УДК 004.048

# МАШИННОЕ ОБУЧЕНИЕ, ЗАДАЧА О МНОГОРУКОМ БАНДИТЕ И ПАКЕТНОЕ ПРАВИЛО UCB

**А.В. Колногоров**

*Новгородский государственный университет им. Ярослава Мудрого*

Россия, 173003, Великий Новгород, Большая Санкт-Петербургская ул., 41

E-mail: Alexander.Kolnogorov@novsu.ru

**С.В. Гарбарь**

*Новгородский государственный университет им. Ярослава Мудрого*

Россия, 173003, Великий Новгород, Большая Санкт-Петербургская ул., 41

E-mail: Sergey.Garbar@novsu.ru

**Ключевые слова:** задача о многоруком бандите, правило UCB, инвариантное описание, близкие распределения, пакетная обработка.

**Аннотация:** Получена оценка максимальных значений функции потерь для некоторой стратегии в задаче о многоруком бандите с гауссовскими распределениями доходов. Рассматриваемая стратегия является асимптотическим обобщением стратегии Дж. Басера (J. Bather), предложенной для задачи о многоруком бандите и использующей правило UCB, т.е. выбора текущего действия, соответствующего максимуму верхней границы доверительного интервала оценки среднего значения одношагового дохода. Результаты получены с помощью инвариантного описания управления на единичном горизонте в области близких распределений, так как именно в области близких распределений функция потерь достигает максимальных значений. Правило UCB широко применяется в машинном обучении. Оно может использоваться для оптимизации пакетной обработки данных, если для обработки имеются два альтернативных метода с различными априори неизвестными эффективностями.

## 1. Введение

Рассматривается задача о многоруком бандите, т.е. об игральном автомате с двумя или более рукоятками, выбор каждой из которых сопровождается случайным доходом игрока, который зависит только от выбранной рукоятки [1]. Цель игрока состоит в максимизации математического ожидания полного дохода. Для этого в процессе игры он должен определить рукоятку, которой соответствует больший ожидаемый доход, и обеспечить ее преимущественный выбор. Проблема известна также как задача о целесообразном поведении [2, 3] и об адаптивном управлении в случайной среде [4, 5]. Задача имеет многочисленные применения в машинном обучении [6, 7].

Ниже рассматривается гауссовский многорукий бандит, который возникает при пакетной обработке данных, если для обработки имеются два альтернативных метода

с различными априори неизвестными эффективностями [8]. Формально это управляемый случайный процесс $\xi_n$, $n = 1, 2, \ldots, N$, значение которого в момент времени $n$ зависит только от текущего выбранного действия $y_n$, интерпретируется как получаемый доход и имеет гауссовское (нормальное) распределение с плотностью $f_D(x|m_\ell) = (2\pi D)^{-1/2} \exp\left(-(x - m_\ell)^2/(2D)\right)$, если $y_n = \ell$, $\ell = 1, \ldots, J$. Дисперсия $D$ предполагается известной, а математические ожидания $m_1, \ldots, m_J$ неизвестными. Требование известности дисперсии можно снять, так как рассматриваемый алгоритм мало чувствителен к значительному изменению дисперсии (например, 5–10%), поэтому ее можно оценить на начальном этапе управления.

Стратегия управления $\sigma$ обеспечивает, вообще говоря, рандомизированный выбор действия $y_n$ на основе имеющейся текущей информации о предыстории процесса. Ниже ограничимся следующей стратегией, предложенной в [9]. Пусть к моменту времени $n$ действие с номером $\ell$ было применено $n_\ell$, раз и пусть $X_\ell$ – соответствующий полный доход ($\ell = 1, \ldots, J$). В этом случае $X_\ell/n_\ell$ является текущей оценкой математического ожидания $m_\ell$. Поскольку целью является максимизация полного дохода, то кажется естественным применять всегда то действие, которому соответствует текущая бóльшая величина $X_\ell/n_\ell$. Однако хорошо известно, что это правило может давать большие потери, так как чисто случайно начальная оценка $X_\ell/n_\ell$, соответствующая наибольшему $m_\ell$, может получить низкое значение и в дальнейшем это действие не будет применяться никогда. Вместо оценок самих $m_\ell$ рассмотрим верхние границы доверительных интервалов этих оценок

$$U_\ell(n) = \frac{X_\ell(n)}{n_\ell} + \frac{aD^{1/2}}{n_\ell^{1/2}}(2 + \zeta_\ell(n)), \tag{1}$$

где $a > 0$, $\{\zeta_l(n)\}$ – независимые одинаково распределенные случайные величины с плотностью $e^{-x}$ при $x > 0$; $\ell = 1, 2, \ldots, J$, $n = 1, 2, \ldots, N$.

Рассматриваемая ниже стратегия предписывает сначала применить все действия по одному разу, а затем в каждый момент времени $n$ выбирать то действие, которому соответствует бóльшая текущая величина $\{U_\ell(n)\}$. Стратегии такого вида называются правилами UCB (Upper Confidence Bound). Данная стратегия при $a = 2/15$ эквивалентна (с точностью до слагаемых порядка $n_\ell^{-1}$) стратегии, предложенной в [9] для бернуллиевского многорукого бандита. В этом случае надо положить $D = 0{,}25$ – максимальное значение дисперсии бернуллиевского одношагового дохода. В [9] отмечается, что при $J = 2$ максимальные приведенные потери (к величине $(DN)^{1/2}$) не превосходят в этом случае 0,72 при больших $N$, однако объяснения этого результата не приведено и, насколько нам известно, в последующем оно также не публиковалось.

Ниже приводится обоснование этого результата, которое получено с помощью инвариантного описания на единичном горизонте управления в области близких распределений, на которых достигаются максимальные потери для задачи о многоруком бандите. Более того, мы рассматриваем пакетную версию стратегии [9] и показываем, что потери определяются только числом обрабатываемых пакетов и инвариантными характеристиками параметра. Отметим, что пакетные (параллельные) стратегии особенно важны там, где время обработки единицы данных является значительным, так как в этом случае полное время обработки определяется количеством пакетов, а не полным числом данных. При этом для пакетного правила максимальные приведенные потери оказались приблизительно равны 0,75, т.е. почти не отличаются от указанных в [9]. Отметим также, что различные модификации правила UCB уже

широко применяются в машинном обучении (см., например, [6, 7]).

## 2. Основные результаты

Рассматриваемый многорукий бандит может быть описан векторным параметром $\theta = (m_1, \ldots, m_J)$. Определим функцию потерь. Если бы параметр был известен, то следовало бы всегда применять действие, соответствующее максимальному значению из $m_1, \ldots, m_J$, при этом математическое ожидание полного дохода равно $N \max(m_1, \ldots, m_J)$. Для реально применяемой стратегии $\sigma$ математическое ожидание полного дохода меньше максимально возможного на величину, называемую функцией потерь и равную

$$L_N(\sigma, \theta) = \mathbb{E}_{\sigma,\theta} \left( \sum_{n=1}^{N} (\max(m_1, \ldots, m_J) - \xi_n) \right). \tag{2}$$

Здесь $\mathbb{E}_{\sigma,\theta}$ обозначает математическое ожидание, вычисленное по мере, порождаемой стратегией $\sigma$ и параметром $\theta$. Нас интересует оценка максимальных потерь, вычисляемых на множестве допустимых значений параметра, которое выберем следующим

$$\Theta = \{m_\ell = m + d_\ell (D/N)^{1/2}; m \in (-\infty, +\infty), |d_\ell| \leqslant C < \infty, \ell = 1, \ldots, J\}.$$

Рассматриваемое множество параметров описывает «близкие» распределения, которые характеризуются тем, что математические ожидания одношаговых доходов различаются на величину порядка $N^{-1/2}$. Именно в этой области достигаются максимальные потери, которые имеют порядок $N^{1/2}$ (см. [10]). Для «удаленных» распределений потери меньше. Например, они имеют порядок $\ln(N)$, если известно, что $\max(m_1, \ldots, m_J)$ отстоит от остальных $\{m_\ell\}$ не меньше, чем на некоторое $\delta > 0$ (см. [11]).

Рассмотрим стратегии управления, которые меняют действия только после применения $M$ раз подряд. Такие стратегии допускают пакетную, в том числе параллельную обработку. Будем считать, что $N = MK$, где $K$ – число пакетов. Для пакетных стратегий верхние границы (1) примут вид

$$U_\ell(k) = \frac{X_\ell(k)}{k_\ell} + \frac{a(MD)^{1/2}}{k_\ell^{1/2}}(2 + \zeta_\ell(k)), \tag{3}$$

где $k$ – номер пакета, а $k_\ell$, $X_\ell(k)$ характеризуют здесь число пакетов, к которым применено действие $\ell$ и соответствующий полный доход после обработки $k$ пакетов $(k = 1, 2, \ldots, K)$. Обозначим через

$$I_\ell(k) = \begin{cases} 1, & \text{если } U_\ell(k) = \max(U_1(k), \ldots, U_J(k)), \\ 0 & \text{в противном случае} \end{cases}$$

– индикатор выбираемого действия для обработки $(k+1)$-го пакета в соответствии с рассматриваемым правилом при $k > J$. Напомним, что при $k \leqslant J$ действия применяются по очереди. Отметим, что с вероятностью 1 только одна из величин $\{I_\ell(k)\}$ равна 1. Для рассматриваемого параметра $\theta$ справедливо представление

$$X_\ell(k) = k_\ell M \left( m + d_\ell \left( \frac{D}{N} \right)^{1/2} \right) + \sum_{i=1}^{k} I_\ell(i).Y_\ell(MD; i), \tag{4}$$

где $\{Y_\ell(MD;i)\}$ независимые нормально распределенные случайные величины с нулевыми математическими ожиданиями и дисперсиями равными $MD$. Введем следующие переменные: $t = kK^{-1}$, $t_\ell = k_\ell K^{-1}$, $\varepsilon = K^{-1}$. Из (3), (4) следует, что

$$U_\ell(k) = Mm + d_l\left(\frac{MD}{K}\right)^{1/2} + \\ + \frac{(MD)^{1/2}\sum_{i=1}^{k} I_\ell(i)Y_\ell(\varepsilon;i)}{t_\ell K^{1/2}} + \frac{a(MD)^{1/2}}{t_\ell^{1/2}K^{1/2}}(2+\zeta_\ell(n)),$$

$\ell = 1,2,\ldots,J$, $k = J+1, J+2,\ldots,K$. После линейного преобразования $u_\ell(t) = (U_\ell(k) - Mm)(MD)^{-1/2}K^{1/2}$, которое не меняет порядок указанных величин, получим верхние границы в инвариантной форме, с горизонтом управления равным 1:

$$u_\ell(t) = d_l + \frac{\sum_{i=1}^{k} I_\ell(i)Y_\ell(\varepsilon;i)}{t_\ell} + \frac{a}{t_\ell^{1/2}}(2+\zeta_\ell(t)), \tag{5}$$

$\ell = 1,2,\ldots,J$; $t = (J+1)\varepsilon, (J+2)\varepsilon,\ldots,1$.

Найдем функцию потерь. Для выбранного параметра без ограничения общности считаем, что $d_1 = \max(d_1,\ldots,d_J)$. Тогда

$$L_N(\sigma,\theta) = (D/N)^{1/2}\sum_{\ell=2}^{J}(d_1-d_\ell)\mathbb{E}_{\sigma,\theta}\left(\sum_{k=1}^{K} MI_\ell(k)\right) = \\ = (DN)^{1/2}\sum_{\ell=2}^{K}(d_1-d_\ell)\mathbb{E}_{\sigma,\theta}\left(\sum_{k=1}^{K}\varepsilon I_\ell(k)\right),$$

откуда для приведенной (к величине $(DN)^{1/2}$) функции потерь имеем выражение

$$(DN)^{-1/2}L_N(\sigma,\theta) = \sum_{\ell=2}^{K}(d_1-d_\ell)\mathbb{E}_{\sigma,\theta}\left(\sum_{k=1}^{K}\varepsilon I_\ell(k)\right), \tag{6}$$

Результаты можно сформулировать в виде теоремы.

**Теорема 1.** *При использовании правила (3), справедливо инвариантное описание на единичном горизонте управления, которое дается формулой (5). Для приведенной (к величине $(DN)^{1/2}$) функции потерь справедливо выражение (6). Выражения (5), (6) определяются только числом обрабатываемых пакетов.*

На рисунке рис. 1 приведены результаты моделирования методом Монте-Карло приведенной функции потерь, соответствующей применению правила UCB при $a = 1/3$ и $K = 2$. Усреднение выполнялось по 10000 симуляциям. При $K = 2$ можно взять $d_1 = -d_2 = 0{,}5d$, значения приведенной функции потерь обозначены $l(d)$. Синяя, красная и зеленая линии получены при значениях горизонта управления $N = 100, 400, 1500$. При этом $\max l(d) \approx 0{,}75$ при $d \approx 3{,}5$. Таким образом, максимальное значение приведенной функции потерь близко к найденному в [9]. А вот значение $a = 1/3$, которое в нашем случае близко к оптимальному, т.е. обеспечивающему минимум максимальных потерь, существенно отличается от $2/15$, найденных в [9]. Последнее может объясняться тем, что в [9] рассматривались сравнительно небольшие горизонты управления, например, $N = 50$. В этом случае добавки порядка $n_\ell^{-1}$, которые присутствуют в правиле, предложенном в [9], существенно влияют на значения $\{U_\ell(n)\}$.

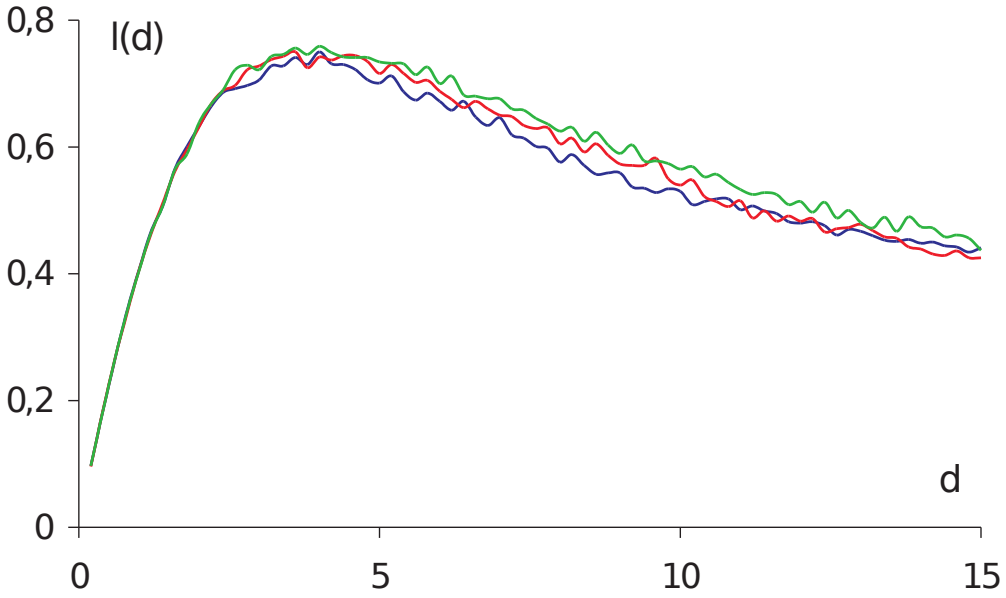


Рис. 1. Приведенные потери

## 3. Заключение

Работа выполнена при поддержке Российского фонда фундаментальных исследований (18-29-16223-мк – первый соавтор), а также Российского фонда фундаментальных исследований и Новгородской области (18-41-530001 – второй соавтор).

## Список литературы

1. Berry D.A., Fristedt B. Bandit Problems: Sequential Allocation of Experiments. London, New York: Chapman and Hall, 1985. 275 p.
2. Цетлин М.Л. Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969. 316 с.
3. Варшавский В.И. Коллективное поведение автоматов. М.: Наука, 1973. 408 с.
4. Срагович В.Г. Адаптивное управление. М.: Наука, 1981. 384 с.
5. Назин А.В., Позняк А.С. Адаптивный выбор вариантов. М.: Наука, 1986. 288 с.
6. Auer P. Using Confidence Bounds for Exploitation-Exploration Trade-offs // Journal of Machine Learning Research. 2002. Vol. 3. P. 397-422.
7. Lugosi G., Cesa-Bianchi N. Prediction, Learning and Games. New York: Cambridge University Press, 2006. 394 p.
8. Колногоров А.В. Робастное параллельное управление в случайной среде (задаче о двуруком бандите) // Автоматика и телемеханика. 2012. № 4. С. 114-130.
9. Bather J.A. The Minimax Risk for the Two-Armed Bandit Problem // Mathematical Learning Models — Theory and Algorithms. Lecture Notes in Statistics. New York: Springer-Verlag, 1983. Vol. 20. P. 1-11.
10. Vogel W. An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem // Ann. Math. Statist. 1960. Vol. 31. P. 444-451.
11. Lai T.L., Levin B., Robbins H., Siegmund D. Sequential Medical Trials (Stopping Rules/Asymptotic Optimality) // Proc. Nati. Acad. Sci. USA. 1980. Vol. 77, No. 6. P. 3135-3138.