Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

To update or not to update? Delayed Nonparametric Bandits with Randomized Allocation

Sakshi Arya*[1] | Yuhong Yang[2]

[1]School of Statistics, University of Minnesota, Ford Hall, Church Street SE, Minneapolis, MN, USA
[2]School of Statistics, University of Minnesota, Ford Hall, Church Street SE, Minneapolis, MN, USA

Correspondence
Email: aryax010@umn.edu

Summary

Delayed rewards problem in contextual bandits has been of interest in various practical settings. We study randomized allocation strategies and provide an understanding on how the exploration-exploitation tradeoff is affected by delays in observing the rewards. In randomized strategies, the extent of exploration-exploitation is controlled by a user-determined exploration probability sequence. In the presence of delayed rewards, one may choose between using the original exploration sequence that updates at every time point or update the sequence only when a new reward is observed, leading to two competing strategies. In this work, we show that while both strategies may lead to strong consistency in allocation, the property holds for a wider scope of situations for the latter. However, for finite sample performance, we illustrate that both strategies have their own advantages and disadvantages, depending on the severity of the delay and underlying reward generating mechanisms.

KEYWORDS:
contextual bandit, delayed reward, histogram, kernel, nonparametric, regret, strong consistency

1 | INTRODUCTION

Contextual bandits provide a natural framework to model a lot of practical sequential decision making problems in various fields. **?** started studying multi-armed bandit problems with side information in a parametric framework, and **?** initiated an investigation from a nonparametric perspective. See **?**;**?** for reviews on general sequential problems and **?** for bandits exclusively. In recent years, bandit problems have gained popularity and have been studied extensively under different names, such as contextual bandits, multi-armed bandits with covariates (MABC), associative bandit problems and multi-armed bandits with side information. For example, when treating patients of a disease, the doctor needs to decide which treatment amongst several competing treatments would be the best for the current patient, given the patient's covariate information and data available from previous patients. Most of the bandit algorithms assume instantaneous observance of rewards, but in most practical situations, rewards are only obtained at some delayed time. For example, it is often the case that several other patients have to be treated before the outcome for the current patient is observed. One way to tackle this problem is to adopt black-box procedures incorporating delayed rewards using the already existing no-delay policies in the stochastic bandits setting. However, we present a case of why it is important to study delays more carefully for contextual bandit strategies based on the context of the problem, rather than always using the already existing no delay bandit strategies in black-box procedures to incorporating delayed rewards. Delays in observing the rewards could affect the performance of bandit algorithms in different ways, depending on the nature of underlying data generating mechanisms and severity of the delays. Thus, it is important to balance the exploration-exploitation trade-off taking these aspects into consideration, in order to utilize most of the available information. We propose two different ϵ-greedy like strategies incorporating delayed reward, which differ in how the exploration probability gets updated with the available information. We illustrate that both strategies can be advantageous in different situations, based on the complexity of the underlying data generating mechanism and the severity of the delays.

2 | SETUP AND RELATED LITERATURE

The setup of stochastic contextual bandits is as follows. Suppose there are $\ell > 1$ competing arms. The covariates are assumed to be random variables generated according to an unknown underlying probability distribution P_X supported in $[0,1]^d$. A bandit strategy or policy is a random function from $[0,1]^d$ to $\{1,2,\ldots,\ell\}$ that decides which arm gets pulled for a given covariate. At time $j \geq 1$, let I_j be the arm allocation made by the bandit strategy based on previous information and present context X_j. We denote $Y_{i,j}$ to be the reward obtained for arm $i = I_j$. Let $f_i(x)$ denote the mean reward for the i^{th} arm with covariate x. We adopt a regression perspective to model the relationship between covariates and rewards,

$$Y_{i,j} = f_i(X_j) + \epsilon_j$$

where ϵ_j's are independent errors with $E(\epsilon_j) = 0$ and $Var(\epsilon_j) < \infty$ for $j \geq 1$.

Now, the problem can be viewed as one of estimating the mean reward functions $f_i(x)$ for $i \in \{1,\ldots,\ell\}$ and allocating arms based on the estimators $\hat{f}_i$. Both parametric and non-parametric approaches for estimating f_i have been well studied, see **?**, **?** for reference. In this work we follow a nonparametric approach with delayed rewards as in **?** adopting modeling techiques similar to the earlier work of **?**, **??**.

In our setup, the rewards can be obtained at some delayed time, which we denote by $\{t_j \in \mathbb{R}^+, j \geq 1\}$. The delay in the reward for pulling arm I_j is given by the random variable, $d_j := t_j - j$. We assume that $\{d_j : d_j \geq 0, j \geq 1\}$ is a sequence of independent random variables. Let the number of rewards obtained at time n be denoted by $\tau_n = \sum_{j=1}^{n} I(t_j \leq n)$, also a random variable.

We devise two sequential allocation strategies η_1 and η_2 in Section 3, incorporating delayed rewards, such that they choose arms sequentially based on previous observations and present covariates. As a measure of performance of each of the strategies, we consider the following ratio,

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

where $(\cdot)$ is used to denote the strategy being considered. Here, $f^*(x) = \max_{1\leq i\leq \ell} f_i(x)$ is the theoretical best mean reward functional value at x, and $i^*(x)$ is the corresponding arm. Then, we establish strong consistency for both strategies for the histogram method in Section 4.1, that is, we show that $R_n(\eta_1) \to 1$ and $R_n(\eta_2) \to 1$ with probability 1, as $n \to \infty$. In addition, from a finite-sample performance perspective, we compare the two allocation strategies and illustrate how both can be advantageous in different situations in Sections 4.3 and 5.

In the stochastic setting, delayed rewards have been studied previously by **?**, **?** where the former considers constant known delay for contextual bandits while the latter provides a more systemic study of online learning problems with random delayed rewards (without covariates). **?** develop meta-algorithms which in a black-box fashion use algorithms developed for the non-delayed case into the ones that can handle delays in a feedback loop. Then, **?** devise a method that guarantees good black-box algorithms when leveraging a prior dataset and incorporating heuristics to help improve empirical performance of the algorithms. **?** use Gaussian process bandits and develop algorithms for parallelizing exploration-exploitation trade-offs. Motivated by delayed conversions in advertising, **??** consider potentially infinite stochastic delays, where the latter deals with the contextual case with a linear regression model and does not assume prior knowledge of delay distribution unlike the former. Recently, **?** design a delay-adaptive algorithm for generalized linear contextual bandits using UCB-style exploration. **?** consider potentially infinite delays in nonparametric bandits and provide strong consistency results for a proposed algorithm. Other works include **?**, **?** where the former considers Gittins procedures for bandits with delayed rewards, while the latter is motivated by applications in music streaming. Apart from the stochastic setting, **????** study delayed rewards in the adversarial setting, while **???** study the delayed anonymous composite feedback setting.

3 | THE PROPOSED STRATEGIES

Define $Z^{n,i}$ to be the set of observations for arm i whose rewards have been observed by time $n-1$, that is, $Z^{n,i} := \{(X_j, Y_{i,j}) : 1 \leq t_j \leq n-1 \text{ and } I_j = i\}$. 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, such that $(\ell - 1)\pi_j < 1$ for all $j \geq 1$. We propose two strategies η_1 and η_2 with a subtle difference in the arm selection step but same structure of the algorithm.

3.1 | Algorithms

Step 1. **Initialize.** Allocate each arm once, $I_1 = 1, I_2 = 2, \ldots, I_\ell = \ell$. 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 $n = m_0 + 1$, based on $Z^{n,i}$, estimate f_i by $\hat{f}_{i,n}$ for $1 \leq i \leq \ell$ using the chosen regression procedure.

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

Step 4. **Select and pull.** Recall, $\tau_n = \sum_{j=1}^{n} I(t_j \leq n)$ is the number of rewards observed by time n.

(a) Strategy η_1: $I_n = \begin{cases} \hat{i}_n, & \text{with probability } 1 - (\ell - 1)\pi_n \\ i, & \text{with probability } \pi_n,\ i \neq \hat{i}_n,\ 1 \leq i \leq \ell. \end{cases}$

(b) Strategy η_2: $I_n = \begin{cases} \hat{i}_n, & \text{with probability} 1 - (\ell - 1)\pi_{\tau_n} \\ i, & \text{with probability } \pi_{\tau_n},\ i \neq \hat{i}_n,\ 1 \leq i \leq \ell. \end{cases}$

Step 5. **Update the estimates.**

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

Step 5b. If no reward is obtained at the n^{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+1} surfaces and so on.

In the algorithms above, Step 1 initializes the allocations by pulling each arm alternatively until we observe at least one reward for each arm. Step 2 estimates the mean reward function for each arm. This could be done using several regression methods, we use kernel regression and histogram method in this work. Steps 3 and 4 enforce an ϵ-greedy type of randomization scheme which prefers the best performing arm so far with some probability and explores with the remaining. The preference is determined by user determined sequence of exploration probability $\{\pi_n, n \geq 1\}$, which for strategy η_2 only gets updated when a new reward is observed, that is, π_{τ_n}. While for strategy η_1, it is updated at every time point irrespective of a reward being observed or not, that is, π_n. Hence, the two strategies differ in the extent of exploration and exploitation that is allowed over time. Finally, in Step 5, the mean reward function estimators are updated if new rewards are observed or they remain the same if no new rewards are observed. For notational convenience, we use $\{\cdot\}$ to denote a user-determined sequence, such as $\{\pi_n\}$, when we only want to refer to the original sequence selected by the user, without distinguishing between when it gets updated.

4 | CONSISTENCY OF THE PROPOSED STRATEGIES

Let $\mathcal{A}_n := \{j : t_j \leq n\}$, denote the time points corresponding to the rewards observed by time n.

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, $||\hat{f}_{i,n} - f_i||_\infty \overset{a.s.}{\to} 0$ as $n \to \infty$ for each $1 \leq i \leq \ell$, where $\hat{f}_{i,n}$ is the estimator based on all previously observed rewards.

Note that, due to the presence of delays, the mean reward function estimators $\hat{f}_{i,n}$ are only updated at the time points where a new reward is observed. Next, we make a mild assumption on the mean reward functions.

Assumption 2. The mean reward functions are continuous and $f_i(x) \geq 0$ such that,

$$A = \sup_{1 \leq i \leq \ell} \sup_{x \in [0,1]^d} (f^*(x) - f_i(x)) < \infty \text{ and } E(f^*(X_1)) > 0.$$

Assumption 3. Let the partial sums of delay distributions satisfy, $E(\tau_n) = \Omega(q(n))$ [1], where $q(n)$ is a sequence that acts as a lower bound to the expected number of observed rewards by time n, and $q(n) \to \infty$ as $n \to \infty$.

Theorem 1. Under Assumptions 1, 2 and 3, the allocation rules η_1 and η_2 are strongly consistent as $n \to \infty$, i.e., $R_n(\eta_1) \to 1$ and $R_n(\eta_2) \to 1$ with probability 1, as $n \to \infty$.

Proof. Note that consistency holds only when the sequence $\{\pi_n, n \geq 1\}$ is chosen such that $\{\pi_n\} \to 0$ as $n \to \infty$. The proof is very similar to the proof in **?** with minor changes for strategy η_2 which are included in the supporting information (Section 6). □

Note that Assumption 1, seemingly natural, is a strong assumption and it requires additional work to verify it for a particular regression setting. We verify this assumption for the histogram method in Section 4.1 and for the kernel method in Section 4.2.

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

4.1 | Histogram method

In this section, we consider the histogram method for the setting with delayed rewards. We assume that the binwidth h is chosen such that $1/h$ is an integer. At time n, partition $[0,1]^d$ into $M = (1/h_{\tau_n})^d$ hyper-cubes with binwidth h_{τ_n}, where τ_n is the number of observed rewards by time n. For some $x \in [0,1]^d$ such that it falls in a hypercube $B(x)$, let $\bar{J}_i(x) = \{j : X_j \in B(x), t_j \le n, I_j = i\}$ and $\bar{N}_i(x)$ be the size of $\bar{J}_i(x)$. Then the histogram estimate for $f_i(x)$ is defined as,

$$\hat{f}_{i,n}(x) = \frac{1}{\bar{N}_i(x)} \sum_{j \in \bar{J}_i(x)} Y_j. \tag{2}$$

For the estimator to behave well, a proper choice of the binwidth, $\{h_n\}$ is necessary. Note that, we only update h_n to h_{n+1} when a new reward is observed, hence we denote it as h_{τ_n}. For notational convenience, when the analysis is focused on a single arm, i is dropped from the subscript of $\hat{f}$, $\bar{N}$ and $\bar{J}$. Next, using the histogram method for estimation, we prove that strong consistency holds for both strategies η_1 and η_2 in Section 3.1.

As already discussed, we only need to verify that Assumption 1 holds for histogram method. Along with Assumptions 2 and 3, we make the following assumptions.

Assumption 4. 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 $\underline{c} < \bar{c}$.

This assumption is needed to make sure that all regions in the covariate space are observed with positive probability, in order to ensure good estimation in all regions.

Assumption 5. The errors satisfy a moment condition that there exists positive constants v and c such that, for all integers $m \ge 2$, the extended Bernstein condition (**??**) is satisfied, that is,

$$E|\epsilon_j|^m \le \frac{m!}{2} v^2 c^{m-2}.$$

This condition on the errors holds in a lot of settings, for example, normal distribution and bounded errors meet this requirement, thus making it useful in a wide range of applications.
The next two assumptions are made on the nature of the delays in observing rewards, so that we could ensure that delays are not being confounded by other factors and we observe a minimum number of rewards with time, so as to ensure proper and effective learning.

Assumption 6. The delays, $\{d_j, j \ge 1\}$, are independent of each other, the choice of arms and also of the covariates.

Along with these assumptions, we define the modulus of continuity that is used in the following results.

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

Lemma 1 (An inequality for Bernoulli trials.)**.** For $1 \le j \le n$, let $\tilde{W}_j$ be Bernoulli random variables, which are not necessarily independent. Assume that the conditional probability of success for $\tilde{W}_j$ given the previous observations is lower bounded by β_j, that is,

$$P(\tilde{W}_j = 1 | \tilde{W}_i, 1 \le i \le j-1) \ge \beta_j \text{ a.s.},$$

for all $1 \le j \le n$. Applying the extended Bernstein's inequality as described in **?**, we have

$$P\left(\sum_{j=1}^{n} \tilde{W}_j \le \left(\sum_{j=1}^{n} \beta_j\right)/2\right) \le \exp\left(-\frac{3\sum_{j=1}^{n}\beta_j}{28}\right). \tag{3}$$

Lemma 2. 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,

$$P^{\eta_1}_{\mathcal{A}_n, X^n}(||\hat{f}_n - f||_\infty \ge \epsilon) \le M \exp\left(-\frac{3\pi_n \min_{1\le b\le M} N_b}{28}\right) + 2M \exp\left(-\frac{\min_{1\le b\le M} N_b \pi_n^2 (\epsilon - w(h_{\tau_n}; f))^2}{8(v^2 + c(\pi_n/2)(\epsilon - w(h_{\tau_n}; f)))}\right), \tag{4}$$

$$P^{\eta_2}_{\mathcal{A}_n, X^n}(||\hat{f}_n - f||_\infty \ge \epsilon) \le M \exp\left(-\frac{3\pi_{\tau_n} \min_{1\le b\le M} N_b}{28}\right) + 2M \exp\left(-\frac{\min_{1\le b\le M} N_b \pi_{\tau_n}^2 (\epsilon - w(h_{\tau_n}; f))^2}{8(v^2 + c(\pi_{\tau_n}/2)(\epsilon - w(h_{\tau_n}; f)))}\right), \tag{5}$$

where $P_{\mathcal{A}_n, X^n}$ denotes conditional probability given design points $X^n = (X_1, \ldots, X_n)$ and $\mathcal{A}_n = \{j : t_j \le n\}$. 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 2 is similar to **?** so we skip it here. For strategy η_1, it is easy to see that a similar lemma with h_n replaced by h_{τ_n} could be derived. For strategy η_2, π_n is replaced by π_{τ_n} and h_n replaced by h_{τ_n}. This is because the result is a conditional probability result, and given $\mathcal{A}_n$ and X^n, τ_n is a known quantity. □

Theorem 2. Suppose Assumptions 2-6 are satisfied.

a) If $\{h_n\}$ and $\{\pi_n\}$ are chosen to satisfy,

$$\frac{h^2_{q(n)}\pi^2_n q(n)}{\log n} \to \infty \text{ as } n \to \infty, \tag{6}$$

then the histogram estimator in (2) is strongly consistent in the L_∞ norm for strategy η_1, hence η_1 is strongly consistent.

b) If $\{h_n\}$ and $\{\pi_n\}$ are chosen to satisfy,

$$\frac{h^2_{q(n)}\pi^2_{q(n)} q(n)}{\log n} \to \infty \text{ as } n \to \infty, \tag{7}$$

then the histogram estimator in (2) is strongly consistent in the L_∞ norm for strategy η_2, hence η_2 is strongly consistent.

Proof. The proofs for a) and b) are quite similar, so we prove b) here and consequently discuss a). Given $\mathcal{A}_n$, the indices corresponding to when rewards were obtained, we know that at time n, the histogram method partitions the unit cube into $M = (1/h_{\tau_n})^d$ small cubes. For each small cube $B_b, 1 \le b \le M$, in the partition, let $N_b = \sum_{j=1}^n I(X_j \in B_b, t_j \le n)$. Note that given $\mathcal{A}_n$, $P_{\mathcal{A}_n}(X_j \in B_b, t_j \le n) = P_{\mathcal{A}_n}(X_j \in B_b) \ge \underline{c}h^d_{\tau_n}$, thus using inequality (3) we have,

$$P_{\mathcal{A}_n}\left(N_b \le \frac{\underline{c}h^d_{\tau_n}\tau_n}{2}\right) \le \exp\left(-\frac{3\underline{c}h^d_{\tau_n}\tau_n}{28}\right) \Rightarrow P_{\mathcal{A}_n}\left(\min_{1\le b\le M} N_b \le \frac{\underline{c}h^d_{\tau_n}\tau_n}{2}\right) \le \exp\left(-\frac{3\underline{c}h^d_{\tau_n}\tau_n}{28}\right). \tag{8}$$

Recall, $\tau_n = \sum_{j=1}^n I\{t_j \le n\}$. First, we show that $\tau_n \overset{\text{a.s.}}{\to} \infty$ as $n \to \infty$ for both strategies, η_1 and η_2. By Assumption 3 and the inequality (3) in Lemma 1 we have that for a large enough n, there exists a positive constant $a_1 > 0$ such that, $E(\tau_n) \ge a_1 q(n)$, therefore,

$$P\left(\tau_n \le \frac{a_1 q(n)}{2}\right) \le P\left(\tau_n \le \frac{E(\tau_n)}{2}\right) \le \exp\left(-\frac{3E(\tau_n)}{28}\right) \le \exp\left(\frac{-3a_1 q(n)}{28}\right).$$

It is easy to see that the upper bound is summable in n under the conditions (6) and (7). By Borel-Cantelli lemma, this implies that event $\{\tau_n > a_1 q(n)/2\}$ happens infinitely often, therefore $\tau_n \overset{\text{a.s.}}{\to} \infty$. Note that, by construction this implies that $h_{\tau_n} \overset{\text{a.s.}}{\to} 0$, and $\pi_{\tau_n} \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$. Let $w(h_{\tau_n}; f_i)$ be the modulus of continuity as in Definition 1. Then, continuity of f_i leads to the conclusion that $w(h_{\tau_n}; f_i) \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$. Thus, for any $\epsilon > 0$, for large enough n, when h_{τ_n} is small enough, $\epsilon - w(h_{\tau_n}; f_i) \ge \epsilon/2$, almost surely. Consider,

$$\begin{aligned} P_{\mathcal{A}_n}\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon\right) &= P_{\mathcal{A}_n}\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon, \min_{1\le b\le M} N_b > \frac{\underline{c}h^d_{\tau_n}\tau_n}{2}\right) + P_{\mathcal{A}_n}\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon, \min_{1\le b\le M} N_b \le \frac{\underline{c}h^d_{\tau_n}\tau_n}{2}\right) \\ &\le E^{X^n} P_{\mathcal{A}_n, X^n}\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon, \min_{1\le b\le M} N_b > \frac{\underline{c}h^d_{\tau_n}\tau_n}{2}\right) + P_{\mathcal{A}_n}\left(\min_{1\le b\le M} N_b \le \frac{\underline{c}h^d_{\tau_n}\tau_n}{2}\right), \end{aligned}$$

where we use law of iterated expectation in the first term and E^{X^n} denotes expectation with respect to X^n. From (5) and (8), we get that,

$$P_{\mathcal{A}_n}\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon\right) \le M\exp\left(-\frac{3\underline{c}\pi_{\tau_n}h^d_{\tau_n}\tau_n}{56}\right) + 2M\exp\left(-\frac{\underline{c}h^d_{\tau_n}\pi^2_{\tau_n}\tau_n(\epsilon - w(Lh_{\tau_n}; f_i))^2}{8(v^2 + c(\pi_{\tau_n}/2)\epsilon)}\right) + M\exp\left(-\frac{3\underline{c}h^d_{\tau_n}\tau_n}{28}\right). \tag{9}$$

Now consider,

$$\begin{aligned} P(||\hat{f}_{i,n} - f_i||_\infty > \epsilon) &\le P\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon, \tau_n > \frac{E(\tau_n)}{2}\right) + P\left(\tau_n \le \frac{E(\tau_n)}{2}\right) \\ &\le E^{\mathcal{A}_n} P_{\mathcal{A}_n}\left(||\hat{f}_{i,n} - f_i||_\infty \ge \epsilon, \tau_n > \frac{E(\tau_n)}{2}\right) + P\left(\tau_n \le \frac{E(\tau_n)}{2}\right). \end{aligned} \tag{10}$$

Let $n_e = \lfloor E(\tau_n)/2 \rfloor$. Then, by using condition (7) and (9) in (10), we have that, for large enough n,

$$\begin{aligned} P(||\hat{f}_{i,n} - f_i||_\infty > \epsilon) &\le M\exp\left(-\frac{3\underline{c}\pi_{n_e}h^d_{n_e}n_e}{56}\right) + 2M\exp\left(-\frac{\underline{c}h^d_{n_e}\pi^2_{n_e}n_e(\epsilon - w(Lh_{n_e}; f_i))^2}{8(v^2 + c(\pi_{n_e}/2)(\epsilon)}\right) + M\exp\left(-\frac{3\underline{c}h^d_{n_e}n_e}{28}\right) + \exp\left(-\frac{3n_e}{14}\right) \\ &\le M\exp\left(-\frac{3\underline{\tilde{c}}\pi_{q(n)}h^d_{q(n)}q(n)}{112}\right) + 2M\exp\left(-\frac{\underline{\tilde{c}}h^d_{q(n)}\pi^2_{q(n)}q(n)(\epsilon - w(Lh_{q(n)}; f_i))^2}{16(v^2 + c(\pi_{q(n)}/2)(\epsilon)}\right) \\ &\quad + M\exp\left(-\frac{3\underline{\tilde{c}}h^d_{q(n)}q(n)}{56}\right) + \exp\left(-\frac{3a_1 q(n)}{28}\right). \end{aligned} \tag{11}$$

where, $\underline{\tilde{c}}$ is a new constant that incorporates functions of a_1 and $\underline{c}$. It can be seen that the above upper bound is summable in n under the condition

$$\frac{h^d_{q(n)}\pi^2_{q(n)}q(n)}{\log n} \to \infty. \tag{12}$$

Since ϵ is arbitrary, by the Borel-Cantelli Lemma, we have that $||\hat{f}_{i,n} - f_i||_\infty \to 0$, almost surely. This is true for all arms $1 \le i \le \ell$. Note that the result a) is similarly obtained by using (4) from Lemma 2 to obtain a result similar to (9) but with π_n instead of π_{τ_n}. Now, we can invoke Theorem 1 to establish strong consistency for both the strategies using the histogram method. □

4.2 | Kernel Regression

We can obtain analogous results for strong consistency of strategy η_1 and η_2 using Nadaraya-Watson estimator. Consider a nonnegative kernel function $\mathsf{K}(\mathsf{u}) : \mathbb{R}^{\mathsf{d}} \to \mathbb{R}$ that satisfies the following Lipschitz and boundedness conditions.

Assumption 7. For some constants $0 < \lambda < \infty$, $|\mathsf{K}(\mathsf{u}) - \mathsf{K}(\mathsf{u}')| \leq \lambda ||\mathsf{u} - \mathsf{u}'||_\infty$, for all $u, u' \in \mathbb{R}^d$.

Assumption 8. $\exists$ constants $\mathsf{L}_1 \leq \mathsf{L}, \mathsf{c}_3 > 0$ and $\mathsf{c}_4 \geq 1$ such that $\mathsf{K}(\mathsf{u}) = 0$ for $||\mathsf{u}||_\infty > \mathsf{L}$, $\mathsf{K}(\mathsf{u}) \geq \mathsf{c}_3$ for $||\mathsf{u}||_\infty \leq \mathsf{L}_1$, and $\mathsf{K}(\mathsf{u}) \leq \mathsf{c}_4$ for all $\mathsf{u} \in \mathbb{R}^{\mathsf{d}}$.

Recall, $\tau_{\mathsf{n}} = \sum_{\mathsf{j}=1}^{\mathsf{n}} \mathsf{I}(\mathsf{t}_{\mathsf{j}} \leq \mathsf{n})$, the number of observed rewards by time n. Define, $\mathsf{J}_{\mathsf{i},\mathsf{n}+1} = \{\mathsf{j} : \mathsf{I}_{\mathsf{j}} = \mathsf{i}, \mathsf{t}_{\mathsf{j}} \leq \mathsf{n}, 1 \leq \mathsf{j} \leq \mathsf{n}\}$, that is, the set of time points corresponding to pulling of arm i whose rewards have been observed by time n. Let $\mathsf{M}_{\mathsf{i},\mathsf{n}+1}$ denote the size of $\mathsf{J}_{\mathsf{i},\mathsf{n}+1}$.

Let $\mathsf{h}_{\tau_{\mathsf{n}}}$ denote the bandwidth, where $\mathsf{h}_{\tau_{\mathsf{n}}} \to 0$ almost surely as $\mathsf{n} \to \infty$. For each arm i, the Nadaraya-Watson estimator of $\mathsf{f}_{\mathsf{i}}(\mathsf{x})$ is defined as,

$$\hat{\mathsf{f}}_{\mathsf{i},\mathsf{n}+1}(\mathsf{x}) = \frac{\sum_{\mathsf{j} \in \mathsf{J}_{\mathsf{i},\mathsf{n}+1}} \mathsf{Y}_{\mathsf{i},\mathsf{j}} \mathsf{K}\left(\frac{\mathsf{x}-\mathsf{X}_{\mathsf{j}}}{\mathsf{h}_{\tau_{\mathsf{n}}}}\right)}{\sum_{\mathsf{j} \in \mathsf{J}_{\mathsf{i},\mathsf{n}+1}} \mathsf{K}\left(\frac{\mathsf{x}-\mathsf{X}_{\mathsf{j}}}{\mathsf{h}_{\tau_{\mathsf{n}}}}\right)}. \tag{13}$$

Theorem 3. Suppose Assumptions 2-8 are satisfied, and,

1. If $\{\mathsf{h}_{\mathsf{n}}\}$ and $\{\pi_{\mathsf{n}}\}$ are chosen to satisfy,

$$\frac{\mathsf{q}(\mathsf{n}) \mathsf{h}_{\mathsf{q}(\mathsf{n})}^{2\mathsf{d}} \pi_{\mathsf{n}}^4}{\log \mathsf{n}} \to \infty,$$

then the Nadaraya-Watson estimator defined in (13) is strongly consistent in L_∞ norm for strategy η_1.

2. If $\{\mathsf{h}_{\mathsf{n}}\}$ and $\{\pi_{\mathsf{n}}\}$ are chosen to satisfy,

$$\frac{\mathsf{q}(\mathsf{n}) \mathsf{h}_{\mathsf{q}(\mathsf{n})}^{2\mathsf{d}} \pi_{\mathsf{q}(\mathsf{n})}^4}{\log \mathsf{n}} \to \infty,$$

then the Nadaraya-Watson estimator defined in (13) is strongly consistent in L_∞ norm for strategy η_2.

Proof. The proof for this theorem can be found in the supporting material (Section 6). □

4.3 | Strategy η_1 versus Strategy η_2

? conduct an analysis for the randomized allocation strategy with $\mathsf{h}_{\mathsf{n}}, \pi_{\mathsf{n}}$, that is, when both sequences are updated at every time point regardless of the delays, and establish its strong consistency. It states that, for $\mathsf{q}(\mathsf{n})$ as in Assumption 6, if $\mathsf{h}_{\mathsf{n}}, \pi_{\mathsf{n}}$ are chosen to satisfy,

$$\frac{\mathsf{h}_{\mathsf{n}}^{\mathsf{d}} \pi_{\mathsf{n}}^2 \mathsf{q}(\mathsf{n})}{\log \mathsf{n}} \to \infty \text{ as } \mathsf{n} \to \infty, \tag{14}$$

then the proposed allocation rule is strongly consistent for the histogram method. Note that, in terms of handling the delays, this allocation rule is in the opposite direction of the black-box approach that simply applies an existing method on the available data (i.e., ignoring all the cases with unobserved rewards at the time of decision). The sharp contrast called for the present investigation of the alternative ways to use π_{n} and h_{n} and understand their relative strengths and weaknesses.

Now if we compare (6), (7) and (14), we see that (14) $\Rightarrow$ (7) $\Rightarrow$ (6), but not vice versa, therefore (6) seems to give more options for the choice of the user-determined sequences, $\{\mathsf{h}_{\mathsf{n}}\}$ and $\{\pi_{\mathsf{n}}\}$, to achieve consistency while there may be a trade-off in the rate of decrease of the average cumulative regret as we will see in the simulations. Note that, we notice a similar relationship in Theorem 3 when using Kernel regression. To understand which choices of hyper-parameter sequences help minimize the cumulative regret, let us consider the regret for a strategy η,

$$\begin{aligned}
\mathsf{R}_{\mathsf{N}}(\eta) = \sum_{\mathsf{j}=\mathsf{m}_0+1}^{\mathsf{N}} (\mathsf{f}^*(\mathsf{X}_{\mathsf{j}}) - \mathsf{f}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}})) &= \sum_{\mathsf{j}=\mathsf{m}_0+1}^{\mathsf{N}} (\mathsf{f}_{\mathsf{i}_{\mathsf{j}}^*}(\mathsf{X}_{\mathsf{j}}) - \hat{\mathsf{f}}_{\mathsf{i}_{\mathsf{j}}^*}(\mathsf{X}_{\mathsf{j}}) + \hat{\mathsf{f}}_{\mathsf{i}_{\mathsf{j}}^*}(\mathsf{X}_{\mathsf{j}}) - \hat{\mathsf{f}}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}}) + \hat{\mathsf{f}}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}}) - \mathsf{f}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}})) \\
&\leq \sum_{\mathsf{j}=\mathsf{m}_0+1}^{\mathsf{N}} (\mathsf{f}_{\mathsf{i}_{\mathsf{j}}^*}(\mathsf{X}_{\mathsf{j}}) - \hat{\mathsf{f}}_{\mathsf{i}_{\mathsf{j}}^*}(\mathsf{X}_{\mathsf{j}}) + \hat{\mathsf{f}}_{\hat{\mathsf{i}}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}}) - \hat{\mathsf{f}}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}}) + \hat{\mathsf{f}}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}}) - \mathsf{f}_{\mathsf{I}_{\mathsf{j}}}(\mathsf{X}_{\mathsf{j}})) \\
&\leq \sum_{\mathsf{j}=\mathsf{m}_0+1}^{\mathsf{N}} 2 \sup_{1 \leq \mathsf{i} \leq \ell} |\mathsf{f}_{\mathsf{i}}(\mathsf{X}_{\mathsf{j}}) - \hat{\mathsf{f}}_{\mathsf{i}}(\mathsf{X}_{\mathsf{j}})| + \mathsf{A}\mathsf{I}\{\mathsf{I}_{\mathsf{j}} \neq \hat{\mathsf{i}}_{\mathsf{j}}\}.
\end{aligned}$$

Thus we can roughly decompose the cumulative regret into estimation error and randomization error. For the no-delay setting, **?** study both these error components in a finite-time setting and show that, $\{\mathsf{h}_{\mathsf{n}}\}$ and $\{\pi_{\mathsf{n}}\}$ can be chosen to achieve an optimal (minimax) rate of convergence for the regret. In their work, the choices of $\{\mathsf{h}_{\mathsf{n}}\}$ and $\{\pi_{\mathsf{n}}\}$ also depend on the smoothness parameter of the mean reward functions. Thus in situations where the mean reward functions are simple and smoother, $\{\mathsf{h}_{\mathsf{n}}\}$ and $\{\pi_{\mathsf{n}}\}$ are chosen to be fast decaying to achieve optimal rates of convergence in no-delay situations. In contrast, for scenarios where the underlying mean reward functions are more complex, they are chosen to be relatively slow decaying in order to guarantee optimal rates. Now the question that arises in the presence of delayed rewards is that, how should sequences $\{h_n\}$ and $\{\pi_n\}$ be updated, so as to minimize the resulting cumulative regret? That is, should one update π_n to π_{n+1} (and h_n to h_{n+1}) at every time point irrespective of observing a reward or only update upon observing a new reward. Let us try to understand the impact of delay and the reward generating mechanisms on the two components of cumulative regret to answer this question.

Different nonparametric methods may be used for estimation purposes, and estimation accuracy largely depends on the complexity of the underlying mean reward functions and the amount of data available for estimation. The binwidth of methods like histogram and kernel regression, usually is a function of the number of data points available for estimation at a given point. Therefore, in the presence of delayed rewards, h_{τ_n} (τ_n being the number of observed rewards until n) seems to be the sensible choice for the binwidth. Choosing h_n may lead to inefficient estimation due to unavailability of data points in some small neighborhood of $[0,1]^d$. Therefore, employing a binwidth sequence that guarantees optimal rates of convergence in the no-delay setting, which updates only when a new reward is obtained, seems to be the right choice from an estimation point of view. Hence, we only consider the policies (η_1 and η_2) that employ h_{τ_n} as the chosen binwidth sequence. It is important to note that from an asymptotic point of view, based on our theoretical results (Theorem 2), estimation will improve with time, but this discussion is from a finite time perspective.

In terms of randomization error, delayed rewards affect this directly through the randomization scheme. This is tied to the exploration-exploitation dilemma which is in turn controlled by the exploration probability $\{\pi_n\}$. In the following illustrations, we try to convey the message of why carefully balancing exploration-exploitation is tied to updating the sequence $\{\pi_n\}$ carefully in the presence of delayed rewards, and the decision to do that can vary in different situations.

Illustration 1. Suppose that the mean reward functions are not too complex and are well-separated. In this setting, it will be easy to get good functional estimates over time, even with less observed data due to presence of large delays. Since the no-delay case is well-studied, for such a setting we could choose an exploration probability sequence $\{\pi_n\}$ that gives the optimal rate of convergence according to **?**. Now, with the delays, we need to decide whether we want to update π_n to π_{n+1} for each n or only when a new reward is observed. In this setting, it would perhaps be advantageous to opt for strategy η_1, which updates at every time step irrespective of whether a reward is obtained or not. This is because using strategy η_2 may lead to excessive exploration which may be unnecessary in such settings even for large delay situations. Thus using η_1 will lead to a smaller randomization error. In order to illustrate that, let $\mathrm{Rand}_j(\eta_1)$ and $\mathrm{Rand}_j(\eta_2)$ denote the indicator $I(I_j \neq \hat{i}_j)$ for η_1 and η_2, respectively. Let $\sigma_t = \min\{\bar{n} : \sum_{j=m_0+1}^{\bar{n}} I(t_j \leq N) \geq t\}$, that is, σ_t is the time index where the t^{th} reward is observed. Then we have that,

$$E_{\mathcal{A}_N}\Big(\sum_{j=m_0+1}^{N} \mathrm{Rand}_j(\eta_2)\Big) = \sum_{j=m_0+1}^{N} P_{\eta_2,\mathcal{A}_N}(I_j \neq \hat{i}_j) = \sum_{t=1}^{\tau_N} (\sigma_{t+1}-\sigma_t)(\ell-1)\pi_t, \tag{15}$$

where $E_{\mathcal{A}_N}$ denotes conditional expectation given $\mathcal{A}_N$, the set of indices when the rewards were observed by time N. Here, $\tau_N = \sum_{j=m_0+1}^{N} I(t_j \leq N)$, number of rewards observed between time m_0 and N. However, for strategy η_1, since the exploration probability $\{\pi_j\}$ does not depend on delays, we have that,

$$E\Big(\sum_{j=m_0+1}^{N} \mathrm{Rand}_j(\eta_1)\Big) = \sum_{j=m_0+1}^{N} P_{\eta_1}(I_j \neq \hat{i}_j) = \sum_{j=1}^{N-m_0-1} (\ell-1)\pi_j. \tag{16}$$

For brevity sake, let us denote $\bar{N} = N - m_0 - 1$ and we start the counting process at $m_0 + 1$. Now, given τ_N, the minimum value that we can get for the R.H.S. in (15) is when all the rewards from $m_0 + 1$ until τ_N are observed instantaneously and after that no reward is observed until we hit the horizon $\bar{N}$. Likewise, an approximate maximum value of R.H.S. in (15) is achieved when the rewards for $(m_0+1)^{th}$ through $(\bar{N}-\tau_N)^{th}$ arms are not observed until time $(\bar{N}-\tau_N)$, and we observe τ_n many, from time $\bar{N}-\tau_N+1$ to $\bar{N}$ respectively. Therefore,

$$\min_{\mathcal{A}_N} E_{\mathcal{A}_N}\Big(\sum_{j=m_0+1}^{N} \mathrm{Rand}_j(\eta_2)\Big) = (\ell-1)\Big[\sum_{t=1}^{\tau_N-1} \pi_t + (\bar{N}-\tau_N)\pi_{\tau_N}\Big],$$

$$\max_{\mathcal{A}_N} E_{\mathcal{A}_N}\Big(\sum_{j=m_0+1}^{N} \mathrm{Rand}_j(\eta_2)\Big) = (\ell-1)\Big[(\bar{N}-\tau_N)\pi_1 + \sum_{t=2}^{\tau_N} \pi_t\Big].$$

For the sake of illustration, assume that we observe a fraction of $\bar{N}$ by time N, that is, $\tau_N = \alpha\bar{N}$, for some $\alpha \in (0,1)$. Then we have that,

$$\min E\Big(\sum_{j=m_0+1}^{N} \mathrm{Rand}_j(\eta_2)\Big) = (\ell-1)\Big[\sum_{t=1}^{\tau_N-1} \pi_t + (1-\alpha)\bar{N}\pi_{\tau_N}\Big], \tag{17}$$

$$\max E\Big(\sum_{j=m_0+1}^{N} \mathrm{Rand}_j(\eta_2)\Big) = (\ell-1)\Big[(1-\alpha)\bar{N}\pi_1 + \sum_{t=2}^{\tau_N} \pi_t\Big]. \tag{18}$$

Notice that the terms $(1-\alpha)\bar{\mathsf{N}}\pi_1$ and $(1-\alpha)\bar{\mathsf{N}}\pi_{\tau_{\mathsf{N}}}$ in the RHS in (17) and (18) can be fairly large and grow as N increases for all reasonably fast choices of $\{\pi_{\mathsf{n}}\}$ such as, $\mathsf{n}^{-1/4}, \log^{-1}\mathsf{n}$. From (16), (17) and (18), we also get that,

$$\sum_{t=\tau_N+1}^{\bar{N}}(\ell-1)(\pi_{\tau_N}-\pi_t) \le E(\sum_{j=m_0+1}^{N}\mathrm{Rand}_j(\eta_2)-\mathrm{Rand}_j(\eta_1)) \le \sum_{t=\tau_N+1}^{\bar{N}}(\ell-1)(\pi_1-\pi_t), \tag{19}$$

where it can be seen that $\sum_{t=\tau_N+1}^{\bar{N}}(\ell-1)(\pi_{\tau_N}-\pi_t) > 0$ for any N and $\sum_{t=\tau_N+1}^{\bar{N}}(\ell-1)(\pi_1-\pi_t) \to \infty$ as $N \to \infty$. Therefore, we see that using strategy η_1, which updates π_n at every time step irrespective of having observed a reward or not, gives a lower randomization error on average as compared to strategy η_2. For example, if we choose $\{\pi_n\} = n^{-1/4}$, $\alpha = 0.25$ (one-fourth of rewards observed) and $m_0 = 30$ (initialization phase), time horizon $N = 10000$, then we get that the average randomization error difference approximately satisfies,

$$0.02(\ell-1) \le \frac{E(\sum_{j=m_0+1}^{N}\mathrm{Rand}_j(\eta_2)-\mathrm{Rand}_j(\eta_1))}{N-(m_0+1)} \le 0.23(\ell-1), \text{ for } N = 10000, m_0 = 30.$$

In situations where mean reward functions are not complex, the randomization error can be quite large and potentially dominate over the estimation error. Thus, using strategy η_1 may reduce the cumulative regret substantially as compared to strategy η_2 in such situations.

Illustration 2. On the other hand, there are situations in which it may be better to use strategy η_2 with π_{τ_n} (updating only when a new reward is observed) as the exploration probability sequence. For example, scenarios where the best arms frequently alternate over regions of covariate space in terms of maximizing reward and it is hard to tell a clear winner with less information available due to presence of large delays. Another such situation is when an arm which is inferior in majority of the covariate space, but is superior with a substantial reward gain in a very small area of the domain and it might be the case that under large delays these under-represented regions remain unexplored. As described, let us assume that the underlying mean reward functions are somewhat complex. In such settings, we would need substantial exploration for a long period of time, specially in the presence of large delays. Here, in the hope of reducing the randomization error, we could employ strategy η_1 and use an exploration probability sequence π_n, which meets the conditions in **?** that ensure optimal convergence rates in no-delay situations. However, this could be disadvantageous in such complex settings. This is because using η_1 may lead to insufficient exploration for the inferior arms. We consider the event that a seemingly inferior arm is chosen at time t, that is, $I(I_t \neq \hat{i}_t)$. Then to ensure enough exploration, we need that this event occurs with a positive probability that is not too small, specially in such complex settings as discussed above. From **?** and **?** for no delay settings, we know that it is necessary to have $\sum_{t=1}^{\infty}\pi_t = \infty$ for the algorithm to perform optimally both asymptotically and in finite time. We also know that $\tau_N \overset{a.s.}{\to} \infty$ as $N \to \infty$. Therefore, using both these facts, the sum of probability of the event $\{I(I_t \neq \hat{i}_t), t \ge 1\}$, over the time points where rewards are observed for strategy η_2 goes to ∞,

$$\sum_{t=1}^{\tau_N} P_{\eta_2}(I_t \neq \hat{i}_t) = \sum_{t=1}^{\tau_N}(\ell-1)\pi_t \overset{a.s.}{\to} \infty, \text{ as } N \to \infty,$$

whereas, for η_1, this sum could actually be summable for large delay situations. Let $\sigma_t = \min\{\bar{n} : \sum_{j=m_0+1}^{\bar{n}} I(t_j \le N) \ge t\}$. Let us assume that the observed rewards are equally spaced, that is, $\sigma_t = tN/\tau_N$, assuming w.l.o.g that N/τ_N is an integer. Then, we have,

$$\sum_{t=1}^{\tau_N} P_{\eta_1}(I_t \neq \hat{i}_t) = \sum_{t=1}^{\tau_N}(\ell-1)\pi_{\sigma_t} = \sum_{t=1}^{\tau_N}(\ell-1)\pi_{tN/\tau_N}.$$

Now, it can be shown that this series is summable for various choices of $\{\pi_n\}$. For example, let $\{\pi_n\} = n^{-1/2}$, then for strategy η_1,

$$\sum_{t=1}^{\tau_N} P_{\eta_1}(I_t \neq \hat{i}_t) = \sum_{t=1}^{\tau_N}(\ell-1)\pi_{tN/\tau_N} = \sum_{t=1}^{\tau_N}\left(\frac{tN}{\tau_N}\right)^{-1/2} = \left(\frac{N}{\tau_N}\right)^{-1/2}\sum_{t=1}^{\tau_N} t^{-1/2} = O\left(\frac{\tau_N}{\sqrt{N}}\right). \tag{20}$$

If the number of observed rewards are small, say $\tau_N = O(\sqrt{N})$, then the series in (20) is summable. Therefore by Borel-Cantelli Lemma, the event $\{I_t \neq \hat{i}_t\}$ occurs only finitely many times out of all instances where the rewards are observed. This will lead to insufficient exploration and may incur large regret in areas that remain unexplored, specially in the more complex settings. Therefore, if we employ strategy η_1 in such settings with large delays, we may end up over-exploiting certain arms and as a result obtain insufficient number of rewards pertaining to a seemingly inferior arm, which may possibly yield higher rewards in some unexplored regions in future. This would adversely affect the performance of the algorithm and lead to high cumulative regret. Therefore, in scenarios like this, it would be advantageous to use strategy η_2.

Note that, η_2 can be thought of as a black-box procedure, in the sense that it only updates at the time points where at least one reward is observed as if there were no delays. From the above discussion, we can conclude that taking the black-box approach might not necessarily be the best in handling delayed rewards in a contextual bandit problem. In the next section, we demonstrate these ideas using four different simulation setups and illustrate the performance of strategies η_1 and η_2 in the four setups respectively. These insights also suggest the need for studying adaptive strategies for updating these parameters in a local fashion, a promising direction to explore in future.

5 | SIMULATIONS

We conduct a simulation study to compare the per-round average regret for strategies η_1 and η_2 under different delayed rewards scenarios. The per-round regret for strategy η is given by,

$$r_n(\eta) = \frac{1}{n}\sum_{j=1}^{n}(f^*(X_j) - f_{I_j}(X_j)).$$

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 $R_n(\eta) \to 1$ a.s. is equivalent to $r_n(\eta) \to 0$ a.s. The data has been generated from the following mean reward functions. We assume $d = 2, \ell = 2$ (or 3) and $x \in [0,1]^2$ and the simulations run until time $N = 8000$ with first 30 rounds of initialization. For each of the setups, we define one-dimensional functions g_1 and g_2, and then for $x_1, x_2 \in [0,1]$, we define, $f_1(x_1, x_2) = g_1(x_1) * x_2$ and $f_2(x_1, x_2) = g_2(x_1) * x_2$.

Setup 1: In this setup, we consider two well-separated sinusoidal functions, where one is a shifted above version of the other.

$$g_1(x) = (-2\sin(20\pi x) + 3),\ g_2(x) = (-2\sin(20\pi x) + 2);\ x \in [0,1].$$

Setup 2: Consider three piecewise-linear functions that are well-separated but over different regions in the covariate space. Then, $f_1(x_1, x_2) = x_2 g_1(x_1), f_2(x_1, x_2) = x_2 g_2(x_1), f_3(x_1, x_2) = x_2 g_3(x_1)$.

$$g_1(x) = \begin{cases} 1 & 0 \le x < 0.5 \\ -10x + 6 & 0.5 \le x < 0.6 \\ 0 & x \ge 0.6 \end{cases},\ g_2(x) = \begin{cases} 0 & 0 \le x < 0.5 \\ 10x - 5 & 0.5 \le x < 0.6 \\ 1 & x \ge 0.6 \end{cases},\ g_3(x) = \begin{cases} 0 & 0 \le x < 0.3 \\ 20x - 6 & 0.3 \le x < 0.4 \\ 2 & 0.4 \le x < 0.6 \\ -20x + 14 & 0.6 \le x < 0.7 \\ 0 & x \ge 0.7. \end{cases}$$

Setup 3: Consider two sinusoidal functions such that the best arm alternates rapidly as the functions oscillate.

$$g_1(x) = 2\cos(5\pi x) + 2,\ \ g_2(x) = -2\sin(5\pi x) + 2,\ \text{for } x \in [0,1].$$

Setup 4: Consider a setup where one arm dominates over majority of the covariate space, except for a small area where it incurs a considerably high regret.

$$g_1(x) = 1,\ \text{for all } x \in [0,1];\ \ g_2(x) = \begin{cases} 0 & 0 \le x < 0.5, 0.505 \le x \le 1 \\ 100000x - 50000 & 0.5 \le x < 0.502 \\ 200 & 0.502 \le x < 0.503 \\ -100000 * x + 50500 & 0.503 \le x < 0.505. \end{cases}$$

We look at both the setups 1) $d = 1$, when $f_1(x) = g_1(x)$ and $f_2(x) = g_2(x)$ and 2) $d = 2$, when $f_1(x_1, x_2) = g_1(x_1) * x_2$ and $f_2(x_1, x_2) = g_2(x_1) * x_2$, but only the results for 2) are displayed in Figure 1. The one dimensional functions g_i for each of these setups are plotted in Figure 1.

5.1 | The simulation process and results

We simulate the data from the above mentioned true mean reward functions as: $Y_{i,j} = f_i(X_j) + 0.5\epsilon_j,\ i \in \{1,2,3\}, j \in \mathbb{N}$, where $\epsilon_j \overset{\text{i.i.d.}}{\sim} N(0,1)$. We use Nadaraya-Watson estimator with Gaussian kernel to estimate the mean reward functions. We run both strategies η_1 and η_2 as in Section 3.1. We consider the following choices of hyper-parameter sequences but in our discussion, we only illustrate a few combinations to make a comparison for the sake of brevity.

$$\pi_n = \{n^{-1/4}, \log^{-1} n, \log^{-2} n; n \ge 1\} \text{ and } h_n = \{n^{-1/4}, n^{-1/6}, \log^{-1} n; n \ge 1\}.$$

Both strategies η_1 and η_2 are run for 60 independent replications (time horizon $N = 8000$). Then the regret is averaged for each time point over the replications, to give a more accurate estimate of the total regret accumulated up to a given time horizon. We create delay scenarios governing when a reward will be observed. We consider the following delay scenarios in the increased order of severity of delays,

No delay; Every reward is observed instantaneously.

Delay 1: Geometric delay with probability of success (observing the reward) $p = 0.3$.

Delay 2: Every 5th reward is not observed by time N and other rewards are obtained with a geometric ($p = 0.3$) delay.

Delay 3: Each case has probability 0.7 to delay and the delay is half-normal with scale parameter, $\sigma = 1500$.

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

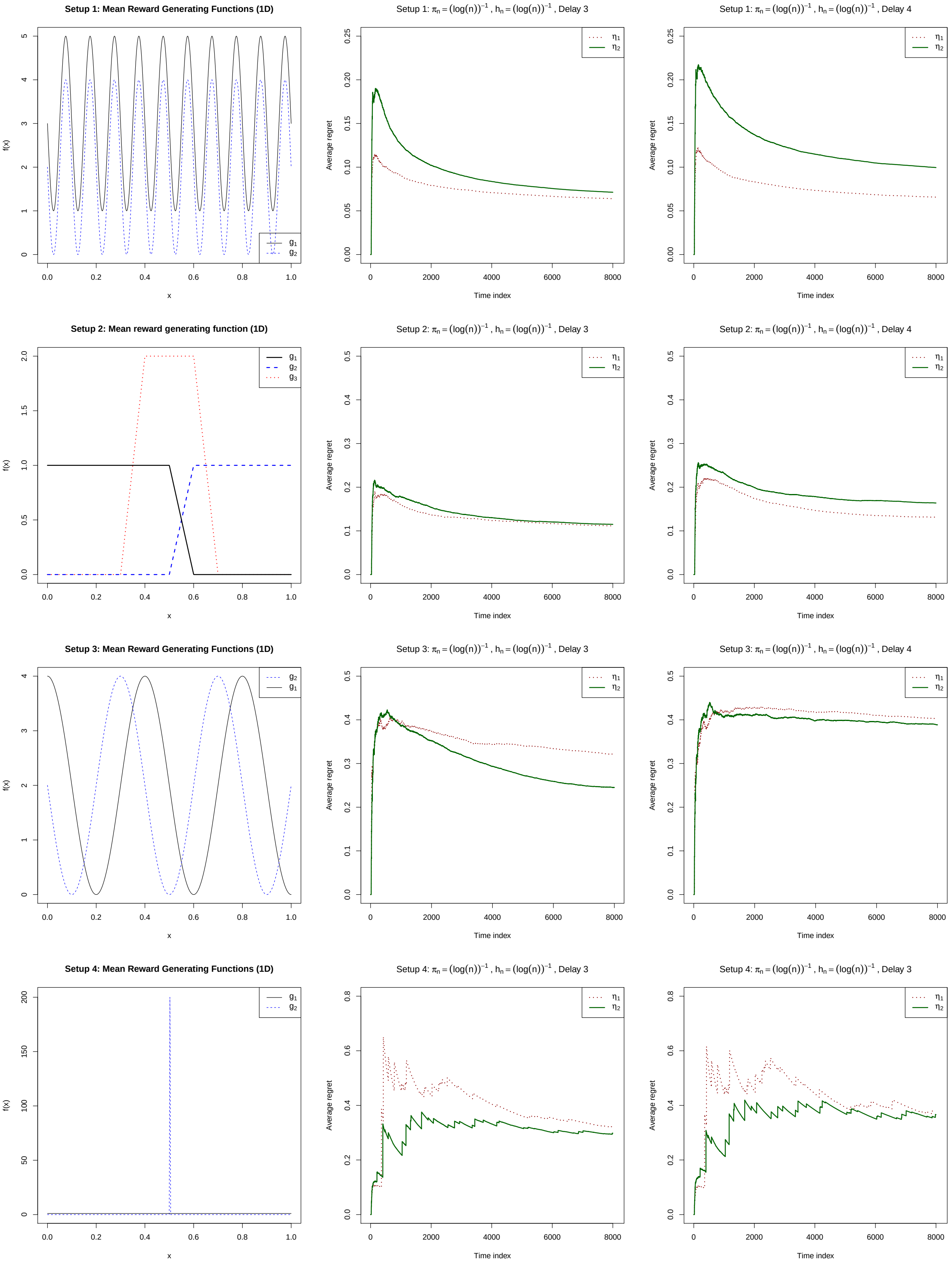

FIGURE 1 Strategy η_1 has lower cumulative average regret in setups 1 and 2 (first two rows) and strategy η_2 has lower cumulative average regret in setups 3 and 4 (rows third and fourth).

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 our simulations, we note that the difference in the cumulative regret is most discernible in the more extreme delay situations, that is, delay 3 and delay 4 in our setup. Therefore, we only illustrate the results on those two delay scenarios. The plots in Figure 1 can be used to compare performance of strategy η_1 and η_2. On the y-axis is the average regret plotted against time on the x-axis. The rows in the figure correspond to the simulation setups and columns 2 and 3 correspond to Delay 3 and Delay 4 respectively. For illustration, we only show the plots corresponding to one choice of hyper-parameter sequences, $\{h_n\} = (\log n)^{-1}$ and $\{\pi_n\} = (\log n)^{-1}$, however results from other combinations show similar trends and are included in the supporting material (section 6).

Note that in setups 1 and 2, η_1 performs better than η_2 in terms of reducing the overall average regret. Both these setups consist of mean reward functions that are well-separated and clear winners in terms of reward gain in substantial portions of the covariate space. Therefore, it is likely that one can get good estimation even in large delay setting when only small amount of observed data is available for estimation. Thus, in these settings, controlling for the randomization error is crucial, which is better achieved by using π_n instead of π_{τ_n}, as illustrated in Section 4.3. On the contrary, in Setup 3 and 4, we notice that strategy η_2 performs better than η_1 in terms of lower average regret. This can be attributed to the fact that under large delay settings, one may require more exploration for a longer period of time to get good estimates for the complex mean reward functions. Therefore, using π_{τ_n} instead of π_n helps improve the mean reward function estimation by exploring for a longer time, leading to a greater chance of exploring the more localized high regret incurring regions of the covariate space. Another interesting observation is that for setups 1 and 2, the average regret curves for strategies η_1 and η_2 are closer with Delay 3 and much separated with Delay 4. Whereas, in setups 3 and 4, an opposite trend is seen, where the difference in the average regret curves for η_1 and η_2 is more pronounced with Delay 3 as compared to Delay 4. A possible reason for this could be that the mean reward functions for setups 1 and 2 are easily distinguishable even with as few observations as with Delay 4, thus fast and continuous exploitation helps reduce the regret. However, the mean reward functions in setups 3 and 4 are harder to distinguish and perhaps with so few observations as in Delay 4, it is hard to do a good job in estimation even while exploring more using π_{τ_n}.

6 | CONCLUSION

In this work, we present a case on the importance of carefully choosing a contextual bandit strategy based on the expected delay situation. Delays are assumed to be independent, but unbounded and could potentially be infinite as long as we expect to see a minimum number of observed rewards in finite time, and have some knowledge of a lower bound to the expected number of observations. We propose two ϵ-greedy like strategies, adopting a nonparametric approach to modeling the mean reward regression functions. In both strategies, the binwidth sequence $\{h_n\}$ is updated only when new rewards are observed, but the difference lies in updating the exploration probability $\{\pi_n\}$. In one strategy, $\{\pi_n\}$ is only updated when a new reward is observed (like a black-box procedure), while in the second strategy, $\{\pi_n\}$ is updated at every time point irrespective of having observed a reward or not. We establish strong consistency for both the strategies and compare the necessary condition required to achieve consistency with the analogous condition that appeared in **?**. Then, using some theoretical illustrations and simulation examples, we show that both these strategies may be advantageous in different settings depending on the underlying data generating scenarios and the severity of the delays in observing rewards. Therefore, based on these empirical results, we recommend that the choice of hyper-parameters $\{h_n\}$ and $\{\pi_n\}$ should depend on the context of the problem, delay scenario, and some broad knowledge of the data generating process. An immediate future direction based on these results is to devise an adaptive strategy which decides whether to update the hyperparameter sequences or not in a more localized way. Conducting a finite-time regret analysis to theoretically prove the insights obtained would help better understand the problem and we hope to address it in future work. It is important to note that optimal arm identifiability and regret minimization may not agree with each other in all problems. It is possible that two different algorithms achieve about the same cumulative regret, despite of one being poor at identifying the best arms as compared to the other, thus is a different problem altogether and requires a different set of tools to address the problem. In our knowledge, best arm identification in delayed rewards for contextual bandits has not been studied so far and would be an interesting future work to consider.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:

Proof for Theorem 1 for strategy η_2 in Section S.1.

Proof for Theorem 3 in Section S.2. Proof of strong consistency result for the proposed strategies when the regression function is estimated using the Nadaraya-Watson estimator.

Additional simulation plots in Section S.3. Figures S.1, S.2 and S.3 corresponding to choices $\{h_n, \pi_n\} = ((\log n)^{-1}, (\log n)^{-2}), (n^{-1/4}, n^{-1/4}), (n^{-1/4}, (\log n)^{-1})$.

□