Dependency of regret on accuracy of variance estimation for different versions of UCB strategy for Gaussian multi-armed bandits

S V Garbar
Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation
E-mail: Sergey.Garbar@novsu.ru

Abstract. We consider two variations of upper confidence bound strategy for Gaussian two-armed bandits. Rewards for the arms are assumed to have unknown expected values and unknown variances. It is demonstrated that expected regret values for both discussed strategies are continuous functions of reward variance. A set of Monte-Carlo simulations was performed to show the nature of the relation between variance estimation and losses. It is shown that 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 and stop this estimation later.

1. Introduction
A multi-armed bandit (MAB) problem is the topic of discussion in this work. Traditionally it is described as a slot machine with \( J \) arms (levers), each of them yields some random reward when chosen (played) by a gambler. A gambler (decision-making agent) chooses an arm multiple times and therefore receives some cumulative income. He or she begins with no prior knowledge about rewards associated with the each of the arms, so it is necessary to simultaneously acquire such a knowledge and optimize decisions based on already acquired knowledge. The goal of the gambler is to maximize the overall expected reward [1]. This is a reinforcement learning problem that exemplifies the exploration–exploitation tradeoff dilemma, so it is also studied in machine learning [2, 3]. This problem is also known as the problem of adaptive control in a random environment [4] and the problem of expedient behavior [5]. MABs have also been used to model problems such as managing research projects in a large organization like a science foundation or a pharmaceutical company [6-9].

Formally MAB is described as a controlled random process \( \xi(n), n = 1,2,\ldots,N \), where \( N \) is called the control horizon, this value can be unknow a priori. Value \( \xi(n) \) at time \( n \) only depends on the chosen arm \( y_n \) and represents the reward. Expected values \( m_1, \ldots, m_J \) of the rewards are assumed to be unknown. Variances \( D_1,\ldots, D_J \) are often assumed to be known, in this case MAB is described by a vector parameter \( \theta_1 = (m_1,\ldots, m_J) \). If variances are unknown, the MAB is described by a parameter \( \theta_2 = (m_1,\ldots, m_J, D_1,\ldots, D_J) \).

If parameter \( \theta \) was known to the gambler, he or she would be able to always choose the arm that yields the highest reward. As \( \theta \) is unknown, the gambler must spend some of the rounds of play to gather information that will allow to estimate the parameter. A control strategy \( \sigma \) determines a choice of action \( y_n \) depending on currently available information about process history.

For applied strategy \( \sigma \) the total expected reward is less than maximally possible due to the rounds that were spent for exploration. Difference between maximally possible expected reward and total...
expected reward according to the choice of arms is called regret (loss function). The regret after \( N \) rounds is defined as the expected value of sum of differences between the reward associated with the optimal strategy and the collected rewards:

\[
L_N(\sigma, \theta) = E_{\sigma, \theta}(\sum_{n=1}^{N}(\max(m_1, \ldots, m_j) - \xi_n)).
\]

\( E_{\sigma, \theta} \) denotes the expected value calculated with respect to measure generated by strategy \( \sigma \) and parameter \( \theta \).

Further we introduce the notation \( D = \max_{i=1..J} D_i \) and analyze the normalized by \( (DN)^{1/2} \) regret

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

Gaussian MABs are of importance in many applications \([10, 11]\). Particularly in batch processing of the arms, when switching the choice of the arm can only be done after the previously selected arm was played for a set number of times, reward for each such batch will be normally distributed according to the central limit theorem. Nowadays that scenario is significant as it allows parallel processing \([12-14]\).

Works \([15-17]\) present invariant descriptions of the batch versions of UCB strategies described below. These descriptions are related to the case of the “close” reward distributions: when difference between expected rewards has the order of \( N^{-1/2} \), the control is the hardest and normalized regret \( (2) \) is the greatest \([18]\). Many of statements in \([15-17]\) are obtained for Gaussian MABs and can be applied to the current work.

It is stated in \([12, 13, 19]\) that the rewards change very little even when the variance is changed moderately, therefore the variance can be estimated during the early stages of control. The goal of this work is to show that assumption that variances are known can be omitted without significant losses for the UCB strategies that we consider further.

2. UCB strategies

As the goal of the gambler is to maximize the sum of rewards during the control time, it is obvious that the arm with the highest mean reward should be chosen at every step to minimize the regret.

Suppose that at the step \( n \) the \( l \)-th arm \((l = 1, \ldots, J)\) was used \( n_l \) times and let \( X_l(n) \) denote cumulative reward for the corresponding arm at that step. In this case \( X_l(n)/n_l \) estimates the expected value of the reward for this arm. It might seem reasonable to apply the action corresponding to currently largest value \( X_l(n)/n_l \). But as there is no or little prior information at the beginning of the control, this estimation can be inaccurate, which can lead to significant losses: the actual best arm will never be used after the incorrect choice. Therefore, on the infinite control horizon the strategy must allow each arm to be played infinitely many times for its mean reward to be estimated correctly.

Instead of estimating mean rewards \( \{m_i\} \) for each of the arms it is proposed to consider the upper bounds of their confidence intervals, and therefore UCB strategies are defined as selection of arm maximizing one of the values (for \( l = 1,2,\ldots,J; n = 1,2,\ldots,N \)):

\[
U_l(n) = \frac{X_l(n)}{n_l} + \frac{\sqrt{\log(n/n_l)}}{\sqrt{n_l}},
\]

for the strategy discussed in \([8]\), which we will further call UCB-Lai;

\[
U_l(n) = \frac{X_l(n)}{n_l} + \frac{\sqrt{n_I\log n}}{\sqrt{n_l}}
\]

for the UCB1 strategy discussed in \([2]\).

This choice of an arm in each of these strategies negotiates the exploration–exploitation trade-off: on one hand it greedily selects an arm with the highest estimated mean (first term in each of the expressions), but on the other hand there is an exploration term that can be derived from an experimental design approach with the notion of information gain, the second term grows with time.
Discussed strategies prescribe to initially apply once every action. Then at each point of time \( n \) it is necessary to choose the action corresponding to the highest value \( \{U_l(n)\} \).

3. Effect of accuracy of variance estimation on normalized regret

We consider a Gaussian two-armed bandit with equal reward variances \( D_1 = D_2 = D \), control horizon is known and equals \( N \).

We assume without loss of generality that the best arm has some positive expected reward \( m_1 > 0 \), and the worse one has zero mean reward \( m_2 = 0 \).

Next we evaluate how losses on next step \((n+1)\) depend on the estimation of the reward variance. Let \( n_1 \) be the number of times first arm was used and \( n_2 = n - n_1 \) is the number of times second arm was used. Reward from using an arm can be decomposed as \( m_l + \sqrt{D} \eta \), where \( \eta \sim N(0,1) \) is a standard normal random variable.

For each of the arms we define an indicator function as \((l = 1,2)\)

\[
I_l(n) = \begin{cases} 1, & \text{if } U_l(k) = \max(U_1(n), U_2(n)), \\ 0, & \text{otherwise}. \end{cases}
\]

Indicators have unit values for numbers of steps when corresponding arm was used.

Cumulative reward for each arm can be expressed as a sum of \( n_l \) Gaussian random variables, i.e. is itself a Gaussian random variable

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

If the variance is known and \( n_1 \) is fixed, then the cumulative losses on step \((n+1)\) corresponding can be expressed as sum of cumulative losses on step \( n \) and as expected value of loss on the step \((n+1)\). If the better arm is chosen, the expected loss on the current step is zero. Otherwise the loss is the difference between arms’ rewards:

\[
L_{n+1}(\sigma) = L_n(\sigma) + 0 \cdot P(U_2(n) < U_1(n)) + m_2 P(U_2(n) \geq U_1(n)).
\]

3.1. UCB1 strategy. For UCB1 described in \([2]\) we estimate the probability of choosing the suboptimal arm as

\[
P(U_2(n) \geq U_1(n)) = P \left( X_1(n) - m_1 + \frac{aD \log n}{\sqrt{n_1}} - X_2(n) - \frac{aD \log n}{\sqrt{n_2}} \leq 0 \right)
\]

\[
= P \left( m_1 - m_2 + \frac{D}{\sqrt{n_1}} \eta_1 - \frac{D}{\sqrt{n_2}} \eta_2 + \frac{aD \log n}{\sqrt{n_1}} - \frac{aD \log n}{\sqrt{n_2}} \leq 0 \right),
\]

where \( \eta_1, \eta_2 \) — i.i.d. standard normal random variables. Their linear combination \( \sqrt{D/n_1} \eta_1 - \sqrt{D/n_2} \eta_2 \) is also normally distributed random variable \( \sqrt{D/n_1} + \frac{D}{n_2} \eta_2 \), therefore

\[
P(U_2(n) \geq U_1(n)) = P \left( \eta \leq -m_1 \sqrt{\frac{n_2}{Dn}} + \frac{aD \log n}{n} \left( \sqrt{n_1} - \sqrt{n_2} \right) \right).
\]

This probability can be expressed using cumulative distribution function of a Gaussian random variable, which we denote as \( \Phi(\cdot) \). Cumulative expected regret is a sum of cumulative expected regret on a previous step and expected regret on a current step. Taking into account the fact that \( n_1 \) can have different values with conditional probabilities \( P(n_1 = l|n) \) depending on the history of control and rewards, we get
This expression represents a continuous function of argument $D$ when $D > 0$ if $L_n(\sigma, \theta)$ is a continuous function. The latter would be correct by the induction if we determine the correctness of the base case.

Algorithm prescribes to use every arm once for the first two steps to make initial estimations of their expected rewards. On the third step and expected regret can be found as

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

which is a continuous function also. Therefore when $n > 3$, (1) is a continuous function of argument $D$.

The obtained results can be expressed as the theorem.

**Theorem 1.** When one uses UCB1 strategy with the bounds described by (4) for Gaussian 2-armed bandit described by the parameter $\theta = (m_1, m_2, D_1, D_2)$, $D_1 = D_2 = D$, the regret (1) is a continuous function of variance estimation $D$.

**Remark 1.** Normalized regret (2) is also a continuous function of $D$.

**Remark 2.** Using an incorrect variance estimation in (4) is effectively the same as using suboptimal value for the parameter $\alpha$. The effect is increasing or decreasing of contribution of value $\sqrt{\alpha D \log n/n_1}$. If its value is too high, suboptimal arm will be played too often, which will diminish the cumulative reward. On the other hand, when it is too low, the gambler gets information about less profitable for him arm less often, and it is by chance actually a more lucrative arm, that also increases the regret.

### 3.2. UCB-Lai strategy

We estimate the probability of choosing the suboptimal arm as

$$P\left(U_2(n) \geq U_1(n)\right) = \left(\frac{x_1(n)}{n_1} + \frac{\sqrt{aD \log n/n_1}}{\sqrt{n_1}} - \frac{x_2(n)}{n_2} - \frac{\sqrt{aD \log n/n_2}}{\sqrt{n_2}} \leq 0\right).$$

Similarly to UCB1, we can estimate this probability as

$$P\left(U_2(n) \geq U_1(n)\right) = \left(\eta \leq -m_1 \sqrt{\frac{n_1 n_2}{D n}} + \frac{a}{n} \left(\sqrt{n_1 \log n/n_2} - \sqrt{n_2 \log n/n_1}\right)\right),$$

which similarly yields a continuous function for regret and we can state it as the theorem.

**Theorem 2.** When one uses UCB-Lai strategy with the bounds described by (3) for Gaussian 2-armed bandit described by the parameter $\theta = (m_1, m_2, D_1, D_2)$, $D_1 = D_2 = D$, the regret (1) is a continuous function of variance estimation $D$.

Remarks 1 and 2 also hold in case of UCB-Lai.

To estimate the value of normalized regret and its relation to variance estimation we use a series of Monte-Carlo simulations.

### 4. Simulation results

We consider how maximum of normalized regret depends on the estimation of variance, i.e. what is the worst-case regret for different variance estimations. Further we study a Gaussian two-armed bandit with equal variances $D_1 = D_2 = D = 1$, control horizon $N = 200$.

Figure 1 shows simulation results for maximum normalized regret vs estimation of variance $\hat{D}$ for UCB1 strategy. Plot was obtained by averaging the maximum expected regret for the given control horizon over 10000 simulations. A case of close reward distributions is considered as it yields the biggest regret.
Figure 1. Normalized regret $(DN)^{1/2}L_N(\sigma, \theta)$ vs variance estimation $\bar{D}$ for UCB1 control strategy for Gaussian 2-armed bandit with reward variances of arms $D_1 = D_2 = 1$.

We can see that when variance estimation is close to correct ($\bar{D} = D = 1$), normalized losses if close to the minimal of all the estimations. When the estimation changes, regret gets bigger, but not in a very significant way: according to the plot, regret grows no more than 5% when variance estimation changes by 25% (both greater and lower than the correct value).

Figure 2 shows the results of similar simulations for UCB-Lai strategy.

Figure 2. Normalized regret $(DN)^{1/2}L_N(\sigma, \theta)$ vs variance estimation $\bar{D}$ for UCB-Lai control strategy for Gaussian 2-armed bandit with reward variances of arms $D_1 = D_2 = 1$.

Conclusions of the previous simulations also hold in case of UCB-Lai.

Estimation of variance takes more time than estimation of expected value, so it might be reasonable to define the permissible level of regret. Based on that level one can define, after how many steps he or she can stop updating variance estimations.
5. Conclusions
We reviewed two versions of UCB strategy and showed that normalized regret is a continuous function of variance estimation. It is observed that using an incorrect estimate is equivalent to using non-optimal parameters of the strategy. It is also demonstrated with a set of Monte-Carlo simulations that 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.

Acknowledgments
The reported study was supported financially by Yaroslav-the-Wise Novgorod State University.

References
[1] Lattimore T and Szepesvari C 2020 Bandit Algorithms (Cambridge University Press)
[2] Auer P 2002 Using confidence bounds for exploitation-exploration trade-offs Journal of Machine Learning Research. 3 397–422
[3] Lugosi G and Cesa-Bianchi N 2006 Prediction, Learning and Games (New York: Cambridge University Press)
[4] Slavnovitch G 2006 Mathematical Theory of Adaptive Control (Singapore: World Scientific Publishing Company)
[5] Tsetlin M L 1973 Automaton Theory and Modeling of Biological Systems (New York: Academic Press)
[6] Berry D A and Fristedt B 1985 Bandit Problems: Sequential Allocation of Experiments, Chapman and Hall (London; New York)
[7] Gittins J C 1989 Multi-armed bandit allocation indices, Wiley-Interscience Series in Systems and Optimization (Chichester: John Wiley & Sons, Ltd)
[8] Lai T L 1987 Adaptive treatment allocation and the multi-armed bandit problem The Annals of Statist 25 1091–1114
[9] Lai T L, Levin B, Robbins H and Siegmund D 1980 Sequential medical trials (Stopping Rules/Asymptotic Optimality) Proc. Nati. Acad. Sci. USA 77 3135–38
[10] Reverdy P B, Srivastava V and Leonard N E 2014 Modeling human decision making in generalized gussian multiarmed bandits Proceedings of the IEEE 102 (4) 544–71 DOI:10.1109/JPROC.2014.2307024
[11] Reverdy P 2016 Gaussian multi-armed bandit problems with multiple objectives American Control Conference (ACC) (Boston: MA) 5263–69 DOI:10.1109/ACC.2016.7526494
[12] Kolnogorov A V 2018 Gaussian two-armed bandit and optimization of batch data processing. Problems of Information Transmission 54 (1) 84–100
[13] Kolnogorov A V 2020 Gaussian two-armed bandit: limiting description Problems of Information Transmission 56 (3) 278–301
[14] Kolnogorov A V and Shiyan D N Parallel 2016 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, August 27-29 pp 241–35
[15] Garbar S V 2020 Invariant description for batch version of UCB strategy for multi-armed bandit J. Phys.: Conf. Ser. 1658 012015 DOI: 10.1088/1742-6596/1658/1/012015
[16] Garbar S and Kolnogorov A 2020 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 ISAST 251–60
[17] Garbar S 2020 Invariant description of UCB strategy for multi-armed bandits for batch processing scenario 24th International Conference on Circuits, Systems, Communications and Computers (CSCC) 75–8 DOI: 10.1109/CSCC49995.2020.00021
[18] Vogel W 1960 An asymptotic minimax theorem for the two-armed bandit problem Ann Math Statist 31 444–451
[19] Kolnogorov A V 2012 Parallel design of robust control in the stochastic environment (the two-armed bandit problem) *Automation and Remote Control* **73** 689–701