MULTI-ARMED BANDIT POLICY UNDER DELAYS FOR THE DESIGN OF CLINICAL TRIALS

A. Dzhoha[1] **and I. Rozora**[2,3]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: andrew.djoga@gmail.com

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: irozora@knu.ua

[3] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine

Abstract

Randomized controlled trials are currently considered to be the gold standard method to evaluate the effectiveness of new drugs or medical procedures. Most trials use a fixed randomization method which does not take into account the individual well-being of patients. To conduct clinical trials with the most health benefits for patients, the collected data can be used dynamically to reassign the groups to give more participants a chance for better care during trials. Such an adaptive design is a great example of using the exploration-exploitation trade-off approach. Thompson (1933) introduced the multi-armed bandit problem for this purpose.

The multi-armed bandit problem is well suited to model sequential resource allocation in the face of uncertainty. In setups like clinical trials, the response to an action is not immediate. Thus, the multi-armed bandit policies need adaptation to delays in order to retain their theoretical guarantees in a not strictly sequential environments.

By conducting simulations using the publicly available dataset The International Stroke Trial (Sandercock, Niewada, Członkowska, and the International Stroke Trial Collaborative Group 2011), we show the importance of the adaptation to delayed feedback. We study the impact on the results of experiments and provide asymptotic analysis. Thompson Sampling policy (Bubeck & Cesa-Bianchi 2012, p. 20) with Bernoulli rewards is considered the main baseline.

As another approach to mitigate the issue of delays, we propose to use the estimation of the effect of treatment as the reward feedback. We assume that such evidence of response to a drug can be collected in a relatively short-term period after the procedure and can be used to represent a certainty of successful treatment. For that, we analyze the Upper Confidence Bound policy (Bubeck & Cesa-Bianchi 2012, p. 11) with beta rewards. This policy has the potential to provide lower regret results which give more patients a chance for better care. Additionally, this policy can be corrected for Bernoulli reward or anticipated estimation error.

Keywords: multi-armed bandit, delayed feedback, Upper Confidence Bound policy.

References

Bubeck S., Cesa-Bianchi N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1–122.

Sandercock PA, Niewada M, Członkowska A, the International Stroke Trial Collaborative Group (2011) The International Stroke Trial database. *Trials*, **12**(1), 101.

Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.