Austrian Journal of Statistics
2023, Volume 52, 26–39.
http://www.ajs.or.at/
doi:10.17713/ajs.v52iSI.1751

Beta Upper Confidence Bound Policy for the Design of Clinical Trials

Andrii Dzhoha
Taras Shevchenko National University of Kyiv

Iryna Rozora
Taras Shevchenko National University of Kyiv,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Abstract

The multi-armed bandit problem is a classic example of the exploration-exploitation trade-off well suited to model sequential resource allocation under uncertainty. One of its typical motivating applications is the adaptive designs in clinical trials which modify the trial's course in accordance with the pre-specified objective by utilizing results accumulating in the trial. Since the response to a procedure in clinical trials is not immediate, the multi-armed bandit policies require adaptation to delays to retain their theoretical guarantees. In this work, we show the importance of such adaptation by evaluating policies using the publicly available dataset The International Stroke Trial of a randomized trial of aspirin and subcutaneous heparin among 19,435 patients with acute ischaemic stroke. In addition to adapted policies, we analyze the Upper Confidence Bound policy with the beta feedback to mitigate delays when the certainty evidence of successful treatment is available in a relatively short-term period after the procedure.

Keywords: multi-armed bandit, upper confidence bound policy, delayed feedback.

1. Introduction

Randomized controlled trials are considered to be the most objective method to evaluate the effectiveness of new drugs or medical treatments. As the name suggests, for such trials participants are randomly assigned to groups, often of the same sizes, each with a different treatment protocol. Comparing the results at the end of a trial provides valuable data for scientific purposes. However, this way of conducting experiments does not consider the individual well-being of participants, which can be crucial in some cases of trials on real patients with serious medical conditions.

The collected data during clinical trials can be used dynamically to reassign the groups or individual treatment to give more patients a chance for better care during trials. Developing such an adaptive design for conducting clinical trials with the most health benefits for participants is a great example of using the exploration-exploitation trade-off approach. For this

purpose, the multi-armed bandit problem was introduced by Thompson (1933).

In many practical applications, the feedback to the decision-maker is not immediate. The classical multi-armed bandit problem does not assume a delay of the feedback in the sequential decision-making process, but in such settings as clinical trials delays are inevitable. Regrettably, the existence of delays is often omitted in simulations, which can make a study not representative.

We consider the most common assumption is that outcomes can be dichotomized as successes or failures to be modeled as a Bernoulli random variable. This binary value as the response to treatment does not become available immediately. In our work, we use the publicly available dataset of The International Stroke Trial (IST) to study the implications of neglecting delays in designing numerical experiments in a clinical trial setting. We adapt policies to the delayed feedback setting and study the impact on the results of experiments and asymptotic analysis.

In addition, we propose a different way of mitigating the issue with delays, when applicable. We assume that the evidence of response to a drug can be collected in a relatively short-term period after the procedure and can represent a certainty of successful treatment. This information we model as a beta random variable and use instead of delayed Bernoulli feedback. We showcase the utility of this model in the numerical experiments using the same IST dataset. Doing asymptotic analysis, we obtain the upper bound for the proposed policy with the help of sub-Gaussian concentration inequalities.

As our secondary contribution, we publish the framework for conducting numerical experiments and analysis as an open-source project (Dzhoha 2023).

Relevant work. We consider Thompson Sampling policy (Robbins 1952) as used in the work of Stirn and Jebara (2018); Varatharajah and Berry (2022) for performing numerical experiments with the IST data. In our work, we adapt this policy to delays that naturally occurred due to the significant number of patients seen per day. For that, we employ the methodology proposed by Joulani, Gyorgy, and Szepesvari (2013) which retains theoretical guarantees of existing policies at the cost of reducing effectiveness in an additive way with respect to delays.

Alternatively, one can consider modifying the policy itself to adapt it to delays in a more efficient way for some cases (Vernade, Cappé, and Perchet 2017; Pike-Burke, Agrawal, Szepesvari, and Grunewalder 2018; Vernade, Carpentier, Lattimore, Zappella, Ermis, and Brückner 2020). But that would imply changes in the underlying algorithm when our aim is to hold on to the original policy for comparison.

For the beta reward case, we adapt the Upper Confidence Bound (UCB) policy (Auer, Cesa-Bianchi, and Fischer 2002a) and provide an asymptotic analysis. As another option, one can consider a variation of UCB based on the Kullback–Leibler divergence, KL-UCB (Garivier and Cappé 2011; Cappé, Garivier, Maillard, Munos, and Stoltz 2013). With the same probability coverage, this policy can have tighter confidence intervals than UCB. Though it comes with the cost of more complicated implementation as one has to solve an additional optimization problem of computing the index with the inequality where Kullback–Leibler divergence is involved.

Another way to model beta rewards is to consider the exponential-weighting policy Exp3 (Auer, Cesa-Bianchi, Freund, and Schapire 2002b) which is based on the importance-weighted estimation. This policy was built for the adversarial setting relaxing the assumption on the stochastic environment, so no assumptions are made about the mechanism that generates the rewards.

The structure of this paper is as follows. We start with a formulation of the problem and an overview of the policies in Section 2. Section 3 provides an asymptotic analysis of the UCB policy with beta feedback. Section 4 introduces the environment with delays. In Section 5 we evaluate the performance of the policies in the environment under delays by conducting numerical experiments with the IST data. Conclusions are given in Section 6.

2. Multi-armed bandit problem

The stationary stochastic multi-armed bandit problem is a sequential interaction between a decision-maker (policy) and an environment over discrete-time horizon T. In each time step $t \in \{1, \ldots, T\}$ the policy chooses an action A_t from an available set of N actions. Each action $i \in \{1, \ldots, N\}$ is associated with the same parametric probability distribution but with different parameters unknown to the policy. In return to action choice A_t at time step t, the environment samples a reward $X_t \in \mathbb{R}_{\geq 0}$ from a corresponding probability distribution with mean μ_{A_t}. These are the reward distributions of the corresponding actions. The most common objective of the policy in this setting is to maximize the expected cumulative reward $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$ over the whole horizon. The sequential action selection depends on the history of previously chosen actions and their results in terms of rewards:

$$H_{t-1} = (A_1, X_1, A_2, X_2, \ldots, A_{t-1}, X_{t-1}).$$

Hence, the parametric stationary stochastic multi-armed bandit model is denoted by a collection of distributions $(P_{\theta_1}, \ldots, P_{\theta_N})$, where θ_i is an unknown parameter of action i.

2.1. Regret

The objective to maximize the cumulative reward is equivalent to the so-called regret minimization introduced by Robbins (1952). The (expected) regret is a common performance measure in the asymptotic analysis of the policies. It is defined as the difference between the expected reward after choosing an optimal action at all times and the policy's expected cumulative reward (choosing the actions with respect to the policy π):

$$R(T) = T \max_{i=1,\ldots,N} \mu_i - \mathbb{E}_\pi \left[\sum_{t=1}^{T} X_t\right],$$

where $\arg\max_{i=1,\ldots,N} \mu_i$ is the optimal action.

As a variation of the regret form defined above, in analysis, it is often more useful to use the regret decomposition which is expressed as a function of the expected number of times each action is chosen multiplied by its suboptimality with respect to the best action mean.

Let $I_B(x)$ denote the indicator function of a set B and it is defined as

$$I_B(x) = \begin{cases} 1 & x \in B \\ 0 & x \notin B. \end{cases}$$

Then, the regret decomposition is defined as follows (Lai and Robbins 1985):

$$R(T) = \sum_{i=1}^{N} \Delta_i \, \mathbb{E}\left[\sum_{t=1}^{T} I_{\{i\}}(A_t)\right], \tag{1}$$

where Δ_i is called a suboptimality gap of action i and expressed as

$$\Delta_i = \max_{j=1,\ldots,N} \mu_j - \mu_i.$$

2.2. Asymptotic analysis

The asymptotic analysis of the stationary stochastic multi-armed bandit model was pioneered by Lai and Robbins (1985). The authors stated that any policy suffers at least logarithmic regret on any parametric problem. One objective in designing an efficient policy is to achieve a sublinear regret in the upper bound:

$$\lim_{T \to \infty} \frac{R(T)}{T} = 0.$$

If the lower bound asymptotically matches the upper bound, a policy is considered asymptotically optimal, formally such that:

$$\sup_{T\to\infty} \frac{R(T)}{\log(T)} \leq \sum_{i=2}^{N} \frac{\Delta_i}{\mathrm{KL}(P_{\theta_i}, P_{\theta_1})},$$

where KL is Kullback-Leibler divergence, and assuming without a loss of generality that the first action ($i = 1$) is optimal. A significant part of the research on the multi-armed bandit problem has focused on finding such policies.

In this paper, we consider two asymptotically optimal policies, Thompson Sampling and Upper Confidence Bound, described and analyzed further.

2.3. Thompson sampling policy

As a baseline in our experiments, we use Thompson Sampling policy with Bernoulli rewards. Kaufmann, Korda, and Munos (2012) showed that this policy is asymptotically optimal and it is considered one of the best performers for a stationary stochastic environment with Bernoulli rewards (see Granmo (2010)).

In this setting, each action draw is considered a Bernoulli trial with the output set $\{0, 1\}$. Rewards $\{ X_t \in \{0, 1\} : t \in \{1, \ldots, T\} \wedge A_t = i \}$ are Bernoulli random variables with the probability of success θ_i of the action i, and $\hat{\theta}_i$ is its estimator by previously observed data.

The policy is based on the Bayesian principles. First, a prior distribution on means $(\mu_i)_i$ is assigned. At each time step, the policy samples $\hat{\theta}_i \sim \pi_i$ for each action i from the posterior distribution π_i on the parameter μ_i:

$$\pi_i = \mathrm{Beta}(\alpha_i, \beta_i),$$

with the beta distribution as a conjugate prior for the Bernoulli distribution. The parameter α_i is the number of successes (event $\{1\}$, meaning successful treatment in our case) when choosing action i, and β_i is the number of failures. Then, the action A_t is chosen in accordance with its posterior probability of being optimal:

$$A_t = \underset{i=1,\ldots,N}{\arg\max}\, \hat{\theta}_i.$$

The posterior concentrates towards the true environment with more data collected.

Agrawal and Goyal (2012) showed that in the stationary stochastic multi-armed bandit model, the distribution-dependent regret of Thompson Sampling policy using a uniform prior satisfies:

$$R(T) \leq \left(\sum_{i=2}^{N} \frac{1}{\Delta_i^2} \right)^2 \log(T), \tag{2}$$

assuming without a loss of generality that the first action is optimal.

The algorithm of Thompson Sampling is summarized in Algorithm 1.

2.4. Upper confidence bound policy

In addition to the Bernoulli rewards setting, we consider a new model with the beta rewards $X_t \in [0, 1]$. Each action i is associated with the beta distribution with the same mean θ_i as in the Bernoulli setting. We model the certainty of successful treatment, which we assume is estimated from the observation of the response to a drug in a relatively short-term period. For simplicity, we assume an accurate and in-time provided estimation. In the discussion of Section 5, we elaborate on how those assumptions can be relaxed to make it practical.

For this setting, we will use Upper Confidence Bound policy. It is based on the principle of "optimism in the face of uncertainty", which makes you assume a "nice" environment where each action is as good as plausibly possible given the observations so far.

Algorithm 1 Thompson Sampling

For each $i \in \{1, \dots, N\}$ set $\alpha_i = 1$, $\beta_i = 1$
for time step $t = 1, \dots, T$ **do**
  for action $i = 1, \dots, N$ **do**
    $\hat{\theta}_i \sim \text{Beta}(\alpha_i, \beta_i)$
  end for
  Draw action $A_t = \arg\max_{i=1,\dots,N} \hat{\theta}_i$ and observe reward X_t
  $\alpha_{A_t} \leftarrow \alpha_{A_t} + X_t$
  $\beta_{A_t} \leftarrow \beta_{A_t} + (1 - X_t)$
end for

Auer *et al.* (2002a) shaped and analyzed this policy for bounded rewards. The authors showed that it is asymptotically optimal.

UCB belongs to the index-based family of policies, as at each time step t they compute an index $U_i(t)$ for each action i. Then, an action with the largest index is chosen. This index represents with high probability an overestimate of the unknown action's mean. It is taken as the upper confidence bound using Hoeffding's inequality to build confidence intervals. The index $U_i(t)$ is computed as a sum of the average reward from observations collected so far and a confidence radius:

$$U_i(t) = \frac{1}{\sum_{p=1}^{t} I_{\{i\}}(A_p)} \sum_{p=1}^{t} I_{\{i\}}(A_p) X_p + \sqrt{\alpha \frac{\log(1/\delta)}{\sum_{p=1}^{t} I_{\{i\}}(A_p)}}, \tag{3}$$

where α and δ are constant parameters to additionally control the exploration-exploitation trade-off. The policy chooses the largest index $U_i(t)$ to select a corresponding action as A_t. The action i can be chosen because of the large average reward which converges towards the true (potentially large) mean with more observations, or because the confidence interval is large (wide margin of error) indicating an insufficient sample size. Hence, the first term in (3) represents the exploitation part, while the second one is the exploration.

The parameter α can be derived directly from Hoeffding's inequality or tuned according to expectations. The confidence level δ is recommended to be chosen a bit smaller than $1/T$.

The algorithm of UCB is summarized in Algorithm 2.

Algorithm 2 Upper Confidence Bound

for time step $t = 1, 2, \dots, N$ **do**
  Draw action $A_t = t$ and observe reward X_t
end for
for time step $t = N + 1, N + 2, \dots, T$ **do**
  for action $i = 1, \dots, N$ **do**
    $U_i(t) = \frac{1}{\sum_{p=1}^{t} I_{\{i\}}(A_p)} \sum_{p=1}^{t} I_{\{i\}}(A_p) X_p + \sqrt{\alpha \frac{\log(1/\delta)}{\sum_{p=1}^{t} I_{\{i\}}(A_p)}}$
  end for
  Draw action $A_t = \arg\max_{i=1,\dots,N} U_i(t)$ and observe reward X_t
end for

3. Beta upper confidence bound analysis

In this section, we show an upper bound for UCB policy with beta rewards with the help of sub-Gaussian concentration inequalities. A random variable X with $\mathbb{E}[X] = 0$ and a positive parameter σ is said to be σ-sub-Gaussian if for all $\lambda \in \mathbb{R}$, it holds that

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

The exponential estimate of σ-sub-Gaussian random variable X for any $\varepsilon \geq 0$ satisfies

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right),$$

and is applicable to any compactly supported random variable such as a centered beta random variable. Let X have a support on the interval $[a, b]$, then for all $\lambda \in \mathbb{R}$ Hoeffding's lemma (Hoeffding 1963) states the following:

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

applying it to a centered beta random variable gives us that X is $1/2$-sub-Gaussian. Then, for independent beta random variables $X_1, X_2, \ldots, X_T$ with mean μ, it holds that the difference between sample and population means

$$\frac{1}{T}\sum_{t=1}^{T} X_t - \mu = \sum_{t=1}^{T}(X_t - \mu)/T \quad \text{(here } X_t - \mu \text{ is } 1/2\text{-sub-Gaussian)}$$

is $1/2\sqrt{T}$-sub-Gaussian. It can be shown by applying additive properties of σ-sub-Gaussian random variables (Kozachenko, Pogorilyak, Rozora, and Tegza 2016), that $\sum_{t=1}^{T} X_t - \mu$ is $\sqrt{T\sigma^2}$-sub-Gaussian.

Therefore, we have bounds on the tails of $\frac{1}{T}\sum_{t=1}^{T} X_t - \mu$ with $(X_t)_{t=1}^{T}$ being independent beta random variables with mean μ as follows:

$$\begin{gathered}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t \geq \mu + \varepsilon\right) \leq \exp\left(-2T\varepsilon^2\right) \\ \text{and} \\ \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t \leq \mu - \varepsilon\right) \leq \exp\left(-2T\varepsilon^2\right).\end{gathered} \tag{4}$$

We will use these sub-Gaussian characteristics to upper bound the expected regret for our use case with beta rewards. The technique of obtaining the upper bound, that follows, is common for such policies and can be found in Bubeck (2012); Lattimore and Szepesvári (2020).

For more details on sub-Gaussian random variables and their properties, we refer the readers to Buldygin and Kozachenko (2000).

Theorem 1. *Consider the stationary stochastic multi-armed bandit problem with beta rewards. Assume without a loss of generality that the first action ($i = 1$) is optimal. For UCB policy with the confidence level $\delta = 1/T^2$, the distribution-dependent regret satisfies*

$$R(T) \leq 2\sum_{i=2}^{N} \Delta_i + \sum_{i=2}^{N} \frac{\log(T)}{2\Delta_i}.$$

Proof. Let $(X_t)_{t=1}^{T}$ be a sequence of independent beta random variables with mean μ. We will use a common deviation form of (4) in the form of the confidence interval. For any $\delta \in [0, 1]$, with probability at least $1 - \delta$, we have

$$\mu \in \left[\frac{1}{T}\sum_{t=1}^{T} X_t - \sqrt{\frac{\log(1/\delta)}{2T}}, \frac{1}{T}\sum_{t=1}^{T} X_t + \sqrt{\frac{\log(1/\delta)}{2T}}\right],$$

which gives us an upper bound for UCB policy with beta rewards to construct the index.

Thus, according to Algorithm 2, we express an upper bound estimate of action i at time step t over subsequence of all rewards $(X_p)_{p=1}^t$ as the following index:

$$U_i(t) = \begin{cases} \infty & \text{if } \sum_{p=1}^t I_{\{i\}}(A_p) = 0 \\ \frac{1}{\sum_{p=1}^t I_{\{i\}}(A_p)} \sum_{p=1}^t I_{\{i\}}(A_p) X_p + \sqrt{\frac{\log(T^2)}{2\sum_{p=1}^t I_{\{i\}}(A_p)}} & \text{otherwise.} \end{cases}$$

To obtain the upper bound, we will use the regret decomposition definition (1) in which our goal is to bound the expected number of times a suboptimal action $i > 1$ gets chosen over the whole horizon T. Recall that the policy's objective is to maximize the total reward $\sum_{t=1}^T X_t$ which leads to identifying the optimal action as early as possible, hence, minimizing the required number of samples $\sum_{t=1}^T I_{\{i\}}(A_t)$ for $i > 1$.

Let B_i define the event when the upper bound estimate of action $i > 1$ after c_i samples is below the optimal mean and the upper bound estimate of the optimal action is never underestimated with respect to its mean:

$$B_i = \left\{ U_i(T) < \mu_1 \mid \sum_{t=1}^T I_{\{i\}}(A_t) = c_i \right\} \cap \left\{ \mu_1 < \min_{t=1,\dots,T} U_1(t) \right\}.$$

We are naturally interested in high-probability event B_i with a small as possible sample size c_i. Then, using the law of total expectation, we can bound the regret decomposition (1) in the following way:

$$\begin{aligned} R(T) &= \sum_{i=2}^N \Delta_i \, \mathbb{E}\left[\sum_{t=1}^T I_{\{i\}}(A_t) \right] \\ &= \sum_{i=2}^N \Delta_i \left(\mathbb{E}\left[\sum_{t=1}^T I_{\{i\}}(A_t) \mid B_i \right] \mathbb{P}(B_i) + \mathbb{E}\left[\sum_{t=1}^T I_{\{i\}}(A_t) \mid B_i^c \right] \mathbb{P}(B_i^c) \right) \\ &\leq \sum_{i=2}^N \Delta_i \left(c_i + \mathbb{P}\left(B_i^c\right) T \right), \end{aligned} \tag{5}$$

where B_i^c is the complement of B_i and by its definition,

$$B_i^c = \left\{ U_i(T) \geq \mu_1 \mid \sum_{t=1}^T I_{\{i\}}(A_t) = c_i \right\} \cup \left\{ \mu_1 \geq \min_{t=1,\dots,T} U_1(t) \right\}. \tag{6}$$

The probability of the second set from (6) is decomposed and bounded by (4):

$$\begin{aligned} \mathbb{P}\left(\mu_1 \geq \min_{t=1,\dots,T} U_1(t) \right) &\leq \sum_{t=1}^T \mathbb{P}\left(\mu_1 \geq U_1(t) \right) \\ &= \sum_{t=1}^T \mathbb{P}\left(\mu_1 \geq \frac{1}{\sum_{p=1}^t I_{\{1\}}(A_p)} \sum_{p=1}^t I_{\{1\}}(A_p) X_p + \sqrt{\frac{\log(T^2)}{2\sum_{p=1}^t I_{\{1\}}(A_p)}} \right) \\ &\leq \sum_{t=1}^T \exp\left(-2 \sqrt{\frac{\log(T^2)}{2\sum_{p=1}^t I_{\{1\}}(A_p)}}^2 \sum_{p=1}^t I_{\{1\}}(A_p) \right) \\ &= \frac{1}{T}, \end{aligned}$$

putting it into (5) gives us

$$R(T) \leq \sum_{i=2}^N \Delta_i \left(c_i + 1 + \mathbb{P}\left(U_i(T) \geq \mu_1 \mid \sum_{t=1}^T I_{\{i\}}(A_t) = c_i \right) T \right). \tag{7}$$

The remaining randomness, the probability of the first set from (6), we bound in a similar way with the help of exponential estimation (4) and then choose c_i. For that, we use the definition of suboptimality gap Δ_i from regret decomposition (1):

$$\begin{aligned}
&\mathbb{P}\left(U_i(T) \geq \mu_1 \mid \sum_{t=1}^{T} I_{\{i\}}(A_t) = c_i\right) \\
&\quad = \mathbb{P}\left(U_i(T) \geq \mu_i + \Delta_i \mid \sum_{t=1}^{T} I_{\{i\}}(A_t) = c_i\right) \\
&\quad = \mathbb{P}\left(\frac{1}{\sum_{t=1}^{T} I_{\{i\}}(A_t)} \sum_{t=1}^{T} I_{\{i\}}(A_t) X_t + \sqrt{\frac{\log(T^2)}{2\sum_{t=1}^{T} I_{\{i\}}(A_t)}} \geq \mu_i + \Delta_i \mid \sum_{t=1}^{T} I_{\{i\}}(A_t) = c_i\right) \\
&\quad \leq \exp\left(-2c_i \left(\Delta_i - \sqrt{\frac{\log(T^2)}{2c_i}}\right)^2\right).
\end{aligned}$$

We know that the regret bound of UCB in general case is logarithmic in T. In order to get rid of linear complexity in (7), let us solve the following equation:

$$\exp\left(-2c_i \left(\Delta_i - \sqrt{\frac{\log(T^2)}{2c_i}}\right)^2\right) = \frac{1}{T},$$

from elementary calculus gives us $c_i = \frac{\left(3 \pm 2\sqrt{2}\right)\log(T)}{2\Delta_i^2}$. We take the smallest integer rounding up to $\log(T)/2\Delta_i^2$. Placing it together with the bound $1/T$ into (7) completes the proof. □

4. Environment with delays

Classical multi-armed bandit policies assume no delays in rewards to retain their theoretical guarantees. The next decision on A_{t+1} happens after receiving a realization of X_t from the environment. Hence, the algorithms need to be adapted to the environment with delays.

Joulani *et al.* (2013) provided the framework to encapsulate the existing policies without their adaptation in the stochastic environment with stochastic delays at the cost of reducing effectiveness in an additive way with respect to delays. The framework is placed between the policy and the environment with delays. When a delay occurs, the framework does not use the policy on the next time step for decision-making. Instead, the framework re-uses the policy's action choice from the previous step. It does so until one of the reward realizations of that action is accessible. With time, the framework accumulates the delayed rewards and uses them further to feed the policy without interaction with the environment. Due to the stochasticity of delays, the order and completeness of the sequence of realizations are not guaranteed. However, the authors showed that such circumstances have no impact on the policy since the rewards in the given subsequence are still independent and identically distributed.

The final upper bound is expected to increase in an additive way with respect to the maximum delay τ_i in time steps occurred per action:

$$R(T) \leq R^{\text{Policy}}(T) + \sum_{i=1}^{N} \Delta_i \, \mathbb{E}\left[\tau_i\right].$$

In our experiments, we use Thompson Sampling policy with Bernoulli rewards by placing it into Joulani et al.'s framework. We consider that our beta UCB policy is not impacted by delays because of an assumption that an accurate estimation of the reward is provided

Table 1: The upper bounds of the variants for the experiments. Assuming without a loss of generality that the first action is optimal.

Policy	Expected Regret Upper Bound
Uniformly at Random	$R(T) = \frac{T}{N}\sum_{i=1}^{N}\Delta_i$
Bernoulli Thompson Sampling with delays	$R(T) \leq \left(\sum_{i=2}^{N}\frac{1}{\Delta_i^2}\right)^2 \log(T) + \sum_{i=1}^{N}\Delta_i\,\mathbb{E}\left[\tau_i\right]$
Beta Upper Confidence Bound	$R(T) \leq 2\sum_{i=2}^{N}\Delta_i + \sum_{i=2}^{N}\frac{\log(T)}{2\Delta_i}$

promptly before the next patient arrives. To simulate the traditional randomized controlled trial, we use a uniform policy that chooses an action uniformly at random. The expected regret is linear:

$$R(T) = \sum_{i=1}^{N}\Delta_i\,\mathbb{E}\left[\sum_{t=1}^{T} I_{\{i\}}(A_t)\right] = \sum_{i=1}^{N}\Delta_i T\,\mathbb{P}\left(A_1 = i\right) = \frac{T}{N}\sum_{i=1}^{N}\Delta_i.$$

The upper bounds of the variants for the experiments are summarized in Table 1.

5. Numerical experiments

In this section, we conduct numerical experiments using data from a randomized clinical trial (Sandercock, Niewada, Członkowska, and the International Stroke Trial Collaborative Group 2011). The International Stroke Trial (IST) studied the effects of aspirin and heparin on stroke victims recording short-term (discharged alive in 14 days) and long-term (fully recovered at 6 months) outcomes. The objective was to establish whether the early administration of aspirin or heparin influenced the clinical course of acute ischemic stroke. The IST dataset contains outcomes for all 19,435 admitted patients. It was stated in the report (IST 1997) that aspirin-allocated patients had significantly fewer recurrent ischaemic strokes within 14 days with a significant reduction in death or non-fatal recurrent stroke, so it was suggested that the aspirin treatment should begin as soon as possible after the onset of ischaemic stroke. For the other findings, we refer the readers to the report itself.

As in the work of Stirn and Jebara (2018); Varatharajah and Berry (2022), we simulate the experiment by empirically computing the true reward parameters across all subjects from the available results and consider the survival outcome as a Bernoulli reward variable. The main difference in the methodology is that we take into account delays that occur because the survival (or not) confirmation depends on the study term (or fatal event date).

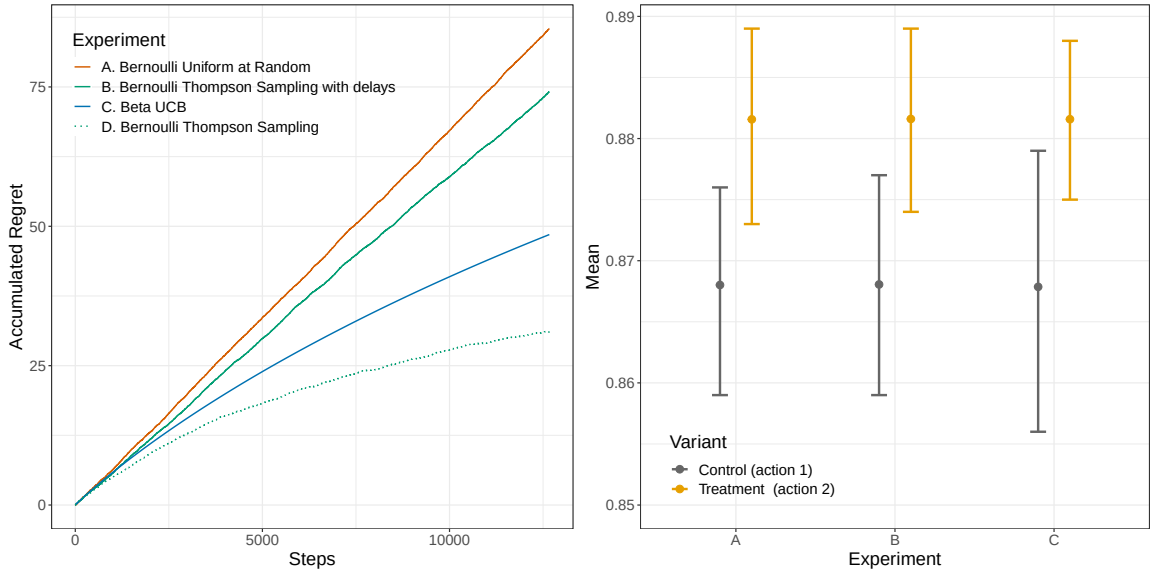
For simplicity and explanatory reasons, we focus on the elderly group of patients analyzing their short-term results (within 14 days) of aspirin treatment (control is no aspirin), which gives us the following means for the Bernoulli multi-armed bandit model:

$$\begin{aligned} \mu_1 &= 0.868, \qquad \text{(control)} \\ \mu_2 &= 0.882. \quad \text{(treatment)} \end{aligned}$$

In the same way, we calculate parameters α_i and β_i for the model with beta rewards:

$$\begin{aligned} \alpha_1 &= 5498, \quad \beta_1 = 836, \qquad \text{(control)} \\ \alpha_2 &= 5584, \quad \beta_2 = 750. \quad \text{(treatment)} \end{aligned}$$

During the trial, there were daily registered on average 11 patients in our groups. Therefore, we use $\tau_i = 11 \cdot 14 = 154$ time steps as a delay for the successful treatment reward $X_t = 1$. For the reward of $X_t = 0$, we sample a delay in accordance with the distribution of the remaining



(a) Regret (b) 95% confidence interval error bars

Figure 1: Results of the experiments using The International Stroke Trial dataset. (a) Accumulative regret of the policies. (b) 95% confidence interval error bars of action parameters per experiments A, B, and C.

lifetime in the case of death provided by data. This distribution can be approximated by the Weibull distribution with the shape parameter $\gamma = 1.2$ and scale parameter $\beta = 11.7$ obtained by performing distribution fitting with maximum likelihood. Lastly, having 6334 patients per group constitutes horizon $T = 12,668$ for our experiments.

We run four experiments in the following multi-armed bandit model and policy settings:

A. Uniform at Random policy with Bernoulli rewards. There is no impact of delays on the policy. It is a simulation of a randomized controlled trial.

B. Thompson Sampling policy with Bernoulli rewards adapted to delays.

C. UCB policy with beta rewards. We assume no delays in the beta rewards setting.

D. Thompson Sampling policy with Bernoulli rewards neglecting delays for illustrative purposes of not-properly designed simulation.

The results of each experiment are averaged over 2000 independent runs and presented in Figure 1. To analyze the results, we use a two-sample t-test for a two-sided test with a significance level set at 5%.

5.1. Results and discussion

Bernoulli rewards. As expected, Thompson Sampling with Bernoulli rewards neglecting delays (D) performed the best in terms of regret. Unfortunately, this setting is not representative for such studies' simulation because of the significant impact of delays that occur(ed) in reality. Thompson Sampling adapted to delays (B) resulted in significantly fewer allocations toward the superior treatment (55% vs 82%). The contributing factor to this difference is the second term in Thompson Sampling regret (see Table 1). Compared to the Uniform at Random policy (A), the adapted Thompson Sampling still brings an improvement in the average number of successes (55% vs 50%) with approximately the same ad-hoc (observed) power (88% vs 90%).

Beta rewards. Both adaptive design methods (B and C) performed well judging solely on patient outcomes compared to the traditional fixed randomized controlled trial approach (A).

The Beta UCB method (C) improved the allocation towards the superior treatment further (71% vs 55%) compared to B but at the cost of reducing the power to detect a significant treatment effect (78% vs 88%), which can be observed in Figure 1 (wider confidence intervals). This is an illustration of the natural tension between the two conflicting goals of maximizing the superior treatment allocation (health of the patients) and maximizing the statistical power to detect a significant treatment effect. In sum, both methods were able to solve the trade-off successfully, in the sense of achieving ∼80% power.

Reviewing the beta UCB setting, we admit that the assumption of an accurate and in-time provided estimation of reward is overconfident. The setting is more to show room for improvement. This assumption can be relaxed, and as a further enhancement, one can consider making a backward correction of the policy's estimations when the true reward gets materialized. We can not change the chosen actions in the past, but nothing stops us from changing observed reward values which are incorporated into the action estimations.

In the case of the contextual multi-armed bandit problem, when a reward is additionally conditioned on side information, we can utilize the beta UCB algorithm per context as described by Dzhoha and Rozora (2023). In clinical trials, such property as context (group) can be well inherited by patients and has a significant impact on the outcome. For example, in the considered IST experiment, one can take into account patient characteristics similar to the work of Varatharajah and Berry (2022).

We publish our framework for the numerical experiments of all studied policies, their results, and the exploratory analysis of the IST dataset as an open-source project (Dzhoha 2023). Python (Van Rossum and Drake Jr 1995) and R (R Core Team 2022) are used.

6. Conclusion

In this paper, we showed the impact of delays on the performance of the policies. A policy, adapted to the delays, can retain its theoretical guarantees but at the cost of reducing effectiveness in an additive way with respect to delays. Evaluating the policies using The International Stroke Trial dataset, we confirmed that the delays can be a significant contributing factor to their performance. To mitigate delays when the certainty evidence of successful treatment is available in a relatively short-term period after the procedure, we analyzed the Upper Confidence Bound policy with beta rewards. This policy can provide benefits in lower regret giving more patients a chance for better care during trials. Additionally, the correction for Bernoulli reward or estimation error can be considered.

References

Agrawal S, Goyal N (2012). "Analysis of Thompson Sampling for the Multi-Armed Bandit Problem." In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings.

Auer P, Cesa-Bianchi N, Fischer P (2002a). "Finite-Time Analysis of the Multiarmed Bandit Problem." *Machine Learning*, **47**(2), 235–256. doi:10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Auer P, Cesa-Bianchi N, Freund Y, Schapire RE (2002b). "The Nonstochastic Multi-armed Bandit Problem." *SIAM Journal on Computing*, **32**(1), 48–77. doi:10.1137/S0097539701398375. https://doi.org/10.1137/S0097539701398375, URL https://doi.org/10.1137/S0097539701398375.

Bubeck S (2012). "Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems." *Foundations and Trends in Machine Learning*, **5**(1), 1–122. ISSN 1935-8237,

1935-8245. doi:10.1561/2200000024. URL http://www.nowpublishers.com/article/Details/MAL-024.

Buldygin VV, Kozachenko YV (2000). *Metric Characterization of Random Variables and Random Processes*, volume 188. American Mathematical Society.

Cappé O, Garivier A, Maillard OA, Munos R, Stoltz G (2013). "Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation." *Annals of Statistics*, **41**(3), 1516–1541. doi:10.1214/13-AOS1119. URL https://hal.science/hal-00738209.

Dzhoha A (2023). "Multi-Armed Bandit (MAB) Problem under Delayed Feedback: Numerical Experiments." https://github.com/djo/delayed-bandit/. Accessed 2023-04-12.

Dzhoha A, Rozora I (2023). "Multi-Armed Bandit Problem with Online Clustering as Side Information." *Journal of Computational and Applied Mathematics*, **427**, 115132. ISSN 0377-0427. doi:https://doi.org/10.1016/j.cam.2023.115132. URL https://www.sciencedirect.com/science/article/pii/S0377042723000766.

Garivier A, Cappé O (2011). "The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond." In SM Kakade, U von Luxburg (eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 359–376. PMLR, Budapest, Hungary. URL https://proceedings.mlr.press/v19/garivier11a.html.

Granmo OC (2010). "Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton." *International Journal of Intelligent Computing and Cybernetics*, **2**(3), 207–234.

Hoeffding W (1963). "Probability Inequalities for Sums of Bounded Random Variables." *Journal of the American Statistical Association*, **58**(301), 13–30. ISSN 01621459. URL http://www.jstor.org/stable/2282952.

IST (1997). "The International Stroke Trial (IST): A Randomised Trial of Aspirin, Subcutaneous Heparin, Both, or Neither Among 19,435 Patients with Acute Ischaemic Stroke." *The Lancet*, **349**(9065), 1569–1581. doi:10.1016/S0140-6736(97)04011-7. URL https://doi.org/10.1016/S0140-6736(97)04011-7.

Joulani P, Gyorgy A, Szepesvari C (2013). "Online Learning under Delayed Feedback." In S Dasgupta, D McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461. PMLR, Atlanta, Georgia, USA. URL https://proceedings.mlr.press/v28/joulani13.html.

Kaufmann E, Korda N, Munos R (2012). "Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis." In *International conference on algorithmic learning theory*, pp. 199–213. Springer.

Kozachenko Y, Pogorilyak O, Rozora I, Tegza A (2016). "1 - The Distribution of the Estimates for the Norm of Sub-Gaussian Stochastic Processes." In Y Kozachenko, O Pogorilyak, I Rozora, A Tegza (eds.), *Simulation of Stochastic Processes with Given Accuracy and Reliability*, pp. 1–70. Elsevier. ISBN 978-1-78548-217-5. doi:https://doi.org/10.1016/B978-1-78548-217-5.50001-5. URL https://www.sciencedirect.com/science/article/pii/B9781785482175500015.

Lai T, Robbins H (1985). "Asymptotically Efficient Adaptive Allocation Rules." *Advances in Applied Mathematics*, **6**(1), 4–22. ISSN 0196-8858. doi:https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Lattimore T, Szepesvári C (2020). *Bandit Algorithms*. Cambridge University Press. doi:10.1017/9781108571401.

Pike-Burke C, Agrawal S, Szepesvari C, Grunewalder S (2018). "Bandits with Delayed, Aggregated Anonymous Feedback." In J Dy, A Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4105–4113. PMLR. URL https://proceedings.mlr.press/v80/pike-burke18a.html.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robbins H (1952). "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society*, **58**(5), 527 – 535.

Sandercock PAG, Niewada M, Członkowska A, the International Stroke Trial Collaborative Group (2011). "The International Stroke Trial Database." *Trials*, **12**(1), 101. doi:10.1186/1745-6215-12-101. URL https://doi.org/10.1186/1745-6215-12-101.

Stirn A, Jebara T (2018). "Thompson Sampling for Noncompliant Bandits." 1812.00856.

Thompson WR (1933). "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples." *Biometrika*, **25**(3-4), 285–294. ISSN 0006-3444, 1464-3510. doi:10.1093/biomet/25.3-4.285. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/25.3-4.285.

Van Rossum G, Drake Jr FL (1995). *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Varatharajah Y, Berry B (2022). "A Contextual-Bandit-Based Approach for Informed Decision-Making in Clinical Trials." *Life*, **12**(8). ISSN 2075-1729. doi:10.3390/life12081277. URL https://www.mdpi.com/2075-1729/12/8/1277.

Vernade C, Cappé O, Perchet V (2017). "Stochastic Bandit Models for Delayed Conversions." In *Conference on Uncertainty in Artificial Intelligence*. Sydney, Australia. URL https://hal.science/hal-01545667.

Vernade C, Carpentier A, Lattimore T, Zappella G, Ermis B, Brückner M (2020). "Linear Bandits with Stochastic Delayed Feedback." In HD III, A Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9712–9721. PMLR. URL https://proceedings.mlr.press/v119/vernade20a.html.

Affiliation:

Andrii Dzhoha
Taras Shevchenko National University of Kyiv
Department of Applied Statistics
64/13 Volodymyrska St., Kyiv, 01601, Ukraine
E-mail: andrew.djoga@gmail.com

Iryna Rozora
Taras Shevchenko National University of Kyiv
Department of Applied Statistics
64/13 Volodymyrska St., Kyiv, 01601, Ukraine
E-mail: irozora@knu.ua
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
Department of Mathematical Analysis and Probability Theory
Peremogy Ave. 37, Kyiv 03056, Ukraine