A2**. $\beta = 1.0$; $\varrho = 0.02$; $M = 100$; $T = 500$; $p = 0.1; 0.5; 0.9$.

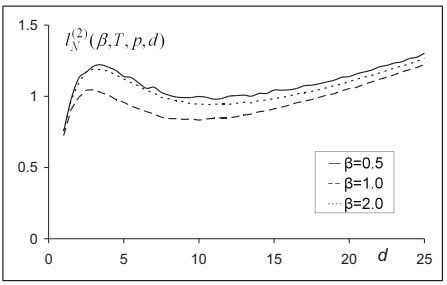

Рисунок 8. Алгоритм **A2**. $\varrho = 0.02$; $M = 100$; $T = 300$; $p = 0.5$; $\beta = 0.5; 1.0; 2.0$.

$$P'^{(2)}(\tau) = \frac{1}{\exp\left(-\dfrac{\varsigma^{(1)}(\tau) - \varsigma^{(2)}(\tau)}{\beta\tau^{1/2}}\right) + 1},$$

$P'^{(1)}(\tau) = 1 - P'^{(2)}(\tau)$, $\overline{P}(\tau) = \mathcal{P}_\varrho\{\overline{P}'(\tau)\}$, $\tau \in [0,1]$. Начальные условия выбираются в виде:

$$\varsigma^{(1)}(0) = \varsigma^{(2)}(0) = 0.$$

А предельная нормированная функция потерь равна

$$\hat{l}^{(2)}(\beta, d) = 2d \int_0^1 \mathbf{E}\left(P^{(2)}(\tau)\right) d\tau.$$

Однако строгого доказательства этого результата пока нет.

Представим результаты моделирования алгоритма **A2**. На рис. 5 представлены значения функции потерь $\hat{l}_N^{(2)}(\beta, T, p, d)$, вычисленные для различных значений M с помощью моделирования Монте-Карло, если $\beta = 1.0$, $p = 0.5$, $\varrho = 0.02$, $T = 100$ и $1 \le d \le 25$. Результаты представлены для $M = 50; 200; 1000$ (соответственно, $T = 200; 50, 10$). Можно видеть, что функция $\hat{l}_N^{(2)}(\beta, T, p, d)$ почти не зависит от размера пакета M.

На рис. 6 представлены значения функции $\hat{l}_N^{(2)}(\beta, T, p, d)$, вычисленные при различных T, если $\beta = 1.0$, $p = 0.5$, $\varrho = 0.02$, $M = 100$ и $1 \le d \le 25$. Результаты представлены для $T = 50; 200; 1000$. Можно видеть, что $\hat{l}_N^{(2)}(\beta, T, p, d)$ демонстрирует сходимость при $T \to \infty$.

Из теоремы 4.1 следует, что $\hat{l}_N^{(2)}(\beta, T, p, d)$ при $N \to \infty$ не зависит от p, если $0 < p < 1$. На рис. 7 представлены значения функции $\hat{l}_N^{(2)}(\beta, T, p, d)$, вычисленные методом Монте-Карло, если $\beta = 1.0$, $M = 100$, $T = 500$, $\varrho = 0.02$ и $1 \le d \le 25$. Результаты представлены для $p = 0.1; 0.5; 0.9$. Можно видеть, что соответствующие кривые являются близкими.

Чтобы определить оптимальное значение β, фиксируем $p = 0.5$, соответствующее максимуму $\hat{D}$, равному 0.25, и вычислим $l_N^{(2)}(\beta, T, p, d)$ методом Монте-Карло, если $M = 100$, $T = 300$, $\varrho = 0.02$ и $0 \le d \le 25$. Результаты представлены на рис. 8 для $\beta = 0.5; 1.0; 2.0$. Можно видеть, что $\beta = 1.0$ является приблизительно оптимальным значением, поскольку минимизирует $l_N^{(2)}(\beta, T, p, d)$, если $d < 20$.

Наконец, вычислим $l_N^{(2)}(\beta, T, p, d)$ при $\beta = 1.0$, $M = 100$, $T = 300$, $\varrho = 0.02$ и $0 \le d \le 25$. Результаты представлены на рис. 9 для $p = 0.1; 0.3; 0.5; 0.7; 0.9$. Можно видеть, что максимальные значения функция $l_N^{(2)}(\beta, T, p, d)$ принимает, если $p = 0.5$. Следовательно, значение $\beta = 1.0$ приблизительно оптимально и

$$r^{(2)} \approx \min_{\beta > 0} \max_{\substack{1 \le d \le 20, \\ 0.1 < p < 0.9}} l_N^{(2)}(\beta, 300, p, d) \approx 1.1.$$

Здесь через $r^{(2)}$ обозначено предельное значение $r^{(2)}(T)$ при $T \to \infty$. Эта оценка даже лучше, чем (3.3). Однако она получена, если $|p_1 - p_2|$ имеет порядок $N^{-1/2}$. Поскольку полные потери на всем горизонте управления не меньше, чем $\varrho |p_2 - p_1| N$, то

$$l_N^{(2)}(\beta, T, p, d) \ge (DN)^{-1/2} (\varrho |p_2 - p_1| N) = \varrho D^{-1/2} |p_2 - p_1| N^{1/2},$$

т.е. $l_N^{(2)}(\beta, T, p, d)$ будет велика при больших $|p_1 - p_2| N^{1/2}$.

5. Еще одна пакетная версия АЗС

Рассмотрим еще одну пакетную версию АЗС, поведение которой близко к поведению обычного алгоритма зеркального спуска.

Алгоритм АЗ.

---

1. Выбрать некоторое $\overline{\zeta}_0$. Положить $\overline{p}_0 = \overline{G}_{\beta_0}(\overline{\zeta}_0)$.

2. **for** $t = 1, 2, \ldots, T$ **do**

(a) i. Положить $\chi_t^{(1)} = \chi_t^{(2)} = 0$.

ii. **for** $n = (t-1)M + 1, \ldots, tM$ **do**

A. Выбрать действие y_n в соответствии с распределением:

$$\Pr(y_n = \ell) = p_{t-1}^{(\ell)}, \ \ \ell = 1, 2;$$

B. Применить действие y_n, получить случайный доход ξ_n в соответствии с распределением:

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

C. и обновить $\chi_t^{(\ell)}$ в соответствии с правилом:

$$\chi_t^{(\ell)} \leftarrow \chi_t^{(\ell)} + (1 - \xi_n),$$

если $y_n = \ell$, $\ell = 1, 2$;

D. **end for**

(b) Вычислить стохастический градиент $\overline{u}_t(\overline{p}_{t-1})$:

$$\overline{u}_t(\overline{p}_{t-1}) = \left(\frac{\chi_t^{(1)}}{p_{t-1}^{(1)}}, \frac{\chi_t^{(2)}}{p_{t-1}^{(2)}} \right);$$

(c) Обновить дуальный вектор и вектор вероятностей

$$\overline{\zeta}_t = \overline{\zeta}_{t-1} + \overline{u}_t(\overline{p}_{t-1}), \ \ \overline{p}_t = \overline{G}_{\beta_t}(\overline{\zeta}_t);$$

(d) **end for**

Непосредственно проверяется, что

$$\chi_t^{(\ell)} = \sum_{n=(t-1)M+1}^{tM} (1 - \hat{\xi}_n^{(\ell)}),$$

где $\{1 - \hat{\xi}_n^{(\ell)}\}$ – независимые одинаково распределенные случайные величины, характеризуемые распределением

$$\Pr\{(1 - \hat{\xi}_n^{(\ell)}) = 1\} = p_{t-1}^{(\ell)} q_\ell, \ \Pr\{(1 - \hat{\xi}_n^{(\ell)}) = 0\} = 1 - p_{t-1}^{(\ell)} q_\ell,$$

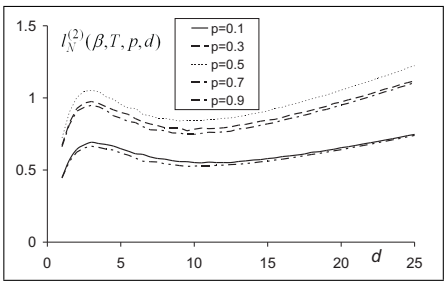

Рисунок 9. Алгоритм **A2**.
$\beta = 1.0;\ \varrho = 0.02;\ M = 100;$
$T = 300;$
$p = 0.1; 0.3; 0.5; 0.7; 0.9.$

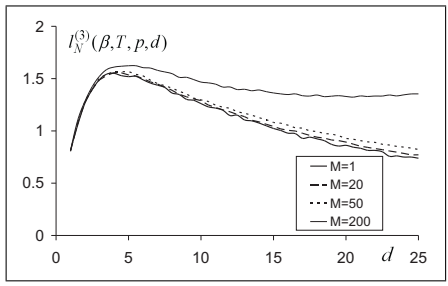

Рисунок 10. Алгоритм **A3**.
$\beta = 2.2;\ N = 10000;$
$p = 0.5;$
$M = 1; 20; 50; 200.$

причем $\{1-\hat{\xi}_n^{(1)}\}$ и $\{1-\hat{\xi}_n^{(2)}\}$ не зависят друг от друга. Следовательно, $\{\chi_t^{(\ell)}\}$ имеют биномиальные распределения, причем

$$\mathbf{E}\chi_t^{(\ell)} = Mp_{t-1}^{(\ell)}q_\ell,\ \mathbf{D}\chi_t^{(\ell)} = Mp_{t-1}^{(\ell)}q_\ell(1-p_{t-1}^{(\ell)}q_\ell),\ \ell = 1,2. \tag{5.1}$$

Рассмотрим функцию потерь $l_N^{(3)}(\beta,T,p,d)$ для алгоритма **A3** с параметром $\beta_t = \beta(DM(t+0.5))^{1/2}$. Справедлива

Теорема 5.1. *Пусть количество этапов пакетной обработки T фиксировано. Тогда при $N = M \times T$ существует предел*

$$l^{(3)}(\beta,T,p,d) = \lim_{M\to\infty} l_N^{(3)}(\beta,T,p,d), \tag{5.2}$$

который зависит от β, T, p и d.

Доказательство. Как и в теореме 4.1 положим

$$\eta_t^{(\ell)} = \left(\chi_t^{(\ell)} - \mathbf{E}(\chi_t^{(\ell)})\right)\left(\mathbf{D}(\chi_t^{(\ell)})\right)^{-1/2}, \tag{5.3}$$

через $f^{(\ell)}(x, I_t^{(\ell)})$ обозначим плотность распределения случайной величины $\eta_t^{(\ell)}$. Плотность $f^{(\ell)}(x, I_t^{(\ell)})$ зависит только от текущего выбранного действия ℓ и объема выборки $I_t^{(\ell)}$ (а, следовательно, $p_t^{(\ell)}$). В силу локальной предельной теоремы Муавра-Лапласа справедливы равенства

$$f^{(\ell)}(x, I_t^{(\ell)}) = \varphi(x)(1+\delta_t^{(\ell)}(x, I_t^{(\ell)})),\quad t = 1,\dots,T,\ \ell = 1,2, \tag{5.4}$$

при $|x| \le A$ и любом $A > 0$, где $\varphi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ – плотность стандартного нормального распределения, а $\delta_t^{(\ell)}(x, I_t^{(\ell)}) \to 0$ равномерно по x, если $I_t^{(\ell)} \to \infty$.

Положим $q_\ell = q + w_\ell(D/N)^{1/2}$, где $w_\ell = (-1)^\ell d$, $\ell = 1, 2$, $d > 0$. Тогда с учетом (5.1), (5.3) имеем

$$\chi_t^{(\ell)} = Mp_{t-1}^{(\ell)}\left(q + w_\ell(D/N)^{1/2}\right) + \left(Mp_{t-1}^{(\ell)}q_\ell(1 - p_{t-1}^{(\ell)}q_\ell)\right)^{1/2}\eta_t^{(\ell)} =$$
$$= qMp_{t-1}^{(\ell)} + \varepsilon w_\ell(DN)^{1/2}p_{t-1}^{(\ell)} + \left(\varepsilon Np_{t-1}^{(\ell)}q_\ell(1 - p_{t-1}^{(\ell)}q_\ell)\right)^{1/2}\eta_t^{(\ell)}.$$

Здесь $\varepsilon = M/N$. Далее, получаем

$$\frac{\chi_t^{(\ell)}}{p_{t-1}^{(\ell)}} = qM + \varepsilon w_\ell(DN)^{1/2} + \left(\frac{\varepsilon Nq_\ell(1 - p_{t-1}^{(\ell)}q_\ell)}{p_{t-1}^{(\ell)}}\right)^{1/2}\eta_t^{(\ell)}.$$

Поэтому

$$\zeta_t^{(\ell)} = qMt + \tau w_\ell(DN)^{1/2} + \sum_{i=1}^{t}\left(\frac{\varepsilon Nq_\ell(1 - p_{i-1}^{(\ell)}q_\ell)}{p_{i-1}^{(\ell)}}\right)^{1/2}\eta_i^{(\ell)}, \qquad (5.5)$$

где $\tau = t/T$. Вспомним, что $\beta_t = \beta(DM(t + 0.5))^{1/2} = \beta(DN(\tau + 0.5\varepsilon))^{1/2}$. Тогда

$$\frac{\zeta_t^{(\ell)}}{\beta_t} = \frac{qMt}{\beta_t} + \frac{1}{\beta}\left(\frac{\tau w_\ell}{(\tau + 0.5\varepsilon)^{1/2}} + X_{t,M}^{(\ell)}\right),$$

где

$$X_{t,M}^{(\ell)} = \sum_{i=1}^{t}\left(\frac{\varepsilon q_\ell(1 - p_{i-1}^{(\ell)}q_\ell)}{D(\tau + 0.5\varepsilon)p_{i-1}^{(\ell)}}\right)^{1/2}\eta_i^{(\ell)},$$

$\ell = 1, 2$. Далее,

$$\frac{\zeta_t^{(1)} - \zeta_t^{(2)}}{\beta_t} = -\frac{2d\tau}{\beta(\tau + 0.5\varepsilon)^{1/2}} + \frac{1}{\beta}\left(X_{t,M}^{(1)} - X_{t,M}^{(2)}\right). \qquad (5.6)$$

При этом

$$p_{t-1}^{(2)} = \frac{1}{\exp\left(-\dfrac{\zeta_t^{(1)} - \zeta_t^{(2)}}{\beta_t}\right) + 1} \qquad (5.7)$$

и $p_{t-1}^{(1)} = 1 - p_{t-1}^{(2)}$, а функция потерь $l_N^{(3)}(\beta, T, p, d)$ имеет вид

$$l_N^{(3)}(\beta, T, p, d) = (DN)^{-1/2}(p_1 - p_2)\sum_{t=1}^{T} M\ \mathbf{E}_{\mathbf{A3}}\left(p_{t-1}^{(2)}\right) = \\ = 2d\sum_{t=1}^{T}\varepsilon\ \mathbf{E}_{\mathbf{A3}}\left(p_{t-1}^{(2)}\right), \tag{5.8}$$

где через $\mathbf{E}_{\mathbf{A3}}$ обозначено математическое ожидание по мере, порожденной алгоритмом **A3**. Поэтому для того, чтобы установить справедливость (5.2) достаточно установить существования предела

$$\lim_{M\to\infty} \mathbf{E}_{\mathbf{A3}}\left(p_t^{(2)}\right) \tag{5.9}$$

при всех $t = 0, \ldots, T-1$. Здесь $\overline{p}_0 = \overline{G}_{\beta_0}(\overline{\zeta}_0)$; в частности, $p_0^{(1)} = p_0^{(2)} = 0.5$ при $\zeta_0^{(1)} = \zeta_0^{(2)} = 0$. При $t \geq 1$ имеем

$$\mathbf{E}_{\mathbf{A3}}\left(p_t^{(2)}\right) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} p_t^{(2)}\left(\overline{x}^t\right) \times \tag{5.10}$$
$$\times\left(\prod_{i=1}^{t} f^{(1)}\left(x_i^{(1)}, I_i^{(1)}(\cdot)\right) f^{(2)}\left(x_i^{(2)}, I_i^{(2)}(\cdot)\right)\right) d\overline{x}^t,$$

где для $\left\{f^{(\ell)}\left(x_i^{(\ell)}, I_i^{(1)}(\cdot)\right)\right\}$ справедливы оценки (5.4). Функции $p_t^{(\ell)}(\overline{x}^t)$ в соответствии с (5.6), (5.7) при $t = 1, \ldots, T-1$ определяются рекуррентно:

$$p_t^{(2)}(\overline{x}^t) = \frac{1}{\exp\left(\dfrac{2d\tau}{\beta(\tau+0.5\varepsilon)^{1/2}} - \dfrac{X_{t,M}^{(1)}(\overline{x}^t) - X_{t,M}^{(2)}(\overline{x}^t)}{\beta}\right) + 1},$$
$$X_{t,M}^{(\ell)}(\overline{x}^t) = \sum_{i=1}^{t}\left(\frac{\varepsilon q_\ell(1 - p_{i-1}^{(\ell)}(\overline{x}^{i-1})q_\ell)}{D(\tau+0.5\varepsilon)p_{i-1}^{(\ell)}(\overline{x}^{i-1})}\right)^{1/2} x_i^{(\ell)}, \quad \ell = 1, 2, \tag{5.11}$$
$$p_t^{(1)}(\overline{x}^t) = 1 - p_t^{(2)}(\overline{x}^t),$$

где $\tau = t/T$, $\varepsilon = 1/T$. Выберем достаточно большое $A > 0$ и рассмотрим множество

$$U_A = \{\overline{x}^{T-1} : |x_t^{(\ell)}| \leq A;\ \ell = 1, 2;\ t = 1, \ldots, T-1\}.$$

Отметим, что все вероятности $\{p_t^{(\ell)}(\overline{x}^t)\}$ непрерывны, как суперпозиции непрерывных функций, и не обращаются в нуль при конечных значениях $\{x_t^{(\ell)}\}$. Так как $q_\ell \to q$ при $\ell = 1, 2$ и $M \to \infty$, то все вероятности $\{p_t^{(\ell)}(\overline{x}^t)\}$ имеют непрерывные пределы при $M \to \infty$, которые также не обращаются в нуль при конечных значениях $\{x_t^{(\ell)}\}$. Поэтому для любого $A > 0$ можно указать такое $\Delta > 0$, что при всех достаточно больших M будут выполнены неравенства

$$p_t^{(\ell)}(\overline{x}^t) \geq \Delta \quad \text{при } \overline{x}^{T-1} \in U_A;\ \ell = 1, 2;\ t = 1, \ldots, T-1. \tag{5.12}$$

Из (5.4) и (5.12) следует, что $\delta_t^{(\ell)}(x_t^{(\ell)}, I_t^{(\ell)}) \to 0$ при $\overline{x}^{T-1} \in U_A$ и $M \to \infty$ равномерно при всех $\ell = 1, 2$ и $t = 1, \ldots, T-1$. При этом интеграл в (5.10), вычисленный по множеству U_A, стремится к его истинному значению при $A \to \infty$. Для вычисления предельного значения интеграла в (5.10) следует положить $q_\ell = q$, $\delta_t^{(\ell)}(x, I_t^{(\ell)}) = 0$ в (5.4), (5.10), (5.11). Отсюда следует существование пределов (5.9) и (5.2), которые зависят от β, T, p и d. □

Замечание 5.1. Можно ожидать сходимости алгоритма **АЗ** при $M \to \infty$, $T \to \infty$. Положим $\varsigma_N^{(\ell)}(\tau) = (DN)^{-1/2}(\zeta_t^{(\ell)} - qMt)$, $P_N^{(\ell)}(\tau) = p_{t-1}^{(\ell)}$, где $\tau = t/T$ ($\ell = 1, 2$). Пусть $\varsigma_N^{(\ell)}(\tau)$ $P_N^{(\ell)}(\tau)$ слабо сходятся к $\varsigma^{(\ell)}(\tau)$, $P^{(\ell)}(\tau)$, а $l^{(3)}(\beta, p, d) = \lim l^{(3)}(\beta, T, p, d)$ при $M \to \infty$, $T \to \infty$. Пусть $W_\ell(\tau)$, $\ell = 1, 2$ обозначают независимые винеровские процессы. С использованием (5.5), (5.7) и (5.8) предельное описание может быть представлено в виде

$$d\varsigma^{(\ell)}(\tau) = w_\ell d\tau + \left(\frac{q(1 - qP^{(\ell)}(\tau))}{DP^{(\ell)}(\tau)}\right)^{1/2} dW_\ell(\tau),\ \ell = 1, 2,$$

$$P^{(2)}(\tau) = \frac{1}{\exp\left(-\dfrac{\varsigma^{(1)}(\tau) - \varsigma^{(2)}(\tau)}{\beta\tau^{1/2}}\right) + 1},$$

где $P^{(1)}(\tau) = 1 - P^{(2)}(\tau)$, $\tau \in [0, 1]$. Начальные условия имеют вид

$$\varsigma^{(1)}(0) = \varsigma^{(2)}(0) = 0.$$

Функция потерь равна

$$l^{(3)}(\beta, p, d) = 2d \int_0^1 \mathbf{E}\left(P^{(2)}(\tau)\right) d\tau.$$

Однако строгого доказательства этого результата пока нет.

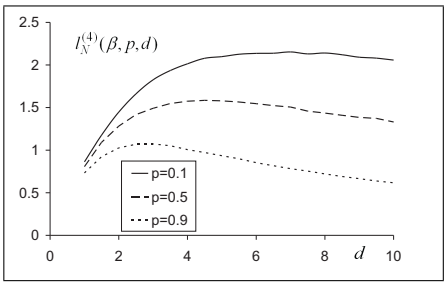

Рисунок 11. Алгоритм **A4**. $\beta = 2.2$; $N = 20000$; $M_0 = 600$; $M = 200$; $p = 0.1; 0.5; 0.9$.

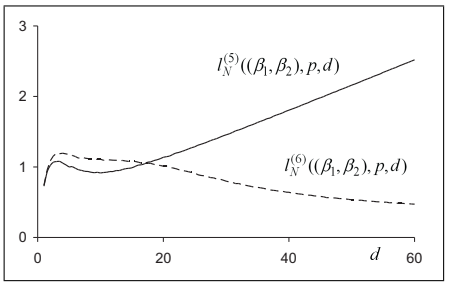

Рисунок 12. Алгоритмы **A5** и **A6**. $\beta_1 = 2.2$; $\beta_2 = 1.0$; $N = 30000$; $p = 0.5$; $M_0 = 900$; $M = 300$, $\varrho = 0.02$, $\kappa = 0.2$.

На рис. 10 представлена функция $l_N^{(3)}(\beta, T, p, d)$, вычисленная для различных M с помощью симуляций Монте-Карло, если $\beta = 2.2$, $p = 0.5$, $N = 10000$ и $1 \le d \le 25$. Результаты представлены для $M = 20, 50, 200$ (соответственно, $T = 500, 200, 50$). Случай $M = 1$ соответствует обычному АЗС и $l_N^{(1)}(\beta, p, d)$. Можно видеть, что $l_N^{(3)}(\beta, T, p, d)$ близка к $l_N^{(1)}(\beta, p, d)$, если T достаточно велико.

Замечание 5.2. Если количество пакетов T велико, а вероятности $\{p_{n-1}^{(\ell)}, \ell = 1, 2\}$, формируемые алгоритмом **A1**, медленно меняются при $n = (t-1)M + 1, \ldots, tM$ и близки к вероятностям $p_{t-1}^{(\ell)}$, $\ell = 1, 2$, формируемым алгоритмом **A3**, то можно ожидать, что **A1** и **A3** будут демонстрировать близкое поведение. Именно такое поведение демонстрирует рис. 10.

6. Комбинированные алгоритмы

Вернемся к рис. 6 и 10. Можно видеть, что для достаточно больших значений d с ростом размера пакета растет и функция потерь. Это является следствием равного применения обоих действий при обработке начального пакета. Чтобы избежать этого эффекта, можно выбирать начальные пакеты меньшего размера. Самым простым решением будет использовать обычный АЗС на коротком начальном этапе, а затем переключиться на пакетную версию. Рассмотрим сна-

чала комбинацию алгоритмов **A1** и **A3**.

Алгоритм A4.

---

1. Применить алгоритм **A1** на начальном горизонте управления $n = 1, \ldots, M_0$. Получить $\overline{\zeta}_{M_0}$ и $\overline{p}_{M_0}$.

2. Применить алгоритм **A3** на заключительном горизонте управления $n = M_0 + 1, \ldots, N$ с начальными данными $\overline{\zeta}_{M_0}$ и $\overline{p}_{M_0}$.

На рис. 11 представлены значения функции $l_N^{(4)}(\beta, p, d)$, вычисленные с помощью симуляций Монте-Карло для алгоритма **A4**, если $\beta = 2.2$; $N = 20000$; $M_0 = 600$; $M = 200$; $p = 0.1; 0.5; 0.9$ (здесь и ниже зависимость функции потерь от числа пакетов опущена). Можно видеть, что эти результаты близки к представленным на рис. 4 для обычного АЗС.

Поскольку алгоритм **A2** обеспечивает лучшую скорость сходимости, чем алгоритм **A3**, рассмотрим следующую комбинацию алгоритмов **A1** и **A2**.

Алгоритм A5.

---

1. Применить алгоритм **A1** на начальном горизонте управления $n = 1, \ldots, M_0$ с параметром $\beta = \beta_1$. Получить $\overline{\zeta}_{M_0}$ и $\overline{p}_{M_0}$.

2. Применить алгоритм **A2** на заключительном горизонте управления $n = M_0 + 1, \ldots, N$ с параметром $\beta = \beta_2$ с начальными данными $\overline{\zeta}_{M_0}$ и $\overline{p}_{M_0}$.

Однако применение алгоритма **A2** приводит к большим потерям, если d достаточно велико (см. рис. 6), поскольку он применяет оба действия с вероятностями не меньшими, чем ϱ. Поэтому рассмотрим следующий комбинированный алгоритм.

Алгоритм A6.

---

1. Применить алгоритм **A1** на начальном горизонте управления $n = 1, \ldots, M_0$ с параметром $\beta = \beta_1$. Получить $\overline{\zeta}_{M_0}$ и $\overline{p}_{M_0}$.

2. Если $\min(p^{(1)}_{M_0}, p^{(2)}_{M_0}) < \kappa$, то применить алгоритм **A3** с параметром $\beta = \beta_1$, в противном случае применить алгоритм **A2** с параметром $\beta = \beta_2$ на заключительном горизонте управления $n = M_0 + 1, \ldots, N$ с начальными данными $\overline{\zeta}_{M_0}$ и $\overline{p}_{M_0}$.

Если величину κ выбрать подходящим образом, то алгоритм **A6** при малых d переключается главным образом на алгоритм **A2**. При больших d он переключается главным образом на алгоритм **A3**. На рис. 12 представлены сравнительные результаты для $l^{(5)}_N((\beta_1, \beta_2), p, d)$ и $l^{(6)}_N((\beta_1, \beta_2), p, d)$, если $\beta_1 = 2.2$; $\beta_2 = 1.0$; $N = 30000$; $p = 0.5$; $M_0 = 900$; $M = 300$, $\varrho = 0.02$, $\kappa = 0.2$. Можно видеть, что $l^{(6)}_N((\beta_1, \beta_2), p, d)$ не растет при больших d.

7. Заключение

Предложены пакетные версии алгоритма зеркального спуска для задачи о двуруком бандите в приложении к обработке данных. Использование пакетных версий АЗС означает, что полное время обработки данных зависит от числа пакетов, на которые они разбиты, а не от полного числа данных. Моделирование Монте-Карло показывает, что максимальные ожидаемые потери для пакетных версий не больше, чем потери для обычной версии, обеспечивающей обработку данных по одному. Однако это верно только для двуруких бандитов, характеризуемых близкими математическими ожиданиями одношаговых доходов. Если математические ожидания значительно различаются, то возникают значительные потери, вызванные равным применением обоих действий при обработке первого пакета. Этих потерь можно избежать, если на начальном достаточно коротком этапе применять обычный АЗС, а затем переключаться на пакетную версию.

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. Гасников А.В., Нестеров Ю.Е., Спокойный В.Г. *Об эффективности одного метода рандомизации зеркального спуска в задачах онлайн оптимизации* // Журнал вычислительной математики и математической физики. 2015. Т. 55. № 4. С. 582–598.

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

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

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

7. Немировский А.С., Юдин Д.Б. *Эффективные методы решения задач выпуклого программирования большой размерности* // Экономика и математические методы. 1979. Т. 15. № 1. С. 135–152.

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

9. 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.

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

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

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

13. Auer P., Cesa-Bianchi N., Fischer P. *Finite-time analysis of the multi-armed bandit problem* // Machine learning. 2002. V. 47. N 2–3. P. 235–256.

14. 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.

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

16. Fabius J., van Zwet W.R. *Some Remarks on the Two-Armed Bandit* // Ann. Math. Stat. 1970. V. 41. P. 1906–1916.

17. Juditsky A., Nazin A.V., Tsybakov A.B., Vayatis N. *Gap-Free Bounds for Stochastic Multi-Armed Bandit* // Proc. 17th World Congress IFAC (Seoul, Korea, July 6–11). 2008. P. 11560–11563.

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

19. Lai T.L., Levin B., Robbins H., Siegmund D. *Sequential medical trials (stopping rules/asymptotic optimality)* // Proc. Nati. Acad. Sci. USA. 1980. V. 77. N 6. P. 3135–3138.

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

21. Lai T.L., Robbins H. *Asymptotically Efficient Adaptive Allocation Rules* // Advances in Applied Mathematics. 1985. V. 6. P. 4–22.

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

23. Robbins H. *Some Aspects of the Sequential Design of Experiments* // Bulletin AMS. 1952. V. 58. N 5. P. 527–535.

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

TWO-ARMED BANDIT PROBLEM AND BATCH VERSION OF THE MIRROR DESCENT ALGORITHM

Alexander V. Kolnogorov, Yaroslav-the-Wise Novgorod State University, Dr.Sc., professor (Alexander.Kolnogorov@novsu.ru),
Alexander V. Nazin, V.A.Trapeznikov Institute of Control Sciences Russian Academy of Sciences, Dr.Sc., professor (nazine@ipu.ru),
Dmitry N. Shiyan, Yaroslav-the-Wise Novgorod State University, postgraduate (dsqq-tm@ya.ru).

Abstract: We consider the minimax setup for the two-armed bandit problem as applied to data processing if there are two alternative processing methods with different a priori unknown efficiencies. One should determine the most efficient method and provide its predominant application. To this end, we use the mirror descent algorithm (MDA). It is well-known that corresponding minimax risk has the order of $N^{1/2}$ with N being the number of processed data and this bound is unimprovable in order. We propose a batch version of the MDA which allows processing data by packets that is especially important if parallel data processing can be provided. In this case, the processing time is determined by the number of batches rather than by the total number of data. Unexpectedly, it turned out that the batch version behaves unlike the ordinary one even if the number of packets is large. Moreover, the batch version provides significantly smaller value of the minimax risk, i.e., it considerably improves a control performance. We explain this result by considering another batch modification of the MDA which behavior is close to behavior of the ordinary version and minimax risk is close as well. Our estimates use invariant descriptions of the algorithms based on Gaussian approximations of incomes in batches of data in the domain of "close" distributions and are obtained by Monte-Carlo simulations.

Keywords: two-armed bandit problem, minimax approach, mirror descent algorithm, EXP3, batch processing.