*The Annals of Statistics*
2002, Vol. 30, No. 1, 100–121

# RANDOMIZED ALLOCATION WITH NONPARAMETRIC ESTIMATION FOR A MULTI-ARMED BANDIT PROBLEM WITH COVARIATES

BY YUHONG YANG[1] AND DAN ZHU[2]

*Iowa State University*

We study a multi-armed bandit problem in a setting where covariates are available. We take a nonparametric approach to estimate the functional relationship between the response (reward) and the covariates. The estimated relationships and appropriate randomization are used to select a good arm to play for a greater expected reward. Randomization helps balance the tendency to trust the currently most promising arm with further exploration of other arms. It is shown that, with some familiar nonparametric methods (e.g., histogram), the proposed strategy is strongly consistent in the sense that the accumulated reward is asymptotically equivalent to that based on the best arm (which depends on the covariates) almost surely.

**1. Introduction.** Multi-armed bandit problems have been extensively studied in probability and statistics in the past few decades and still receive considerable interest in these and related fields. Readers are referred to Berry and Fristedt (1985) and Gittins (1989) and the references cited therein for the history, many elegant results and applications in clinical trials, scheduling and other industrial problems. Some recent developments have been reported in, for example, Lai and Yakowitz (1995) who considered the situation in which observations are dependent, Berry, Chen, Zame, Heath and Shepp (1997) who considered infinitely many arms and Auer, Cesa-Bianchi, Freund and Schapire (1995) who considered worst-case performances.

In classic bandit problems, each arm sequentially generates rewards based on a distribution with some unknown parameters, and one needs to sequentially select one arm to play for the maximum expected reward. In a majority of the earlier work, no auxiliary information beyond the observed rewards was considered when selecting an arm to play. Exceptions include Woodroofe (1979), Sarkar (1991) and Clayton (1989) who considered one-armed bandit problems with covariates. The first two papers studied Bayesian sequential allocation in non-Bernoulli bandit models with parametric frameworks and showed that the myopic rule is asymptotically optimal. The third studied Bernoulli bandit problems

Received February 2001; revised August 2001.

[1]Supported in part by NSF Grant DMS-00-94323.

[2]Supported in part by the Chinese National Science Foundation.

*AMS 2000 subject classifications.* 62L05, 62C25.

*Key words and phrases.* Multi-armed bandits, sequential allocation, randomized allocation, concomitant variable, nonparametric regression.

with covariates using a link function to relate the probability of success with the covariate. As covariates are often available in many potential applications, incorporating such information in decision making is desirable for a better performance. In this work, we consider continuous rewards with covariates available and propose a method for arm selection (allocation) with a proven asymptotic property. Our approach employs nonparametric regression procedures for estimating the dependence of the rewards on the covariates for the arms and uses a randomized allocation scheme to control the trade-off between the tendency to use the currently most promising arm and further exploration to find the arm that is truly the best (which depends on covariates in general). The use of nonparametric approaches has the advantage of more flexibility with a wider range of applications.

The rest of the paper is organized as follows. In Section 2, we set up the bandit problem to be studied. We propose our strategy in Section 3. Section 4 gives some examples of regression procedures that can be applied to the bandit problem. Consistency of the proposed strategy is established in Section 5. Sections 6 and 7 show the consistency with histogram and nearest neighbor methods, respectively. Conclusions are in Section 8. Some supporting technical results and proofs are given in the Appendix.

**2. Problem setup.** Assume that there are $\mathcal{l}$, $\mathcal{l} \geq 2$, arms available for playing. After pulling an arm, a random reward is generated. Each time before deciding which arm to pull, a $d$-dimensional characteristic or covariate (concomitant variable) $x \in R^d$ is observed. The goal is to maximize the total reward after a number of plays. We assume that characteristics (or covariates) are continuous variables and take values in a hypercube taken as $[0, 1]^d$ without loss of generality. By pulling the $i$th arm, the mean reward with the given covariate $x$ is denoted as $f_i(x)$, $1 \leq i \leq \mathcal{l}$. The actual reward with covariate $x$ of pulling the $i$th arm is modeled as

$$f_i(x) + \varepsilon,$$

where $\varepsilon$ denotes random error with mean 0 and a finite variance.

The functions $f_i$, $1 \leq i \leq \mathcal{l}$, are assumed to be unknown and not necessarily of a known parametric form. Ideally, if the $f_i$'s (but not the errors) were known, with the observed covariate $x$, one would pull the arm with the largest mean reward at $x$; that is, one would choose arm $i^*$ with $f_{i^*}(x) \geq f_i(x)$ for any other $i \neq i^*$. Note that, in general, for different $x$, the best arm is different. For the purpose of pursuing the highest possible reward, finding the arm with the best overall performance is not sufficient and the information in the covariates should be incorporated when choosing an arm to pull. Let $i^*(x)$ denote this optimal choice of arm. (The optimal choice may not be unique and one may break ties, if any, with any reasonable rule.) The corresponding mean reward, $f^*(x) = \max_{1 \leq i \leq \mathcal{l}} f_i(x)$, is the ideal mean reward with given covariate $x$.

Let $X_1, \ldots, X_n, \ldots$ be a sequence of covariates independently generated from a population supported in $[0,1]^d$. Let $P_X$ denote the underlying probability distribution, which is also assumed unknown in this work. At each time $j \geq 1$, when the covariate $X_j$ is observed, we need to select an arm to pull based on $X_j$ and the previous data. We assume that, as is typically the case in bandit problems, only one arm can be pulled at a time and therefore we will *not* observe the rewards from other arms. In addition to efficiency considerations, this is also a realistic setting for many applications where it is impossible or impractical to pull multiple arms simultaneously (e.g., applying medical procedures or testing quality of a product in a destructive way). Let $Y_{i,j}$ denote the reward of pulling the $i$th arm when the covariate $X_j$ is presented. As mentioned previoulsy, for each $j$, only one $Y_{i,j}$ will be observed. Let $\varepsilon_{ij}$ denote the error (deviation from the mean) that occurs when arm $i$ is pulled. In this work, we assume that errors associated with different realizations of the covariates and/or different arms are independent and are all independent of the $X_i$'s.

Let $\delta$ be a sequential allocation rule. Let $I_1, I_2, \ldots, I_n, \ldots$ be the chosen arm at time $1, 2, \ldots$ based only on $X_1$, on $X^2 = (X_1, X_2)$, $I_1$ and $Y_{I_1,1}, \ldots$ on $X^n = (X_1, \ldots, X_n)$, $I_1, \ldots, I_{n-1}$ and $(Y_{I_1,1}, \ldots, Y_{I_{n-1},n-1}), \ldots$ respectively. With the allocation rule, given the previous observations and $X_j$, the mean reward (averaging out the present error) at the given $X_j$ is $f_{I_j}(X_j)$ for $j \geq 1$. The total of this mean reward up to time $n$ is $\sum_{j=1}^n f_{I_j}(X_j)$. Clearly, without knowing the random errors, the ideal performance occurs when the choices $I_1, \ldots, I_n$ match $i^*(X_1), \ldots, i^*(X_n)$, yielding the optimal total (conditional) reward $\sum_{j=1}^n f^*(X_j)$. It is thus of interest to study the quantity

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

Obviously, $R_n$ is a random variable no bigger than 1. It measures the performance of the allocation rule relative to the ideal one with the optimal arm known for each $x$.

DEFINITION. An allocation rule $\delta$ is said to be strongly consistent if $R_n(\delta) \to 1$ with probability 1.

REMARK 1. It is also reasonable to study the ratio

$$\frac{\sum_{j=1}^n Y_{I_j,j}}{\sum_{j=1}^n Y_{i^*(X_j),j}}$$

as a measure of performance for allocation. As can be easily seen (see the Appendix), the two measures are basically the same.

REMARK 2. 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 $\frac{1}{n}\sum_{j=1}^{n}(f_{I_j}(X_j) - f^*(X_j)) \to 0$ a.s.

REMARK 3. In some contexts such as problem-solving in machine learning [e.g., Gratch, DeJong and Yang (1994)], one naturally wants to minimize (rather than maximize) the response (e.g., time used to solve a problem). The preceding definition and the subsequent results to be presented are also suitable for such cases with some straightforward modifications.

Clearly, consistency is a desirable property. Allocation rules will be constructed and will be shown to be strongly consistent.

Not surprisingly, efficient allocation requires that the individual functions $f_i$ be estimated to some extent. We will apply nonparametric techniques to estimate the mean reward functions and then base the selection of an arm on a consideration that involves a comparison of the estimates of $f_i$'s. To achieve efficiency, one needs to appropriately balance the tendency to use the currently most promising arm with the desire to try other arms. We use a randomization technique to automatically balance the two competing tendencies.

Our work is very different from previous results on bandit problems with covariates by Woodroofe (1979), Clayton (1989) and Sarkar (1991). The main differences are as follows: (1) we consider multi-arm problems while previous work considered one-arm problems; (2) modeling of the dependence of $Y$ on the covariates is different—we use a nonparametric regression framework and previous work assumed parametric relationships; (3) no discounting is considered in our performance measure; (4) unlike previous work, our result is not in a Bayesian framework and we study a strong consistency property of allocation; and (5) we use a randomized allocation rule and the previous work used deterministic rules. Our approach seems quite realistic for many applications.

**3. Proposed strategy.** There are two main ingredients in our approach on selecting an arm: (1) nonparametric estimation of the individual functions $f_i$ and (2) a proper allocation scheme to control the trade-off between the two competing tendencies mentioned previously.

For estimating the $f_i$'s, consider a nonparametric regression procedure, for example, histogram, nearest neighbor, kernel or local polynomial regression. At each time $n \geq 1$, let $Z^{n,i}$ denote the set of observations $(X_j, Y_{I_j,j})$, $1 \leq j \leq n$, to which the $i$th arm is pulled (i.e., $I_j = i$). Let $\hat{f}_{i,n}$ denote the regression estimator of $f_i$ based on the data $Z^{n,i}$.

The following is our proposed strategy of allocation. Let $\{\pi_j,\ j \geq 1\}$ be a sequence of positive numbers decreasing to 0.

STEP 1. Initialize. Give each arm a small number of applications. We here take $I_1 = 1, I_2 = 2, \ldots, I_\ell = \ell$ (i.e., give each arm a try).

STEP 2. Estimate the individual functions $f_i$. For $n = \ell + 1$, based on the current data $Z^{n,i}$, 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 the next covariate $X_{n+1}$, let $\hat{\imath}_{n+1}(X_{n+1})$ be the maximizer of $\hat{f}_{i,n}(X_{n+1})$ over $1 \le i \le \ell$ (if there is a tie, any tie-breaking rule can be used).

STEP 4. Select and pull. Randomly select an arm, with probability $1 - (\ell - 1)\pi_{n+1}$ for $i = \hat{\imath}_{n+1}$ (the currently most promising choice) and with probability $\pi_{n+1}$ for each of the remaining arms. [Here it is assumed that $(\ell - 1)\pi_{n+1} < 1$.] Let $I_{n+1}$ denote the selected arm. Pull the arm $I_{n+1}$ to receive the reward.

STEP 5. Update the estimates. After the new observation $X_{n+1}$, $I_{n+1}$, $Y_{I_{n+1},n+1}$, update the function estimate of $f_i$ for $i = I_{n+1}$.

STEP 6. Repeat Steps 3–5 when the next covariate $X_{n+2}$ surfaces and so on.

Note that in Step 4 a randomized selection is used. With high probability, we select the currently projectedly "best" arm (based on the estimates of the $f_i$'s), but still give other arms some chance. Since $\pi_n$ decreases to 0, with more and more data, the chance gets smaller and smaller. When the variances of the errors are large with $n$ being small or moderate, the estimates of the $f_i$'s are not very accurate, and therefore $\hat{\imath}_{n+1}$ is not very reliable. In such cases, it is better to choose a $\pi_n$ that is not too small so as to reach a sound comparison among the arms more rapidly. The same argument applies when the individual functions are not very smooth and can change rapidly. In some sense, the speed at which $\pi_n \to 0$ reflects our confidence in the accuracy of the estimates of the functions $f_i$. Like the bandwidth in kernel regression, the choice of $\pi_n$ affects the final performance of our proposed strategy. In this work, we will not pursue the issue of finding an automated choice of $\pi_n$.

The proposed allocation rule will be denoted $\delta_\pi$.

**4. Examples of regression procedures.** Various regression procedures can be used to estimate the individual mean reward function $f_i$'s. We mention a few below. We do not address design issues here and assume that the covariate values are given.

Consider the regression model

$$Y_j = f(x_j) + \varepsilon_j, \qquad 1 \le j \le n,$$

where $x_1, \ldots, x_n \in [0, 1]^d$ are given design points and the $\varepsilon_j$'s are independent errors with mean 0 and finite variance. One needs to estimate the regression function $f$.

4.1. *Histogram method.* 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 each $x$, let $J(x) = \{j : 1 \le j \le n,\ x_j$ and $x$ belong to the same cube$\}$ indicate the design points that fall in the same cube as $x$. Let $N(x)$ denote the size of $J(x)$. Then define $\hat{f}(x) = \frac{1}{N(x)} \sum_{j \in J(x)} Y_j$ to be the average of the $Y$ values in the cube [if $N(x) = 0$, define $\hat{f}(x)$ as any chosen positive constant]. For the estimator to behave well, a proper choice of $h = h_n$ is necessary. For convergence results, see, for example, Stone (1977), Devroye and Györfi (1985) and Nobel (1996) and the references therein. When applying the histogram method to our problem, one may use different widths $h_{i,n}$ for estimating the function $f_i$'s.

4.2. *Nearest neighbor method.* Consider the use of the $N_n$ nearest neighbor method for estimating the function $f$. Let $d$ be a distance on $[0, 1]^d$. A natural choice is the Euclidean distance $d(x, y) = \|x - y\| = \sqrt{\sum_{j=1}^d (x_j - y_j)^2}$. For a chosen integer $N = N_n$ and $x \in [0, 1]^d$, let $J(x; N) = \{j : 1 \le j \le n$ and $x_j$ is among the $N$ closest points to $x\}$ indicate observations with $x_j$ being the $N$ closest to $x$ in distance $d$. Then let $\hat{f}(x) = \frac{1}{N} \sum_{j \in J(x;N)} Y_j$ be the average of the $Y$ values of the $N$ nearest neighboring points to $x$. As is well known, roughly speaking, when $N$ is large, the variance of $\hat{f}$ is small but the bias of $\hat{f}$ can be large; conversely, when $N$ is small, the variance of $\hat{f}$ is large but the bias of $\hat{f}$ is small. An appropriate choice of $N$ is needed to reach an overall good performance. In general, $N = N_n$ should be chosen to increase in $n$ for the estimation risk to converge to 0. See, for example, Devroye, Györfi, Krzyżak and Lugosi (1994) and the references cited therein for convergence results on nearest neighbor methods.

4.3. *Kernel method.* When the underlying regression function $f$ is very smooth, estimates that are smoother than the histogram or nearest neighbor estimate may improve performance. Kernel and local polynomial regression techniques have been widely studied [see, e.g., Fan and Gijbels (1996)].

One can also consider other estimation methods such as polynomial or trigonometric expansion, spline methods and neural nets (for high-dimensional settings).

## 5. Consistency of the proposed strategy.

ASSUMPTION A. The regression procedure is strongly consistent in $L_\infty$ norm for all individual mean functions $f_i$ under the proposed allocation scheme, that is, $\|\hat{f}_{i,n} - f_i\|_\infty \to 0$ a.s. for each $1 \le i \le \mathcal{l}$ as $n \to \infty$.

ASSUMPTION B. The mean functions satisfy $f_i(x) \ge 0$, $A = \sup_{1 \le i \le \mathcal{l}} \sup_{x \in [0,1]^d} (f^*(x) - f_i(x)) < \infty$ and $E(f^*(X_1)) > 0$.

The first assumption requires the estimators of the individual mean functions $f_i$ to become more and more accurate as $n \to \infty$ and the conditions in the second assumption are natural in our setup of the problem or mild.

THEOREM 1. *Under Assumptions* A *and* B, *the allocation rule* $\delta_\pi$ *given in Section* 3 *is strongly consistent*.

PROOF. Since the ratio $R_n(\delta_\pi)$ is always upper bounded by 1, we only need to work on the lower bound direction. Corresponding to Step 1, define $\hat{I}_j = I_j = j$ for $1 \le j \le \mathit{l}$. Note that

$$R_n(\delta_\pi) = \frac{\sum_{j=1}^n f_{\hat{\imath}_j}(X_j)}{\sum_{j=1}^n f^*(X_j)} + \frac{\sum_{j=1}^n (f_{I_j}(X_j) - f_{\hat{\imath}_j}(X_j))}{\sum_{j=1}^n f^*(X_j)}$$
$$\ge \frac{\sum_{j=1}^n f_{\hat{\imath}_j}(X_j)}{\sum_{j=1}^n f^*(X_j)} - \frac{\frac{1}{n}\sum_{j=1}^n A I_{\{I_j \ne \hat{\imath}_j\}}}{\frac{1}{n}\sum_{j=1}^n f^*(X_j)},$$

where the inequality follows from Assumption B. [Note that, since $E(f^*(X_1)) > 0$, the denominator $\sum_{j=1}^n f^*(X_j)$ is eventually positive with probability 1.] Let $U_j = I_{\{I_j \ne \hat{\imath}_j\}}$. Since $\frac{1}{n}\sum_{j=1}^n f^*(X_j)$ converges a.s. to $Ef^*(X) > 0$, the second term in the above inequality converges to 0 almost surely if $\frac{1}{n}\sum_{j=1}^n U_j \to 0$ a.s. Note that, for $j \ge \mathit{l}+1$, the $U_j$'s are independent Bernoulli random variables with success probability $(\mathit{l}-1)\pi_j$. Since

$$\sum_{j=\mathit{l}+1}^{\infty} \mathrm{Var}\Big(\frac{U_j}{j}\Big) = \sum_{j=\mathit{l}+1}^{\infty} \frac{(\mathit{l}-1)\pi_j(1-(\mathit{l}-1)\pi_j)}{j^2} < \infty,$$

we have $\sum_{j=\mathit{l}+1}^{\infty}((U_j - (\mathit{l}-1)\pi_j)/j)$ converging almost surely. It then follows by Kronecker's lemma that $\frac{1}{n}\sum_{j=1}^n (U_j - (\mathit{l}-1)\pi_j) \to 0$ a.s. Observing that $\frac{1}{n}\sum_{j=1}^n (\mathit{l}-1)\pi_j \to 0$ since $\pi_j \to 0$ as $j \to \infty$, we thus know $\frac{1}{n}\sum_{j=1}^n U_j \to 0$ a.s.

From above, to show $R_n(\delta_\pi) \to 1$ a.s., it remains to show $\sum_{j=1}^n f_{\hat{\imath}_j}(X_j)/\sum_{j=1}^n f^*(X_j) \to 1$ a.s. or, equivalently, $\sum_{j=1}^n (f_{\hat{\imath}_j}(X_j) - f^*(X_j))/\sum_{j=1}^n f^*(X_j) \to 0$ a.s. By the definition of $\hat{\imath}_j$, for $j \ge \mathit{l}+1$, we have $\hat{f}_{\hat{\imath}_j, j-1}(X_j) \ge \hat{f}_{i^*(X_j), j-1}(X_j)$ and thus

$$\begin{aligned} f_{\hat{\imath}_j}(X_j) - f^*(X_j) &= f_{\hat{\imath}_j}(X_j) - \hat{f}_{\hat{\imath}_j, j-1}(X_j) + \hat{f}_{\hat{\imath}_j, j-1}(X_j) - \hat{f}_{i^*(X_j), j-1}(X_j) \\ &\quad + \hat{f}_{i^*(X_j), j-1}(X_j) - f_{i^*(X_j)}(X_j) \\ &\ge f_{\hat{\imath}_j}(X_j) - \hat{f}_{\hat{\imath}_j, j-1}(X_j) + \hat{f}_{i^*(X_j), j-1}(X_j) - f_{i^*(X_j)}(X_j) \\ &\ge -2 \sup_{1 \le i \le \mathit{l}} \|\hat{f}_{i, j-1} - f_i\|_\infty. \end{aligned}$$

For $1 \le j \le \ell$, we have $f_{\hat{i}_j}(X_j) - f^*(X_j) \ge -A$. Based on Assumption A, $\|\hat{f}_{i,j-1} - f_i\|_\infty \to 0$ a.s. as $j \to \infty$ for each $i$, and thus $\sup_{1\le i\le \ell} \|\hat{f}_{i,j-1} - f_i\|_\infty \to 0$ a.s. It follows that for, $n > \ell$,

$$\frac{\sum_{j=1}^n (f_{\hat{i}_j}(X_j) - f^*(X_j))}{\sum_{j=1}^n f^*(X_j)} \ge \frac{-A\ell/n - (2/n)\sum_{j=\ell+1}^n \sup_{1\le i\le \ell} \|\hat{f}_{i,j-1} - f_i\|_\infty}{(1/n)\sum_{j=1}^n f^*(X_j)}.$$

Clearly the right-hand side converges to 0 almost surely. By the definition of $f^*$, the left-hand side is upper bounded by 0. The conclusion follows. This completes the proof of Theorem 1. □

Although Assumption A for Theorem 1 seems quite natural, it is somewhat heavy since it imposes a condition in terms of both the estimation procedure and the allocation scheme. It may be difficult to check in general. In the next two sections, we verify it for two cases, namely, histogram and nearest neighbor procedures.

**6. Allocation with histogram estimates.** In this section, we show that the histogram regression procedure described in Section 4 together with the allocation scheme in Section 3 leads to strong consistency under some reasonable conditions on the random errors, design distribution and the individual mean functions $f_i$.

ASSUMPTION 1. The functions $f_i$ are nonnegative and continuous on $[0, 1]^d$ and $Ef^*(X_1) > 0$.

ASSUMPTION 2. 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 $\underline{c} \le p(x) \le \overline{c}$ for some positive constants $\underline{c} < \overline{c}$.

ASSUMPTION 3. The errors satisfy a moment condition that there exist positive constants $v$ and $c$ such that, for all $m \ge 2$,

$$E|\varepsilon_{ij}|^m \le \frac{m!}{2} v^2 c^{m-2}. \tag{1}$$

Note that the condition does not require identical distribution of errors. The condition is often called the (refined) Bernstein condition [see, e.g., van der Vaart and Wellner (1996), Lemma 2.2.11, and Birgé and Massart (1998), Lemma 8].

For simplicity, for each $n$, the same side width $h_n$ is used for histogram estimations of the different functions $f_i$, $1 \le i \le \ell$, based on the data prior to the next covariate $X_{n+1}$.

THEOREM 2. *Suppose Assumptions* 1–3 *are satisfied. If* $h = h_n$ *and* $\pi_n$ *are chosen to satisfy*

$$\frac{nh^d\pi_n^2}{\log n} \to \infty,$$

*then the allocation rule* $\delta_\pi$ *is strongly consistent.*

REMARK. If there are no available covariates or the available covariates are irrelevant, then $f_i(x) \equiv c_i$ for some positive constants $c_i$ for $1 \le i \le \mathcal{I}$. In this case, the best arm $i^*(X)$ does not depend on $X$. By Theorem 2, the allocation rule asymptotically does as well as the arm $i^*$. Consistency (in expectation instead of a.s.) for a two-armed bandit problem without covariates was first obtained by Robbins (1952). Lai and Robbins (1985) moved a step forward by constructing asymptotically efficient allocation rules.

Note that, for estimating the individual functions $f_i$, the observations so far are divided into $\mathcal{I}$ subsamples according to the arms pulled. Intuitively, if the covariate values in each subsample are eventually dense in $[0, 1]^d$, the histogram estimators should become more and more accurate. Technically speaking, however, it is quite nontrivial to verify Assumption A. Strong consistency of histogram estimators (under $L_\infty$ norm) in a regular regression setting does not readily imply the satisfaction of Assumption A. A major difficulty in the analysis arises from the fact that the $Y$'s in each subsample are no longer independent since the allocation rule ties them together.

PROOF OF THEOREM 2. By Theorem 1, since Assumption B is clearly satisfied, we only need to verify Assumption A, that is, show that the histogram method is strongly consistent in $L_\infty$ norm for estimating $f_i$'s under the allocation scheme given in Section 3.

The histogram technique partitions the unit cube into $M = (1/h)^n$ small cubes. Under Assumption 2, from Section A.5, for each small cube $C_l$, $1 \le l \le M$, in the partition, the number of observations $X_j$, $1 \le j \le n$, that fall in the cube, denoted by $N_l$, is unlikely to be very small relative to $nh^d$ as shown in the inequality:

$$P\left(N_l \le \frac{\underline{c}nh^d}{2}\right) \le \exp\left(-\frac{3\underline{c}nh^d}{28}\right).$$

It follows that

$$P\left(\min_{1\le l\le M} N_l \le \frac{\underline{c}nh^d}{2}\right) \le M\exp\left(-\frac{3\underline{c}nh^d}{28}\right). \tag{2}$$

Now condition on a realization of the design variables $X_1 = x_1, \ldots, X_n = x_n$. Consider the estimation of $f_i(x)$ for a fixed $i$ in $\{1, \ldots, \mathcal{I}\}$. Let $W_1, \ldots, W_n$ be Bernoulli random variables indicating whether the $i$th arm is selected ($W_j = 1$) for the characteristic $X_j$, $1 \le j \le n$, or not ($W_j = 0$). Note that, conditional on the

previous observations and $X_j$, the probability of $W_j = 1$ is almost surely lower bounded by $\pi_j \geq \pi_n$ for $1 \leq j \leq n$ (since $\pi_j$ is nonincreasing). Let $\omega(h; f_i)$ be the modulus of continuity, as defined in (5) in Section A.2, of the function $f_i$. Under the continuity assumption on $f_i$, we have $\omega(h; f_i) \to 0$ as $h \to 0$. Thus, for any $\epsilon > 0$, when $h$ is small enough, $\epsilon - \omega(h; f_i) \geq \epsilon/2$. From Lemma 1 in Section A.2, we have that, given the design points, for any $\epsilon > 0$, when $h = h_n$ is small enough,

$$
\begin{aligned}
P_{x^n}\bigl(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon\bigr) &\leq M \exp\biggl(-\frac{3\pi_n \min_{1\leq l\leq M} N_l}{28}\biggr) \\
&\quad + 2M \exp\biggl(-\frac{\min_{1\leq l\leq M} N_l \pi_n^2 \bigl(\epsilon - \omega(h; f_i)\bigr)^2}{8\bigl(v^2 + c(\pi_n/2)\bigl(\epsilon - \omega(h; f_i)\bigr)\bigr)}\biggr).
\end{aligned} \tag{3}
$$

Here for applying Lemma 1, we used the observation that, for each $j \geq 1$, $W_j$ is independent of the $\epsilon_{i,l}$'s for all $l \geq j$ since $W_j$ depends only on the previous observations and $X_j$. From (3), we have

$$
\begin{aligned}
&P_{x^n}\biggl(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon, \min_{1\leq l\leq M} N_l \geq \frac{\underline{c} n h^d}{2}\biggr) \\
&\quad \leq \begin{cases}
M \exp\biggl(-\dfrac{3\underline{c} n h^d \pi_n}{56}\biggr) + 2M \exp\biggl(-\dfrac{\underline{c} n h^d \pi_n^2 \bigl(\epsilon - \omega(h; f_i)\bigr)^2}{16\bigl(v^2 + c(\pi_n/2)\bigl(\epsilon - \omega(h; f_i)\bigr)\bigr)}\biggr), \\
\qquad \text{when } \min_{1\leq l\leq M} N_l \geq \dfrac{\underline{c} n h^d}{2}, \\
0, \qquad \text{otherwise.}
\end{cases}
\end{aligned}
$$

Together with (2), we have

$$
\begin{aligned}
&P\bigl(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon\bigr) \\
&\quad = P\biggl(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon, \min_{1\leq l\leq M} N_l < \frac{\underline{c} n h^d}{2}\biggr) \\
&\qquad + P\biggl(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon, \min_{1\leq l\leq M} N_l \geq \frac{\underline{c} n h^d}{2}\biggr) \\
&\quad \leq P\biggl(\min_{1\leq l\leq M} N_l < \frac{\underline{c} n h^d}{2}\biggr) + E\, P_{X^n}\biggl(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon, \min_{1\leq l\leq M} N_l \geq \frac{\underline{c} n h^d}{2}\biggr) \\
&\quad \leq M \exp\biggl(-\frac{3\underline{c} n h^d}{28}\biggr) + M \exp\biggl(-\frac{3\underline{c} n h^d \pi_n}{56}\biggr) \\
&\qquad + 2M \exp\biggl(-\frac{\underline{c} n h^d \pi_n^2 \bigl(\epsilon - \omega(h; f_i)\bigr)^2}{16\bigl(v^2 + c(\pi_n/2)\bigl(\epsilon - \omega(h; f_i)\bigr)\bigr)}\biggr).
\end{aligned}
$$

It is straightforward to show that, under the condition $nh^d\pi_n^2/\log n \to \infty$, the above upper bound is summable in $n$ and thus

$$\sum_{n=1}^{\infty} P\big(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon\big) < \infty.$$

Since $\epsilon$ is arbitrary, by the Borel–Cantelli lemma, $\|\hat{f}_{i,n} - f_i\|_\infty \to 0$. This completes the proof of Theorem 2. $\square$

**7. Allocation with nearest neighbor estimates.** Like the histogram approach, nearest neighbor estimators can be used to achieve strong consistency.

For estimating the functions $f_i$ based on the information accumulated before the next covariate $X_{n+1}$, choose an integer $N = N_n$. For $x \in [0,1]^d$, let $J(x; N) = \{j: 1 \leq j \leq n$ and $X_j$ is among the $N$ closest points to $x\}$. For $1 \leq i \leq \mathcal{I}$, let $J_i(x; N)$ be the subset of $J(x; N)$ that corresponds to pulling the $i$th arm. Let $N_i(x)$ denote the size of $J_i(x; N)$. Then, for $1 \leq i \leq \mathcal{I}$, let

$$\hat{f}_{i,n}(x) = \frac{1}{N_i(x)} \sum_{j \in J_i(x;N)} Y_{i,j}$$

be the estimator of $f_i(x)$. [If $N_i(x) = 0$, define $\hat{f}_{i,n}(x)$ to be any chosen positive constant.]

THEOREM 3. *Suppose Assumptions* 1–3 *are satisfied. With the $N_n$ nearest neighbor estimators defined previously, if $N_n$ and $\pi_n$ are chosen to satisfy*

$$\frac{N_n\pi_n^2}{\log n} \to \infty,$$

$$\frac{N_n}{n} \to 0,$$

*then the allocation rule $\delta_\pi$ is strongly consistent.*

PROOF. Again, by Theorem 1, since Assumption B is satisfied, we only need to show that the nearest neighbor method is strongly consistent in $L_\infty$ norm for estimating $f_i$'s under the allocation scheme given in Section 3.

Fix $1 \leq i \leq \mathcal{I}$. For $x = (x_1, \ldots, x_d) \in [0,1]^d$ and $X_j = (X_{j,1}, \ldots, X_{j,d})$, define

$$r(x) = \sup_{j \in J(x;N_n)} \sup_{1 \leq l \leq d} |x_l - X_{j,l}|.$$

Let $\omega(h; f_i)$ denote the modulus of continuity of $f_i$. For $\epsilon > 0$, let $\eta_\epsilon = \sup\{t : \omega(t; f_i) \leq \epsilon\}$. From Lemma 3 in Section A.6, we have

$$\begin{aligned} &P\big(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon\big) \\ &\quad \leq P\Big(\sup_x r(x) \geq \eta_{\epsilon/4}\Big) \\ &\qquad + (n^{d+2}+1)\left(\exp\Big(-\frac{3N_n\pi_n}{28}\Big) + \exp\Big(-\frac{N_n\pi_n^2\epsilon^2}{16(v^2 + c\pi_n\epsilon/4)}\Big)\right). \end{aligned}$$

For an analysis of the first term in the preceding upper bound, consider partitioning the unit cube into smaller cubes of side length $h$ as in the histogram estimation. There are $M = (1/h)^d$ many such cubes. Fix a small cube. If there are at least $N_n$ points in the cube, then $r_i(x) \le h$ for $x$ in the small cube. Thus, the set $\{(X_1, \ldots, X_n) : \sup_x r(x) \ge \eta_{\epsilon/4}\}$ is contained in the set where there exists at least one small cube of side length $h < \eta_{\epsilon/4}$ with the number of observations corresponding to the $i$th arm less than $N_n$. It follows that

$$P\Big(\sup_x r(x) \ge \eta_{\epsilon/4}\Big) \le Mp,$$

where $p$ is the probability that there are less than $N_n$ observations in a given small cube. For $h$ such that $\underline{c}nh^d/2 \ge N_n$, by Section A.5, we have

$$p \le \exp\Big(-\frac{3\underline{c}nh^d}{28}\Big) \le \exp\Big(-\frac{3N_n}{14}\Big).$$

From all the above, with $h < \eta_{\epsilon/4}$ and $N_n \le \underline{c}nh^d/2$,

$$\begin{aligned} &P\big(\|\hat{f}_{i,n} - f_i\|_\infty \ge \epsilon\big) \\ &\quad \le M \exp\Big(-\frac{3N_n}{14}\Big) \\ &\qquad + (n^{d+2}+1)\Big(\exp\Big(-\frac{3N_n\pi_n}{28}\Big) + \exp\Big(-\frac{N_n\pi_n^2\epsilon^2}{16(v^2 + c\pi_n\epsilon/4)}\Big)\Big). \end{aligned} \tag{4}$$

Under the continuity assumption, we have $\omega(h; f_i) \to 0$ as $h \to 0$. Thus, for any $\epsilon > 0$, $\eta_\epsilon > 0$. Then, if we take $h_n \to 0$, eventually we have $h_n \le \eta_{\epsilon/4}$. For $N_n$ such that $N_n\pi_n^2/\log n \to \infty$ and $N_n = o(n)$, we can choose $h_n \to 0$ satisfying $h_n \ge (2N_n/(\underline{c}n))^{1/d}$ as needed for (4). As a consequence, for each $\epsilon > 0$, the upper bound in (4) is summable in $n$ and thus

$$\sum_{n=1}^{\infty} P\big(\|\hat{f}_{i,n} - f_i\|_\infty \ge \epsilon\big) < \infty.$$

By the Borel–Cantelli lemma, since $\epsilon$ is arbitrary, $\|\hat{f}_{i,n} - f_i\|_\infty \to 0$ with probability 1. This completes the proof of Theorem 3. □

**8. Conclusions.** Multi-armed bandit problems have applications in various practical settings, including clinical trials, scheduling and automated problem solving in machine learning. In most of these situations, some covariates or concomitant variables are available, that, when utilized appropriately, can be helpful in selecting a good arm and thus obtaining a high reward. However, in the vast majority of previous work, such auxiliary information was not considered.

In this work, we model the relationship between the reward generated by an arm and the covariate in a nonparametric regression framework. With the covariate

observed, the best arm (with the highest mean reward) depends on its value and, of course, is unknown. An allocation rule is said to be strongly consistent if the total reward it receives up to time $n$ is almost surely asymptotically equivalent to that obtained by always pulling the best arm. Nonparametric regression techniques can be used to estimate the functional relationship (the mean reward function) between the reward and the covariate for each arm based on the past observations. At the next observed covariate value, based on the estimated mean reward functions of the arms, one arm is projected to give the highest reward. Due to uncertainty in the estimators of the mean reward functions, one faces the challenge of two conflicting tendencies: on one hand, one tends to pull the projected best arm for high reward based on the currently available information; on the other hand, one wants to try other arms to accumulate more information so that a comparison based on the estimated mean reward functions becomes more reliable. We propose a randomization method to automatically balance the two competing tendencies. The allocation rule is shown to be strongly consistent when the regression procedure used to estimate the mean reward functions satisfies certain conditions, which are shown to hold for familiar histogram and nearest neighbor methods.

Our approach and results on bandit problems with covariates differ significantly from earlier work reported in the literature. Unlike earlier results, no discounting of later rewards is considered in our work. In contrast to earlier parametric approaches, we use a more flexible nonparametric framework to model relationships between the rewards and covariates. In addition, the allocation rule in our work is randomized instead of deterministic.

There are several directions for future work. The property of strong consistency does not address the issue of how quickly the total reward based on the allocation rule approaches the ideal one. It is important to study the optimal rate of convergence (e.g., in terms of the smoothness of the mean reward functions) and find allocation rules that achieve the optimal rate. A more challenging task is to construct an adaptive allocation rule that automatically achieves the optimal rate of convergence without the knowledge of, for example, smoothness of the mean reward functions. It is also of interest to incorporate discounting (e.g., geometric discounting) in the definition of consistency in our framework and investigate the corresponding properties. Another interesting and important issue is design of covariates. In this work, covariates are assumed to be independent and identically distributed. In some practical settings such as in clinical trials, one needs to choose a sampling scheme of the covariates as well. For better performance, nonuniform and adaptive schemes should be considered. For instance, in a region of covariates where the mean reward functions barely change, one should sample less frequently. A difficulty here is that a good sampling scheme requires some knowledge of the unknown mean reward functions. Work in these directions will find more realistic applications.

## APPENDIX

**A.1. A slightly different measure of performance.** The definition of consistency given in Section 2 involves the mean reward averaged over random errors. It is essentially the same as the following defined in terms of the observed rewards.

DEFINITION. An allocation rule is said to be strongly consistent if

$$\frac{\sum_{j=1}^{n} Y_{I_j,j}}{\sum_{j=1}^{n} Y_{i^*(X_j),j}} \to 1 \quad \text{with probability 1.}$$

Note that

$$\frac{\sum_{j=1}^{n} Y_{I_j,j}}{\sum_{j=1}^{n} Y_{i^*(X_j),j}} = \frac{\frac{1}{n}\sum_{j=1}^{n} f_{I_j}(X_j) + \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{I_j,j}}{\frac{1}{n}\sum_{j=1}^{n} f_{i^*(X_j)}(X_j) + \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{i^*(X_j),j}}.$$

By the strong law of large numbers, $\frac{1}{n}\sum_{j=1}^{n} \varepsilon_{I_j,j} \to 0$ and $\frac{1}{n}\sum_{j=1}^{n} \varepsilon_{i^*(X_j),j} \to 0$ almost surely. Thus, the two measures are essentially the same.

One may consider a more ambitious goal, to asymptotically achieve the performance obtainable only when one knows the realization of rewards in advance. That is, one wants to have an allocation rule such that

$$\frac{\sum_{j=1}^{n} Y_{I_j,j}}{\sum_{j=1}^{n} \underline{Y}_j} \to 1 \quad \text{a.s.,}$$

where $\underline{Y}_j = \max\{Y_{i,j} : 1 \le i \le \mathcal{A}\}$. It is not hard to show that this is impossible to achieve in general.

**A.2. A probability bound on the performance of the histogram method.** Consider the regression model

$$Y_j = f(x_j) + \varepsilon_j, \qquad 1 \le j \le n,$$

where $x_1, \ldots, x_n \in [0,1]^d$ are given design points and the $\varepsilon_j$'s are independent errors satisfying the moment condition in Assumption 3 in Section 6. Let $W_1, \ldots, W_n$ be Bernoulli random variables with success probability lower bounded by $\pi_j$, $1 \le j \le n$, that decide if $Y_j$ is observed ($W_j = 1$) or not ($W_j = 0$). Assume, for each $1 \le j \le n$, $W_j$ is independent of $\{\varepsilon_k : k \ge j\}$. Let $\hat{f}_n$ be the histogram estimator of $f$ as defined in Section 4. Let $\omega(h; f)$ denote a modulus of continuity defined by

$$\omega(h; f) = \sup\{|f(x_1) - f(x_2)| : |x_{1i} - x_{2i}| \le h \text{ for all } 1 \le i \le d\}. \tag{5}$$

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

$$P_{x^n}\big(\|\hat{f}_n - f\|_\infty \geq \epsilon\big) \leq M \exp\left(-\frac{3\pi_n \min_{1\leq l\leq M} N_l}{28}\right) + 2M \exp\left(-\frac{\min_{1\leq l\leq M} N_l \pi_n^2 \big(\epsilon - \omega(h; f)\big)^2}{8\big(v^2 + c(\pi_n/2)\big(\epsilon - \omega(h; f)\big)\big)}\right).$$

*Here the probability (denoted by $P_{x^n}$) is conditioned on the design points.*

PROOF. Note that the preceding inequality trivially holds if $\min_{1\leq l\leq M} N_l = 0$. Thus, we assume that $\min_{1\leq l\leq M} N_l > 0$. Let $N(x)$ denote the number of $x_i$'s that fall in the same cube as $x$ and let $J(x)$ denote the set of indices $1 \leq j \leq n$ of such design points. Let $\overline{J}(x)$ denote the subset of $J(x)$, where $W_j$ takes value 1, and let $\overline{N}(x)$ denote the size of the set. Note

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{\overline{N}(x)} \sum_{j\in\overline{J}(x)} Y_j \\ &= f(x) + \frac{1}{\overline{N}(x)} \sum_{j\in\overline{J}(x)} \big(f(x_j) - f(x)\big) + \frac{1}{\overline{N}(x)} \sum_{j\in\overline{J}(x)} \varepsilon_j.\end{aligned}$$

It follows that

$$|\hat{f}_n(x) - f(x)| \leq \omega(h; f) + \left|\frac{1}{\overline{N}(x)} \sum_{j\in\overline{J}(x)} \varepsilon_j\right|.$$

Consequently, for any $\epsilon > \omega(h; f)$, with the given design points,

$$P\big(\|\hat{f}_n - f\|_\infty \geq \epsilon\big) \leq P\left(\sup_x \left|\frac{1}{\overline{N}(x)} \sum_{j\in\overline{J}(x)} \varepsilon_j\right| \geq \epsilon - \omega(h; f)\right).$$

Note that $N(x)$, $\overline{N}(x)$, $J(x)$ and $\overline{J}(x)$ are the same for $x$ in the same small cube $C$, respectively. Let $x_0$ be a fixed point in $C$. Then

$$\begin{aligned}&P\left(\sup_{x\in C} \left|\frac{1}{\overline{N}(x)} \sum_{j\in\overline{J}(x)} \varepsilon_j\right| \geq \epsilon - \omega(h; f)\right) \\ &\quad = P\left(\left|\sum_{j\in\overline{J}(x_0)} \varepsilon_j\right| \geq \overline{N}(x_0)\big(\epsilon - \omega(h; f)\big)\right) \\ &\quad = P\left(\left|\sum_{j\in J(x_0)} W_j \varepsilon_j\right| \geq N(x_0)\frac{\overline{N}(x_0)}{N(x_0)}\big(\epsilon - \omega(h; f)\big)\right)\end{aligned}$$

$$
\begin{aligned}
&= P\left(\left|\sum_{j\in J(x_0)} W_j\varepsilon_j\right| \ge N(x_0)\frac{\overline{N}(x_0)}{N(x_0)}\bigl(\epsilon-\omega(h;f)\bigr), \frac{\overline{N}(x_0)}{N(x_0)} \le \frac{\pi_n}{2}\right)\\
&\quad + P\left(\left|\sum_{j\in J(x_0)} W_j\varepsilon_j\right| \ge N(x_0)\frac{\overline{N}(x_0)}{N(x_0)}\bigl(\epsilon-\omega(h;f)\bigr), \frac{\overline{N}(x_0)}{N(x_0)} > \frac{\pi_n}{2}\right)\\
&\le P\left(\frac{\overline{N}(x_0)}{N(x_0)} \le \frac{\pi_n}{2}\right) + P\left(\left|\frac{1}{N(x_0)}\sum_{j\in J(x_0)} W_j\varepsilon_j\right| \ge \frac{\pi_n}{2}\bigl(\epsilon-\omega(h;f)\bigr)\right)\\
&\le \exp\left(-\frac{3N(x_0)\pi_n}{28}\right) + 2\exp\left(-\frac{N(x_0)(\pi_n/2)^2\bigl(\epsilon-\omega(h;f)\bigr)^2}{2\bigl(v^2+c(\pi_n/2)\bigl(\epsilon-\omega(h;f)\bigr)\bigr)}\right),
\end{aligned}
$$

where the last inequality follows from inequality (8) in Section A.4 and Lemma 2 in Section A.3. Therefore, we have

$$
\begin{aligned}
P_{x^n}\bigl(\|\hat{f}_n - f\|_\infty \ge \epsilon\bigr) &\le M\exp\left(-\frac{3\pi_n \min_{1\le l\le M} N_l}{28}\right)\\
&\quad + 2M\exp\left(-\frac{\min_{1\le l\le M} N_l(\pi_n/2)^2\bigl(\epsilon-\omega(h;f)\bigr)^2}{2\bigl(v^2+c(\pi_n/2)\bigl(\epsilon-\omega(h;f)\bigr)\bigr)}\right).
\end{aligned}
$$

The conclusion follows. This completes the proof of Lemma 1. □

**A.3. A probability inequality for sums of certain random variables.** Let $\varepsilon_1, \varepsilon_2, \ldots$ be independent random variables satisfying the refined Bernstein condition (1) in Assumption 3. Let $I_1, I_2, \ldots$ be Bernoulli random variables such that $I_j$ is independent of $\{\varepsilon_l : l \ge j\}$ for all $j \ge 1$.

LEMMA 2. *For any* $\epsilon > 0$,

$$
P\left(\sum_{j=1}^{n} I_j\varepsilon_j \ge n\epsilon\right) \le \exp\left(-\frac{n\epsilon^2/2}{v^2+c\epsilon}\right).
$$

*Particularly, by taking* $I_1 = I_2 = \cdots = I_n$ *with probability* 1, *we have*

$$
P\left(\sum_{j=1}^{n} \varepsilon_j \ge n\epsilon\right) \le \exp\left(-\frac{n\epsilon^2/2}{v^2+c\epsilon}\right). \tag{6}
$$

REMARK. The second inequality in (6) above is called the (refined) Bernstein inequality [see, e.g., van der Vaart and Wellner (1996), Lemma 2.2.11, and Birgé and Massart (1998), Lemma 8].

PROOF OF LEMMA 2. Following a standard argument, we have, for any $t > 0$,

$$\begin{aligned}
&P\left(\sum_{j=1}^{n} I_j\varepsilon_j \geq n\epsilon\right) \\
&\quad \leq e^{-nt\epsilon} E\exp\left(t\sum_{j=1}^{n} I_j\varepsilon_j\right) \\
&\quad = e^{-nt\epsilon} E\left(E\exp\left(t\sum_{j=1}^{n} I_j\varepsilon_j\right)\Big|\, \varepsilon_1,\ldots,\varepsilon_{n-1}, I_1,\ldots,I_{n-1}\right) \\
&\quad = e^{-nt\epsilon} E\left[\exp\left(t\sum_{j=1}^{n-1} I_j\varepsilon_j\right) E\big(e^{tI_n\varepsilon_n}|\varepsilon_1,\ldots,\varepsilon_{n-1}, I_1,\ldots,I_{n-1}\big)\right].
\end{aligned}$$

Since $\varepsilon_n$ is independent of $\varepsilon_1,\ldots,\varepsilon_{n-1}$, $I_1,\ldots,I_{n-1}$ and $I_n$, we have

$$\begin{aligned}
&E\big(e^{tI_n\varepsilon_n}|\varepsilon_1,\ldots,\varepsilon_{n-1}, I_1,\ldots,I_{n-1}\big) \\
&\quad = E\big(e^{t\varepsilon_n}\big) P\big(I_n = 1|\varepsilon_1,\ldots,\varepsilon_{n-1}, I_1,\ldots,I_{n-1}\big) \\
&\qquad + 1 - P\big(I_n = 1|\varepsilon_1,\ldots,\varepsilon_{n-1}, I_1,\ldots,I_{n-1}\big).
\end{aligned}$$

Under the Bernstein condition on the errors,

$$E\big(e^{t\varepsilon_n}\big) \leq \exp\left(\frac{v^2t^2}{2(1-tc)}\right)$$

for $t < 1/c$. Since $\exp(v^2t^2/(2(1-tc))) > 1$ when $t < 1/c$,

$$E\big(e^{tI_n\varepsilon_n}|\varepsilon_1,\ldots,\varepsilon_{n-1}, I_1,\ldots,I_{n-1}\big) \leq \exp\left(\frac{v^2t^2}{2(1-tc)}\right).$$

By induction, we have, for $t < 1/c$,

$$E\exp\left(t\sum_{j=1}^{n} I_j\varepsilon_j\right) \leq \exp\left(\frac{nv^2t^2}{2(1-tc)}\right)$$

and, consequently,

$$P\left(\sum_{j=1}^{n} I_j\varepsilon_j \geq n\epsilon\right) \leq \exp\left(-nt\epsilon + \frac{nv^2t^2}{2(1-tc)}\right).$$

Minimizing the exponent of the upper bound over $t$ [as in Birgé and Massart (1998), Lemma 8] gives the claimed inequality. □

**A.4. An inequality for Bernoulli trials.** For $1 \le j \le n$, let $W_j$ be independent Bernoulli random variables with success probability $\beta_i$. Applying Bernstein's inequality [see, e.g., Pollard (1984), page 193], we have

$$P\left(\sum_{j=1}^{n} W_j \le \left(\sum_{j=1}^{n} \beta_j\right)\Big/2\right) \le \exp\left(-\frac{3\sum_{j=1}^{n}\beta_j}{28}\right). \tag{7}$$

In a somewhat more complicated setting, for $1 \le j \le n$, let $\widetilde{W}_j$ be Bernoulli random variables, which are not necessarily independent. Assume that the conditional probability of success for $\widetilde{W}_j$ given the previous observations is lower bounded by $\beta_j$, that is,

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

for all $1 \le j \le n$. Then it can be easily shown that $\sum_{j=1}^{n} \widetilde{W}_j$ is stochastically no smaller than $\sum_{j=1}^{n} W_j$ with the $W_j$'s defined earlier in this section. Therefore, it follows that

$$P\left(\sum_{j=1}^{n} \widetilde{W}_j \le \left(\sum_{j=1}^{n} \beta_j\right)\Big/2\right) \le \exp\left(-\frac{3\sum_{j=1}^{n}\beta_j}{28}\right). \tag{8}$$

**A.5. Number of observations in a small cube for histogram estimation.** Let $X_1, \ldots, X_n$ be i.i.d. random variables in $[0,1]^d$. Assume that the design distribution $P_X$ has a density $p(x)$ with respect to Lebesgue measure and $p(x)$ satisfies $\underline{c} \le p(x) \le \overline{c}$ for some positive constants $\underline{c} \le \overline{c}$. Let $N$ be the number of observations falling in a fixed cube of side width $h$. It is easily seen that $N$ has a binomial distribution with success probability $\beta \ge \underline{c}h^d$. From the inequality (7), we have

$$P\left(N \le \frac{\underline{c}nh^d}{2}\right) \le \exp\left(-\frac{3\underline{c}nh^d}{28}\right).$$

**A.6. A probability bound on the performance of the nearest neighbor method.** Consider the nearest neighbor estimators of the functions $f_i$ as defined in Section 7. Now fix $i$ in $\{1, \ldots, \mathcal{l}\}$. Let $W_1, \ldots, W_n$ be the Bernoulli random variables that decide if the $i$th arm is pulled for $X_j$ ($W_j = 1$) or not ($W_j = 0$) for $1 \le j \le n$. From the description of the allocation scheme in Section 3, it is clear that, for each $1 \le j \le n$, $W_j$ is independent of $\{\varepsilon_{i,k} : k \ge j\}$. Note also that, conditional on the previous observations and $X_j$, the probability of $W_j = 1$ is almost surely lower bounded by $\pi_j \ge \pi_n$ for $1 \le j \le n$.

Let $J(x) = J(x; N)$ and $J_i(x) = J_i(x; N)$ be defined as in Section 7.

For $x = (x_1, \ldots, x_d) \in [0,1]^d$ and $X_j = (X_{j,1}, \ldots, X_{j,d})$, define $r(x) = \sup_{j \in J(x)} \sup_{1 \le l \le d} |x_l - X_{j,l}|$. Let $\omega(h; f_i)$ denote the modulus of continuity of $f_i$. For $\epsilon > 0$, let $\eta_\varepsilon = \sup\{t : \omega(t; f_i) \le \epsilon\}$.

LEMMA 3. *Suppose that Assumptions* 1–3 *are satisfied. Then the $N_n$ nearest neighbor estimator $\hat{f}_{i,n}$ satisfies*

$$P\big(\|\hat{f}_{i,n} - f_i\|_\infty \geq \epsilon\big) \leq P\Big(\sup_x r(x) \geq \eta_{\epsilon/4}\Big) + (n^{d+2}+1)\left(\exp\left(-\frac{3N_n\pi_n}{28}\right) + \exp\left(-\frac{N_n\pi_n^2\epsilon^2}{16(v^2 + c\pi_n\epsilon/4)}\right)\right).$$

PROOF. Note that

$$\hat{f}_{i,n}(x) = \frac{1}{N_i(x)} \sum_{j \in J_i(x)} Y_{i,j} = f(x) + \frac{1}{N_i(x)} \sum_{j \in J_i(x)} \big(f_i(X_j) - f_i(x)\big) + \frac{1}{N_i(x)} \sum_{j \in J_i(x)} \varepsilon_{i,j}.$$

It follows that

$$|\hat{f}_{i,n}(x) - f_i(x)| \leq \omega\big(r(x);\, f_i\big) + \left|\frac{1}{N_i(x)} \sum_{j \in J_i(x)} \varepsilon_{i,j}\right|.$$

Consequently, for any $\epsilon > 0$,

$$\begin{aligned} P\big(\|\hat{f}_n - f\|_\infty \geq \epsilon\big) &\leq P\left(\sup_x \omega\big(r(x);\, f_i\big) \geq \frac{\epsilon}{2}\right) + P\left(\sup_x \left|\frac{1}{N_i(x)} \sum_{j \in J_i(x)} \varepsilon_{i,j}\right| \geq \frac{\epsilon}{2}\right) \\ &\leq P\Big(\sup_x r(x) \geq \eta_{\epsilon/4}\Big) + P\left(\sup_x \left|\frac{1}{N_i(x)} \sum_{j \in J_i(x)} W_j \varepsilon_{i,j}\right| \geq \frac{\epsilon}{2}\right), \end{aligned}$$

where, for the second inequality, we used the fact that $\omega(\cdot;\, f_i)$ is nondecreasing [since if $r(x) < \eta_{\epsilon/4}$ then $\omega(r(x);\, f_i) \leq \varepsilon/4$]. We handle below the second term of the aforementioned second inequality.

Now condition on the design points. Note that, for different $x$, $J(x)$ may be the same. Let $L$ be the total number of choices that $J(x)$ can take for $x \in [0,1]^d$ and let $t_1, \ldots, t_L$ be any chosen representatives for these distinct values. Observing that $L$ depends only on the design points, we have that, conditional on $X_1 = x_1, \ldots, X_n = x_n$,

$$P_{x^n}\left(\sup_x \left|\frac{1}{N_i(x)} \sum_{j \in J(x)} W_j \varepsilon_{i,j}\right| \geq \frac{\epsilon}{2}\right) \leq \sum_{l=1}^{L} P_{x^n}\left(\left|\sum_{j \in J(t_l)} W_j \varepsilon_{i,j}\right| \geq N_i(t_l)\frac{\epsilon}{2}\right).$$

Applying Lemma 2 in Section A.3, we have that, for any $\delta > 0$,

$$P_{x^n}\left(\left|\sum_{j\in J(t_l)} W_j\varepsilon_{i,j}\right| \geq N_n\delta\right) \leq \exp\left(-\frac{N_n\delta^2/2}{v^2 + c\delta}\right).$$

By (8), we have

$$P_{x^n}\left(N_i(t_l) \leq \frac{N_n\pi_n}{2}\right) \leq \exp\left(-\frac{3N_n\pi_n}{28}\right).$$

It follows then

$$\begin{aligned}
&P_{x^n}\left(\left|\sum_{j\in J(t_l)} W_j\varepsilon_{i,j}\right| \geq \frac{N_i(t_l)\epsilon}{2}\right)\\
&\quad= P_{x^n}\left(\left|\sum_{j\in J(t_l)} W_j\varepsilon_{i,j}\right| \geq \frac{N_i(t_l)\epsilon}{2}, N_i(t_l) \leq \frac{N_n\pi_n}{2}\right)\\
&\qquad+ P_{x^n}\left(\left|\sum_{j\in J(t_l)} W_j\varepsilon_{i,j}\right| \geq \frac{N_i(t_l)\epsilon}{2}, N_i(t_l) > \frac{N_n\pi_n}{2}\right)\\
&\quad\leq P_{x^n}\left(N_i(t_l) \leq \frac{N_n\pi_n}{2}\right)\\
&\qquad+ P_{x^n}\left(\left|\sum_{j\in J(t_l)} W_j\varepsilon_{i,j}\right| \geq \frac{N_n\pi_n\epsilon}{4}, N_i(t_l) > \frac{N_n\pi_n}{2}\right)\\
&\quad\leq P_{x^n}\left(N_i(t_l) \leq \frac{N_n\pi_n}{2}\right) + P_{x^n}\left(\left|\sum_{j\in J(t_l)} W_j\varepsilon_{i,j}\right| \geq \frac{N_n\pi_n\epsilon}{4}\right)\\
&\quad\leq \exp\left(-\frac{3N_n\pi_n}{28}\right) + \exp\left(-\frac{N_n\pi_n^2\epsilon^2}{16(v^2 + c\pi_n\epsilon/4)}\right).
\end{aligned}$$

Now we upper-bound the constant $L$. Let $D(x)$ be the collection of $x_j$, $1 \leq j \leq n$, with $j \in J(x)$. Note that, with probability 1 (since the design distribution has a Lebesgue density), $D(x)$ is of the form $D(x) = \{x_j : \|x_j - x\| \leq r_x\}$ for some $r_x > 0$, where $\|\cdot\|$ denotes the Euclidean norm on $R^d$. Thus, $L$ is upper bounded (with probability 1) by the size of the collection of $x_j$'s that are in any balls. This is then bounded above by a quantity involving the VC dimension of the set of all balls. Let $v_B$ be the VC dimension of the set of balls in $R^d$. From Devroye, Györfi and Lugosi [(1996), Corollary 13.2], $v_B \leq d + 2$. It follows from the VC lemma [see, e.g., Devroye, Györfi and Lugosi (1996), Theorem 13.2] that

$$L \leq n^{d+2} + 1.$$

From all the above,

$$P_{x^n}\left(\sup_x\left|\frac{1}{N_i(x)}\sum_{j\in J(x)}W_j\varepsilon_{i,j}\right|\geq\frac{\epsilon}{2}\right)$$
$$\leq (n^{d+2}+1)\left(\exp\left(-\frac{3N_n\pi_n}{28}\right)+\exp\left(-\frac{N_n\pi_n^2\epsilon^2}{16(v^2+c\pi_n\epsilon/4)}\right)\right).$$

Since the upper bound does not depend on $x^n$, it also upper-bounds the unconditional probability

$$P\left(\sup_x\left|\frac{1}{N_i(x)}\sum_{j\in J(x)}W_j\varepsilon_{i,j}\right|\geq\frac{\epsilon}{2}\right).$$

The conclusion of Lemma 3 follows. This completes the proof of Lemma 3. □

**Acknowledgments.** The authors thank two anonymous reviewers and the Associate Editor for their helpful comments for improving the presentation of the paper.

## REFERENCES

AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (1995). Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *36th Annual Symposium on Foundations of Computer Science* 322–331. IEEE Computer Society Press, Los Alamitos, CA.

BERRY, D. A., CHEN, R. W., ZAME, A., HEATH, D. C. and SHEPP, L. A. (1997). Bandit problems with infinitely many arms. *Ann. Statist.* **25** 2103–2116.

BERRY, D. A. and FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, New York.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

CLAYTON, M. K. (1989). Covariate models for Bernoulli bandits. *Sequential Anal.* **8** 405–426.

DEVROYE, L. and GYÖRFI, L. (1985). Distribution-free exponential bounds on the $l_1$ error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics* (F. Konecny, J. Mogyoródi and W. Wertz, eds.) 67–76. Akadémiai Kiadó, Budapest.

DEVROYE, L., GYÖRFI, L., KRZYŻAK, A. and LUGOSI, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.* **22** 1371–1385.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, New York.

GITTINS, J. C. (1989). *Multi-armed Bandit Allocation Indices*. Wiley, New York.

GRATCH, J., DEJONG, G. and YANG, Y. (1994). Rational learning: finding a balance between utility and efficiency. *Selecting Models from Data: Artificial Intelligence and Statistics. Lecture Notes in Statist.* **89** 11–20. Springer, New York.

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22.

LAI, T. L. and YAKOWITZ, S. (1995). Machine learning and nonparametric bandit theory. *IEEE Trans. Automat. Control* **40** 1199–1209.

NOBEL, A. (1996). Histogram regression estimation using data-dependent partitions. *Ann. Statist.* **24** 1084–1105.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535.

SARKAR, J. (1991). One-armed bandit problems with covariates. *Ann. Statist.* **19** 1978–2002.

STONE, C. S. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

WOODROOFE, M. (1979). A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74** 799–806.

DEPARTMENT OF STATISTICS AND
DEPARTMENT OF LOGISTICS,
OPERATIONS AND MIS
IOWA STATE UNIVERSITY
AMES, IOWA
E-MAIL: yyang@iastate.edu