

ВЛИЯНИЕ ОЦЕНКИ НЕИЗВЕСТНОЙ ДИСПЕРСИИ ДОХОДА НА ПОТЕРИ ПРИ ИСПОЛЬЗОВАНИИ СТРАТЕГИИ UCB ДЛЯ ГАУССОВСКОГО ДВУРУКОГО БАНДИТА

С.В.Гарбарь

RELATION BETWEEN REWARD VARIANCE ESTIMATION AND LOSSES FOR UCB STRATEGY FOR GAUSSIAN TWO-ARMED BANDIT

S.V.Garbar

Новгородский государственный университет имени Ярослава Мудрого, Sergey.Garbar@novsu.ru

Рассматривается задача о двуруком бандите с гауссовским распределением доходов действий с неизвестными математическим ожиданием и дисперсией. Гауссовские двурукие бандиты могут быть использованы при рассмотрении пакетной обработки, когда имеются два возможных метода обработки. Показано, что при использовании стратегии UCB1 величина ожидаемых потерь непрерывно зависит от оценки дисперсии доходов действий. С использованием метода Монте-Карло установлен характер зависимости. Отмечается, что использование некорректной оценки равносильно неоптимальному выбору параметров алгоритма, но величина потерь расчета незначительна при достаточно большой ошибке, что дает возможность оценить величину дисперсии на начальном этапе управления.

Ключевые слова: управление в случайной среде, задача о двуруком бандите, UCB

Для цитирования: Гарбарь С.В. Влияние оценки неизвестной дисперсии дохода на потери при использовании стратегии UCB для гауссовского двурукого бандита // Вестник НовГУ. Сер.: Технические науки. 2021. №2(123). С.17-20. DOI: [https://doi.org/10.34680/2076-8052.2021.2\(123\).17-20](https://doi.org/10.34680/2076-8052.2021.2(123).17-20)

Gaussian two-armed bandit problem is considered. Awards are assumed to have unknown expected values and unknown variances. Gaussian two-armed bandits may prove useful in a batch processing scenario, when there are two methods available. It is demonstrated that expected regret value is a continuous function of reward variance when using UCB1 strategy. Monte-Carlo simulations are used to show the nature of the relation between variance estimation and losses. It is shown that using an incorrect estimate is equivalent to using non-optimal parameters of the strategy, but the regret grows only slightly when the estimation error is fairly large, which allows to estimate the variance during the initial steps of the control.

Keywords: control in a random environment, two-armed bandit problem, UCB

For citation: Garbar S.V. Relation between reward variance estimation and losses for UCB strategy for Gaussian two-armed bandit // Vestnik NovSU. Issue: Engineering Sciences. 2021. №2(123). P.17-20. DOI: [https://doi.org/10.34680/2076-8052.2021.2\(123\).17-20](https://doi.org/10.34680/2076-8052.2021.2(123).17-20)

1. Введение

Рассмотрим задачу о многоруком бандите, который традиционно описывается как игровой автомат с J рукоятками, одну из которых на очередном шаге выбирает игрок (данный выбор называется действием) [1]. Формально многорукий бандит представляет собой управляемый случайный процесс $\xi(n), n=1,2,\dots,N$; количество раз, которые можно выбрать действие, N называется горизонтом управления, он может быть известен или не определен заранее. Значение $\xi(n)$ на шаге n зависит только от выбранного действия y_n и называется доходом. Математические ожидания доходов $m_1, \dots, m_J$ предполагаются неизвестными.

В связи с тем, что игрок не имеет информации об ожидаемых доходах, соответствующих каждому из действий, то ему необходимо одновременно получать эту информацию и принимать оптимальные решения, исходя из уже полученных знаний, чтобы максимизировать общий полученный доход.

Зачастую для простоты исследований делается предположение и об известности дисперсий $D_1, \dots, D_J$, в таком случае рассматриваемый многорукий бандит

описывается векторным параметром $\theta_1 = (m_1, \dots, m_J)$.

Если же дисперсии $D_1, \dots, D_J$ неизвестны, то для описания бандита используется векторный параметр $\theta_2 = (m_1, \dots, m_J, D_1, \dots, D_J)$.

Целью игрока является максимизация общего полученного дохода, но, так как отсутствуют априорные знания о параметре θ , задачей является также приобретение этой информации. Для используемой стратегии σ общий ожидаемый доход будет меньше, чем максимально возможный, поскольку часть раундов будет использована для получения информации о θ , на величину, называемую потерями. Потери после N раундов определяются как математическое ожидание разности между наибольшим возможным доходом и фактически полученным

$$L_N(\sigma, \theta) = E_{\sigma, \theta} \left(\sum_{n=1}^N (\max(m_1, \dots, m_J) - \xi_n) \right) \quad (1)$$

где $E_{\sigma, \theta}(\cdot)$ — математическое ожидание, найденное относительно меры, порожденной σ и θ .

Введем обозначение $D = \max_{l=1 \dots J} D_l$ и будем также анализировать величину нормализованных к $(DN)^{1/2}$ потерь

$$L'_N(\sigma, \theta) = (DN)^{1/2} L_N(\sigma, \theta). \quad (2)$$

Предположение об известности дисперсий на практике можно опустить при использовании многих стратегий, так как при небольших изменениях дисперсии потери будут меняться незначительно, что было представлено в работах [2,3]. Целью данной работы является обоснование этого наблюдения для стратегии UCB и проведение соответствующих численных экспериментов, устанавливающих влияние величины ошибки в оценке дисперсии на приведенные потери при заданном горизонте управления.

При неизвестных дисперсиях доходов действительный возможен их оценка на первых этапах управления или постоянная переоценка в процессе управления, и решение о том, насколько точная оценка требуется, также можно принимать на основании анализов результатов данной работы.

2. Стратегии UCB

Ввиду того, что целью является минимизация потерь, то при известном значении параметра многорукого бандита θ следовало бы выбирать действие с наибольшим ожидаемым доходом. Поскольку параметр является неизвестным, то возникает дилемма «информация или управление», сущность которой заключается в необходимости использовать некоторое число шагов действий для исследования распределения доходов действий, что влечет за собой получение менее чем максимально возможного дохода. При этом возможно получение оценок параметров многорукого бандита, наибольшую важность из которых имеют математические ожидания доходов различных действий.

Если на n -м шаге действие с номером l было выбрано n_l раз, от чего получен доход $X_l(n)$, то величина $X_l(n)/n_l$ будет являться точечной оценкой для ожидаемого дохода m_l . Может показаться разумным выбирать действие, соответствующее наибольшему значению данной оценки, однако это может привести к значительным потерям, если в силу случайности наилучшее действие на первых шагах принесло небольшой доход. Другими словами, для корректной оценки m_l необходимо, чтобы при $N \rightarrow \infty$ каждое действие было выбрано бесконечное число раз.

Стратегии UCB [4-6] предполагают вместо точечных оценок рассматривать верхние границы доверительных интервалов их интервальных оценок, например, UCB1 [6] предписывает выбирать действие, максимизирующее значение

$$U_l(n) = \frac{X_l(n)}{n_l} + \frac{\sqrt{2D \log n}}{\sqrt{n_l}}, l=1,2,\dots,J; n=1,2,\dots,N. \quad (3)$$

При этом предполагается сначала однократно выбрать каждое действие для формирования начальных значений (3) для каждого из действий. Рост величины $\sqrt{2D \log n}/\sqrt{n_l}$ с номером шага n гарантирует, что каждое из действий почти наверное будет выполнено бесконечное число раз на бесконечном горизонте управления. С точки зрения получения наименьшего значения наихудших нормализованных потерь вместо данной величины оказывается рациональным использовать параметризованную $\sqrt{aD \log n}/\sqrt{n_l}$.

3. Влияние оценки дисперсии на потери

Рассмотрим гауссовского двурукого бандита с равными дисперсиями доходов $D_1 = D_2 = D$, пусть N — горизонт управления. Предположим, не умаляя общности, что худшее действие имеет нулевые средние ожидаемые доходы, соответственно, лучшее — положительные ожидаемые доходы, т. е. $m_1 > 0, m_2 = 0$.

Гауссовские двурукие бандиты играют важную роль во многих приложениях [7,8]. В частности, при пакетной обработке данных, когда выбор действия может быть изменен лишь после того, как оно было использовано заданное число раз, доходы от обработки одного пакета будут иметь нормальное распределение в силу центральной предельной теоремы. Это особенно важно, так как при этом возможна параллельная обработка [2,3,9].

В работах [10,11] построены инвариантные описания пакетных версий различных вариантов стратегий UCB с горизонтом, равным единице, на множестве близких распределений доходов — когда разница между математическими ожиданиями имеет порядок $N^{-1/2}$ (что является наиболее сложным для управления случаем, так как при этом величины нормализованных потерь (2) принимают наибольшие значения [12]). Многие из приведенных в работах [10,11] утверждений относительно пакетной обработки данных справедливы также и для гауссовских двуруких бандитов.

Рассмотрим, каким образом ожидаемые доходы при выборе очередного действия зависят от оценки дисперсии.

Обозначим через $(n+1)$ номер очередного шага, и пусть n_1 — количество применений первого действия на текущий момент, соответственно, $n_2 = n - n_1$ — количество применений второго действия.

Доход от однократного применения каждого действия для гауссовского запишем как $m_l + \sqrt{D}\eta$, где η — стандартная нормально распределенная случайная величина.

Доход от применения l -го действия можно записать с использованием индикаторных функций $I_l(n)$, которые определены для каждого имеющегося действия и принимают единичное значение для тех шагов, на которых те были выбраны:

$$I_l(n) = \begin{cases} 1, & \text{если } U_l(n) = \max(U_1(n), U_2(n)), \\ 0, & \text{в противном случае.} \end{cases} \quad (4)$$

Ожидаемый доход для l -го действия записывается как сумма n_l гауссовских случайных величин, т. е. также является нормально распределенной случайной величиной

$$X_l(n) = n_l m_l + \sum_{i=1}^n I_l(i) \sqrt{D} \eta = n_l m_l + \sqrt{n_l D} \eta. \quad (5)$$

Общие ожидаемые потери на шаге $(n+1)$ в случае известной дисперсии при фиксированном n_1 по определению можно найти как математическое ожидание потерь при выборе действия. Если выбрано первое действие с более высоким ожидаемым доходом, то потери нулевые, в противном случае равны разнице ожидаемых доходов от действий

$$L_{n+1}(\sigma, \theta) = L_n(\sigma, \theta) + 0 \cdot P(U_2(n) < U_1(n)) + m_1 P(U_2(n) \geq U_1(n)). \quad (6)$$

Оценим вероятность выбора неоптимального действия

$$P(U_2(n) \geq U_1(n)) = P\left(\frac{X_1(n)}{n_1} + \frac{\sqrt{aD \log n}}{\sqrt{n_1}} - \frac{X_2(n)}{n_2} - \frac{\sqrt{aD \log n}}{\sqrt{n_2}} \leq 0\right) = P\left(m_1 - m_2 + \sqrt{D/n_1} \eta_1 - \sqrt{D/n_2} \eta_2 + \frac{\sqrt{aD \log n}}{\sqrt{n_1}} - \frac{\sqrt{aD \log n}}{\sqrt{n_2}} \leq 0\right), \quad (7)$$

где η_1, η_2 — независимые стандартные нормально распределенные случайные величины. Их линейная комбинация $\sqrt{D/n_1} \eta_1 - \sqrt{D/n_2} \eta_2$ также будет нормально распределенной случайной величиной $\sqrt{D(n_1 + n_2)/n_1 n_2} \eta$, соответственно

$$P(U_2(n) \geq U_1(n)) = P\left(\eta \leq -m_1 \sqrt{\frac{n_1 n_2}{Dn}} + \sqrt{\frac{a \log n}{n}} (\sqrt{n_1} - \sqrt{n_2})\right). \quad (8)$$

Вероятность (8) может быть записана с использованием функции распределения гауссовской случайной величины — функции Лапласа $\Phi(\cdot)$. Ожидаемые потери можно записать как сумму потерь на предыдущем шаге и математическое ожидание потерь на текущем шаге, учитывая, что n_1 может принимать различные значения с вероятностями $P(n_1 = i|n)$, $i = 1, \dots, n$, при этом данная вероятность является зависящей от

$$L_{n+1}(\sigma, \theta) = L_n(\sigma, \theta) + m_1 \sum_{i=1}^n \Phi\left(\frac{\sqrt{a \log n}}{n} (\sqrt{n_1} - \sqrt{(n-n_1)_2}) - m_1 \frac{\sqrt{n_1(n-n_1)}}{Dn}\right) \times P(n_1 = i|n). \quad (9)$$

Выражение (9) является непрерывной функцией от дисперсии D при положительных значениях аргумента (как композиция непрерывных функций), если $L_n(\sigma, \theta)$ является непрерывной функцией от D . Последнее будет верно по математической индукции, однако нужно определить базу индукции.

В соответствии с рассматриваемым алгоритмом первые два шага каждое из действий выбирается единожды для формирования начальных оценок ожидаемых доходов. Соответственно, на третьем шаге, когда $n_1 = n_2 = 1$, ожидаемые потери равны

$$L_3(\sigma, \theta) = m_1 \Phi\left(-\frac{m_1}{\sqrt{2D}}\right), \quad (10)$$

что также является непрерывной функцией от D .

Значит, при $n > 3$ $L_n(\sigma, \theta)$ является непрерывной функцией от D .

Вывод из вышеприведенных рассуждений можно сформулировать в виде теоремы.

Теорема 1. При использовании стратегии UCSB с границами (3) для двурукого бандита, описываемого параметром $\theta = (m_1, m_2, D_1, D_2)$, $D_1 = D_2 = D$, потери (1) являются непрерывной функцией от оценки дисперсии D .

Заметим, что анализируемая далее величина нормализованных потерь (2) также непрерывно зависит от D .

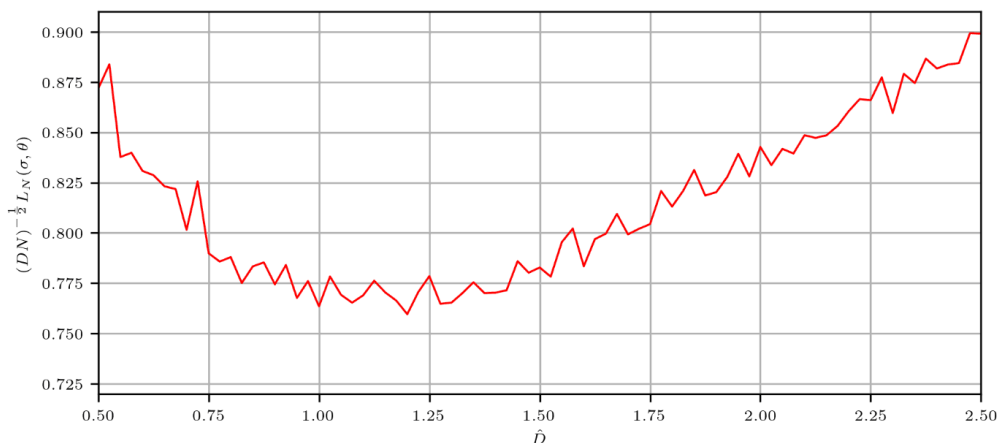
Заметим также, что использование некорректной оценки для дисперсии в формуле (3) имеет тот же эффект, что использование неоптимального значения для параметра a : растет или уменьшается вклад в оценку верхней границы доверительного интервала величины $\sqrt{aD \log n / n_i}$. При ее чересчур малых значениях алгоритм предписывает меньше использовать действие, в среднем принесшее меньший доход, и в худшем случае игрок не получает достаточно информации о нем. При слишком большом значении в худшем случае менее выгодное действие будет применено излишнее количество раз.

Для оценки величины нормализованных потерь в зависимости от оценки дисперсии воспользуемся численным моделированием по методу Монте-Карло.

4. Численное моделирование

Проверим, как на практике максимальные нормализованные потери зависят от величины оценки дисперсии, т. е. какими в худшем случае будут потери для различных оценок дисперсии. Рассмотрим гауссовского двурукого бандита с равными дисперсиями доходов ($D_1 = D_2 = D = 1$). Положим $a = 1$, горизонт управления $N = 200$.

На рисунке представлены результаты численного моделирования: максимальные нормализованные потери при различных величинах оценки дисперсии $\hat{D}$. График получен усреднением максимальных ожидаемых потерь для выбранного горизонта по 10000 моделирований. Рассмотрен случай близких



Зависимость нормализованных потерь $(DN)^{1/2} L_N(\sigma, \theta)$ при использовании стратегии UCSB для гауссовского двурукого бандита с дисперсией доходов $D = 1$ от оценки дисперсии $\hat{D}$

распределений доходов как наихудший с точки зрения величины нормализованных потерь.

Видно, что при корректной оценке дисперсии величина потерь для данного значения параметра a является близкой к минимальной, однако при ее изменении потери возрастают. Также можно заметить, что выбранное значение параметра $a = 1$ не является оптимальным для данного алгоритма. Однако нормализованные потери меняются незначительно (не более чем на 5% от оптимального значения при оптимальном выборе значения параметра a) при изменении оценки дисперсии на 25% как в сторону увеличения, так и в сторону уменьшения.

В связи с тем, что оценка дисперсии является вычислительно более сложной задачей, чем оценка математического ожидания, то при наличии ограничений на время вычислений возможно с заданной достоверностью определить, в течение скольких действий необходимо уточнять оценку дисперсии для достижения требуемых величин потерь в худшем случае. Для этого потребуется построение доверительного интервала для дисперсии нормально распределенной случайной величины с заданным уровнем значимости.

5. Заключение

Показано, что потери при использовании стратегии UCB для гауссовского двурукого бандита являются непрерывной функцией от оценки дисперсии доходов действий. С использованием численного моделирования по методу Монте-Карло установлен характер этой зависимости.

Исследование проведено при финансовой поддержке НовГУ.

1. Lattimore T., Szepesvari C. Bandit Algorithms. Cambridge University Press, 2020. 536 p.
2. Колногоров А.В. Гауссовский двурукий бандит и оптимизация групповой обработки данных // Пробл. передачи информ. 2018. Т.54. №1. С.93-111.
3. Колногоров А.В., Гауссовский двурукий бандит: предельное описание // Пробл. передачи информ. 2020. Т.56. №3. С.86-111.
4. Bather J.A. The Minimax Risk for the Two-Armed Bandit Problem // Mathematical Learning Models — Theory and Algorithms. Lecture Notes in Statistics. 1983. V.20. P.1-11. DOI: https://doi.org/10.1007/978-1-4612-5612-0_1
5. Lai T.L. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem // The Annals of Statist. 1987. V.25. P.1091-1114.
6. Auer P. Using Confidence Bounds for Exploitation-Exploration Trade-offs // Journal of Machine Learning Research. 2002. V.3. P.397-422.
7. Reverdy P.B., Srivastava V., Leonard N.E. Modeling Human Decision Making in Generalized Gaussian Multiarmed Bandits // Proceedings of the IEEE. 2014. V.102. №4. P.544-571. DOI: <https://doi.org/10.1109/JPROC.2014.2307024>
8. Reverdy P. Gaussian multi-armed bandit problems with multiple objectives // Proceedings of the American Control

Conference (ACC). Boston, USA, 2016. P.5263-5269. DOI: 10.1109/ACC.2016.7526494

9. Kolnagorov A.V., Shiyani D.N. Parallel Version of the Mirror Descent Algorithm for the Two-Armed Bandit Problem // Proceedings of the 2016 Third International Conference on Mathematics and Computers in Sciences and in Industry (MCSI 2016). Chania, Crete. 27-29 August 2016. P.241-245. DOI: <https://doi.org/10.1109/MCSI.2016.052>
10. Garbar S.V. Invariant description for batch version of UCB strategy for multi-armed bandit // Journal of Physics: Conference Series. 2020. Vol.1658. Article number: 012015. DOI: <https://doi.org/10.1088/1742-6596/1658/1/012015>
11. Garbar S., Kolnagorov A. 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 (SMTDA). Barcelona, Spain, 2-5 June 2020. P.251-260.
12. Vogel W. An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem // Ann. Math. Statist. 1960. V.31. P.444-451.

References

1. Lattimore T., Szepesvari C. Bandit Algorithms. Cambridge University Press, 2020.
2. Kolnagorov A.V. Gaussovskiy dvurukiy bandit i optimizatsiya gruppovoy obrabotki dannykh [Gaussian Two-Armed Bandit and Optimization of Batch Data Processing]. Problems of Information Transmission, 2018, vol.54, no. 1, pp.27-44.
3. Kolnagorov A.V. Gaussovskiy dvurukiy bandit: predel'noe opisanie [Gaussian Two-Armed Bandit: Limiting Description]. Problems of Information Transmission, 2020, vol.56, iss.3, pp.86-111
4. Bather J.A. The Minimax Risk for the Two-Armed Bandit Problem. Mathematical Learning Models — Theory and Algorithms. Lecture Notes in Statistics. 1983, vol.20, pp.1-11. DOI: https://doi.org/10.1007/978-1-4612-5612-0_1
5. Lai T.L. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. The Annals of Statist., 1987, vol.25, pp.1091-1114.
6. Auer P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. Journal of Machine Learning Research, 2002, vol.3, pp.397-422.
7. Reverdy P.B., Srivastava V., Leonard N.E. Modeling Human Decision Making in Generalized Gaussian Multiarmed Bandits. Proceedings of the IEEE, 2014, v.102, no.4, pp.544-571. DOI: <https://doi.org/10.1109/JPROC.2014.2307024>
8. Reverdy P. Gaussian multi-armed bandit problems with multiple objectives. Proceedings of the American Control Conference (ACC). Boston, USA, 2016, pp.5263-5269. DOI: 10.1109/ACC.2016.7526494
9. Kolnagorov A.V., Shiyani D.N. Parallel Version of the Mirror Descent Algorithm for the Two-Armed Bandit Problem. Proceedings of the 2016 Third International Conference on Mathematics and Computers in Sciences and in Industry (MCSI 2016), Chania, Crete. 27-29 August 2016, pp.241-245. DOI: <https://doi.org/10.1109/MCSI.2016.052>
10. Garbar S.V. Invariant description for batch version of UCB strategy for multi-armed bandit. Journal of Physics: Conference Series, 2020, vol.1658. Article number: 012015. DOI: <https://doi.org/10.1088/1742-6596/1658/1/012015>
11. Garbar S., Kolnagorov A. 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 (SMTDA). Barcelona, Spain, 2-5 June 2020, pp.251-260.
12. Vogel W. An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem. Ann. Math. Statist., 1960, vol.31, pp.444-451.