УДК 519.832, 519.245

СТРАТЕГИИ UCB В ЗАДАЧЕ О ГАУССОВСКОМ ДВУРУКОМ БАНДИТЕ

А.В. Колногоров

Новгородский государственный университет имени Ярослава Мудрого

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

E-mail: Alexander.Kolnogorov@novsu.ru

Ключевые слова: гауссовский двурукий бандит, минимаксный подход, правило UCB, инвариантное описание, моделирование Монте-Карло.

Аннотация: Рассматривается оптимизация пакетной обработке данных, если для обработки имеются два альтернативных метода с фиксированными, но априори неизвестными эффективностями. В процессе обработки требуется определить более эффективный метод и обеспечить его преимущественное применение. Данная проблема рассматривается в рамках задачи о гауссовском двуруком бандите, одношаговые доходы которого характеризуются неизвестными математическими ожиданиями и дисперсиями. Для управления используются простые в реализации стратегии UCB. Для этих стратегий справедливо инвариантное описание управления, зависящее только от количества пакетов, но не от полного числа обрабатываемых данных. Численные эксперименты показывают высокую эффективность предложенных стратегий.

1. Введение

Рассматривается задача о двуруком бандите [1–3], математической моделью которой является игральный автомат с двумя рукоятками (в дальнейшем именуемыми действиями), выбор каждой из которых сопровождается случайным доходом игрока. Распределения доходов фиксированы в процессе игры, но неизвестны игроку. Количество игр предполагается известным. Требуется, наблюдая статистику игры, определить действие, которому соответствует большее математическое ожидание одношагового дохода и обеспечить его преимущественное применение. Задача имеет приложения в моделировании поведения [4], адаптивного управления в случайной среде [5, 6], медицине, интернет-технологиях, обработке данных [3, 7, 8].

В данной статье задача рассматривается в приложении к пакетной обработке данных, если для обработки имеются два альтернативных метода, эффективности которых фиксированы, но априори неизвестны. Под эффективностью метода можно понимать, например, математическое ожидание количества успешно обработанных данных пакета. В этом случае методы обработки соответствуют действиям, а количество успешно обработанных данных – доходам. Если размеры пакетов достаточно велики, то эти доходы имеют приблизительно нормальное (гауссовское)

распределение; ниже удобно считать, что эти распределения являются в точности нормальными.

Пусть полное число обрабатываемых данных равно $N = MK$, где M– количество данных в пакете, K – количество пакетов. Формально рассматриваемый ниже гауссовский двурукий бандит – это управляемый случайный процесс ξ_k, $k = 1, 2, \ldots, K$, значения которого интерпретируются как доходы, зависят только от текущих выбираемых действий (методов обработки) y_k и имеют нормальную плотность распределения $f_{\tilde{D}_\ell}(x|\tilde{m}_\ell) = (2\pi\tilde{D}_\ell)^{-1/2}\exp\left(-(x-\tilde{m}_\ell)^2/(2\tilde{D}_\ell)\right)$, если для управления выбрано действие $y_k = \ell$, $\ell = 1, 2$. Здесь $\tilde{m}_\ell = Mm_\ell$, $\tilde{D}_\ell = MD_\ell$, где m_ℓ, D_ℓ – математическое ожидание и дисперсия дохода за обработку единицы данных с использованием ℓ-го действия. Такой двурукий бандит описывается параметром $\theta = (m_1, D_1, m_2, D_2)$. Значение параметра фиксировано в процессе управления, но неизвестно. Известным предполагаем допустимое множество параметров Θ, которое будет определено ниже.

Стратегия управления σ при обработке пакета с номером $k+1$ осуществляет выбор действия в зависимости от текущей предыстории процесса, которая в данном случае полностью описывается достаточными статистиками, включающими полные доходы и s^2-статистики за применение обоих действий. Функция потерь

$$L_N(\sigma, \theta) = N\max(m_1, m_2) - \mathbf{E}_{\sigma,\theta}\left(\sum_{k=1}^{K}\xi_k\right) \tag{1}$$

характеризует математическое ожидание потерь полного дохода относительно максимально возможной величины вследствие неполноты информации. Здесь $\mathbf{E}_{\sigma,\theta}$ – знак математического ожидания по вероятностной мере, соответствующей выбранным σ и θ. По функции потерь определяется минимаксный риск

$$R_N(\Theta) = \inf_{\{\sigma\}}\sup_{\Theta} L_N(\sigma, \theta). \tag{2}$$

В [9] установлена асимптотическая оценка минимаксного риска, имеющая порядок $N^{1/2}$. Пакетная обработка в задаче о многоруком бандите первоначально была предложена для лечения пациентов альтернативными лекарствами с различными эффективностями; хороший обзор, библиография и новые результаты имеются в [7]. В приложении к пакетной обработке данных задача о двуруком бандите рассматривалась в [8]. Важное свойство пакетной обработки состоит в том, что она практически не увеличивает минимаксный риск в сравнении с оптимальной обработкой по-одному, если количество данных и количество пакетов, на которые они разбиты, достаточно велики. Например, пакетная обработка данных, разбитых на 50 пакетов приводит к росту минимаксного риска только на 2% в сравнении с оптимальной обработкой по-одному.

Нахождение минимаксных стратегий и риска является трудоемким процессом. В соответствии с основной теоремой теории игр они ищутся как байесовские, вычисленные относительно наихудшего априорного распределения (на котором байесовский риск достигает максимума), и требуют решения уравнения динамического программирования. Поэтому возникает необходимость в простых стратегиях, которые обеспечивали бы величину максимальных потерь близкую к значению минимаксного риска. Именно такими являются рассматриваемые ниже стратегии UCB (Upper Confidence Bound – верхняя граница доверительного интервала).

2. Стратегии UCB

Ранее использование стратегий UCB рассматривалось для однорукого бандита, т.е для двурукого бандита, с известными параметрами одношагового дохода за применение первого действия [10]. Далее мы рассматриваем обобщение этих результатов на случай, если неизвестными являются характеристики одношаговых доходов обоих действий. Обозначим через $x_{\ell i}$ доход за i-е применение ℓ-го действия к пакету данных. При этом возможна ситуация, когда эти пакеты сами разбиваются на $M_2 \geqslant 2$ равных малых пакетов объема M_1, так что $M = M_1 M_2$. Если разбиения на малые пакеты нет, то считаем, что $M_2 = 1$, $M_1 = M$. Обозначим доход j-го малого пакета $x'_{\ell ij}$, тогда $x_{\ell i} = \sum_{j=1}^{M_2} x'_{\ell ij}$. Если $M_2 > 1$, то при обработке малых пакетов можно вычислить s^2-статистику для большого пакета по формуле $s_{\ell i} = \sum_{j=1}^{M_2}(x'_{\ell ij} - x_{\ell i}/M_2)^2 = \sum_{j=1}^{M_2}(x'_{\ell ij})^2 - x_{\ell i}^2/M_2$. Пусть после обработки k пакетов первое и второе действия были применены k_1 и k_2 раз. Тогда достаточными статистиками являются $X_\ell(k_\ell) = \sum_{i=1}^{k_\ell} x_{\ell i}$, $S_\ell(k_\ell) = \sum_{i=1}^{k_\ell}\sum_{j=1}^{M_2}(x'_{\ell ij} - X_\ell(k_\ell)/(M_2 k_\ell))^2 = \sum_{i=1}^{k_\ell}\sum_{j=1}^{M_2}(x'_{\ell ij})^2 - X_\ell^2(k_\ell)/(M_2 k_\ell)$, $\ell = 1, 2$. В [10] показано, что достаточные статистики можно пересчитывать рекуррентно. В этом случае $X_\ell(0) = 0$ и далее $X_\ell(k_\ell + 1) = X_\ell(k_\ell) + x_{\ell,k_\ell+1}$, $k_\ell \geqslant 0$. Пересчет $S_\ell(k_\ell)$ зависит от M_2. При $M_2 = 1$ выполнено $S_\ell(0) = S_\ell(1) = 0$ и далее $S_\ell(k_\ell + 1) = S_\ell(k_\ell) + \left(k_\ell x_{\ell,k_\ell+1} - X_\ell(k_\ell)\right)^2/(k_\ell(k_\ell + 1))$, $k_\ell \geqslant 1$. При $M_2 > 1$ выполнено $S_\ell(0) = 0$, $S_\ell(1) = s_{\ell 1}$ и далее $S_\ell(k_\ell + 1) = S_\ell(k_\ell) + s_{\ell,k_\ell+1} + \left(k_\ell x_{\ell,k_\ell+1} - X_\ell(k_\ell)\right)^2/(M_2 k_\ell(k_\ell + 1))$, $k_\ell \geqslant 1$. По достаточным статистикам делаются текущие оценки математического ожидания и дисперсии дохода за обработку одного пакета данных по формулам $\hat{m}_\ell = X_\ell(k_\ell)/k_\ell$, $\hat{D}_\ell = M_2 S_\ell(k_\ell)/(k_\ell M_2 - 1)$, $\ell = 1, 2$.

Стратегия UCB в начале управления для набора статистики каждое действие применяет равное количество k_0 раз, причем $k_0 \geqslant 2$ при $M_2 = 1$ и $k_0 \geqslant 1$ при $M_2 > 1$. Далее на шаге с номером $k + 1$, где $k = k_1 + k_2$, выбирается действие, которому соответствует бóльшая из величин

$$Q_\ell(k_\ell) = \hat{m}_\ell + B_\ell(\hat{\gamma}_1, \hat{\gamma}_2, k_\ell)\left(\hat{D}_\ell/k_\ell\right)^{1/2}, \quad \ell = 1, 2, \tag{3}$$

в случае их равенства для определенности выбираем первое действие. Легко видеть, что $Q_1(k_1)$, $Q_2(k_2)$ являются верхними границами доверительных интервалов для оценок $\hat{m}_1$, $\hat{m}_2$. Здесь $B_\ell(\hat{\gamma}_1, \hat{\gamma}_2, k_\ell)$ зависит от стратегии. Для пакетного аналога стратегии Басера $B_\ell^B(\hat{\gamma}_1, \hat{\gamma}_2, k_\ell) = a_\ell^B(\hat{\gamma}_1, \hat{\gamma}_2)(2 + \zeta_\ell(k))$, где $\zeta_\ell(k)$ – последовательность независимых одинаково распределенных случайных величин с экспоненциальной плотностью распределения равной e^{-x} при $x \geqslant 0$ и 0 при $x < 0$. Для пакетного аналога стратегии Ауера-Сеза-Бьянки-Фишера (АСБФ) $B_\ell^{ACBF}(\hat{\gamma}_1, \hat{\gamma}_2, k_\ell) = a_\ell^{ACBF}(\hat{\gamma}_1, \hat{\gamma}_2)\ln^{1/2}(k)$. Такие $Q_1(k_1)$, $Q_2(k_2)$ не только стимулируют выбор действия, которому соответствует большее из значений $\hat{m}_1$, $\hat{m}_2$, но и гарантируют достаточно большое количество применений обоих действий k_1, k_2.

Положительные функции $a_\ell^B(\hat{\gamma}_1, \hat{\gamma}_2)$, $a_\ell^{ACBF}(\hat{\gamma}_1, \hat{\gamma}_2)$ выбираются так, чтобы минимизировать максимальные значения функции потерь (1). В случае известных параметров первого действия, рассмотренном в [10], эти функции были положительными числовыми константами; в нашем случае они зависят от отношения оценок дисперсий. Положим $D = \max(D_1, D_2)$, $\gamma_1 = (D_1/D)^{1/2}$, $\gamma_2 = (D_2/D)^{1/2}$. При $0{,}5 \leqslant \gamma_\ell \leqslant 1$, $\ell = 1, 2$, численными методами были

найдены близкие к оптимальным функции $a_\ell^B(\gamma_1,\gamma_2)$, $a_\ell^{ACBF}(\gamma_1,\gamma_2)$ в случае известных дисперсий D_1, D_2, если $m_1 = m + d(D/N)^{1/2}$, $m_2 = m - d(D/N)^{1/2}$ (здесь m может выбираться произвольно). Для аналога стратегии Басера функции имеют вид $a_1^B(1,\gamma) = 0{,}3 - 0{,}1(1-\gamma)$, $a_2^B(1,\gamma) = 0{,}329 - 0{,}029/\gamma$, $a_1^B(\gamma,1) = a_2^B(1,\gamma)$, $a_2^B(\gamma,1) = a_1^B(1,\gamma)$. Для аналога стратегии АСБФ они равны $a_1^{ACBF}(1,\gamma) = 0{,}93 - 0{,}5(1-\gamma)$, $a_2^{ACBF}(1,\gamma) = 1{,}112 - 0{,}182/\gamma$, $a_1^{ACBF}(\gamma,1) = a_2^{ACBF}(1,\gamma)$, $a_2^{ACBF}(\gamma,1) = a_1^{ACBF}(1,\gamma)$. Эти функции и были использованы в оценках (3), где $\hat\gamma_1 = (\hat D_1/\hat D)^{1/2}$, $\hat\gamma_2 = (\hat D_2/\hat D)^{1/2}$ при $\hat D = \max(\hat D_1, \hat D_2)$,.

Результаты численного моделирования методом Монте-Карло для стратегий Басера и АСБФ при обработке $K = 50$ пакетов данных, разбитых на $M_2 = 4$ малых пакетов, приведены на рис. 1 и 2 соответственно. Множество параметров Θ выбрано из условия $D = D_1$, $0{,}5 \leqslant (D_2/D)^{1/2} \leqslant 1$, $m_1 = m + d(D/N)^{1/2}$, $m_2 = m - d(D/N)^{1/2}$, где $|d| \leqslant 20$, а m может быть произвольным (в этом случае $\gamma_1 = 1$, $0{,}5 \leqslant \gamma_2 \leqslant 1$). Как и в случае, рассмотренном в [10], для управления имеет место инвариантное описание, которое зависит только от K и M_2, но не от полного числа данных N. На рис. 1 и 2 представлены нормированные значения функции потерь (1), вычисляемые по формуле $l_K(\sigma,d) = (DN)^{-1/2} L_N(\sigma,\theta)$. Линии 1, 2 и 3 соответствуют значениям $\gamma_2 = 1$, $\gamma_2 = 0{,}8$ и $\gamma_2 = 0{,}5$, причем жирные сплошные линии описывают полные потери, а тонкие пунктирные – потери без обработки пакетов по очереди на двух начальных этапах (при $k_0 = 1$). Видно, что при больших d потери определяются применением обоих действий на начальных этапах. Для их уменьшения следует или увеличить число этапов K или уменьшить размеры пакетов на начальных этапах Максимальные значения $l_K(\sigma,d)$ достигаются при равных дисперсиях и приблизительно равны 0.77 для стратегии Басера и 0.76 для стратегии АСБФ. Поскольку принципиальной нижней границей для нормированного к величине $(DN)^{1/2}$ минимаксного риска (2) в случае известных равных дисперсий является значение 0,637, можно говорить, что простые в реализации пакетные стратегии UCB обеспечивают высокое качество управления.

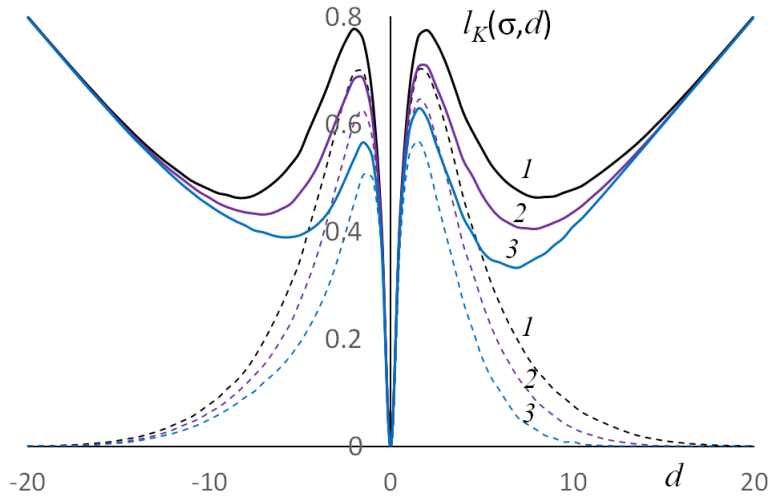

Рис. 1. Нормированные потери для стратегии Басера

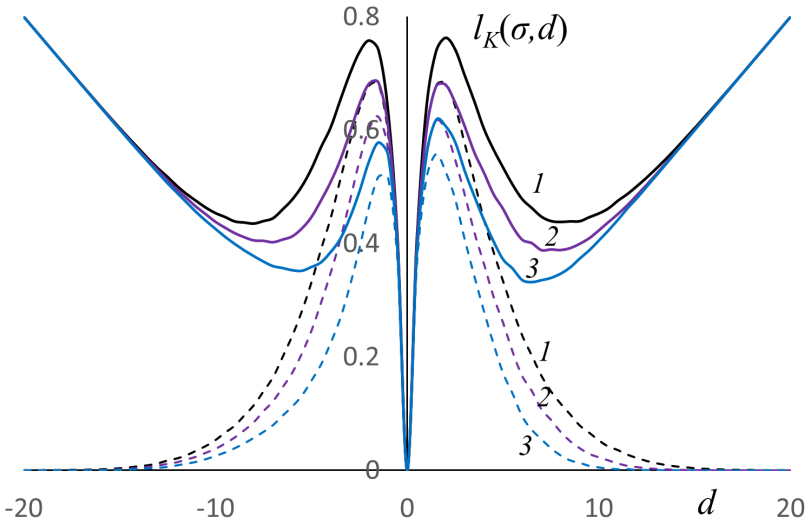

Рис. 2. Нормированные потери для стратегии АСБФ

3. Заключение

Предложенные стратегии просты в реализации и обеспечивают значения максимальных нормированных потерь приблизительно равные 0.77 для стратегии Басера и 0.76 для стратегии АСБФ. Поскольку принципиальной нижней границей для нормированного минимаксного риска является значение 0,637, указанные стратегии обеспечивают высокое качество управления.

Исследование выполнено за счет гранта Российского научного фонда № 23-21-00447, https://rscf.ru/project/23-21-00447/.

Список литературы

1. Berry D.A., Fristedt B. Bandit Problems: Sequential Allocation of Experiments. London, New York: Chapman and Hall, 1985. 275 p.
2. Пресман Э.Л., Сонин И.М. Последовательное управление по неполным данным. М.: Наука, 1982. 256 с.
3. Lattimore T., Szepesvari C. Bandit Algorithms. Cambridge: Cambridge University Press, 2020. 588 p.
4. Цетлин М.Л. Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969. 316 с.
5. Срагович В.Г. Адаптивное управление. М.: Наука, 1981. 384 с.
6. Назин А.В., Позняк А.С. Адаптивный выбор вариантов. М.: Наука, 1986. 288 с.
7. Perchet V., Rigollet P., Chassang S., Snowberg E. Batched Bandit Problems // Annals of Statistics. 2016. Vol. 44, No. 2, P. 660–681.
8. Колногоров А.В. Робастное параллельное управление в случайной среде и оптимизация обработки данных // Автоматика и телемеханика. 2014. № 12. С. 42–55.
9. Vogel W. An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem // Ann. Math. Stat. 1960. Vol. 31, P. 444–451.
10. Гарбарь С.В., Колногоров А.В. Лазутченко А.Н. Стратегии UCB и оптимизация пакетной обработки в задаче об одноруком бандите // Математическая теория игр и ее приложения. 2023. Т. 15. Вып. 4. С. 3–27.