Invariant description for batch version of UCB strategy for multi-armed bandit

S V Garbar
Yaroslav-the-Wise Novgorod State University, 41, ul. B. St. Petersburgskaya, Veliky Novgorod, Russian Federation
E-mail: Sergey.Garbar@novsu.ru

Abstract. We consider a variation of upper confidence bound strategy for multi-armed bandit in batch processing setting. Invariant descriptions with the unit control horizon are obtained for upper bounds in the strategy and for regret. A set of Monte-Carlo simulations are performed for different settings of MABs to determine the minimax regret for multi-armed bandits with different configurations.

1. Introduction
In the paper we consider multi-armed bandit (MAB) problem, which is traditionally presented as a slot machine that has two or more arms (levers) [1]. A gambler (decision-making agent) chooses an arm and consequently receives some random income. As he or she begins with no prior knowledge about rewards associated with the arms, it is necessary to simultaneously acquire new knowledge (exploration) and optimize decisions based on existing knowledge (exploitation). The goal of the gambler is to maximize the overall expected reward.

This problem is also known as the problem of adaptive control in a random environment [2] and the problem of expedient behavior [3]. This is a reinforcement learning problem that exemplifies the exploration–exploitation tradeoff dilemma, so it is also studied in machine learning ([4], [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 [1], [6].

Further we consider Gaussian MAB with \( J \) arms. It can be described formally as a controlled random process \( \xi(n) \), \( n = 1, 2, ... N \). Value \( \xi(n) \) at time (round of play) \( n \) only depends on the chosen arm \( y_n \) and can be interpreted as reward. \( \xi(n) \) is normally distributed with probability density function

\[
f_D(x|m_l) = \left(2\pi D\right)^{-1/2} e^{-\frac{(x-m_l)^2}{2D}},
\]

if chosen arm \( y_n = l \), and \( l = 1, ..., J \). Expected values \( m_1, ..., m_J \) of the rewards are assumed to be unknown. Variance is assumed to be known and equals \( D \) for all the arms. Requirement of the prior knowledge of variance can be omitted as the rewards for using the strategy that we consider further change very little when the variance is changed moderately (e.g. 5–10% change). Hence the variance can be estimated during the initial control stage.

Studied MAB is therefore described with a vector parameter \( \theta = (m_1, ..., m_J) \).

Gaussian MAB problem arises when batch data processing is optimized and there are two or more processing methods available with different a priori unknown efficiencies [7].

If the gambler knew \( \theta \) his or her best strategy would be to exclusively use the arm with the highest associated mean reward. But as there is no such knowledge available, it is required to devote some of
the rounds to gathering such data. A control strategy determines a choice of action \( y_n \) depending of currently available information about process history.

For applied strategy \( \sigma \) the total expected reward is less than maximally possible by the value which is called the regret. The regret after \( N \) rounds is defined as the expected difference between the reward sum associated with the optimal strategy and the sum of the collected rewards:

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

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

Suppose that at the step \( n \) the \( l \)-th arm was chosen \( n_l \) times and let \( X_l(n) \) denote cumulative reward for the corresponding \( l \)-th arm \( (l = 1, ..., J) \). Then \( X_l(n)/n_l \) estimates the expected value of the reward \( m_l \) for this arm. As the goal of control is to maximize the overall expected reward, it might seem reasonable to apply the action corresponding to currently largest value \( X_l(n)/n_l \). However, such rule can result in a significant losses due to the fact that initial estimate \( X_l(n)/n_l \), corresponding to the largest \( m_l \), can by chance take a lower value and consequently this action will be never applied, which can entail significant losses.

Instead of estimates of \( \{m_l\} \) per se it is proposed to consider the upper bounds of their confidence intervals, and therefore UCB strategy (as described by Lai [8]) is described as selection of arm maximizing the value

\[
U_l(n) = \frac{X_l(n)}{n_l} + \sqrt{2D \log(n/n_l)},
\]

for \( l = 1, 2, ..., J; n = 1,2, ..., N \). Strategy prescribes to initially apply every action once. And then it is necessary to select the arm with the highest value of \( U_l(n) \). We see that confidence bound, \( \sqrt{2D \log(n/n_l)} / n_l \), grows slowly as gambler plays more rounds (i.e. as \( l \) increases), ensuring that he or she never stops playing any arm of the MAB.

Strategies of that kind are called UCB (upper confidence bound) rules. Choice of an arm negotiates the exploration—exploitation trade-off the following way: strategy strives to greedily choose an arm with the highest estimated mean, but also there is an exploration term that makes it possible to play an arm that was not explored thoroughly enough which grows with time.

We consider the batch version of the strategy described in [8] and aim to build an invariant description of the control strategy on the unit horizon in the domain of “close” distributions, as in case of close distributions the maximum values of expected regret are attained. Note that batch (parallel) strategies are important when processing time of the arm is significant, because in this case the total processing time depends on the number of batches rather than on the total number of plays.

We aim to estimate the upper bound of the maximum regret calculated over the set of acceptable values of a parameter which is chosen as follows

\[
\Theta = \{ m_l = m + d_l(D/N)^{1/2}; m \in (-\infty, +\infty), |d_l| \leq C < \infty, l = 1, ..., J \}.
\]

This set of parameters describes “close” distributions. Their definitive feature is the difference between expected values of the order \( N^{-1/2} \).

Maximal normalized regrets are observed on that domain and have the order \( N^{1/2} \) (see [9]). For “distant” distributions the normalized regrets have smaller values. For example, they have order \( \log N \) if \( \max(m_1, ..., m_j) \) exceeds all other \( \{m_l\} \) by some \( \delta > 0 \) (see [10]).

2. Batch version of UCB strategy for multi-armed bandit

Further we consider a version of UCB strategy that can change the arm only after using it \( M \) times in a row. These strategies allow batch (and also parallel) processing. We assume for simplicity that \( N = MK \), so the number of batches is \( K \). The upper bounds take form
\[ U_l(k) = \frac{X_l(k)}{k_l} + \frac{\sqrt{aMD\log(k/k_l)}}{\sqrt{k_l}}, \]

where \( k \) is the number of processed batches, \( k_l \) is the number of batches for which \( l \)-th arm was chosen and \( X_l(k) \) is the corresponding cumulative reward after processing \( k \) batches \((k = 1, 2, ..., K)\). Note that we changed the value of constant coefficient in upper bound estimate to variable \( a \) as its optimal value is still needed to be found.

We assume that each processed batch has a reward with a normal distribution as a corollary to a central limit theorem. Therefore, the following representation for reward of \( k \)-th batch is valid:

\[ \xi(k) = M \cdot \left( m + d_l \sqrt{\frac{D}{MK}} \right) + \sqrt{MD}\eta, \]

i.e. reward can be represented as its mean value with added Gaussian random variable \( \eta \sim N(0,1) \) with zero mean and variance \( MD \).

3. Finding an invariant description

For further reasoning we denote by \( I_l(k) \) the indicator of chosen action for processing the \((k + 1)\)-th batch according to the rule:

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

For \( k \leq J \) recall that every arm is chosen once for a batch. Also note that with probability 1 only one of the values of \( \{I_l(k)\} \) is equal to 1.

In this notation, the cumulative reward for each arm can be written out as the sum of its rewards for the number of plays it was selected

\[ X_l(k) = k_l M m_l + \sum_{i=1}^{k} I_l(i) \eta_{l,i}(MD), \]

where \( \eta_{l,i}(MD) \sim N(0,\sqrt{MD}) \) are i.i.d. normally distributed random variables with zero means and variances equal to \( MD \).

Note that for arm \( l \) indicator equals 1 exactly \( k_l \) times, so \( \sum_{i=1}^{k} I_l(i) \eta_{l,i}(MD) \) is the sum of \( k_l \) Gaussian random variables, which can be presented as a standard normal random variable scaled by a standard deviation. Therefore, under the assumption of arms having close distributions

\[ X_l(k) = k_l M \cdot \left( m + d_l \sqrt{\frac{D}{MK}} \right) + \sqrt{k_l MD}\eta, \]

where \( \eta \sim N(0,1) \) is a standard normal random variable.

The upper bounds for batch version of the strategy take form:

\[ U_l(k) = mM + d_l \frac{\sqrt{MD}}{\sqrt{K}} + \frac{\sqrt{MD}}{\sqrt{k_l}} \eta + \frac{aMD \log(k/k_l)}{\sqrt{k_l}}, \]

for \( l = 1, 2, ..., J; k = J + 1, J + 2, ..., K \).

Next, we apply the linear transformation that does not change the arrangement of the bound:

\[ u_l(t) = (U_l(k) - Mm)\sqrt{K/MD}. \]

To convert the description of the strategy to one with unit control horizon we introduce the following notation:

\[ t = kK^{-1}, t_l = k_l K^{-1}. \]

The expression for transformed upper bounds in considered UCB strategy with the unit control horizon, i.e. in the invariant form, is

\[ u_l(k) = d_l + \frac{\eta}{\sqrt{t_l}} + \frac{a \log(t/t_l)}{\sqrt{t_l}}, \quad l = 1, 2, ..., J. \]

Next, we find the expression for regret in invariant form. Let us assume without loss of generality, that \( d_1 = \max(d_1, ..., d_J) \). We find the expected regret as the sum of losses occurred due to selecting each non-optimal arm:
\[ L_N(\sigma, \theta) = (D/N)^{1/2} \sum_{l=2}^{J} (d_1 - d_l) E_{\sigma, \theta} \left( \sum_{k=1}^{K} M_l(k) \right) = (D/N)^{1/2} \sum_{l=2}^{J} M(d_1 - d_l) E_{\sigma, \theta}(k_l) \]

\[ = (DN)^{1/2} \sum_{l=2}^{J} (d_1 - d_l) E_{\sigma, \theta}(t_l). \]

Obtained results prove the following theorem.

Theorem 1. For Gaussian multi-armed bandit with \( J \) arms, fixed known variance \( D \) and unknown expected values \( m_1, \ldots, m_J \) the usage of the batch version of UCB strategy with bounds

\[ U_l(k) = \frac{X_l(k)}{k_l} + \frac{\sqrt{aMD \log(k/k_l)}}{\sqrt{k_l}}, \]

results in invariant description on the unit control horizon described by bounds

\[ u_l(k) = d_l + \frac{\eta}{\sqrt{t_l}} + \frac{\sqrt{a \log(t/t_l)}}{\sqrt{t_l}}, \]

where \( \eta \sim N(0,1) \) is a standard normal random variable.

For minimax regret the following expression holds:

\[ (DN)^{-1/2} L_N(\sigma, \theta) = \sum_{l=2}^{J} (d_1 - d_l) E_{\sigma, \theta}(t_l). \]

4. Simulation results

As we obtained an invariant description of the batch version of UCB strategy for the MAB, we can justify using Monte-Carlo method for studying the minimax regret.

To study the normalized regret values, the following tasks are completed: first we determine the optimal value for the parameter \( a \) for the strategy, then we find the minimax regrets for cases of 2-armed and 3-armed bandits.

In all the following simulations we take \( d_1 = 0 \) and \( d_J \) is always shown on the horizontal axis of the figures. In this case we can also regard \( d_J \) as a difference between mean rewards of bandit’s best (most profitable) and worst arms.

Figure 1 shows regret for different values of parameter \( a \) of the strategy in interval \( d_2 \in [1.8, 5.0] \) where regret is the highest. Horizon of size 200 was considered and regret was averaged over 10000 simulations. \( d_2 \) is shown on the horizontal axis.

![Figure 1. Minimax regret vs difference between mean rewards for different values of \( a \) for batch UCB strategy for 2-arm bandit.](image-url)
Optimal value (the one that yields the lowest minimax regret) for $\alpha$ can be found in the interval $[0.95, 1.05]$, which is different from $\alpha = 2$ that is present in original UCB-Lai strategy. Hence, for the following simulations value $\alpha = 1$ is used.

Figure 2 shows a relation between difference in mean rewards of arms $d_2$ for 2-armed bandit and normalized regret averaged over 10000 simulations for 2-armed bandit ($J = 2$). We consider horizons $N = 100, 400, 1500$ shown by different lines on the plot. Maximum normalized regret is approximately equal to 0.73 and is reached when $d_2 = d_2 - d_1 \approx 3.3$.

Figure 2. Minimax regret vs difference in mean rewards for different sizes of control horizons for batch UCB strategy for 2-arm bandit.

Figure 3 shows that batch size does not affect the minimax regret much for close distributions of the rewards of 2-armed bandit. Results for the plot were averaged over 5000 simulations, we considered 400 batches of sizes $M = 20, 50, 100$. We see that for close distributions regret does not depend on the batch size, but, in case of bigger differences between means, bigger batch sizes entail higher regret, as in UCB rules each arm is never ceased to be used, and in its batch variations we play an arm for pre-defined number of times.

Figure 3. Minimax regret vs difference between mean rewards for different batch sizes for batch UCB strategy for 2-arm bandit.
Next, we study 3-armed bandit to determine what is the worst case for the regret and how it depends on the relation between means of rewards of different arms. We set $d_1 = 0$ and then consider different values for $d_2$: figure 4 shows cases when $d_2 = 0, 2, 4$ as different lines, and $d_3$ is shown on the horizontal axis. Each point is calculated as an average regret over 10000 simulations, control horizon was chosen $N = 400$. Worst case for normalized regret of 1.23 was reached when $d_1 = d_2 = 0$ and $d_3 \approx 4.5$. These results can be easily explained with the invariant description that was obtained earlier:
\
\[ (DN)^{-1/2} L_N(\sigma, \theta) = \sum_{l=1}^{2} (d_3 - d_l) E_{\sigma, \theta} \lambda_l \] for regret is the highest when $d_1 = d_2 = 0$: more profitable strategy has not one but two competitor strategies that have close distributions.
\
Figure 4. Minimax regret vs mean rewards of 3-armed bandit, $d_2$ is shown with different lines, $d_3$ is on the horizontal axis.

Figure 5 shows relation between mean rewards for worst-case configuration of MAB and minimax regret. Maximum regret of 1.23 is reached when $d_1 = d_2 = 0$, $d_3 \approx 4.25$. Different numbers of processed batches are shown with different lines. Data is averaged over 10000 simulations.

Figure 5. Minimax regret vs mean rewards of 3-armed bandit, $d_1 = d_2 = 0$, $d_3$ is on the horizontal axis for different numbers of processed batches.

5. Conclusion
We reviewed a variant of UCB strategy proposed in [8] and applied it in batch processing scenario for Gaussian MAB. An invariant description and minimax regrets for considered strategy were found.
Values for minimax regrets were estimated by a set of Monte-Carlo simulations with fairly large control horizons and different batch sizes.

**Acknowledgments**

The reported study was funded by Russian Foundation for Basic Research, project no. 20-01-00062. Author would like to thank professor Alexander Kolnogorov for valuable advice and fruitful discussions.

**References**

[1] Berry D A and Fristedt B 1985 Bandit Problems: Sequential Allocation of Experiments *Chapman and Hall* (London, New York)

[2] Sragovich V G 2006 Mathematical Theory of Adaptive Control *World Sci.* (Singapore)

[3] Tsetlin M L 1973 Automaton Theory and Modeling of Biological Systems *Academic Press* (New York)

[4] Auer P 2002 Using Confidence Bounds for Exploitation-Exploration Trade-offs *Journal of Machine Learning Research* 3 397–422

[5] Lugosi G. and Cesa-Bianchi N 2006 Prediction, Learning and Games *University Press* (New York, Cambridge)

[6] Gittins J C 1989 Multi-armed bandit allocation indices *Wiley-Interscience Series in Systems and Optimization* (Chichester: John Wiley & Sons, Ltd.)

[7] 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

[8] Lai T L 1987 Adaptive Treatment Allocation and the Multi-Armed Bandit Problem *The Annals of Statist.* Vol. 25 1091–1114

[9] Vogel W 1960 An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem *Ann. Math. Statist.* 31 444–451

[10] Lai T L, Levin B, Robbins H and Siegmund D 1980 Sequential Medical Trials (Stopping Rules/Asymptotic Optimality) *Proc. Nati. Acad. Sci. USA* 77 3135–3138