УДК 519.814
https://doi.org/10.17721/1812-5409.2021/1.2

А. С. Джога, *аспірант*

Багаторукий бандит з розподілом Бернуллі в середовищі з затримками

Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова 4д,
e-mail: andrew.djoga@gmail.com

A. S. Dzhoha, *PhD student*

Bernoulli multi-armed bandit problem under delayed feedback

Taras Shevchenko National University of Kyiv, 83000, Kyiv, 4d Glushkova str.,
e-mail: andrew.djoga@gmail.com

Останнім часом все більше уваги приділяється онлайновому навчанню машин з відкладеним зворотнім зв'язком. Навчання з затримками є доцільнішим в більшості практичних застосувань, оскільки зворотній зв'язок від навколишнього середовища не є миттєвим. Наприклад, в клінічних випробуваннях, результати яких ми використовуємо в даній роботі, прояв реакції на ліки може зайняти деякий час. У даній роботі розглядається проблема стаціонарного стохастичного багаторукого бандита в середовищі з затримками, де кожна дія задається розподілом Бернуллі, параметри якого не відомі заздалегідь. Головною метою моделі у представленому середовищу є максимізація сукупної винагороди на скінченному горизонті, що еквівалентно мінімізації сукупних втрат. Розглядається стратегія Explore-First для даного випадку, яка визначається кількістю разів кожна дія буде обрана для дослідження. Наводиться асимптотичний аналіз ефективності алгоритму і вивчається вплив затримок у середовищі. Отримані теоретичні результати використовуються для розробки програмного забезпечення для проведення чисельних експериментів.

Ключові слова: проблема багаторукого бандита, стохастичне середовище з затримками, чисельні експерименти.

Online learning under delayed feedback has been recently gaining increasing attention. Learning with delays is more natural in most practical applications since the feedback from the environment is not immediate. For example, the response to a drug in clinical trials could take a while. In this paper, we study the multi-armed bandit problem with Bernoulli distribution in the environment with delays by evaluating the Explore-First algorithm. We obtain the upper bounds of the algorithm, the theoretical results are applied to develop the software framework for conducting numerical experiments.

Key Words: multi-armed bandit problem, stochastic environment with delays, numerical experiments.

Communicated by Associate Prof. Rozora I. V.

1 Introduction

The multi-armed bandit (MAB, or just bandit) problem is an example of a sequential decision-making process under uncertainty in real-time and represents a dilemma between exploration of new possibilities and exploitation of already known actions with predicted results. This dilemma is known as the exploration-exploitation trade-off. The bandit problem was introduced by Thompson [1] while studying clinical trials for developing the response-adaptive design methods for sequential allocation of different possible medical treatments.

The essential part of the research of new drugs and medical treatments is the randomized controlled trial when patients are randomly divided into two or more groups with different treatment protocols prescribed and at the end of trials the results from each group are compared. This way of conducting a clinical trial may satisfy scientific purposes but doesn't consider subjects' well-being as a significant amount of patients would suffer from a lack of efficient treatment during the trial. The bandit problem as a methodology can be used to improve such situations by developing an adaptive strategy that allows to correct the treatment protocol according to the already observed effectiveness during the clinical trial.

In this article, we study the bandit problem in

© А.С. Джога, 2021

the clinical trial setting and conduct experiments by using data from The International Stroke Trial simulating delays in the environment. The classical bandit problem assumes no delays of results in the sequential decision-making process. But in settings like clinical trials delays are inevitable and there is a constant need for the next decision-making action with no observed results yet for the previous one. In this setup, existing algorithms need an adaptation to retain theoretical guarantees. One of the possible solutions is the meta-algorithm designed by Joulani et al. [2] which allows the use of existing algorithms in environments with delays with retaining theoretical guarantees by the cost of reducing effectiveness in an additive way with respect to delays.

We provide an asymptotic analysis of one of the algorithms (Explore-First) for the MAB problem with Bernoulli distribution, we analyze an impact of the environment with delays on the algorithm and provide numerical experiments supporting obtained theoretical results. The results are applied to develop the software framework for conducting numerical experiments.

2 The multi-armed bandit problem and its solutions review

The multi-armed bandit problem is a sequential game between an algorithm and an environment. The game is being played over positive natural n time steps called the horizon. On each time step $t \in [n]$ the algorithm chooses an action A_t from a given set $\mathcal{A}$, then a reward $X_t \in \mathbb{R}$ is revealed by the environment. The action choice depends on the history of the previously chosen actions and their results in terms of rewards $H_{t-1} = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. The objective of the algorithm is a sequential selection of actions in order to maximize the sum of the rewards accumulated over n time steps $\sum_{t=1}^{n} X_t$.

Mathematically, the bandit problem is defined by the reward process associated with each action. Depending on the assumed nature of the reward process there are three fundamental formalizations: stochastic (stationary) bandits, Markov bandits, and adversarial bandits. In the stochastic MAB setting the reward process associated with each action is assumed to be described as an identically distributed random variable. In Markov MAB each action is associated with a Markov chain that evolves on each time step. The adversarial MAB setting has no assumptions about the origin of the considered reward process.

For the purpose of this article, we consider the stochastic multi-armed bandit problem with Bernoulli distribution for the reward process. This model is described as a system consisting of a collection of k actions denoted by $\mu = (\mu_1, ..., \mu_k)$, where each action is a Bernoulli distribution with parameter μ_i. The mean vector $\mu \in [0, 1]^k$ is not known in advance. In this model X_t is a realization of the random variable drawn from the distribution μ_{A_t} associated with the chosen action A_t on time step t, i.e $X_t \sim \text{Bernoulli}(\mu_{A_t})$ is a reward on a given time step.

As a metric of the effectiveness of the algorithms, it's common to use the notion of regret, which was introduced by Robbins [3]. The regret is the difference between the expected accumulated reward when choosing the optimal action (with the highest mean among others) during the whole horizon and the expected accumulated reward obtained by the algorithm following its policy. The regret after n time steps is represented as

$$R_n = n \max_{i \in \mathcal{A}} \mu_i - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right].$$

The regret can be rewritten as a function of the number of draws of each sub-optimal action. Let $\Delta_i = \max_{i \in \mathcal{A}} \mu_i - \mu_i$ denote a suboptimality gap of action i which has been chosen $N_i(n)$ times, then the regret decomposition [4] is defined by

$$R_n = \sum_{i=1}^{k} \Delta_i \, \mathbb{E}\left[N_i(n)\right]. \tag{2.1}$$

Policy, which maximizes the expected accumulated reward is equivalent to minimizing the regret. Asymptotic analysis of the algorithms consists of obtaining the regret bounds.

In this article, we consider the Explore-First algorithm as the main policy in the context of the Bernoulli bandit. This algorithm is characterized by the number of exploration m of each of the given k actions in the first phase, and exploiting the empirically best action in the second phase. Let $N_i(t)$ denote the number of times action i has been chosen by the algorithm after t time steps, then the average reward of action i after t time

steps can be defined as follows:

$$\hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{j=1}^{t} \mathbf{1}_{(A_j=i)} X_j.$$

Hence the chosen action A_t on time step t can be described in the following form:

$$A_t = \begin{cases} (t \bmod k) + 1 & \text{if } t \leqslant mk \\ \arg\max_{i \in \mathcal{A}} \left(\hat{\mu}_i(mk)\right) & \text{if } t > mk. \end{cases} \quad (2.2)$$

The protocol of the Explore-First algorithm:

1) Exploration phase (first mk time steps): choose each action exactly m times
2) Exploitation phase (the rest $n - mk$ time steps): use an empirically best action $\arg\max_{i \in \mathcal{A}} \left(\hat{\mu}_i(mk)\right)$.

This algorithm appears in the paper by Robbins [3], where the author formulated the stochastic multi-armed bandit problem. There was shown that the regret of Explore-First is sublinear in the general case:

$$\lim_{n \to \infty} \frac{R_n}{n} = 0.$$

Later Anscombe [5] considers this algorithm in the context of clinical trials and gives an asymptotic analysis for the MAB with Gaussian rewards, the author highlights many of the important considerations. For our work we use the results of the upper bound obtained by Slivkins [6] for Bernoulli MAB:

$$\mathbb{E}\left[R_n\right] \leqslant n^{2/3} \times O\left(k \log n\right)^{1/3}, \quad (2.3)$$

where the upper bound is minimized by choosing the following m:

$$m = (n/k)^{2/3} \cdot O\left(\log n\right)^{1/3}.$$

For the other algorithms overview, we refer the readers to Bubeck et al. [7].

3 Asymptotic analysis of the Explore-First algorithm

In order to obtain a stronger upper bound than in the inequality (2.3), we introduce a dependency on the MAB instance's mean vector. With help of concentration inequality, we show that a centered Bernoulli random variable is 1/2-sub-Gaussian, which allows us to utilize an additive property of sub-Gaussian random variables in the tail probability estimation for the MAB with a number of actions $k > 1$. The theory of the sub-Gaussian variables is presented in the monograph [8], in addition, we refer the readers to Kozachenko et al. [9]. Here we list only basic facts needed for obtaining the upper bound.

Definition 3.1. A random variable X with σ parameter is said to be σ-sub-Gaussian if for all $\lambda \in \mathbb{R}$, it holds that

$$\mathbb{E}\left[\exp(\lambda X)\right] \leqslant \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

For σ-sub-Gaussian random variable X we can write the exponential estimate of its tail probability, for any $\varepsilon \geqslant 0$, it holds that

$$\mathbb{P}(X \geqslant \varepsilon) \leqslant \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \quad (3.1)$$

In addition, sub-Gaussian random variables possess an additive property [9].

Lemma 3.1. *If X is a centered Bernoulli random variable with parameter p ($\mathbb{P}(X = 1 - p) = p, \mathbb{P}(X = -p) = 1 - p$), then X is 1/2-sub-Gaussian.*

Proof. Hoeffding's lemma [10] states that for a random variable X supported on the interval $[a, b]$ for all $\lambda \in \mathbb{R}$, holds that

$$\mathbb{E}\left[\exp(\lambda X)\right] \leqslant \exp\left(\lambda\, \mathbb{E}[X] + \frac{\lambda^2 (b-a)^2}{8}\right),$$

applying it to a centered Bernoulli random variable we get

$$\mathbb{E}\left[\exp(\lambda X)\right] \leqslant \exp\left(\frac{\lambda^2((1-p) - (-p))^2}{8}\right) = \exp\left(\frac{\lambda^2 (1/2)^2}{2}\right),$$

which shows that X is 1/2-sub-Gaussian random variable by definition. □

Theorem 3.1. *For the stochastic multi-armed bandit with Bernoulli rewards and $1 \leqslant m \leqslant n/k$, the regret of Explore-First algorithm satisfies:*

$$R_n \leqslant m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-m\Delta_i^2\right). \quad (3.2)$$

Proof. In order to obtain the upper bound, we use the regret decomposition (2.1), where we bound the expected number of times an action i has been chosen by the algorithm after n time steps $\mathbb{E}\left(N_i(n)\right)$. We follow similar steps as in [11] for the Gaussian use case.

Let $i^* = \arg\max_{i\in\mathcal{A}}\left(\mu_i\right)$ denote the optimal action. The probability of action i having the highest average reward received after mk time steps is $\mathbb{P}\left(\hat{\mu}_i(mk) \geqslant \max_{j\neq i}\hat{\mu}_j(mk)\right)$, then according to the algorithm's protocol (2.2) each action is chosen exactly m times during exploration and $n-mk$ times with the probability of having the highest average during the exploitation:

$$\begin{aligned}\mathbb{E}\left[N_i(n)\right] &\leqslant m + (n-mk)\,\mathbb{P}\left(\hat{\mu}_i(mk) \geqslant \max_{j\neq i}\hat{\mu}_j(mk)\right)\\ &\leqslant m + (n-mk)\,\mathbb{P}\left(\hat{\mu}_i(mk) \geqslant \hat{\mu_{i^*}}(mk)\right)\\ &= m + (n-mk)\,\mathbb{P}\left(\hat{\mu}_i(mk) - \hat{\mu_{i^*}}(mk) \geqslant \Delta_i - (\mu_{i^*}-\mu_i)\right)\\ &= m + (n-mk)\,\mathbb{P}\left(\hat{\mu}_i(mk) - \mu_i - (\hat{\mu_{i^*}}(mk) - \mu_{i^*}) \geqslant \Delta_i\right).\end{aligned} \tag{3.3}$$

Next, we show that $\hat{\mu}_i(mk) - \mu_i - (\hat{\mu_{i^*}}(mk) - \mu_{i^*})$ is $1/\sqrt{2m}$-sub-Gaussian by using Lemma 3.1 which states that a centered Bernoulli random variable is 1/2-sub-Gaussian. Let Y_i denote a sampled reward from action i. By definition of the algorithm the exploration phase takes mk time steps, where each action gets chosen exactly m times, we have $\hat{\mu}_i(mk) = \frac{1}{m}\sum_{j=1}^{m}\left(Y_i\right)_j$, hence

$$\begin{aligned}\hat{\mu}_i(mk) - \mu_i - (\hat{\mu_{i^*}}(mk) - \mu_{i^*}) &= \frac{1}{m}\sum_{j=1}^{m}\left(Y_i\right)_j - \mu_i - \left(\frac{1}{m}\sum_{j=1}^{m}\left(Y_{i^*}\right)_j - \mu_{i^*}\right)\\ &= \frac{1}{m}\sum_{j=1}^{m}\Big(\left(Y_i\right)_j - \mu_i\Big) - \frac{1}{m}\sum_{j=1}^{m}\Big(\left(Y_{i^*}\right)_j - \mu_{i^*}\Big),\end{aligned}$$

where $\left(Y_i\right)_j - \mu_i$ is a centered Bernoulli random variable, which is 1/2-sub-Gaussian. The additive property of σ-sub-Gaussian random variables gives us that $\hat{\mu}_i(mk) - \mu_i - (\hat{\mu_{i^*}}(mk) - \mu_{i^*})$ is $1/\sqrt{2m}$-sub-Gaussian. Applying the exponential estimate (3.1) of $1/\sqrt{2m}$-sub-Gaussian random variable to the probability on the right-hand side of inequality (3.3) we obtain

$$\mathbb{P}\left(\hat{\mu}_i(mk) - \mu_i - (\hat{\mu_{i^*}}(mk) - \mu_{i^*}) \geqslant \Delta_i\right) \leqslant \exp\left(-\frac{\Delta_i^2}{2\left(1/\sqrt{2m}\right)^2}\right) = \exp\left(-m\Delta_i^2\right).$$

Hence the inequality (3.3) takes the following form:

$$\mathbb{E}\left[N_i(n)\right] \leqslant m + (n-mk)\exp\left(-m\Delta_i^2\right),$$

substituting into the regret decomposition (2.1) completes the proof. □

4 Multi-armed bandit with Bernoulli distribution under delays

Classical algorithms of multi-armed bandit problem assume no delays in rewards from the environment, that is to say, that realization of random variable X_t must be accessible by the algorithm at the same time step t when corresponding action has been chosen. The meta-algorithm [2] provides the framework to encapsulate the classical (base) algorithms in order to use them in the stochastic environment with delayed rewards without a need for an adaptation of the base algorithms to delays. When a delay of observation of the reward realization happens, the meta-algorithm doesn't use the base algorithm on the next step for the decision making, instead, it re-uses the action chosen by the base algorithm on the previous step (sends it to the environment) until one of the realizations of the chosen action is accessible. The meta-algorithm accumulates such delayed rewards when available in the internal buffer and uses them in the further time steps to feed the base algorithm without interaction with the environment. Assume that the first action has been chosen by the algorithm $I = A_1$, then the protocol of the meta-algorithm

is defined as follows:

1) If realization of the chosen action is available in the buffer then feed it to the base algorithm and ask for the next action I

2) Send the chosen action I to the environment

3) Receive all available (delayed) realizations from the environment and put them into the buffer; go to step 1.

Due to the stochasticity of delays, order and completeness of the sequence of realizations available for the algorithm are not guaranteed. The authors show that under such circumstances there is no impact on the base algorithm in the stochastic environment, that the independent and identically distributed property of the given subsequence is preserved:

Lemma 4.1. *[2] (see Lemma 4) Consider the multi-armed bandit problem with the stochastic environemnt under delays which are independent of the rewards. For any action i, for any $s \in \mathbb{N}$ let $Y_{i,s}$ denote the s^{th} reward the algorithm observes for selecting an action i. Then the sequence $\{Y_{i,s}\}_{s\in\mathbb{N}}$ is an i.i.d. sequence with the same distribution as the sequence of the rewards $\{X_{i,t}\}_{t\in\mathbb{N}}$.*

Also, it's shown that delays increase the upper bound of regret in an additive way with respect to the maximum delay $\tau^*_{i,n}$ of action i after n time steps [2] (see *Theorem 6*):

$$\mathbb{E}\left[R_n\right] \leqslant \mathbb{E}\left[R_n^{\text{Base}}\right] + \sum_{i=1}^{k} \Delta_i \, \mathbb{E}\left[\tau^*_{i,n}\right], \qquad (4.1)$$

where R_n^{Base} is the upper bound of the base algorithm. Hence, using the obtained inequality (3.2) we can write the upper bound of the Explore-First algorithm in the stochastic environment under delays in the scope of meta-algorithm:

$$R_n \leqslant m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-m\Delta_i^2\right) + \sum_{i=1}^{k} \Delta_i \, \mathbb{E}\left[\tau^*_{i,n}\right]. \qquad (4.2)$$

5 Numerical experiments

In this chapter, we present the results of the empirical tests of the Explore-First algorithm after we implemented software to simulate the multi-armed bandit in the stochastic environment with delays.

In the first experiment, we use the instance of the MAB with Bernoulli rewards with two actions of expectations 0.77 and 0.8 respectively, this structure is sufficient to study the delay impact, not the MAB problem itself. The experiment uses a horizon of $n = 1000$ time steps and is aggregated over 1000 independent runs. Results of numerical experiments and obtained theoretical upper bounds are shown in Figure 1.

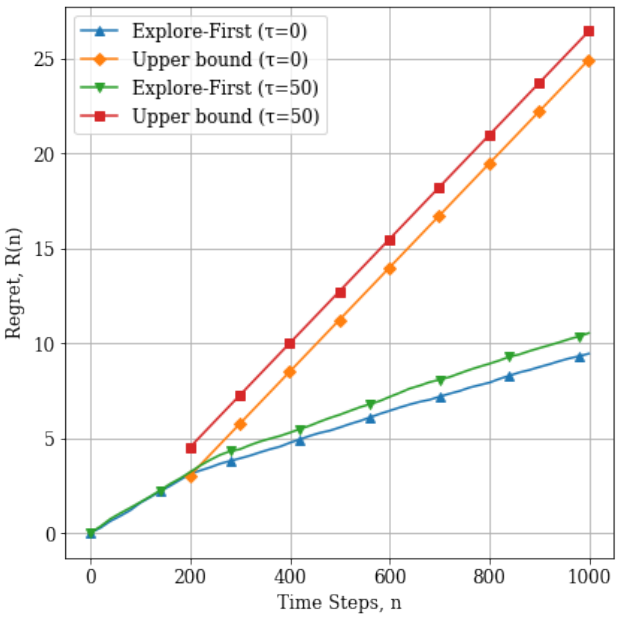

Figure 1: An empirical test of the Explore-First algorithm with delays $\tau = 0$ and $\tau = 50$ for a time horizon $n = 1000$, with the upper bounds (3.2) and (4.2)

There is a notable gap in an approximation of the upper bounds given by obtained inequalities (3.2) and (4.2) and empirical results, however, this gap is significantly smaller than in the case of inequality (2.3).

In the second experiment, we use data from a randomized clinical trial, The International Stroke Trial [12], that studied the effects of Heparin and Aspirin as combinations $H \times A$, where

$$H = \{\text{high heparin dose}, \text{low dose}, \text{no dose}\}$$
$$A = \{\text{with aspirin dose}, \text{no dose}\}.$$

19,435 patients took part in The International Stroke Trial studies, where long-term recovery (after 6 months) was being recorded. In this context, we consider recovery cases as a Bernoulli random variable, where were empirically computed the reward parameters of the original

study to obtain the mean vector across all subjects to simulate the experiment under the Bernoulli multi-armed bandit problem:

$$\mu = \{0.182, 0.178, 0.182, 0.201, 0.208, 0.206\}\,.$$

We simulate the experiment using the data given above to compare the Explore-First algorithm to the randomized controlled trial in the stochastic environment with delays. Results are aggregated over 500 independent runs and shown in Figure 2.

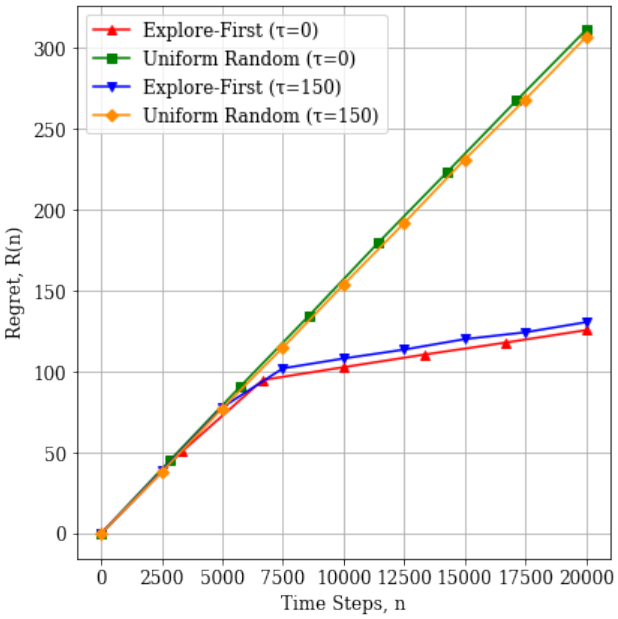

Figure 2: An empirical test of the algorithms Explore-First and Uniform Random (simulation of randomized controlled trial) with delays $\tau = 0$ and $\tau = 150$ for a time horizon $n = 20000$

The response-adaptive design has the potential to increase the effectiveness of the medical treatment within the clinical trial but often can reduce the statistical power. Another obstacle could be the delayed feedback, the existing algorithms need adaptation to such settings.

6 Conclusion

In this article, we considered the multi-armed bandit problem with Bernoulli distribution in the environment under delays. We presented an asymptotic analysis of the Explore-First algorithm in the considered settings by utilizing the exponential estimate of sub-Gaussian tail probability. Numerical experiments showed that the algorithm adapted to the environment with delays retains its theoretical guarantees by decreasing the effectiveness in an additive way with respect to the maximum delay. The developed software for simulation and numerical experiments have been published as an open-source library [13].

Список використаних джерел

1. *Thompson W. R.* On the likelihood that one unknown probability exceeds another in view of the evidence of two samples / W. R. Thompson // Biometrika. — 1933. — Vol. 25. — No. 3-4. — P. 285-294.

2. *Joulani P.* Online learning under delayed feedback / P. Joulani, A. Gyorgy, C. Szepesvri // PMLR. — 2013. — P. 1453-1461.

3. *Robbins H.* Some aspects of the sequential design of experiments / H. Robbins // Bulletin of the American Mathematical Society. — 1952. — Vol. 58. — No. 5. — P. 527-535.

4. *Lai T. L.* Asymptotically efficient adaptive allocation rules / T. L. Lai, H. Robbins // Advances in applied mathematics. — 1985. — Vol. 6 —No. 1. — P. 4–22.

References

1. THOMPSON, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*. 25 (3/4). p. 285-294.

2. JOULANI, P., GYORGY, A., & SZEPESVARI, C. (2013) Online learning under delayed feedback. In *International Conference on Machine Learning*. p. 1453-1461. PMLR.

3. ROBBINS, H. (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*. 58 (5). p. 527-535.

4. LAI, T. L., & ROBBINS, H. (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*. 6 (1). p. 4-22.

5. *Anscombe F. J.* Sequential medical trials / F. J. Anscombe // Journal of the American Statistical Association. — 1963. — Vol. 58. — No. 302. — P. 365–383.

6. *Slivkins A.* Introduction to multi-armed bandits / A. Slivkins // Foundations and Trends in Machine Learning. — 2019. — Vol. 12. — No. 1-2. — P. 1–286.

7. *Bubeck S.* Regret analysis of stochastic and nonstochastic multi-armed bandit problems / S. Bubeck, N. Cesa-Bianchi // Foundations and Trends in Machine Learning. — 2012. — Vol. 5 — No. 1. — P. 1–122.

8. *Булдыгин В. В.* Метрические характеристики случайных величин и процессов / В. В. Булдыгин, Ю. В. Козаченко. — К.: ТВіМС, 1998. — 290 с.

9. *Kozachenko Yu. V.* Simulation of stochastic processes with given accuracy and reliability / Yu. V. Kozachenko, O. O. Pogorilyak, I. V. Rozora, A. M. Tegza. — Elsevier, 2016.

10. *Hoeffding W.* Probability Inequalities for Sums of Bounded Random Variables / W. Hoeffding // Journal of the American Statistical Association. — 1963. — Vol. 58. — No. 301. — P. 13–30.

11. *Lattimore T.* Bandit algorithms / T. Lattimore, C. Szepesvari. — Cambridge University Press, 2020. — 537 p.

12. *Sandercock P.* International stroke trial collaborative Group / P. Sandercock, M. Niewada, A. Czlonkowska // The international stroke trial database. Trials. — 2011. — Vol. 12. — No. 1. — P. 101.

13. *Dzhoha A.* Multi-armed bandit problem under delayed feedback: numerical experiments [Електронний ресурс] / A. Dzhoha. — 2021. — Режим доступу до ресурсу: https://github.com/djo/delayed-bandit.

5. ANSCOMBE, F. J. (1963) Sequential medical trials. *Journal of the American Statistical Association*. 58 (302). p. 365–383.

6. SLIVKINS, A. (2019) Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*. 12 (1–2). p. 1–286.

7. BUBECK, S., & CESA-BIANCHI, N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*. 5 (1). p. 1–122.

8. BULDYGIN, V. V., KOZACHENKO, YU. V. (2000) *Metric Characterization of Random Variables and Random Processes*. AMS, Providence, RI, 257 p.

9. KOZACHENKO, YU. V., POGORILYAK, O. O., ROZORA, I. V., & TEGZA, A. M. (2016) *Simulation of stochastic processes with given accuracy and reliability*. Elsevier.

10. HOEFFDING, W. (1963) Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*. 58 (301). p. 13–30.

11. LATTIMORE, T., & SZEPESVARI, C. (2020) *Bandit algorithms*. Cambridge University Press, 537 p.

12. SANDERCOCK, P., NIEWADA, M., & CZLONKOWSKA, A. (2011) International stroke trial collaborative Group. *The international stroke trial database. Trials*. 12 (1). p. 101.

13. DZHOHA, A. (2021) *Multi-armed bandit problem under delayed feedback: numerical experiments*. [Online] Available from: https://github.com/djo/delayed-bandit.

Received: 19.02.2021