Statistics and Probability Letters 164 (2020) 108818

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards

Sakshi Arya *, Yuhong Yang

School of Statistics, University of Minnesota, Ford Hall, Church St SE, Minneapolis, MN, United States

ARTICLE INFO

Article history:
Received 12 January 2019
Received in revised form 2 September 2019
Accepted 10 May 2020
Available online 15 May 2020

Keywords:
Multi-armed bandit with covariates
Delayed rewards
Histogram method
Strong consistency

ABSTRACT

We study a multi-armed bandit problem with covariates in a setting where there is a possible delay in observing the rewards. Under some reasonable assumptions on the probability distributions for the delays and using an appropriate randomization to select the arms, the proposed strategy is shown to be strongly consistent.

1. Introduction

Multi-armed bandits were first introduced in the landmark paper by Robbins (1952). The development of multi-armed bandit methodology has been partly motivated by clinical trials with the aim of balancing two competing goals, (1) to effectively identify the best treatment (exploration) and (2) to treat patients as effectively as possible during the trial (exploitation).

The classic formulation of the multi-armed bandit problem in the context of clinical practice is as follows: there are $\ell \geq 2$ treatments (arms) to treat a disease. The doctor (decision maker) has to choose for each patient, one of the ℓ available treatments, which result in a reward (response) of improvement in the condition of the patient. The goal is to maximize the cumulated rewards as much as possible. In the classic multi-armed bandit terminology, this is achieved by devising a policy for sequentially pulling arms out of the ℓ available arms, with the goal of maximizing the total cumulative reward, or minimizing the regret. Substantial amount of work has been done both on standard context-free bandit problems (Gittins, 1979; Berry and Fristedt, 1985; Lai and Robbins, 1985; Auer et al., 2002) and on contextual bandits or multi-armed bandits with covariates (MABC) (Woodroofe, 1979; Sarkar, 1991; Yang and Zhu, 2002; Langford and Zhang, 2008; Li et al., 2010; Slivkins, 2014). The MABC problems have been studied in both parametric and nonparametric frameworks. Our work follows nonparametric framework of MABC in Yang and Zhu (2002) where the randomized strategy is an annealed ϵ-greedy strategy, which is a popular heuristic in bandits literature (Sutton and Barto, 2018, Chapter 2). Some of the other notable work in studying finite time analysis for MABC problems in a nonparametric framework are Perchet and Rigollet (2013), Qian and Yang (2016a,b). Some insightful overviews and bibliographic remarks can be found in Bubeck and Cesa-Bianchi (2012), Cesa-Bianchi and Lugosi (2006) and Lattimore and Szepesvári (2018).

In most multi-armed bandit settings it is assumed that the rewards related to each treatment allocation are achieved before the next patient arrives. This is not realistic since in most cases the treatment effect is seen at some delayed time

* Corresponding author.
E-mail address: aryax010@umn.edu (S. Arya).

https://doi.org/10.1016/j.spl.2020.108818
0167-7152/

after the treatment is provided. Most often, it would be the case that while waiting for treatment results of one patient, other patients would have to be treated. In such a situation, all past patient information and feedback is not yet available to make the best treatment choices for the patients being treated at present.

While an overwhelming amount of work has been done assuming instantaneous observations in both contextual and non-contextual multi-armed bandit problems, not much work has been done for the case with delayed rewards. The importance of considering delays was highlighted by Anderson (1964) and Suzuki (1966). They used Bayesian multi-armed bandits to devise optimal policies. Thompson sampling (Agrawal and Goyal, 2012; Russo et al., 2018) is another commonly used Bayesian heuristic. Chapelle and Li (2011) conducted an empirical study to illustrate robustness of Thompson sampling in the case of constant delayed feedback. Most of the work that has been done in the recent years is motivated by reward delays in online settings like advertisement and news article recommendations. Dudik et al. (2011) considered a constant known delay which resulted in an additional additive penalty in the regret for the setting with covariates. Joulani et al. (2013) propose some black box multi-armed bandit algorithms that use the algorithms for the non-delayed case to handle the delayed case. Their finite time results show an additive increase in the regret for stochastic multi-armed bandit problems. More recently, Pike-Burke et al. (2018) proposed a variant of delayed bandits with aggregated anonymous feedback. They show that with their proposed algorithm and with the knowledge of the expected delay, an additive regret increase like in Joulani et al. (2013) can still be maintained. Some other work related to delayed bandits can be found in Mandel et al. (2015), Cesa-Bianchi et al. (2016) and Vernade et al. (2017).

In our knowledge, there does not seem to be any work on delayed MABCs using a nonparametric framework. In this work, we propose an algorithm accounting for delayed rewards with optimal treatment decision making as the motivation. We use nonparametric estimation to estimate the functional relationship between the rewards and the covariates. We show that the proposed algorithm is strongly consistent in that the cumulated rewards almost surely converge to the optimal cumulated rewards.

2. Problem setup

Assume that there are $\ell \geq 2$ arms available for allocation. Each arm allocation results in a reward which is obtained at some random time after the arm allocation. For each time $j \geq 1$, a treatment I_j is allotted based on the data observed previously and the covariate X_j. We assume that the covariates are d-dimensional continuous random variables and take values in the hypercube $[0, 1]^d$. Since the rewards can be obtained at some delayed time, we denote $\{t_j \in \mathbb{R}^+, j \geq 1\}$ to be the observation time for the rewards for arms $\{I_j, j \geq 1\}$ respectively. Let $Y_{i,j}$ be the reward obtained at time $t_j \geq j$ for arm $i = I_j$. The mean reward with covariate X_j for the ith arm is denoted as $f_i(X_j)$, $1 \leq i \leq \ell$. The observed reward with covariate X_j by pulling the ith arm is modeled as, $Y_{i,j} = f_i(X_j) + \epsilon_{i,j}$, where $\epsilon_{i,j}$ denotes random error with $\mathrm{E}(\epsilon_{i,j}) = 0$ and $\mathrm{Var}(\epsilon_{i,j}) < \infty$ for all $1 \leq i \leq \ell$ and $j \in \mathbb{N}$. The functions f_i are assumed to be unknown and not of any given parametric form.

The rewards are observed at delayed times t_j; the delay in the reward for arm I_j pulled at the jth time is given by a random variable $d_j := t_j - j$. Assume that these delays are mutually independent, independent of the covariates, and could be drawn from different distributions. That is, let $\{d_j, j \geq 1\}$ be a sequence of independent random variables with probability density functions $\{g_j, j \geq 1\}$ and the cumulative distribution functions $\{G_j, j \geq 1\}$, respectively.

Let $\{X_j, j \geq 1\}$ be a sequence of covariates independently generated according to an unknown underlying probability distribution P_X, from a population supported in $[0, 1]^d$. Let δ be a sequential allocation rule, which for each time j chooses an arm I_j based on the previous observations and X_j. The total mean reward up to time n is $\sum_{j=1}^n f_{I_j}(X_j)$. To evaluate the performance of the allocation strategy, let $i^*(x) = \mathrm{argmax}_{1 \leq i \leq \ell} f_i(x)$ and $f^*(x) = f_{i^*(x)}(x)$. Without the knowledge of the random errors, the ideal performance occurs when the choices of arms selected $I_1, \ldots, I_n$ match the optimal arms $i^*(X_1), \ldots, i^*(X_n)$, yielding the optimal total reward $\sum_{j=1}^n f^*(X_j)$. The ratio of these two quantities is the quantity of interest,

$$R_n(\delta) = \frac{\sum_{j=1}^n f_{I_j}(X_j)}{\sum_{j=1}^n f^*(X_j)}. \tag{1}$$

It can be seen that R_n is a random variable no bigger than 1.

Definition 1. An allocation rule δ is said to be strongly consistent if $R_n(\delta) \to 1$ with probability 1, as $n \to \infty$.

In Section 3, we propose an allocation rule which takes into account reward delays. Then in Sections 3.1 and 4.1, we discuss the consistency of the proposed allocation rule under some assumptions and then validate those assumptions when the histogram method is used to estimate the regression functions respectively.

3. The proposed strategy

Let $Z^{n,i}$ denote the set of observations for arm i whose rewards have been obtained up to time n, that is, $Z^{n,i} := \{(X_j, Y_{i,j}) : 1 \leq t_j \leq n \text{ and } I_j = i\}$. Let $\hat{f}_{i,n}$ denote the regression estimator of f_i based on the data $Z^{n,i}$. Let $\{\pi_j, j \geq 1\}$ be a sequence of positive numbers in $[0, 1]$ decreasing to zero.

Step 1. **Initialize.** Allocate each arm once, w.l.o.g., we can have $I_1 = 1, I_2 = 2, \ldots, I_\ell = \ell$. Since the rewards are not immediately obtained for each of these ℓ arms, we continue these forced allocations until we have at least one reward observed for each arm. Suppose, that happens at time m_0.

Step 2. **Estimate the individual functions** f_i. For $n = m_0$, based on $Z^{n,i}$, estimate f_i by $\hat{f}_{i,n}$ for $1 \le i \le \ell$ using the chosen regression procedure.

Step 3. **Estimate the best arm.** For X_{n+1}, let $\hat{i}_{n+1}(X_{n+1}) = \arg\max_{1\le i\le \ell} \hat{f}_{i,n}(X_{n+1})$.

Step 4. **Select and pull.** Randomly select an arm with probability $1 - (\ell - 1)\pi_{n+1}$ for $i = \hat{i}_{n+1}$ and with probability π_{n+1}, for all other arms, $i \neq \hat{i}_{n+1}$. Let I_{n+1} denote this selected arm.

Step 5. **Update the estimates.**

Step 5a. If a reward is obtained at the $(n+1)$th time (could be one or more rewards corresponding to one or more arms I_j, $1 \le j \le (n+1)$), update the function estimates of f_i for the respective arm (or arms) for which the reward (or rewards) are obtained at $(n+1)$th time.

Step 5b. If no reward is obtained at the $(n+1)$th time, use the previous function estimators, i.e. $\hat{f}_{i,n+1} = \hat{f}_{i,n}\ \forall\ i \in \{1, \ldots, \ell\}$.

Step 6. **Repeat.** Repeat steps 3–5 when the next covariate X_{n+2} surfaces and so on.

The choice of π_n in the randomization Step 4 is crucial in determining how much exploration and exploitation is done at any phase of the trial. To emphasize the role of π_n, we may use δ_π to denote the allocation rule. In order to select the best arm as time progresses, π_n needs to decrease to zero but the rate of decrease will play a key role in determining how well the allocations work. For example, if in our set-up we have large delays for some arms then it might be beneficial to decrease π_n at a slower rate so that there is enough exploration and the accuracy of our estimates is not affected in the long run. We use a user-determined choice of π_n in this work, that is, the sequence π_n does not adapt to the data.

3.1. Consistency of the proposed strategy

Let $A_n := \{j : t_j \le n\}$, denote the time points for which rewards were obtained by time n. If A_n is known, then the total number of observed rewards until time n, denoted by N_n, is also known. Recall that it is possible to observe multiple rewards at the same time point. Given A_n, let $\{s_k, k = 1, \ldots, N_n\}$ be the reordered sequence of these observed reward timings, $\{t_k, k \in A_n\}$, arranged in a non-decreasing order.

Assumption 1. The regression procedure is strongly consistent in L_∞ norm for all individual mean functions f_i under the proposed allocation scheme. That is, $\|\hat{f}_{i,n} - f_i\|_\infty \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$ for each $1 \le i \le \ell$. As described in the allocation strategy in Section 3, $\hat{f}_{i,n}$ is the estimator based on all previously observed rewards. That is, after initialization, the mean reward function estimators are only updated at the time points $\{s_k, k = 1, \ldots N_n\}$ where N_n is the number of rewards observed by time n. Therefore, this condition is equivalent to saying $\|\hat{f}_{i,s_n} - f_i\|_\infty \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

Assumption 2. Mean functions satisfy $f_i(x) \ge 0$, $A = \sup_{1\le i\le \ell} \sup_{x\in[0,1]^d}(f^*(x) - f_i(x)) < \infty$ and $\mathrm{E}(f^*(X_1)) > 0$.

Theorem 1. *Under Assumptions 1 and 2, the allocation rule δ_π is strongly consistent as $n \to \infty$.*

Proof. Note that consistency holds only when the sequence $\{\pi_n, n \ge 1\}$ is chosen such that $\pi_n \to 0$ as $n \to \infty$. The proof is very similar to the proof in Yang and Zhu (2002). The details can be found in the supplementary material (see *Appendix A.1* in Appendix A). □

Note that Assumption 1, seemingly natural, is a strong assumption and it requires additional work to verify this assumption for a particular regression setting. We verify this assumption for the histogram method in Section 4.1. On the other hand, Assumption 2 does not involve the estimation procedure and does not require any verification.

4. The histogram method

In this section, we explain the histogram method for the setting with delayed rewards. Partition $[0, 1]^d$ into $M = (1/h)^d$ hyper-cubes with side width h, assuming h is chosen such that $1/h$ is an integer. For some $x \in [0, 1]^d$, let $J(x)$ denote the set of time points, for which the corresponding design points observed until time n fall in the same cube as x, say $B(x)$, and for which the corresponding rewards are observed by time n. Let $N(x)$ denote the size of $J(x)$. That is, let $J(x) = \{j : X_j \in B(x), t_j \le n\}$ and $N(x) = \sum_{j=1}^n I\{X_j \in B(x), t_j \le n\}$. Furthermore, let $\bar{J}_i(x)$ be the subset of $J(x)$ corresponding to arm i and $\bar{N}_i(x)$ is the number of such time points, that is, $\bar{J}_i(x) = \{j \in J(x) : I_j = i\}$ and $\bar{N}_i(x) = \sum_{j=1}^n I\{I_j = i, X_j \in B(x), t_j \le n\}$. Then the histogram estimate for $f_i(x)$ is defined as,

$$\hat{f}_{i,n}(x) = \frac{1}{\bar{N}_i(x)} \sum_{j\in\bar{J}_i(x)} Y_j.$$

For the estimator to behave well, a proper choice of the bandwidth, $h = h_n$ is necessary. Although one could choose different widths $h_{i,n}$ for estimating different f_i's, for simplicity, the same bandwidth h_n is used in the following sections. For notational convenience, when the analysis is focused on a single arm, i is dropped from the subscript of $\hat{f}$, $\bar{N}$ and $\bar{J}$.

Other nonparametric methods like nearest-neighbors, kernel method, spline fitting and wavelets can also be considered for estimation. Assumption 1 could be verified for these methods using the same broad approach as illustrated in the following sections for the Histogram method, along with some method specific mathematical tools and assumptions.

4.1. Allocation with histogram estimates

Here, we show that the histogram estimation method along with the allocation scheme described in Section 3, leads to strong consistency under some reasonable conditions on random errors, design distribution, mean reward functions and delays. As already discussed in Section 3.1, we only need to verify that Assumption 1 holds for histogram method estimators. Along with Assumption 2, we make the following assumptions.

Assumption 3. The design distribution P_X is dominated by the Lebesgue measure with a density $p(x)$ uniformly bounded above and away from 0 on $[0, 1]^d$; that is, $p(x)$ satisfies $\underline{c} \le p(x) \le \bar{c}$ for some positive constants $\underline{c} < \bar{c}$.

Assumption 4. The errors satisfy a moment condition that there exists positive constants v and c such that, for all $m \ge 2$, the Bernstein condition is satisfied, that is, $\mathrm{E}|\epsilon_{ij}|^m \le \frac{m!}{2} v^2 c^{m-2}$.

Assumption 5. The delays, $\{d_j, j \ge 1\}$, are independent of each other, the choice of arms and also of the covariates.

Assumption 6. Let the partial sums of delay distributions satisfy, $\sum_{j=1}^{n} G_j(n-j) = \Omega(n^\alpha \log^\beta n)$[1] for some $\alpha > 0$, $\beta \in \mathbb{R}$ or for $\alpha = 0$ and $\beta > 1$.

Note that, the choice $n^\alpha \log^\beta n$ could be generalized to a sub-linear function $q(n)$ with a growth rate faster than $\log n$.

Definition 2. Let $x_1, x_2 \in [0, 1]^d$. Then $w(h; f)$ denotes a modulus of continuity defined by, $w(h; f) = \sup\{|f(x_1) - f(x_2)| : |x_{1k} - x_{2k}| \le h \text{ for all } 1 \le k \le d\}$.

4.2. Number of observations in a small cube for histogram estimation

From Assumptions 3 and 5, we have that for a fixed cube B with side width h_n at time n, $P(X_j \in B, t_j \le n) = P(X_j \in B)P(t_j \le n) \ge \underline{c} h_n^d G_j(n-j)$. Let N be the number of observations that fall in B and are observed by time n, that is $N = \sum_{j=1}^{n} I_{\{X_j \in B, t_j \le n\}}$. It is easily seen that N is a random variable with expectation $\beta \ge \sum_{j=1}^{n} \underline{c} h_n^d G_j(n-j)$. From the extended Bernstein inequality (see *Appendix A.3* in Appendix A), we have

$$\mathrm{P}\left(N \le \frac{\underline{c} h_n^d \sum_{j=1}^{n} G_j(n-j)}{2}\right) \le \exp\left(-\frac{3\underline{c} h_n^d \sum_{j=1}^{n} G_j(n-j)}{28}\right). \tag{2}$$

Lemma 1. *Let $\epsilon > 0$ be given. Suppose that h is small enough such that $w(h; f) < \epsilon$. Then the histogram estimator $\hat{f}_n$ satisfies,*

$$P_{A_n, X^n}(\|\hat{f}_n - f\|_\infty \ge \epsilon) \le M \exp\left(-\frac{3\pi_n \min_{1 \le b \le M} N_b}{28}\right) + 2M \exp\left(-\frac{\min_{1 \le b \le M} N_b \pi_n^2 (\epsilon - w(h; f))^2}{8(v^2 + c(\pi_n/2)(\epsilon - w(h; f)))}\right),$$

where the probability P_{A_n, X^n} denotes conditional probability given design points $X^n = (X_1, X_2, \ldots, X_n)$ and $A_n = \{j : t_j \le n\}$. Here, N_b is the number of design points for which the rewards have been observed by time n such that they fall in the bth small cube of the partition of the unit cube at time n.

Proof. The proof of Lemma 1 is included in the supplementary materials (*Appendix A.2* in Appendix A). □

Theorem 2. *Suppose Assumptions 2–6 are satisfied. If for some $\alpha > 0$ and $\beta \in \mathbb{R}$ or $\alpha = 0$ and $\beta > 1$, h_n and π_n are chosen to satisfy,*

$$n^\alpha (\log n)^{\beta - 1} h_n^d \pi_n^2 \to \infty, \tag{3}$$

then the allocation rule δ_π is strongly consistent.

[1] $f(n) = \Omega(g(n))$ if for some positive constant c, $f(n) \ge c g(n)$ when n is large enough.

Proof of Theorem 2. The histogram technique partitions the unit cube into $M = (1/h)^d$ small cubes. For each small cube B_b, $1 \le b \le M$, in the partition of the unit cube, let N_b denote the number of time points, for which the corresponding design points fall in the cube B_b and corresponding arm rewards are observed by time n. In other words, $N_b = \sum_{j=1}^{n} I_{\{X_j \in B_b, t_j \le n\}}$. Using inequality (2) we have,

$$P\left(N_b \le \frac{\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{2}\right) \le \exp\left(-\frac{3\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{28}\right)$$
$$\Rightarrow P\left(\min_{1\le b\le M} N_b \le \frac{\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{2}\right) \le M \exp\left(-\frac{3\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{28}\right). \tag{4}$$

Let $W_1, \ldots, W_n$ be Bernoulli random variables indicating whether the ith arm is selected ($W_j = 1$) for time point j, or not ($W_j = 0$). Note that, conditional on the previous observations and X_j, the probability of $W_j = 1$ is almost surely bounded below by $\pi_j \ge \pi_n$ for $1 \le j \le n$. Let $w(h_n; f_i)$ be the modulus of continuity as in Definition 2. Note that, under the continuity assumption of f_i, we have $w(h_n; f_i) \to 0$ as $h_n \to 0$. Thus, for any $\epsilon > 0$, when h_n is small enough, $\epsilon - w(h_n; f_i) \ge \epsilon/2$. Consider,

$$\begin{aligned} P(\|\hat{f}_{i,n} - f_i\|_\infty > \epsilon) &= P\left(\|\hat{f}_{i,n} - f_i\|_\infty > \epsilon, \min_{1\le b\le M} N_b \ge \frac{\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{2}\right) \\ &\quad + P\left(\|\hat{f}_{i,n} - f_i\|_\infty > \epsilon, \min_{1\le b\le M} N_b < \frac{\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{2}\right) \\ &\le \mathrm{E} P_{A_n, X^n}\left(\|\hat{f}_{i,n} - f_i\|_\infty > \epsilon, \min_{1\le b\le M} N_b \ge \frac{\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{2}\right) \\ &\quad + P\left(\min_{1\le b\le M} N_b < \frac{\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{2}\right), \end{aligned}$$

where P_{A_n, X^n} denotes conditional probability given the design points until time n, $X^n = \{X_1, X_2, \ldots, X_n\}$ and the event, $A_n := \{j : t_j \le n\}$.

From Lemma 1, we have that given the design points and the time points for which rewards were observed, for any $\epsilon > 0$, when h is small enough,

$$\mathrm{P}_{A_n, X^n}(\|\hat{f}_n - f\|_\infty \ge \epsilon) \le M \exp\left(-\frac{3\pi_n \min_{1\le b\le M} N_b}{28}\right) + 2M \exp\left(-\frac{\min_{1\le b\le M} N_b \pi_n^2 (\epsilon - w(h_n; f))^2}{8(v^2 + c(\pi_n/2)(\epsilon - w(h_n; f)))}\right).$$

Using the above inequality and (4), we have,

$$\begin{aligned} P(\|\hat{f}_{i,n} - f_i\|_\infty > \epsilon) &\le 2M \exp\left(-\frac{\underline{c} h_n^d (\sum_{j=1}^n G_j(n-j)) \pi_n^2 (\epsilon - w(h_n; f_i))^2}{16(v^2 + c\pi_n/2(\epsilon - w(h_n; f_i)))}\right) \\ &\quad + M \exp\left(-\frac{3\underline{c} h_n^d \pi_n \sum_{j=1}^n G_j(n-j)}{56}\right) + \exp\left(-\frac{3\underline{c} h_n^d \sum_{j=1}^n G_j(n-j)}{28}\right). \end{aligned}$$

It can be shown that the above upper bound is summable in n under the condition,

$$\frac{h_n^d \pi_n^2 \sum_{j=1}^n G_j(n-j)}{\log n} \to \infty. \tag{5}$$

It is easy to see that this follows from Assumption 6 and (3).

Since ϵ is arbitrary, by the Borel–Cantelli lemma, we have that $\|\hat{f}_{i,n} - f_i\|_\infty \to 0$. This is true for all arms $1 \le i \le \ell$. Hence, this completes the proof of Theorem 2. □

4.3. Effects of reward delay distributions

As one would expect, the amount of delay in observing the rewards will have a considerable effect on the speed of sequential learning. In terms of treatment allocation, if there are substantial delays in observing patient responses for a particular treatment, the learning for that treatment will slow down and as a result the efficiency of the allocation strategy will decrease. Therefore, Assumption 6 imposes some restrictions on the delay distributions to ensure that at least a small proportion of rewards will be obtained in finite time. It is of interest to see how the delay distribution affects the rate at which π_n and h_n are allowed to decrease. This relationship can be understood by examining condition (3) for Theorem 2.

Note that Assumption 6 and (3) in Theorem 2 can be generalized to include any function $q(x)$ with at least a growth rate faster than logarithmic growth rate. We assume $\sum_{j=1}^{n} G_j(n-j) = \Omega(q(n))$ where $q(n)$ satisfies, $q(n)/\log(n) \to \infty$ as $n \to \infty$. Then it is easy to see that h_n and π_n can be chosen such that,

$$\frac{h_n^d \pi_n^2 q(n)}{\log(n)} \to \infty \text{ as } n \to \infty.$$

which implies condition (5) holds. A possible advantage of this is that we allow a wide range of possible delay distributions with mild restrictions on the delays. Below, we consider some cases of the delay distributions and see how they effect exploration (π_n) and bandwidth (h_n) of the histogram estimator as time progresses.

1. In condition (3), $q(n) = n^{\alpha}\log^{\beta} n$ for $\alpha > 0$ and $\beta \in \mathbb{R}$ or $\alpha = 0$ and $\beta > 1$. Let us first consider the case when $\alpha = 0$ and $\beta > 1$, we have $q(n) = \log^{\beta} n$ for $\beta > 1$ and we want $\sum_{j=1}^{n} G_j(n-j) = \Omega(\log^{\beta} n)$. Consider, $\pi_n = (\log n)^{-(\beta-1)/(2+d)}$ for $n > m_0$ and $\beta > 1$, then for (5) to hold we need the bandwidth h_n also to be of order $\Omega((\log n)^{-(\beta-1)/(2+d)})$. For example, $h_n = (\log n)^{-(\beta-1)/\beta(2+d)}$ would guarantee consistency. Notice that with these π_n and h_n, one would spend a lot of time in exploration and the bandwidth would also decay very slowly which would effect the accuracy of the reward function estimates until n is sufficiently large.
   Notice that the restriction of partial sum of probability distributions for the delays, being at least of the order $\log^{\beta} n$ gives the possibility of modeling cases with extremely large delays. For example, in clinical studies when the outcome of interest is survival time and we want to administer treatments for a disease such that the survival time is maximized. With the unprecedented advances in drug development, the life expectancy of patients is more likely to increase, hence the survival time for a patient given any treatment would be large. Therefore, the assumption that partial sums of probability distributions for the delays until time n need only be at least $\log^{\beta} n$ seems to be quite reasonable when the expected waiting times (in this case survival times) are long. For example, diseases like diabetes and hypertension which have a long survival time, since they cannot be cured, but can be controlled with medications. These diseases also have fairly high prevalence, so a large sample size to be able to get close to optimality would not be a problem. For such diseases, assuming that one would only observe the responses (survival times) of a small fraction of patients in finite time seems reasonable.
2. For the case when $\alpha > 0$ and $\beta \in \mathbb{R}$, we have that $\sum_{j=1}^{n} G_j(n-j) = \Omega(n^{\alpha}\log^{\beta} n)$. Consider, $\pi_n = n^{-\alpha/(2+d)}$ for $n > m_0$, then for the condition (5) to hold we need h_n to also be of order $\Omega(n^{-\alpha/(2+d)})$. For example, $h_n = n^{-\alpha/2(2+d)}$ results in $h_n^d \pi_n^2 n^{\alpha}\log^{\beta-1}(n) = n^{\alpha d/2(2+d)}\log^{\beta-1}(n) \to \infty$ as n $\to \infty$, irrespective of the value of β. Here the lower bound on the partial sums of probability distributions for the delays can grow faster than the previous case, depending on the values of α and β.
   This restriction of order $n^{\alpha}(\log^{\beta} n)$ can model cases with moderately large delays. From a clinical point of view, one could model diseases in which treatments show their effect in a short to moderate duration of time, for examples diseases like diarrhea, common cold, headache, and nutritional deficiencies. Here the response of interest would be improvement in the condition of a patient as a result of a treatment. For such diseases, one can expect to see the treatment effects on patients in a short period of time. Hence, the delay in observing treatment results will not be too long. If the response considered was survival (survived or not), then stroke could also fall in this category because of high mortality.
   Note that, Assumption 6 only restricts on the proportion of rewards expected to be observed in the long run. Therefore, it is possible for strong consistency to be achieved even when there is infinite delay in observing the rewards of some arms (non-observance of some rewards).

5. Simulation study

We conduct a simulation study to compare the effect of different delay scenarios on the per-round average regret of our proposed strategy. The per-round regret is given by, $r_n(\delta) = \frac{1}{n}\sum_{j=1}^{n}(f^*(X_j) - f_{l_j}(X_j))$.

Note that if $\frac{1}{n}\sum_{j=1}^{n} f^*(X_j)$ is eventually bounded above and away from 0 with probability 1, then $R_n(\delta) \to 1$ a.s. is equivalent to $r_n(\delta) \to 0$ a.s.

5.1. Simulation setup

Consider number of arms, $\ell = 3$, and the covariate space to be two-dimensional, $d = 2$. Let $X_n = (X_{n1}, X_{n2})$ where $X_{ni} \overset{i.i.d.}{\sim}$ Unif(0, 1). We assume that the errors $\epsilon_n \sim 0.5\text{N}(0,1)$. The first 30 rounds were used for initialization. The following true mean reward functions are used,

$$f_1(\mathbf{x}) = 0.7(x_1 + x_2),\ f_2(\mathbf{x}) = 0.5x_1^{0.75} + \sin(x_2),\ f_3(\mathbf{x}) = \frac{2x_1}{0.5 + (1.5 + x_2)^{1.5}}.$$

We consider the following delay scenarios and run simulations until $N = 10\,000$. (1) *No delay*; (2) *Delay 1:* Geometric delay with probability of success (observing the reward) $p = 0.3$; (3) *Delay 2:* Every 5th reward is not observed by time N and other rewards are obtained with a geometric ($p = 0.3$) delay; (4) *Delay 3:* Each case has probability 0.7 to delay and

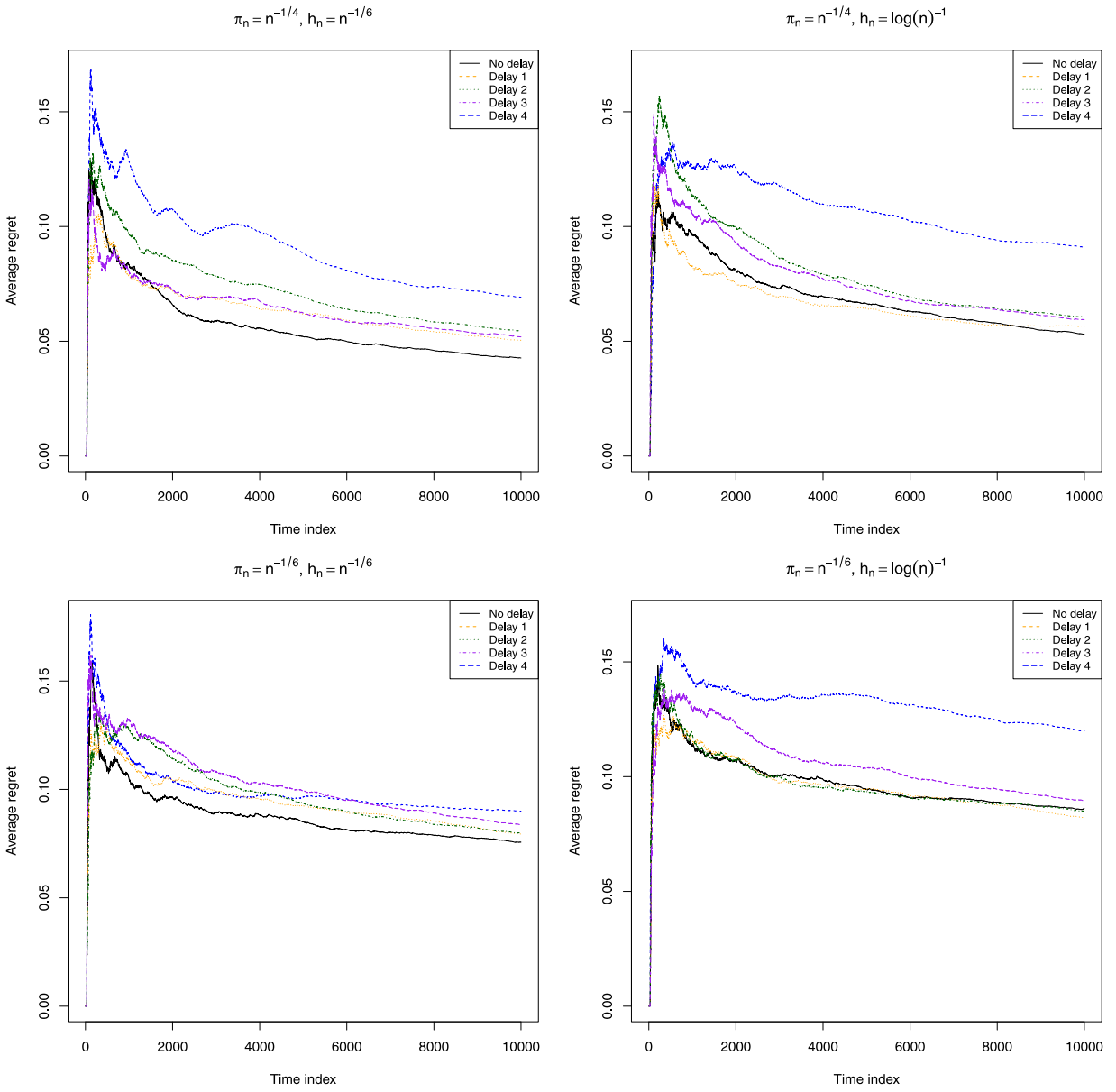

Fig. 1. Per-round regret for the proposed strategy for different delay scenarios. The grid of plots represent 4 different combination of choices for $\{\pi_n\}$ and $\{h_n\}$. For a given row, π_n remains fixed and h_n varies and vice versa for columns.

the delay is half-normal with scale parameter, $\sigma = 1500$; (5) *Delay 4:* In this case we increase the number of non-observed rewards. Divide the data into four equal consecutive parts (quarters), such that, in part 1, we only observe every 10th (with Geom(0.3) delay) observation by time N and not observe the remaining; in part 2, we only observe every 15th observation; in part 3, only observe every 20th observation; in part 4, only observe every 25th observation.

In Fig. 1, we plot the per-round regret vs time by delay type for four combinations of π_n and h_n. As one would expect (see Fig. 1), the severity of delay has a clear effect on the regret, and for delay scenarios where a large number of rewards are not observed in finite time, the regret is comparatively higher. Note that most delay scenarios for which a substantial number of rewards can be obtained in finite time, tend to converge in quite similar patterns.

Choice of $\{\pi_n\}$ **and** $\{h_n\}$: According to Theorem 2, if π_n and h_n are chosen such that condition (3) is met, consistency of the allocation rule follows. Therefore, for the case with $d = 2$, which is the case of the simulation setting, we have to choose sequences slower than ($\pi_n = n^{-1/2}$, $h_n = n^{-1/2}$), even in the case of no delays. Keeping this in mind, we chose two different choices of sequences for π_n ($n^{-1/4}$, $n^{-1/6}$) and two choices of $h_n((\log n)^{-1}, n^{-1/6})$. Note that, in Fig. 1, for a given row, π_n remains fixed while h_n varies and vice versa for columns. It can be seen that the regret gets worse when h_n decays

too fast (in our range of n as $N = 10\,000$), specially for the scenario (Delay 4) with increasing number of non-observed rewards, possibly because of violation of condition (3). Also notice that, slow decaying π_n has higher regret (last row). This could be because of large randomization error that leads to high exploration price. In general, there are a large pool of choices for h_n and π_n that satisfy equation (3) as can be seen from Fig. 1. However, a thorough understanding of the finite-time regret rates and further research would be needed to evaluate optimal choices of $\{\pi_n\}$ and $\{h_n\}$ for a given scenario.

6. Conclusion

In this work we develop an allocation rule for multi-armed bandit problem with covariates when there is delay in observing rewards. We show that strong consistency can be established for the proposed allocation rule using the histogram method for estimation, under reasonable restrictions on the delay distributions and also illustrate that using a simulation study. Our approach on modeling reward delays is different from the previous work done in this field because, (1) we use nonparametric estimation technique to estimate the functional relationship between the rewards and covariates and (2) we allow for delays to be unbounded with some assumptions on the delay distributions. The assumptions impose mild restrictions on the delays in the sense that they allow for the possibility of non-observance of some rewards as long as a certain proportion of rewards are obtained in finite time. With this general setup, it is possible to model many different situations including the one with no delays. The conditions on the delay distributions easily allow for large delays as long as they grow at a certain minimal rate. This obviously will result in slower rate of convergence because of longer time spent in exploration. Ideally, we would like our allocation scheme to devise the optimal treatments sooner, for which we would need to impose stricter conditions on the delay distributions. Therefore, working on finite-time analysis for the setting with delayed rewards seems to be an immediate future direction. In addition, we assume some knowledge on the delay distributions, so for situations where there is little understanding of the delays, a different approach might be needed, such as a methodology which adaptively updates the delay distributions.

Acknowledgments

We thank the editor and two anonymous reviewers for their constructive comments which have helped us to improve the manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2020.108818.

References

Agrawal, S., Goyal, N., 2012. Analysis of thompson sampling for the multi-armed bandit problem. In: Conference on Learning Theory (COLT).
Anderson, T., 1964. Sequential analysis with delayed observations. J. Amer. Statist. Assoc. 59 (308), 1006–1015.
Auer, P., Cesa-Bianchi, N., Fischer, P., 2002. Finite-time analysis of the multiarmed bandit problem. Mach. Learn. 47 (2–3), 235–256.
Berry, D.A., Fristedt, B., 1985. Bandit Problems: Sequential Allocation of Experiments. In: Monographs on Statistics and Applied Probability, vol. 5, Chapman and Hall, London, pp. 71–87.
Bubeck, S., Cesa-Bianchi, N., 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Found. Trends Mach. Learn. 5 (1), 1–122.
Cesa-Bianchi, N., Gentile, C., Mansour, Y., Minora, A., 2016. Delay and cooperation in nonstochastic bandits. J. Mach. Learn. Res. 49 (1), 613–650.
Cesa-Bianchi, N., Lugosi, G., 2006. Prediction, Learning, and Games. Cambridge University Press.
Chapelle, O., Li, L., 2011. An empirical evaluation of thompson sampling. In: Advances in Neural Information Processing Systems. pp. 2249–2257.
Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., Zhang, T., 2011. Efficient optimal learning for contextual bandits. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. AUAI Press.
Gittins, J.C., 1979. Bandit processes and dynamic allocation indices. J. R. Stat. Soc. Ser. B Stat. Methodol. 41 (2), 148–164.
Joulani, P., Gyorgy, A., Szepesvári, C., 2013. Online learning under delayed feedback. In: International Conference on Machine Learning. pp. 1453–1461.
Lai, T.L., Robbins, H., 1985. Asymptotically efficient adaptive allocation rules. Adv. Appl. Math. 6 (1), 4–22.
Langford, J., Zhang, T., 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In: Advances in Neural Information Processing Systems. pp. 817–824.
Lattimore, T., Szepesvári, C., 2018. Bandit Algorithms. Cambridge University Press.
Li, L., Chu, W., Langford, J., Schapire, R.E., 2010. A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 661–670.
Mandel, T., Liu, Y.-E., Brunskill, E., Popović, Z., 2015. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In: Twenty-Ninth AAAI Conference on Artificial Intelligence.
Perchet, V., Rigollet, P., 2013. The multi-armed bandit problem with covariates. Ann. Statist. 41 (2), 693–721.
Pike-Burke, C., Agrawal, S., Szepesvári, C., Grunewalder, S., 2018. Bandits with delayed, aggregated anonymous feedback. In: International Conference on Machine Learning.
Qian, W., Yang, Y., 2016a. Kernel estimation and model combination in a bandit problem with covariates. J. Mach. Learn. Res. (1), 5181–5217.
Qian, W., Yang, Y., 2016b. Randomized allocation with arm elimination in a bandit problem with covariates. Electron. J. Stat. 10 (1), 242–270.
Robbins, H., 1952. Some aspects of the sequential design of experiments. Bull. Amer. Math. Soc. 58 (5), 527–535.
Russo, D.J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., 2018. A tutorial on thompson sampling. Found. Trends Mach. Learn. 11 (1), 1–96.
Sarkar, J., 1991. One-armed bandit problems with covariates. Ann. Statist. 19 (4), 1978–2002.
Slivkins, A., 2014. Contextual bandits with similarity information. J. Mach. Learn. Res. 15 (1), 2533–2568.

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction. MIT press.
Suzuki, Y., 1966. On sequential decision problems with delayed observations. Ann. Inst. Statist. Math. 18 (1), 229–267.
Vernade, C., Cappé, O., Perchet, V., 2017. Stochastic bandit models for delayed conversions. In: Conference on Uncertainty in Artificial Intelligence.
Woodroofe, M., 1979. A one-armed bandit problem with a concomitant variable. J. Amer. Statist. Assoc. 74 (368), 799–806.
Yang, Y., Zhu, D., 2002. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. Ann. Statist. (1), 100–121.