Journal of Machine Learning Research 17 (2016) 1-37 Submitted 5/13; Revised 9/15; Published 10/16

Kernel Estimation and Model Combination in A Bandit Problem with Covariates

Wei Qian WXQSMA@RIT.EDU
School of Mathematical Sciences
Rochester Institute of Technology
Rochester, NY 14623, USA

Yuhong Yang YYANG@STAT.UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Editor: Gábor Lugosi

Abstract

Multi-armed bandit problem is an important optimization game that requires an exploration-exploitation tradeoff to achieve optimal total reward. Motivated from industrial applications such as online advertising and clinical research, we consider a setting where the rewards of bandit machines are associated with covariates, and the accurate estimation of the corresponding mean reward functions plays an important role in the performance of allocation rules. Under a flexible problem setup, we establish asymptotic strong consistency and perform a finite-time regret analysis for a sequential randomized allocation strategy based on kernel estimation. In addition, since many nonparametric and parametric methods in supervised learning may be applied to estimating the mean reward functions but guidance on how to choose among them is generally unavailable, we propose a model combining allocation strategy for adaptive performance. Simulations and a real data evaluation are conducted to illustrate the performance of the proposed allocation strategy.

Keywords: contextual bandit problem, exploration-exploitation tradeoff, nonparametric regression, regret bound, upper confidence bound

1. Introduction

Following the seminal work by Robbins (1954), multi-armed bandit problems have been studied in multiple fields. The general bandit problem involves the following optimization game: A gambler is given l gambling machines, and each machine has an "arm" that the gambler can pull to receive the reward. The distribution of reward for each arm is unknown and the goal is to maximize the total reward over a given time horizon. If we define the regret to be the reward difference between the optimal arm and the pulled arm, the equivalent goal of the bandit problem is to minimize the total regret. Under a standard setting, it is assumed that the reward of each arm has fixed mean and variance throughout the time horizon of the game. Some of the representative work for standard bandit problem includes Lai and Robbins (1985), Berry and Fristedt (1985), Gittins (1989) and Auer et al. (2002).

See Cesa-Bianchi and Lugosi (2006) and Bubeck and Cesa-Bianchi (2012) for bibliographic remarks and recent overviews on bandit problems.

Different variants of the bandit problem motivated by real applications have been studied extensively in the past decade. One promising setting is to assume that the reward distribution of each bandit arm is associated with some common external covariate. More specifically, for an l-armed bandit problem, the game player is given a d-dimensional external covariate $x \in R^d$ at each round of the game, and the expected reward of each bandit arm given x has a functional form $f_i(x)$, $i = 1 \cdots, l$. We call this variant **m**ulti-**a**rmed **b**andit problem with **c**ovariates, or MABC for its abbreviation (MABC is also referred to as CMAB for **c**ontextual **m**ulti-**a**rmed **b**andit problem in the literature). The consideration of external covariates is potentially important in applications such as personalized medicine. For example, before deciding which treatment arm to be assigned to a patient, we can observe the patient prognostic factors such as age, blood pressure or genetic information, and then use such information for adaptive treatment assignment for best outcome. It is worth noting that the consideration of external covariate is recently further generalized to partial monitoring by Bartók and Szepesvári (2012).

The MABC problems have been studied under both parametric and nonparametric frameworks with various types of algorithms. The first work in a parametric framework appears in Woodroofe (1979) under a somewhat restrictive setting. A linear response bandit problem in more flexible settings is recently studied under a minimax framework (Goldenshluger and Zeevi, 2009; Goldenshluger and Zeevi, 2013). Empirical studies are also reported for parametric UCB-type algorithms (e.g., Li et al., 2010). The regret analysis of a special linear setting is given in e.g., Auer (2002), Chu et al. (2011) and Agrawal and Goyal (2013), in which the linear parameters are assumed to be the same for all arms while the observed covariates can be different across different arms.

MABC problems with the nonparametric framework are first studied by Yang and Zhu (2002). They show that with histogram or K-nearest neighbor estimation, the function estimation is uniformly strongly consistent, and consequently, the cumulative reward of their randomized allocation rule is asymptotically equivalent to the optimal cumulative reward. Their notion of reward strong consistency has been recently established for a Bayesian sampling method (May et al., 2012). Notably, under the Hölder smoothness condition and a margin condition, the recent work of Perchet and Rigollet (2013) establishes a regret upper bound by arm elimination algorithms with the same order as the minimax lower bound of a two-armed MABC problem (Rigollet and Zeevi, 2010). A different stream of work represented by, e.g., Langford and Zhang (2007) and Dudik et al. (2011) imposes neither linear nor any smoothness assumption on the mean reward function; instead, they consider a class of (finitely many) policies, and the cumulative reward of the proposed algorithms is compared to the best of the policies. Interested readers are also referred to Bubeck and Cesa-Bianchi (2012, Section 4) and its bibliography remarks for studies from numerous different perspectives.

Another important line of development in the bandit problem literature (closely related to, but different from the setting of MABC) is to consider the arm space as opposed to the covariate space in MABC. It is assumed that there are infinitely many arms, and at each round of the game, the player has the freedom to play one arm chosen from the arm space. Like MABC, the setting with the arm space can be studied from both parametric linear and nonparametric frameworks. Examples of the linear parametric framework include Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010) and Abbasi-Yadkori et al. (2011). Notable examples of the nonparametric framework (also known as the continuum-armed bandit problem) under the local or global Hölder and Lipchitz smoothness conditions are Kleinberg (2004), Auer et al. (2007), Kleinberg et al. (2007) and Bubeck et al. (2011). Abbasi-Yadkori (2009) studies a forced exploration algorithm over the arm space, which is applied to both parametric and nonparametric frameworks. Interestingly, Lu et al. (2010) and Slivkins (2011) consider both the arm space and the covariate space, and study the problem by imposing Lipschitz conditions on the joint space of arms and covariates.

Our work in this paper follows the nonparametric framework of MABC in Yang and Zhu (2002) and Rigollet and Zeevi (2010) with finitely many arms. One contribution in this work is to show that kernel methods enjoy estimation uniform strong consistency as well, which leads to strongly consistent allocation rules. Note that due to the dependence of the observations for each arm by the nature of the proposed randomized allocation strategy, it is difficult to apply the well-established kernel regression analysis results of i.i.d. or weak dependence settings (e.g., Devroye, 1978; Härdle and Luckhaus, 1984; Hansen, 2008). New technical tools and arguments such as "chaining" are developed in this paper.

In addition, with the help of the Hölder smoothness condition, we provide a deeper understanding of the proposed randomized allocation strategy via a finite-time regret analysis. Compared with the result in Rigollet and Zeevi (2010) and Perchet and Rigollet (2013), our finite-time result remains sub-optimal in the minimax sense. Indeed, given Hölder smoothness parameter κ and total time horizon N, our expected cumulative regret upper bound is $\tilde{O}(N^{1-\frac{1}{3+d/\kappa}})$ as compared to $O(N^{1-\frac{1}{2+d/\kappa}})$ of Perchet and Rigollet (2013) (without the extra margin condition). The slightly sub-optimal rate can also be shown to apply to the histogram based randomized allocation strategy proposed in Yang and Zhu (2002). We tend to think that this rate is the best possible for these methods, reflecting to some extent the theoretical limitation of the randomized allocation strategy. In spite of this sub-optimality, our result explicitly shows both the bias-variance tradeoff and the exploration-exploitation tradeoff, which reflects the underlying nature of the proposed algorithm for the MABC problem. With a model combining strategy and dimension reduction technique to be introduced later, the kernel estimation based randomized allocation strategy can be quite flexible with wide potential practical use. Moreover, in Appendix A, we incorporate the kernel estimation into a UCB-type algorithm with randomization and show that its regret rate becomes minimax optimal up to a logarithmic factor.

One natural and interesting issue in the randomized allocation strategy in MABC is how to choose the modeling methods among numerous nonparametric and parametric estimators. The motivation of such a question shares the flavor of model aggregation/combining in statistical learning (see, e.g., Audibert, 2009; Rigollet and Tsybakov, 2012; Wang et al., 2014 and references therein). In the bandit problem literature, model combining is also quite relevant to the adversary bandit problem (e.g., Cesa-Bianchi and Lugosi, 2006; Auer et al., 2003). As a recent example, Maillard and Munos (2011) study the history-dependent adversarial bandit to target the best among a pool of history class mapping strategies.

As an empirical solution to the difficulty in choosing the best estimation method for each arm in the randomized allocation strategy for MABC, we introduce a fully data-driven model combining technique motivated by the AFTER algorithm, which has shown success both theoretically (Yang, 2004) and empirically (e.g., Zou and Yang, 2004; Wei and Yang, 2012). We integrate a model combining step by AFTER for reward function estimation into the randomized allocation strategy for MABC. Preliminary simulation results of combining various dimension reduction methods are reported in Qian and Yang (2012). However, no theoretical justification is given there. As another contribution of this paper, we present here new theoretical and numerical results on the proposed combining algorithm. In particular, the strong consistency of the model combining allocation strategy is established.

The rest of this paper is organized as follows. We present a general and flexible problem setup for MABC in Section 2. We describe the algorithm in Section 3 and study the strong consistency and the finite-time regret analysis of kernel estimation methods in Section 4. We also introduce a dimension reduction sub-procedure to handle the situation that the covariate dimension is high. The asymptotic results of the model combining allocation strategy is established in Section 5. We show in Section 6 and Section 7 the numerical performance of the proposed allocation strategy using simulations and a web-based news article recommendation data set, respectively. A brief conclusion is given in Section 8. The kernel estimation based UCB-type algorithm with randomization is described in Appendix A, all technical lemmas and proofs are given in Appendix B, and additional numerical results of the implemented algorithms are left in Appendix C.

2. Problem Setup

Suppose a bandit problem has l ($l \geq 2$) candidate arms to play. At each time point of the game, a d-dimensional covariate x is observed before we decide which arm to pull. Assume that the covariate x takes values in the hypercube $[0,1]^d$. Also assume the (conditional) mean reward for arm i given x, denoted by $f_i(x)$, is uniformly upper bounded and unknown to game players. The observed reward is modeled as $f_i(x) + \varepsilon$, where ε is a random error with mean 0.

Let $\{X_n, \, n \geq 1\}$ be a sequence of independent covariates generated from an underlying probability distribution P_X supported in $[0,1]^d$. At each time $n \geq 1$, we need to apply a sequential allocation rule η to decide which arm to pull based on X_n and the previous observations. We denote the chosen arm by I_n and the observed reward of pulling the arm $I_n = i$ at time n by $Y_{i,n}$, $1 \leq i \leq l$. As a result, $Y_{I_n,n} = f_{I_n}(X_n) + \varepsilon_n$, where ε_n is the random error with $E(\varepsilon_n | X_n) = 0$. Different from Yang and Zhu (2002), the error ε_n may be dependent on the covariate X_n. Consider the simple scenario of online advertising where the response is binary (click: $Y = 1$; no click: $Y = 0$). Given an arm i and covariate $x \in [0,1]$, suppose the mean reward function satisfies e.g., $f_i(x) = x$. Then it is easy to see that the distribution of the random error ε depends on x. In case of a continuous response, it is also well-known that heteroscedastic errors commonly occur.

By the previous definitions, we know that at time n, an allocation strategy chooses the arm I_n based on X_n and $(X_j, \, I_j, \, Y_{I_j,j})$, $1 \leq j \leq n-1$. To evaluate the performance of the allocation strategy, let $i^*(x) = \mathrm{argmax}_{1 \leq i \leq l} f_i(x)$ and $f^*(x) = f_{i^*(x)}(x)$ (any tie-breaking rule can be applied if there are ties). Without the knowledge of random error ε_j, the optimal performance occurs when $I_j = i^*(X_j)$, and the corresponding optimal cumulative reward given $X_1, \cdots, X_n$ can be represented as $\sum_{j=1}^n f^*(X_j)$. The cumulative mean reward of the applied allocation rule can be represented as $\sum_{j=1}^n f_{I_n}(X_j)$. Thus we can measure the performance of an allocation rule η by the cumulative regret

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

We say the allocation rule η is strongly consistent if $R_n(\eta) = o(n)$ with probability one. Also, $R_n(\eta)$ is commonly used for finite-time regret analysis. In addition, define the per-round regret $r_n(\eta)$ by

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

To maintain the readability for the rest of this paper, we use i only for bandit arms, j and n only for time points, r and s only for reward function estimation methods, and t and T only for the total number of times a specific arm is pulled.

3. Algorithm

In this section, we present the model-combining-based randomized allocation strategy. At each time $n \geq 1$, denote the set of past observations $\{(X_j, I_j, Y_{I_j,j}) : 1 \leq j \leq n-1\}$ by Z^n, and denote the arm i associated subset $\{(X_j, I_j, Y_{I_j,j}) : I_j = i,\ 1 \leq j \leq n-1\}$ by $Z^{i,n}$. For estimating the f_i's, suppose we have m candidate regression estimation procedures (e.g., histogram, kernel estimation, etc.), and we denote the class of these candidate procedures by $\Delta = \{\delta_1, \cdots, \delta_m\}$. Let $\hat{f}_{i,n,r}$ denote the regression estimate of procedure δ_r based on $Z^{i,n}$, and let $\hat{f}_{i,n}$ denote the weighted average of $\hat{f}_{i,n,r}$'s, $1 \leq r \leq m$, by the model combining algorithm to be given. Let $\{\pi_n,\ n \geq 1\}$ be a decreasing sequence of positive numbers approaching 0, and assume that $(l-1)\pi_n < 1$ for all $n \geq 1$. The model combining allocation strategy includes the following steps.

STEP 1. Initialize with forced arm selections. Give each arm a small number of applications. For example, we may pull each arm n_0 times at the beginning by taking $I_1 = 1$, $I_2 = 2, \cdots I_l = l$, $I_{l+1} = 1, \cdots, I_{2l} = l, \cdots, I_{(n_0-1)l+1} = 1, \cdots, I_{n_0 l} = l$.

STEP 2. Initialize the weights and the error variance estimates. For $n = n_0 l + 1$, initialize the weights by

$$W_{i,n,r} = \frac{1}{m}, \quad 1 \leq i \leq l,\ 1 \leq r \leq m,$$

and initialize the error variance estimates by e.g.,

$$\hat{v}_{i,n,r} = 1,\ \hat{v}_{i,n} = 1, \quad 1 \leq i \leq l,\ 1 \leq r \leq m.$$

STEP 3. Estimate the individual functions f_i for $1 \leq i \leq l$. For $n = n_0 l + 1$, based on the current data $Z^{i,n}$, obtain $\hat{f}_{i,n,r}$ using regression procedure δ_r, $1 \leq r \leq m$.

STEP 4. Combine the regression estimates and obtain the weighted average estimates

$$\hat{f}_{i,n} = \sum_{r=1}^{m} W_{i,n,r}\hat{f}_{i,n,r}, \quad 1 \leq i \leq l.$$

STEP 5. Estimate the best arm, select and pull. For the covariate X_n, define $\hat{i}_n = \text{argmax}_{1 \leq i \leq l} \hat{f}_{i,n}(X_n)$ (If there is a tie, any tie-breaking rule may apply). Choose an arm, with probability $1-(l-1)\pi_n$ for arm $\hat{i}_n$ (the currently most promising choice) and with probability π_n for each of the remaining arms. That is,

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

Then pull the arm I_n to receive the reward $Y_{I_n,n}$.

STEP 6. Update the weights and the error variance estimates. For $1 \leq i \leq l$, if $i \neq I_n$, let $W_{i,n+1,r} = W_{i,n,r}$, $1 \leq r \leq m$, $\hat{v}_{i,n+1,r} = \hat{v}_{i,n,r}$, $1 \leq r \leq m$, and $\hat{v}_{i,n+1} = \hat{v}_{i,n}$. If $i = I_n$, update the weights and the error variance estimates by

$$W_{i,n+1,r} = \frac{\dfrac{W_{i,n,r}}{\hat{v}_{i,n,r}^{1/2}} \exp\left(-\dfrac{(\hat{f}_{i,n,r}(X_n) - Y_{i,n})^2}{2\hat{v}_{i,n}}\right)}{\displaystyle\sum_{k=1}^{m} \dfrac{W_{i,n,k}}{\hat{v}_{i,n,k}^{1/2}} \exp\left(-\dfrac{(\hat{f}_{i,n,k}(X_n) - Y_{i,n})^2}{2\hat{v}_{i,n}}\right)}, \quad 1 \leq r \leq m,$$

$$\hat{v}_{i,n+1,r} = \frac{\displaystyle\sum_{k=n_0 l+1}^{n} (Y_{I_k,k} - \hat{f}_{I_k,k,r}(X_k))^2 I(I_k = i)}{\displaystyle\sum_{k=n_0 l+1}^{n} I(I_k = i)} \vee \underline{v}, \quad 1 \leq r \leq m,$$

and

$$\hat{v}_{i,n+1} = \sum_{r=1}^{n} W_{i,n+1,r} \hat{v}_{i,n+1,r},$$

where $I(\cdot)$ is the indicator function and $\underline{v}$ is a small positive constant (to ensure that $\hat{v}_{i,n+1,r}$ is nonzero). In practice, we set $\underline{v} = 10^{-16}$.

STEP 7. Repeat steps 3 - 6 for $n = n_0 l + 2, n_0 l + 3, \cdots$, and so on.

In the allocation strategy above, step 1 and step 2 initialize the game and pull each arm the same number of times. Step 3 and step 4 estimate the reward function for each arm using several regression methods, and combine the estimates by a weighted average scheme. Clearly, the importance of these regression methods are differentiated by their corresponding weights. Step 5 performs an enforced randomization algorithm, which gives preference to the arm with the highest reward estimate. This type of arm randomization is also known as the ϵ-greedy algorithm. Step 6 is the key to the model combining algorithm, which updates the weights for the recently played arm. Its weight updating formula implies that if the estimated reward from a regression method turns out to be far away from the observed reward, we penalize this method by decreasing its weight, while if the estimated reward turns out to be accurate, we reward this method by increasing its weight. Note that our combining approach has few tuning parameters except for what is already included in the individual regression procedures.

4. Kernel Regression Procedures

In this section, we consider the special case that kernel estimation is used as the only modeling method. The primary goals include: 1) establishing the uniform strong consistency of kernel estimation under the proposed allocation strategy; 2) performing the finite-time regret analysis. To extend the applicability of kernel methods, a dimension reduction subprocedure is described in Section 4.3.

4.1 Strong Consistency

We focus on the Nadaraya-Watson regression and study its strong consistency under the proposed allocation strategy. Given a regression method $\delta_r \in \Delta$ and an arm i, we say it is strongly consistent in L_∞ norm for arm i if $\|\hat{f}_{i,n,r} - f_i\|_\infty \to 0$ a.s. as $n \to \infty$.

Assumption 0. *The errors satisfy a (conditional) moment condition that there exist positive constants v and c such that for all integers $k \geq 2$ and $n \geq 1$,*

$$E(|\varepsilon_n|^k|X_n) \leq \frac{k!}{2}v^2c^{k-2}$$

almost surely.

Assumption 0 means that the error distributions, which could depend on the covariates, satisfy a moment condition known as refined Bernstein condition (e.g., Birgé and Massart, 1998, Lemma 8). Normal distribution, for instance, satisfies the condition. Bounded errors trivially meet the requirement. Therefore, Assumption 0 is met in a wide range of real applications, and will also be used in the next section for understanding strong consistency of model combining procedures. Note that heavy-tailed distributions are also possible for bandit problems (Bubeck et al., 2013).

Given a bandit arm $1 \leq i \leq l$, at each time point n, define $J_{i,n} = \{j : I_j = i, 1 \leq j \leq n-1\}$, the set of past time points at which arm i is pulled. Let $M_{i,n}$ denote the size of the set $J_{i,n}$. For each $u = (u_1, u_2, \cdots, u_d) \in R^d$, define $\|u\|_\infty = \max\{|u_1|, |u_2|, \cdots, |u_d|\}$. Consider two natural conditions on the mean reward functions and the covariate density as follows.

Assumption 1. *The functions f_i are continuous on $[0,1]^d$ with $A =: \sup_{1\leq i\leq l}\sup_{x\in[0,1]^d} (f^*(x) - f_i(x)) < \infty$.*

Assumption 2. *The design distribution P_X is dominated by the Lebesgue measure with a continuous density $p(x)$ uniformly bounded above and away from 0 on $[0,1]^d$; that is, $p(x)$ satisfies $\underline{c} \leq p(x) \leq \overline{c}$ for some positive constants $\underline{c} \leq \overline{c}$.*

In addition, consider a multivariate nonnegative kernel function $K(u) : R^d \to R$ that satisfies Lipschitz, boundedness and bounded support conditions.

Assumption 3. *For some constants $0 < \lambda < \infty$,*

$$|K(u) - K(u')| \leq \lambda\|u - u'\|_\infty$$

for all u, $u' \in R^d$.

Assumption 4. *There exist constants $L_1 \leq L$, $c_3 > 0$ and $c_4 \geq 1$ such that $K(u) = 0$ for $\|u\|_\infty > L$, $K(u) \geq c_3$ for $\|u\|_\infty \leq L_1$, and $K(u) \leq c_4$ for all $u \in R^d$.*

Let h_n denote the bandwidth, where $h_n \to 0$ as $n \to \infty$. The Nadaraya-Watson estimator of $f_i(x)$ is

$$\hat{f}_{i,n+1}(x) = \frac{\sum_{j \in J_{i,n+1}} Y_{i,j} K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)}. \tag{1}$$

Theorem 1. *Suppose Assumptions 0-4 are satisfied. If the bandwidth sequence $\{h_n\}$ and the decreasing sequence $\{\pi_n\}$ are chosen to satisfy $h_n \to 0$, $\pi_n \to 0$ and*

$$\frac{n h_n^{2d} \pi_n^4}{\log n} \to \infty,$$

then the Nadaraya-Watson estimators defined in (1) are strongly consistent in L_∞ norm for the functions f_i.

Note that since checking L_∞ norm strong consistency of kernel methods is more challenging than that of histogram methods, new technical tools are necessarily developed to establish the strong consistency (as seen in the proof of Lemma 3 and Theorem 1 in Appendix B).

4.2 Finite-Time Regret Analysis

Next, we provide a finite-time regret analysis for the Nadaraya-Watson regression based randomized allocation strategy. To understand the regret cumulative rate, define a modulus of continuity $\omega(h; f_i)$ by

$$\omega(h; f_i) = \sup\{|f_i(x_1) - f_i(x_2)| : \|x_1 - x_2\|_\infty \leq h\}.$$

For technical convenience of guarding against the situation that the denominator of (1) is extremely small (which might occur with a non-negligible probability due to arm selection), in this subsection, we replace $K(\cdot)$ in (1) with the uniform kernel $I(\|u\|_\infty \leq L)$ when

$$\sum_{j \in J_{i,n+1}} K(\frac{x - X_j}{h_n}) < c_5 \sum_{j \in J_{i,n+1}} I(\|x - X_j\|_\infty \leq L h_n) \tag{2}$$

for some small positive constant $0 < c_5 < 1$. Given $0 < \delta < 1$ and the total time horizon N, we define a special time point $\tilde{n}_\delta$ by

$$n_\delta = \min\Big\{n > n_0 l : \sqrt{\frac{16 v^2 \log(8 l N^2/\delta)}{\underline{c} n (2 L h_n)^d \pi_n}} \leq \frac{c_5 v^2}{c} \text{ and } \exp\Big(-\frac{3 \underline{c} n (2 L h_n)^d \pi_n}{56}\Big) \leq \frac{\delta}{4 l N}\Big\}. \tag{3}$$

Under the condition that $\lim_{n \to \infty} n h_n^d \pi_n / \log n = \infty$, we can see from (3) that $n_\delta / N \to 0$ as $N \to \infty$. As a result, if the total time horizon is long enough, we have $N > n_\delta$.

Theorem 2. *Suppose Assumptions 0-2 and 4 are satisfied and $\{\pi_n\}$ is a decreasing sequence. Assume $N > n_\delta$ and the kernel function is chosen as described in* (2)*. Then with probability larger than $1-2\delta$, the cumulative regret satisfies*

$$R_N(\eta) < An_\delta + \sum_{n=n_\delta}^{N}\Big(2\max_{1\le i\le l}\omega(Lh_n;f_i) + \frac{C_{N,\delta}}{\sqrt{nh_n^d\pi_n}} + (l-1)\pi_n\Big) + A\sqrt{\frac{N}{2}}\log\big(\frac{1}{\delta}\big), \tag{4}$$

where $C_{N,\delta} = \sqrt{16c_4^2v^2\log(8lN^2/\delta)/c_5^2\underline{c}(2L)^d}$.

It is interesting to see from the right hand side of (4) that the regret upper bound consists of several terms that make intuitive sense. The first term An_δ comes from the initial rough exploration. The second term has three essential components: $\max_{1\le i\le l}\omega(Lh_n;f_i)$ is associated with the estimation bias, $C_{N,\delta}/\sqrt{nh_n^d\pi_n}$ conforms with the notion of estimation standard error, and $(l-1)\pi_n$ is the randomization error. The third term reflects the fluctuation of the randomization scheme. Such an upper bound explicitly illustrates both the bias-variance tradeoff and the exploration-exploitation tradeoff, which reflects the underlying nature of the proposed algorithm for the MABC problem.

Now we consider a smoothness assumption of the mean reward functions as follows.

Assumption 5. *There exist positive constants ρ and $\kappa \le 1$ such that for each reward function f_i, the modulus of continuity satisfies*

$$\omega(h;f_i) \le \rho h^\kappa.$$

Clearly, when $\kappa = 1$, Assumption 5 becomes Lipschitz continuity. As an immediate consequence of Theorem 2 and Assumption 5, we obtain the following result if we choose $h_n = \frac{1}{L}n^{-\frac{1}{3\kappa+d}}$ and $\pi_n = \frac{1}{l-1}n^{-\frac{1}{3+d/\kappa}}$.

Corollary 1. *Suppose the same conditions as in Theorem 2 are satisfied. Further assume Assumption 5 holds. Let $h_n = \frac{1}{L}n^{-\frac{1}{3\kappa+d}}$, $\pi_n = \frac{1}{l-1}n^{-\frac{1}{3+d/\kappa}}$ and $N > n_\delta$. Then with probability larger than $1-2\delta$, the cumulative regret satisfies*

$$R_N(\eta) < An_\delta + 2(2\rho + C^*_{N,\delta} + 1)N^{1-\frac{1}{3+d/\kappa}} + A\sqrt{\frac{N}{2}}\log\big(\frac{1}{\delta}\big),$$

where $C^*_{N,\delta} = \sqrt{16c_4^2v^2(l-1)\log(8lN^2/\delta)/2^dc_5^2\underline{c}}$.

In Corollary 1, the first term of the regret upper bound is dominated by the second term. Therefore, with high probability, the cumulative regret $R_N(\eta)$ increases at rate no faster than the order of $N^{1-\frac{1}{3+d/\kappa}}\log^{1/2}N$. This result can be seen more explicitly in Corollary 2, which gives an upper bound for the mean of $R_N(\eta)$. Note that by the definition of n_δ, the condition $N > n_{\delta^*}$ in Corollary 2 is satisfied if N is large enough.

Corollary 2. *Suppose the same conditions as in Theorem 2 are satisfied. Further assume Assumption 5 holds. Let $h_n = \frac{1}{L}n^{-\frac{1}{3\kappa+d}}$, $\pi_n = \frac{1}{l-1}n^{-\frac{1}{3+d/\kappa}}$ and $N > n_{\delta^*}$, where $\delta^* = N^{-\frac{1}{3+d/\kappa}}$. Then there exists a constant $C^* > 0$ (not dependent on N) such that the mean of cumulative regret satisfies*

$$ER_N(\eta) < C^*N^{1-\frac{1}{3+d/\kappa}}\log^{1/2}N.$$

As mentioned in Section 1, the derived regret cumulative rate in Corollary 2 is slightly slower than the minimax rate $N^{1-\frac{1}{2+d/\kappa}}$ obtained by Perchet and Rigollet (2013) (without assuming any extra margin condition). We tend to think this shows a limitation of the ϵ-greedy type approach. Nevertheless, with the help of the aforementioned model combining strategy along with the dimension reduction technique to be introduced in the next subsection, the kernel method based allocation can be quite flexible with potential practical use.

4.3 Dimension Reduction

When the covariate dimension is high, the Nadaraya-Watson estimation cannot be applied due to the curse of dimensionality. Next, we describe a dimension reduction subprocedure to handle this situation, which is also discussed in Qian and Yang (2012). Different from the method there, a sparse dimension reduction technique will be included to handle cases with higher-dimensional covariates.

Recall that Z^n is the set of observations $\{(X_j, I_j, Y_{I_j,j}),\, 1 \leq j \leq n-1\}$, and $Z^{i,n}$ is the subset of Z^n where $I_j = i$. Then $M_{i,n}$ is the number of observations in $Z^{i,n}$. Let $X^{i,n}$ be the $M_{i,n} \times d$ design matrix consisting of all covariates in $Z^{i,n}$, and let $Y^{i,n} \in R^{M_{i,n}}$ be the observed reward vector corresponding to $X^{i,n}$. It is known that kernel methods do not perform well when the dimension of covariates is high. We want to apply some dimension reduction methods (see, e.g., Li, 1991; Chen et al., 2010) to ($X^{i,n}$, $Y^{i,n}$) first to obtain lower dimensional covariates before using kernel estimation.

Specifically, suppose for each arm i, there exits a reduction function $s_i : R^d \to R^{r_i}$ ($r_i < d$), such that $f_i(x) = g_i(s_i(x))$ for some function $g_i : R^{r_i} \to R$. Clearly, if the reduction function s_i is known, $s_i(x)$ can be treated like the new lower-dimensional covariate, with which the kernel methods can be applied to find the estimate of g_i, and hence f_i. However, s_i is generally unknown in practice, and it is necessary to first obtain the estimate of s_i. In addition, we assume that s_i is a linear reduction function in the sense that $s_i(x) = B_i^T x$, where $B_i \in R^{d\times r_i}$ is a dimension reduction matrix. It is worth mentioning that s_i is not unique, i.e., $s_i(x) = \tilde{A}B_i^T x$ is a valid reduction function for any full rank matrix $\tilde{A} \in R^{r_i\times r_i}$. Therefore, it suffices to estimate the dimension reduction subspace $\texttt{span}(B_i)$ spanned by the columns of B_i, and obtain $\hat{s}_{i,n}(x) = \hat{B}_{i,n}^T x$, where $\hat{B}_{i,n} \in R^{d\times r_i}$ is one basis matrix of the estimated subspace at time n, and $\hat{s}_{i,n}$ is the estimate of s_i.

Dimension reduction methods such as sliced inverse regression (also known as SIR, see Li, 1991) can be applied to ($X^{i,n}$, $Y^{i,n}$) to obtain $\hat{B}_{i,n}$. In practice, it is convenient to have $X^{i,n}$ work on the standardized scale (i.e., the sample mean is zero and the sample covariance matrix is the identity matrix; Li, 1991; Cook, 2007). Suppose the Nadaraya-Watson estimation is used with $K_i(u) : R^{r_i} \to R$ being a multivariate symmetric kernel function for arm i. Recall $J_{i,n} = \{j : I_j = i,\, 1 \leq j \leq n-1\}$ is the set of past time points at which arm i is pulled. Then, we can obtain $\hat{f}_{i,n}$ with the following steps.

Step 1. Transform $X^{i,n}$ to the standardized-scale matrix $X_*^{n,i}$: transform the original covariates X_j's by $X_j^* = \hat{\Sigma}_{i,n}^{-1/2}(X_j - \bar{X}_{i,n})$ for every $j \in J_{i,n}$, where $\bar{X}_{i,n}$ and $\hat{\Sigma}_{i,n}$ are the sample mean vector and the sample covariance matrix of $X^{i,n}$, respectively.

Step 2. Apply a dimension reduction method to $(X_*^{i,n}, Y^{i,n})$ to obtain the estimated $d \times r_i$ dimension reduction matrix $\hat{B}_{i,n}^*$, where $\hat{B}_{i,n}^{*T}\hat{B}_{i,n}^* = I_{r_i}$. For example, we can apply SIR (Li, 1991) to obtain $\hat{B}_{i,n}^*$ by using the MATLAB package LDR (Cook et al., 2011, available at `https://sites.google.com/site/lilianaforzani/ldr-package`)

Step 3. Given $x \in R^d$, let $x^* = \hat{\Sigma}_{i,n}^{-1/2}(x - \bar{X}_{i,n})$ be the transformed x at the standardized scale. The Nadaraya-Watson estimator of $f_i(x)$ is

$$\hat{f}_{i,n}(x) = \frac{\displaystyle\sum_{j \in J_{i,n}} Y_{i,j} K_i\left(\frac{\hat{B}_{i,n}^{*T}x^* - \hat{B}_{i,n}^{*T}X_j^*}{h_{n-1}}\right)}{\displaystyle\sum_{j \in J_{i,n}} K_i\left(\frac{\hat{B}_{i,n}^{*T}x^* - \hat{B}_{i,n}^{*T}X_j^*}{h_{n-1}}\right)}. \tag{5}$$

In addition to estimating the reward function for each arm, it is sometimes of interest to know which variables contribute to the reward for each arm, and some sparse dimension reduction techniques can be applied. In particular, Chen et al. (2010) propose the coordinate-independent sparse estimation (CISE) to give sparse dimension reduction matrix such that the estimated coefficients of some predictors are zero for all reduction directions (i.e., some row vectors in $\hat{B}_{i,n}^*$ become $\mathbf{0}$). When the SIR objective function is used, the corresponding CISE method is denoted by CIS-SIR. To obtain $\hat{B}_{i,n}^*$ in Step 2 above using CIS-SIR, we can apply the MATLAB package CISE (Chen et al., 2010, available at `http://www.stat.nus.edu.sg/~stacx/`).

The simulation example in Section 6 and the real data example in Section 7 both use the algorithms described here. The simulation example is implemented in MATLAB and the real data example is implemented in C++. The major source code illustrating the proposed algorithms is available upon request.

5. Strong Consistency in Model Combining Based Allocation

Next, we consider the general case that multiple function estimation methods are used for model combining. In general, it is technically difficult to verify strong consistency in L_∞ norm for a regression method. Also, practically, it is likely that some methods may give good estimation for only a subset of the arms, but performs poorly for the rest. Not knowing which methods work well for which arms, we proposed the combining algorithm in Section 3 to address this issue. We will show that even in the presence of bad-performing regression methods, the strong consistency of our allocation strategy still holds if for any given arm, there is at least one good regression method included for combining.

Given an arm i, let $N_t^{(i)} = \inf\{n : \sum_{j=n_0 l+1}^{n} I(I_j = i) \geq t\}$, $t \geq 1$, be the earliest time point where arm i is pulled exactly t times after the forced sampling period. For notation brevity, we use N_t instead of $N_t^{(i)}$ in the rest of this section. Consider the assumptions as follows.

Assumption A. *Given any arm $1 \leq i \leq l$, the candidate regression procedures in Δ can be categorized into one of the two subsets denoted by Δ_{i1} (non-empty) and Δ_{i2}. All procedures* *in Δ_{i1} are strongly consistent in L_∞ norm for arm i, while procedures in Δ_{i2} are less well-performing in the sense that for each procedure δ_s in Δ_{i2}, there exist a procedure δ_r in Δ_{i1} and some constants $b > 0.5$, $c_1 > 0$ such that*

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T}\big(\hat{f}_{i,N_t,s}(X_{N_t}) - f_i(X_{N_t})\big)^2 - \sum_{t=1}^{T}\big(\hat{f}_{i,N_t,r}(X_{N_t}) - f_i(X_{N_t})\big)^2}{\sqrt{T}(\log T)^b} > c_1$$

with probability one.

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

Assumption C. *$\|\hat{f}_{i,n,r} - f_i\|_\infty$ is upper bounded by a constant c_2 for all $1 \le i \le l$, $n \ge n_0 l + 1$ and $1 \le r \le m$.*

Assumption D. *The variance estimates $\hat{v}_{i,n}$ are upper bounded by a positive constant q with probability one for all $1 \le i \le l$ and $n \ge n_0 l + 1$.*

Assumption E. *The sequence $\{\pi_n, n \ge 1\}$ satisfies that $\sum_{n=1}^{\infty}\pi_n$ diverges.*

Note that Assumption A is automatically satisfied if all the regression methods happen to be strongly consistent (i.e., Δ_{i2} is empty). When a bad-performing method does exist, Assumption A requires that the difference of the mean square errors between a good-performing method and a bad-performing method decreases slower than the order of $(\log T)^b/\sqrt{T}$. If a parametric method δ_s in Δ is based on a wrong model, $\sum_{t=1}^{T}\big(\hat{f}_{i,N_t,s}(X_{N_t}) - f_i(X_{N_t})\big)^2$ is of order T, and then the requirement in Assumption A is met. For an inefficient nonparametric method, the enlargement of the mean square error by the order larger than $(\log T)^b/\sqrt{T}$ is natural to expect. Assumption B is a natural condition in the context of our bandit problem. Assumptions C and D are immediately satisfied if the response is bounded and the estimator is, e.g., a weighted average of some previous observations. Assumption E ensures that N_t is finite as shown in Lemma 5 in the Appendix. As implied in Lemma 5, if we are allowed to play the game infinitely many times, each arm will be pulled beyond any given integer. This guarantees that each "inferior" arm can be pulled reasonably often to ensure enough exploration.

Theorem 3. *Under Assumption 0 and Assumptions A-E, the model combining allocation strategy is strongly consistent.*

With Theorem 3, one is safe to explore different models or methods in estimating the mean reward functions that may or may not work well for some or all arms. The resulting per-round regret can be much improved if good methods (possibly different for different arms) are added in.

6. Simulations

In this section, we intend to illustrate the dimension reduction function estimation procedures described in Section 4.3 for bandit problem with multivariate covariates. Two

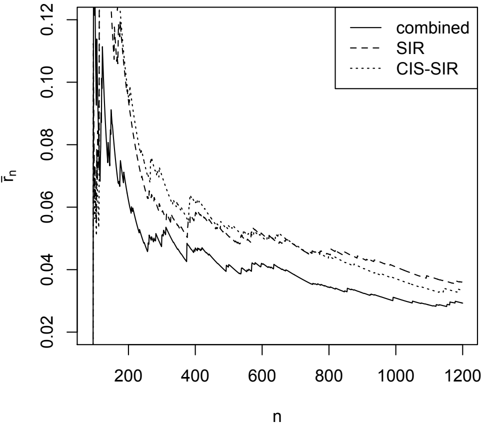

Figure 1: Averaged per-round regret from combining SIR and CIS-SIR.

readily available MATLAB packages for dimension reduction are used: LDR package (Cook et al., 2011) for SIR, and CISE package (Chen et al., 2010) for CIS-SIR. The kernel used is the Gaussian kernel

$$K(t) = \exp(-\frac{\|t\|_2^2}{2}).$$

We consider a three-arm bandit model with $d = 10$. Assume that at each time n, the covariate is $X_n = (X_{n1}, X_{n2}, \cdots, X_{nd})^T$, and X_{ni}'s ($i = 1, \cdots, d$) are i.i.d random variables from uniform(0,1). Assume the error $\epsilon_n \sim 0.5N(0, 1)$. Consider the mean reward functions

$$\begin{aligned}
f_1(X_n) =& 0.5(X_{n1} + X_{n2} + X_{n3}),\\
f_2(X_n) =& 0.4(X_{n3} + X_{n4})^2 + 1.5\sin(X_{n1} + 0.25X_{n4}),\\
f_3(X_n) =& \frac{2X_{n3}}{0.5 + (1.5 + X_{n3} + X_{n4})}.
\end{aligned}$$

We set the reduction dimensions for the three arms by $r_1 = 1$, $r_2 = 2$ and $r_3 = 2$. Given the time horizon $N = 1200$, the first 90 rounds of the game are the forced sampling period. Let the "inferior" arm sampling probability be $\pi_n = \frac{1}{(\log_2 n)^2}$, and the kernel bandwidth for arm i be $h = n^{-1/(2+r_i)}$, $i = 1, 2, 3$. Dimension reduction methods SIR, CIS-SIR as well as their combining strategy are run separately, and their per-round regret r_n is summarized in Figure 1 (right panel), which shows that the combining strategy performs the best. Since the second arm ($i = 2$) is played the most (for SIR, 1022 times; for CIS-SIR, 1026 times), we show the estimated dimension reduction matrix for the second arm at the last time point $n = N$ in Table 1. As expected, CIS-SIR results in a sparse dimension reduction matrix with rows 1, 3 and 4 being non-zero.

It is worth mentioning that in the simulation above, we assume the reduction dimensions for all arms are already known. In cases where the reduction dimensions are unknown, we may apply model selection procedures to choose them, which will be investigated in the future.

	SIR		CIS-SIR	
1	-0.658	-0.599	-0.611	0.681
2	0.011	-0.091	0	0
3	-0.469	0.601	-0.491	0.071
4	-0.582	0.219	-0.620	-0.728
5	-0.001	0.075	0	0
6	0.071	0.232	0	0
7	0.013	-0.340	0	0
8	-0.019	0.087	0	0
9	-0.029	-0.194	0	0
10	0.016	0.030	0	0

Table 1: Comparing the estimated dimension reduction matrix $\hat{B}^*_{2,N}$ for the second arm between SIR and CIS-SIR.

7. Web-Based Personalized News Article Recommendation

In this section, we use the Yahoo! Front Page Today Module User Click Log data set (Yahoo! Academic Relations, 2011) to evaluate the proposed allocation strategy. The complete data set contains about 46 million web page visit interaction events collected during the first ten days in May 2009. Each of these events has four components: (1) five variables constructed from the Yahoo! front page visitor's information; (2) a pool of about 10-14 editor-picked news articles; (3) one article actually displayed to the visitor (it is selected uniformly at random from the article pool); (4) the visitor's response to the selected article (no click: 0, click: 1). Since different visitors may have different preferences for the same article, it is reasonable to believe that the displayed article should be selected based on the visitor associated variables. If we treat the articles in the pool as the bandit arms, and the visitor associated variables as the covariates, this data set provides the necessary platform to test a MABC algorithm.

One remaining issue before algorithm evaluation is that the complete data set is long-term in nature and the pool of articles is dynamic, i.e., some outdated articles are dropped out as people's interest in these articles fades away, and some breaking-news articles can appear and be added to the pool. Our current problem setup, however, assumes stationary mean reward functions with a fixed set of arms. To avoid introducing biased evaluation results, we focus on short-term performance where people's interest on a particular article does not change too much and the pool of articles remains stable. Therefore, we consider only one day's data (May 1, 2009). Also, we choose four articles ($l = 4$) as the candidate bandit arms (article id 109511 - 109514), and keep only the events where the four articles are included in the article pool and one of the four articles is actually displayed to the visitor. A similar screening treatment of the data set is used in May et al. (2012) for MABC algorithm evaluation purposes. With the above, we obtain a reduced data set containing 452,189 interaction events for subsequent use.

Another challenge in evaluating a MABC algorithm comes from the intrinsic nature of bandit problem: for every visitor interaction event, only one article is displayed, and we only have this visitor's response to the displayed article, while his/her response to other articles is not available, causing a difficulty if the actually displayed article does not match the article selected by a MABC algorithm. To overcome this issue caused by limited feedback, we apply the unbiased offline evaluation method proposed by Li et al. (2010). Briefly, for each encountered event, the MABC algorithm uses the previous "valid" data set (history) to estimate the mean reward functions and propose an arm to pull. If the proposed arm matches the actually displayed arm, this event is kept as a "valid" event, and the "valid" data set (history) is updated by adding this event. On the other hand, if the proposed arm does not match the displayed arm, this event is ignored, and the "valid" data set (history) is unchanged. This process is run sequentially over all the interaction events to generate the final "valid" data set, upon which a MABC algorithm can be evaluated by calculating the click-through rate (CTR, the proportion of times a click is made). Under the fact that in each interaction event, the displayed arm was selected uniformly at random from the pool, it can be argued that the final "valid" data set is like being obtained from running the MABC algorithm over a random sample of the underlying population.

With the reduced data set and the unbiased offline evaluation method, we evaluate the performance of the following algorithms.

random: an arm is selected uniformly at random.

ϵ-greedy: The randomized allocation strategy is run naively without consideration of covariates. A simple average is used to estimate the mean reward for each arm.

SIR-kernel: The randomized allocation strategy is run using SIR-kernel method to estimate the mean reward functions. Three sequences of bandwidth choices are considered: $h_{n1} = n^{-1/6}$, $h_{n2} = n^{-1/8}$ and $h_{n3} = n^{-1/10}$.

model combining: Model combining based randomized allocation strategy described in Section 3 is run with SIR-kernel method ($h_{n3} = n^{-1/10}$) and the naive simple average method (ϵ-greedy) as two candidate modeling methods.

The ϵ-greedy, SIR-kernel and model combining algorithms described above all take the first 1000 time points to be the forced sampling stage and use $\pi_n = n^{-1/4}/6$. Also, for any given arm, the SIR-kernel method limits the history time window for reward estimation to have maximum sample size of 10,000 (larger history sample size does not give us noticeable difference in performance). In addition, we consider the following parametric algorithm:

LinUCB: LinUCB employs Bayesian logistic regression to estimate the mean reward functions. The detailed implementation procedures are described in Algorithm 3 of Chapelle and Li (2011).

Each of the algorithms listed above is run 100 times over the reduced data set with the unbiased offline evaluation method. For each of the 100 runs, the algorithm starts at a position randomly chosen from the first 10,000 events of the reduced data set. The resulting CTRs are divided by the mean CTR of the random algorithm to give the normalized CTRs, and their boxplots are shown in Fig. 2.

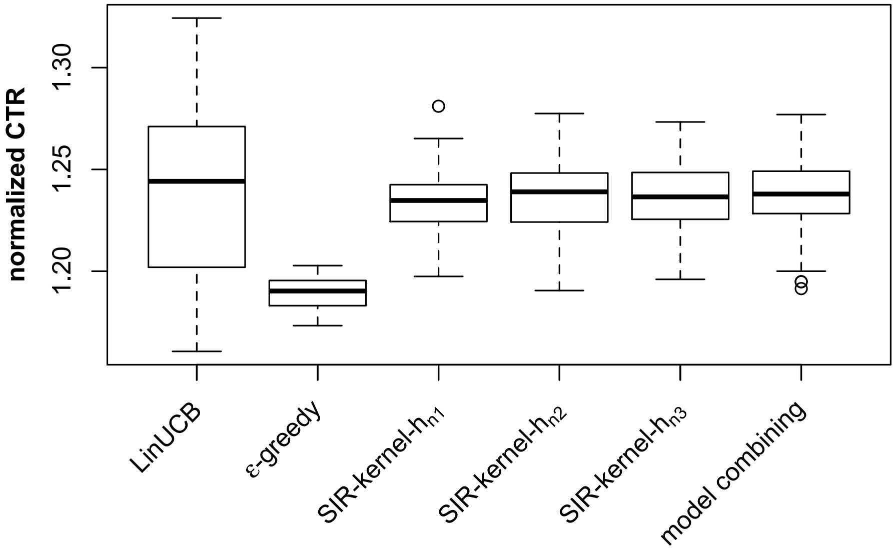

Figure 2: Boxplots of normalized CTRs of various algorithms on the news article recommendation data set. Algorithms include (from left to right): LinUCB, ϵ-greedy, SIR-kernel (h_{n1}), SIR-kernel (h_{n2}), SIR-kernel (h_{n3}), model combining with SIR-kernel (h_{n3}) and ϵ-greedy. CTRs are normalized with respect to the random algorithm.

It appears that the SIR-kernel methods with different candidate bandwidth sequences have very similar performance. The naive ϵ-greedy algorithm, however, clearly underperforms due to its failure to take advantage of the response-covariate association. When the naive simple average estimation (ϵ-greedy) is used together with SIR-kernel method (h_{n3}) in the model combining algorithm, the overall performance does not seem to deteriorate with the existence of this naive estimation method, showing once again that the model combining algorithm allows us to safely explore new modeling methods by automatically selecting the appropriate modeling candidate. Given that the covariates in the news article recommendation data set are constructed with logistic regression related methods (Li et al., 2010), it is satisfactory to observe that SIR-kernel algorithm can have similar performance with relatively small variation when compared with the LinUCB algorithm.

8. Conclusions

In this work, we study the kernel estimation based randomized allocation strategy in a flexible setting for MABC that works for both continuous and discrete response, and establish both the strong consistency and a finite-time regret upper bound. Allowing dependence of the covariate and the error, we rely on new technical tools to add kernel methods to the family of strongly consistent methods, which can potentially improve estimation efficiency for smooth reward functions. Although the finite-time regret upper bound is slightly sub-optimal for the investigated randomized allocation strategy in the minimax sense (Perchet and Rigollet, 2013), the flexibility in estimation of the mean reward functions can be very useful in applications. In that regard, we integrate a model combination technique into the allocation strategy to share the strengths of different statistical learning methods for reward function estimation. It is shown that with the proposed data-driven combination of estimation methods, our allocation strategy can remain strongly consistent.

In Appendix A, we also show that by resorting to an alternative UCB-type criterion for arm comparison, the regret rate of the modified randomized allocation algorithm is improved to be minimax optimal up to a logarithmic factor. It remains to be seen if the UCB modification can be incorporated to construct a model combination algorithm with adaptive minimax rate. Moreover, as an important open question raised by a reviewer, it would be interesting to see whether the cumulative regret of the model combination strategy is comparable to that of the candidate model with the smallest regret in the sense of an oracle-type inequality similar to that of, e.g., Audibert (2009).

Acknowledgments

We would like to thank the Editor and two anonymous referees for their very constructive comments that help improving this manuscript significantly. This research was supported by the US NSF grant DMS-1106576

Appendix A. A Kernel Estimation Based UCB Algorithm

In this section, we modify the randomized allocation strategy to give a UCB-type algorithm that results in an improved rate of the cumulative regret. Similar to Section 4, we consider the Nadaraya-Watson estimation as the only modeling method, that is,

$$\hat{f}_{i,n}(x) = \frac{\sum_{j\in J_{i,n}} Y_{i,j} K\left(\frac{x-X_j}{h_{n-1}}\right)}{\sum_{j\in J_{i,n}} K\left(\frac{x-X_j}{h_{n-1}}\right)}.$$

We slightly revise step 5 of the proposed randomized allocation strategy:

STEP 5′. Estimate the best arm, select and pull. For the covariate X_n, define

$$\tilde{i}_n = \operatorname{argmax}_{1\leq i\leq l} \hat{f}_{i,n}(X_n) + U_{i,n}(X_n), \tag{6}$$

where $U_{i,n}(x) = \frac{\tilde{c}\sqrt{(\log N)\sum_{j\in J_{i,n}} K^2\left(\frac{x-X_j}{h_{n-1}}\right)}}{\sum_{j\in J_{i,n}} K\left(\frac{x-X_j}{h_{n-1}}\right)}$ and $\tilde{c}$ is some positive constant (if there is a tie, any tie-breaking rule may be applied). Choose an arm, with probability $1-(l-1)\pi_n$ for arm $\tilde{i}_n$ (the currently most promising choice) and with probability π_n for each of the remaining arms. That is,

$$I_n = \begin{cases} \tilde{i}_n, & \text{with probability } 1-(l-1)\pi_n, \\ i, & \text{with probability } \pi_n,\ i \neq \tilde{i}_n,\ 1\leq i\leq l. \end{cases}$$

Clearly, (6) shows a UCB-type algorithm that naturally extends from the UCB1 of Auer et al. (2002) and the UCBogram of Rigollet and Zeevi (2010). Indeed, given the uniform kernel $K(u) = I(\|u\|_\infty \leq 1)$, we have $U_{i,n}(x) = \tilde{c}\sqrt{\frac{\log N}{N_{i,n}(x)}}$, where $N_{i,n}(x)$ is the number of times arm i gets pulled inside the cube that centers at x with bin width $2h_{n-1}$. For presentation clarity, we assume $K(\cdot)$ is the uniform kernel, but the results can be generalized to kernel functions that satisfy Assumption 4. As shown in Theorem 4 below, the finite-time regret upper bound of the UCB-type algorithm achieves the minimax rate up to a logarithmic factor.

Theorem 4. *Suppose Assumptions 0-1 hold and the uniform kernel function is used. Then for the modified algorithm, if $n_0 = lNh^\kappa$, $\tilde{c} > \max\{2\sqrt{3}v, 12c\}$, $h = h_n = 1/\lceil (\frac{N}{\log N})^{\frac{1}{2\kappa+d}} \rceil$ and $\pi_n \leq \frac{1}{l} \wedge \frac{1}{\underline{c}}(\frac{\log N}{n})^{\frac{1}{2+d/\kappa}}$, the mean of cumulative regret satisfies*

$$ER_N(\eta) < \tilde{C}^* N^{1-\frac{1}{2+d/\kappa}} (\log N)^{\frac{1}{2+d/\kappa}}. \tag{7}$$

It is worth noting that despite the seemingly minor algorithmic modification, the proof techniques used by Theorem 2 and Theorem 4 are quite different. The key difference is that: the UCB-type criterion enables us to provide upper bounds (with high probability) for the number of times the "inferior" arms are selected, and these bounds are dependent on the reward difference between the "optimal" and the "inferior" arms; for the algorithm before modification, we have to rely on studying the estimation errors of the reward functions and the UCB-type arguments do not apply. It is not settled yet as to whether the suboptimal rate of the ϵ-greedy type algorithm is intrinsic to the method or is the limitation of the proof techniques. But we tend to think that the rate given for the ϵ-greedy type algorithm is intrinsic to the method. Also, although the UCB-type algorithm leads to an improved regret rate, it is not yet clear how it could be used to construct a model combination algorithm.

Appendix B. Lemmas and Proofs

B.1 Proof of Theorem 1

Lemma 1. *Suppose $\{\mathcal{F}_j, j = 1, 2, \cdots\}$ is an increasing filtration of σ-fields. For each $j \geq 1$, let ε_j be an $\mathcal{F}_{j+1}$-measurable random variable that satisfies $E(\varepsilon_j|\mathcal{F}_j) = 0$, and let T_j be an $\mathcal{F}_j$-measurable random variable that is upper bounded by a constant $C > 0$ in absolute value almost surely. If there exist positive constants v and c such that for all $k \geq 2$ and $j \geq 1$, $E(|\varepsilon_j|^k|\mathcal{F}_j) \leq k!v^2c^{k-2}/2$, then for every $\epsilon > 0$ and every integer $n \geq 1$,*

$$P\Big(\sum_{j=1}^{n} T_j\varepsilon_j \geq n\epsilon\Big) \leq \exp\Big(-\frac{n\epsilon^2}{2C^2(v^2 + c\epsilon/C)}\Big).$$

Proof of Lemma 1. Note that

$$\begin{aligned} P\Big(\sum_{j=1}^{n} T_j\varepsilon_j \geq n\epsilon\Big) &\leq e^{-tn\epsilon}E\Big[\exp\Big(t\sum_{j=1}^{n}T_j\varepsilon_j\Big)\Big] \\ &= e^{-tn\epsilon}E\Big[E\Big(\exp(t\sum_{j=1}^{n}T_j\varepsilon_j)|\mathcal{F}_n\Big)\Big] \\ &= e^{-tn\epsilon}E\Big[\exp\Big(t\sum_{j=1}^{n-1}T_j\varepsilon_j\Big)E(e^{tT_n\varepsilon_n}|\mathcal{F}_n)\Big]. \end{aligned}$$

By the moment condition on ε_n and Taylor expansion, we have

$$\begin{aligned} \log E(e^{tT_n\varepsilon_n}|\mathcal{F}_n) &\leq E(e^{tT_n\varepsilon_n}|\mathcal{F}_n) - 1 \\ &\leq tT_nE(\varepsilon_n|\mathcal{F}_n) + \sum_{k=2}^{\infty}\frac{t^k|T_n|^k}{k!}E(|\varepsilon_n|^k|\mathcal{F}_n) \\ &\leq \frac{v^2C^2t^2}{2}(1 + cCt + (cCt)^2 + \cdots) \\ &= \frac{v^2C^2t^2}{2(1-cCt)} \end{aligned}$$

for $t < 1/cC$. Thus, it follows by induction that

$$\begin{aligned} P\Big(\sum_{j=1}^{n} T_j\varepsilon_j \geq n\epsilon\Big) &\leq \exp\Big(-tn\epsilon + \frac{nv^2C^2t^2}{2(1-cCt)}\Big) \\ &\leq \exp\Big(-\frac{n\epsilon^2}{2C^2(v^2 + c\epsilon/C)}\Big), \end{aligned}$$

where the last inequality is obtained by minimization over t. This completes the proof of Lemma 1. □

Lemma 2. *Suppose $\{\mathcal{F}_j, j = 1, 2, \cdots\}$ is an increasing filtration of σ-fields. For each $j \geq 1$, let W_j be an $\mathcal{F}_j$-measurable Bernoulli random variable whose conditional success probability satisfies*

$$P(W_j = 1|\mathcal{F}_{j-1}) \geq \beta_j$$

for some $0 \leq \beta_j \leq 1$. Then given $n \geq 1$,

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

Lemma 2 is known as an extended Bernstein inequality (see, e.g., Yang and Zhu (2002), Section A.4.). For completeness, we give a brief proof here.

Proof of Lemma 2. Suppose $\tilde{W}_j$, $1 \leq j \leq n$ are independent Bernoulli random variables with success probability β_j, and are assumed to be independent of $\mathcal{F}_n$. By Bernstein's inequality (e.g., Cesa-Bianchi and Lugosi, 2006, Corollary A.3),

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

Also, $\sum_{j=1}^{n} W_j$ is stochastically no smaller than $\sum_{j=1}^{n} \tilde{W}_j$, that is, for every t, $P(\sum_{j=1}^{n} W_j > t) \geq P(\sum_{j=1}^{n} \tilde{W}_j > t)$. Indeed, noting that $P(W_n > t|\mathcal{F}_{n-1}) \geq P(\tilde{W}_n > t)$ for every t, we have

$$P(W_1 + \cdots + W_n > t|\mathcal{F}_{n-1}) \geq P(W_1 + \cdots + W_{n-1} + \tilde{W}_n > t|\mathcal{F}_{n-1}).$$

Similarly, by $P(W_{n-1} > t|\mathcal{F}_{n-2}) \geq P(\tilde{W}_{n-1} > t)$ for every t and the independence of $\tilde{W}_j$'s,

$$P(W_1 + \cdots + W_{n-1} + \tilde{W}_n > t|\mathcal{F}_{n-2}, \tilde{W}_n) \geq P(W_1 + \cdots + W_{n-2} + \tilde{W}_{n-1} + \tilde{W}_n > t|\mathcal{F}_{n-2}, \tilde{W}_n).$$

Continuing the process above, we can see that $P(\sum_{j=1}^{n} W_j > t) \geq P(\sum_{j=1}^{n} \tilde{W}_j > t)$ holds. □

Lemma 3. *Under the settings of the kernel estimation in Section 4.1, given arm i and a cube $A \subset [0,1]^d$ with side width h, if Assumptions 0, 3 and 4 are satisfied, then for any $\epsilon > 0$,*

$$\begin{aligned} &P\Big(\sup_{x \in A} \sum_{j \in J_{i,n+1}} \varepsilon_j K\Big(\frac{x - X_j}{h_n}\Big) > \frac{n\epsilon}{1 - 1/\sqrt{2}}\Big) \\ &\leq \exp\Big(-\frac{n\epsilon^2}{4c_4^2 v^2}\Big) + \exp\Big(-\frac{n\epsilon}{4c_4 c}\Big) + \sum_{k=1}^{\infty} 2^{kd} \exp\Big(-\frac{2^k n\epsilon^2}{\lambda^2 v^2}\Big) + \sum_{k=1}^{\infty} 2^{kd} \exp\Big(-\frac{2^{k/2} n\epsilon}{2\lambda c}\Big). \end{aligned}$$

Proof of Lemma 3. At each time point j, let $W_j = 1$ if arm i is pulled (i.e., $I_j = i$), and $W_j = 0$ otherwise. Denote $G(x) = \sum_{j=1}^{n} \varepsilon_j W_j K(\frac{x - X_j}{h_n})$. Then, to find an upper bound for $P(\sup_{x \in A} G(x) > n\epsilon/(1 - 1/\sqrt{2}))$, we use a "chaining" argument. For each $k \geq 0$, let $\gamma_k = h_n/2^k$, and we can partition the cube A into 2^{kd} bins with bin width γ_k. Let F_k denote the set consisting of the center points of these 2^{kd} bins. Clearly, $\text{card}(F_k) = 2^{kd}$, and F_k is a $\gamma_k/2$-net of A in the sense that for every $x \in A$, we can find a $x' \in F_k$ such that $\|x - x'\|_\infty \leq \gamma_k/2$. Let $\tau_k(x) = \text{argmin}_{x' \in F_k} \|x - x'\|_\infty$ be the closest point to x in the net F_k. With the sequence $F_0, F_1, F_2, \cdots$ of $\gamma_0/2, \gamma_1/2, \gamma_2/2, \cdots$ nets in A, it is easy to see that for every $x \in A$, $\|\tau_k(x) - \tau_{k-1}(x)\|_\infty \leq \gamma_k/2$ and $\lim_{k \to \infty} \tau_k(x) = x$. Thus, by the continuity of the kernel function, we have $\lim_{k \to \infty} G(\tau_k(x)) = G(x)$. It follows that

$$G(x) = G(\tau_0(x)) + \sum_{k=1}^{\infty} \big[G(\tau_k(x)) - G(\tau_{k-1}(x))\big].$$

Thus,

$$
\begin{aligned}
&P\Big(\sup_{x\in A} G(x) > \frac{n\epsilon}{1-1/\sqrt{2}}\Big)\\
&= P\Big(\sup_{x\in A}\big\{G(\tau_0(x)) + \sum_{k=1}^{\infty}\big[G(\tau_k(x)) - G(\tau_{k-1}(x))\big]\big\} > \sum_{k=0}^{\infty}\frac{n\epsilon}{2^{k/2}}\Big)\\
&\le P\Big(\sup_{x\in A} G(\tau_0(x)) > n\epsilon\Big) + \sum_{k=1}^{\infty} P\Big(\sup_{x\in A}\big[G(\tau_k(x)) - G(\tau_{k-1}(x))\big] > \frac{n\epsilon}{2^{k/2}}\Big)\\
&\le P\Big(\sup_{x\in F_0} G(x) > n\epsilon\Big) + \sum_{k=1}^{\infty} P\Big(\sup_{\substack{x_2\in F_k,\, x_1\in F_{k-1}\\ \|x_2-x_1\|_\infty\le\gamma_k/2}}\big[G(x_2) - G(x_1)\big] > \frac{n\epsilon}{2^{k/2}}\Big)\\
&\le \mathrm{card}(F_0)\max_{x\in F_0} P\big(G(x) > n\epsilon\big)\\
&\qquad + \sum_{k=1}^{\infty} 2^d \mathrm{card}(F_{k-1}) \max_{\substack{x_2\in F_k,\, x_1\in F_{k-1}\\ \|x_2-x_1\|_\infty\le\gamma_k/2}} P\Big(G(x_2) - G(x_1) > \frac{n\epsilon}{2^{k/2}}\Big), \qquad (8)
\end{aligned}
$$

where the last inequality holds because for each $x_1 \in F_{k-1}$, there are only 2^d such points $x_2 \in F_k$ that can satisfy $\|x_2 - x_1\|_\infty \le \gamma_k/2$. Given $x \in F_0$, since $|W_j K(\frac{x-X_j}{h})| \le c_4$ almost surely for all $j \ge 1$, it follows by Lemma 1 that

$$
P\big(G(x) > n\epsilon\big) \le \exp\Big(-\frac{n\epsilon^2}{2c_4^2(v^2 + c\epsilon/c_4)}\Big). \qquad (9)
$$

Similarly, given $x_2 \in F_k$, $x_1 \in F_{k-1}$ and $\|x_2 - x_1\|_\infty \le \gamma_k$, since

$$
\Big|K\Big(\frac{x_2 - X_j}{h_n}\Big) - K\Big(\frac{x_1 - X_j}{h_n}\Big)\Big| \le \frac{\lambda\|x_2 - x_1\|_\infty}{h_n} \le \frac{\lambda\gamma_k}{2h_n} = \frac{\lambda}{2^{k+1}}
$$

almost surely, it follows by Lemma 1 that

$$
\begin{aligned}
P\Big(G(x_2) - G(x_1) > \frac{n\epsilon}{2^{k/2}}\Big) &= P\Big(\sum_{j=1}^{n}\epsilon_j W_j\Big[K\Big(\frac{x_2 - X_j}{h}\Big) - K\Big(\frac{x_1 - X_j}{h}\Big)\Big] > \frac{n\epsilon}{2^{k/2}}\Big)\\
&\le \exp\Big(-\frac{2^{k+2} n\epsilon^2}{2\lambda^2(v^2 + 2^{k/2+1}c\epsilon/\lambda)}\Big). \qquad (10)
\end{aligned}
$$

Thus, by (8), (9) and (10),

$$
\begin{aligned}
&P\Big(\sup_{x\in A} G(x) > \frac{n\epsilon}{1-1/\sqrt{2}}\Big)\\
&\le \exp\Big(-\frac{n\epsilon^2}{2c_4^2(v^2 + c\epsilon/c_4)}\Big) + \sum_{k=1}^{\infty} 2^{kd}\exp\Big(-\frac{2^{k+2} n\epsilon^2}{2\lambda^2(v^2 + 2^{k/2+1}c\epsilon/\lambda)}\Big)\\
&\le \exp\Big(-\frac{n\epsilon^2}{4c_4^2 v^2}\Big) + \exp\Big(-\frac{n\epsilon}{4c_4 c}\Big) + \sum_{k=1}^{\infty} 2^{kd}\exp\Big(-\frac{2^k n\epsilon^2}{\lambda^2 v^2}\Big) + \sum_{k=1}^{\infty} 2^{kd}\exp\Big(-\frac{2^{k/2} n\epsilon}{2\lambda c}\Big).
\end{aligned}
$$

This completes the proof of Lemma 3. □

Proof of Theorem 1. Recall that $M_{i,n} = |J_{i,n}|$, $\underline{c}$ is the covariate density lower bound, and L, L_1, c_3 are constants defined in Assumption 4 for the kernel function $K(\cdot)$, and. Note that for each $x \in R^d$,

$$|\hat{f}_{i,n+1}(x) - f_i(x)| = \left| \frac{\sum\limits_{j \in J_{i,n+1}} Y_{i,j} K\left(\dfrac{x - X_j}{h_n}\right)}{\sum\limits_{j \in J_{i,n+1}} K\left(\dfrac{x - X_j}{h_n}\right)} - f_i(x) \right|$$

$$= \left| \frac{\sum\limits_{j \in J_{i,n+1}} (f_i(X_j) + \varepsilon_j) K\left(\dfrac{x - X_j}{h_n}\right)}{\sum\limits_{j \in J_{i,n+1}} K\left(\dfrac{x - X_j}{h_n}\right)} - f_i(x) \right|$$

$$= \left| \frac{\sum\limits_{j \in J_{i,n+1}} (f_i(X_j) - f_i(x)) K\left(\dfrac{x - X_j}{h_n}\right)}{\sum\limits_{j \in J_{i,n+1}} K\left(\dfrac{x - X_j}{h_n}\right)} + \frac{\sum\limits_{j \in J_{i,n+1}} \varepsilon_j K\left(\dfrac{x - X_j}{h_n}\right)}{\sum\limits_{j \in J_{i,n+1}} K\left(\dfrac{x - X_j}{h_n}\right)} \right|$$

$$\le \sup_{\{x,y: \|x-y\|_\infty \le L h_n\}} |f_i(x) - f_i(y)| + \left| \frac{\dfrac{1}{M_{i,n+1} h_n^d} \sum\limits_{j \in J_{i,n+1}} \varepsilon_j K\left(\dfrac{x - X_j}{h_n}\right)}{\dfrac{1}{M_{i,n+1} h_n^d} \sum\limits_{j \in J_{i,n+1}} K\left(\dfrac{x - X_j}{h_n}\right)} \right|, \tag{11}$$

where the last inequality follows from the bounded support assumption of kernel function $K(\cdot)$. By uniform continuity of the function f_i,

$$\lim_{n \to \infty} \sup_{\{x,y: \|x-y\|_\infty \le L h_n\}} |f_i(x) - f_i(y)| = 0.$$

To show that $\|\hat{f}_{i,n} - f_i\|_\infty \to 0$ as $n \to \infty$, we only need

$$\sup_{x \in [0,1]^d} \left| \frac{\dfrac{1}{M_{i,n+1} h_n^d} \sum\limits_{j \in J_{i,n+1}} \varepsilon_j K\left(\dfrac{x - X_j}{h_n}\right)}{\dfrac{1}{M_{i,n+1} h_n^d} \sum\limits_{j \in J_{i,n+1}} K\left(\dfrac{x - X_j}{h_n}\right)} \right| \to 0 \quad \text{as } n \to \infty. \tag{12}$$

First, we want to show

$$\inf_{x \in [0,1]^d} \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} K\Big(\frac{x - X_j}{h_n}\Big) > \frac{c_3 \underline{c} L_1^d \pi_n}{2}, \tag{13}$$

almost surely for large enough n. Indeed, for each $n \ge n_0 l + 1$, we can partition the unit cube $[0,1]^d$ into $\tilde{B}$ bins with bin width $L_1 h_n$ such that $\tilde{B} \le 1/(L_1 h_n)^d$. We denote these bins by $\tilde{A}_1, \tilde{A}_2, \cdots, \tilde{A}_{\tilde{B}}$. Given an arm i and $1 \leq k \leq \tilde{B}$, for every $x \in \tilde{A}_k$, we have

$$\begin{aligned}\sum_{j \in J_{i,n+1}} K\Big(\frac{x - X_j}{h_n}\Big) &= \sum_{j=1}^{n} I(I_j = i) K\Big(\frac{x - X_j}{h_n}\Big) \\ &\geq \sum_{j=1}^{n} I(I_j = i, X_j \in \tilde{A}_k) K\Big(\frac{x - X_j}{h_n}\Big) \\ &\geq c_3 \sum_{j=1}^{n} I(I_j = i, X_j \in \tilde{A}_k),\end{aligned}$$

where the last inequality follows by Assumption 4. Consequently,

$$\begin{aligned}&P\Big(\inf_{x \in \tilde{A}_k} \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} K\Big(\frac{x - X_j}{h_n}\Big) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\Big) \\ \leq &P\Big(\inf_{x \in \tilde{A}_k} \frac{1}{n h_n^d} \sum_{j \in J_{i,n+1}} K\Big(\frac{x - X_j}{h_n}\Big) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\Big) \\ \leq &P\Big(\frac{c_3}{n h_n^d} \sum_{j=1}^{n} I(I_j = i, X_j \in \tilde{A}_k) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\Big) \\ = &P\Big(\sum_{j=1}^{n} I(I_j = i, X_j \in \tilde{A}_k) \leq \frac{\underline{c} n (L_1 h_n)^d \pi_n}{2}\Big). \qquad (14)\end{aligned}$$

Noting that $P(I_j = i, X_j \in \tilde{A}_k | Z^j) \geq \underline{c}(L_1 h_n)^d \pi_j$ for $1 \leq j \leq n$, we have by Lemma 2 that

$$P\Big(\sum_{j=1}^{n} I(I_j = i, X_j \in \tilde{A}_k) \leq \frac{\underline{c} n (L_1 h_n)^d \pi_n}{2}\Big) \leq \exp\Big(-\frac{3 \underline{c} n (L_1 h_n)^d \pi_n}{28}\Big). \qquad (15)$$

Therefore,

$$\begin{aligned}&P\Big(\inf_{x \in [0,1]^d} \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} K\Big(\frac{x - X_j}{h_n}\Big) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\Big) \\ \leq &\sum_{k=1}^{\tilde{B}} P\Big(\inf_{x \in \tilde{A}_k} \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} K\Big(\frac{x - X_j}{h_n}\Big) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\Big) \\ \leq &\tilde{B} \exp\Big(-\frac{3 \underline{c} n (L_1 h_n)^d \pi_n}{28}\Big),\end{aligned}$$

where the last inequality follows by (14) and (15). With the condition $nh^{2d}\pi_n^4/\log n \to \infty$, we immediately obtain (13) by Borel-Cantelli lemma.

By (13), it follows that (12) holds if

$$\sup_{x \in [0,1]^d} \Big| \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\Big(\frac{x - X_j}{h_n}\Big) \Big| = o(\pi_n). \qquad (16)$$

In the rest of the proof, we want to show that (16) holds. For each $n \geq n_0 l + 1$, we can partition the unit cube $[0,1]^d$ into B bins with bin length h_n such that $B \leq 1/h_n^d$. At each time point j, let $W_j = 1$ if arm i is pulled (i.e., $I_j = i$), and $W_j = 0$ otherwise. Then given $\epsilon > 0$,

$$\begin{aligned}
&P\Big(\sup_{x\in[0,1]^d}\Big|\frac{1}{M_{i,n+1}h_n^d}\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big| > \pi_n\epsilon\Big)\\
\leq& B\max_{1\leq k\leq B} P\Big(\sup_{x\in A_k}\Big|\frac{1}{M_{i,n+1}h_n^d}\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big| > \pi_n\epsilon\Big)\\
\leq& BP\Big(\frac{M_{i,n+1}}{n}\leq\frac{\pi_n}{2}\Big)\\
&\quad + B\max_{1\leq k\leq B} P\Big(\sup_{x\in A_k}\Big|\frac{1}{M_{i,n+1}h_n^d}\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big| > \pi_n\epsilon,\ \frac{M_{i,n+1}}{n}>\frac{\pi_n}{2}\Big)\\
\leq& BP\Big(\frac{M_{i,n+1}}{n}\leq\frac{\pi_n}{2}\Big) + B\max_{1\leq k\leq B} P\Big(\sup_{x\in A_k}\Big|\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big| > \frac{n\pi_n^2 h_n^d\epsilon}{2}\Big)\\
\leq& B\exp\Big(-\frac{3n\pi_n}{28}\Big) + B\max_{1\leq k\leq B} P\Big(\sup_{x\in A_k}\Big|\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big| > \frac{n\pi_n^2 h_n^d\epsilon}{2}\Big), \qquad (17)
\end{aligned}$$

where the last inequality follows by Lemma 2. Note that by Lemma 3,

$$\begin{aligned}
&P\Big(\sup_{x\in A_k}\Big|\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big| > \frac{n\pi_n^2 h_n^d\epsilon}{2}\Big)\\
\leq& P\Big(\sup_{x\in A_k}\sum_{j\in J_{i,n+1}}\varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big) > \frac{n\pi_n^2 h_n^d\epsilon}{2}\Big)\\
&\quad + P\Big(\sup_{x\in A_k}\sum_{j\in J_{i,n+1}}(-\varepsilon_j) K\Big(\frac{x-X_j}{h_n}\Big) > \frac{n\pi_n^2 h_n^d\epsilon}{2}\Big).\\
\leq& 2\exp\Big(-\frac{(\sqrt{2}-1)^2 n\pi_n^4 h_n^{2d}\epsilon^2}{32c_4^2v^2}\Big) + 2\exp\Big(-\frac{(\sqrt{2}-1)n\pi_n^2 h_n^d\epsilon}{8\sqrt{2}c_4c}\Big)\\
&\quad + 2\sum_{k=1}^{\infty}2^{kd}\exp\Big(-\frac{(\sqrt{2}-1)^2 2^k n\pi_n^4 h_n^{2d}\epsilon^2}{8\lambda^2v^2}\Big) + 2\sum_{k=1}^{\infty}2^{kd}\exp\Big(-\frac{(\sqrt{2}-1)2^{k/2} n\pi_n^2 h_n^d\epsilon}{4\sqrt{2}\lambda c}\Big). \qquad (18)
\end{aligned}$$

Thus, by (17), (18) and the condition that $nh_n^{2d}\pi_n^4/\log n \to \infty$, (16) is an immediate consequence of Borel-Cantelli lemma. This completes the proof of Theorem 1. □

B.2 Proofs of Theorem 2 and Corollary 2

Given $x \in [0,1]^d$, $1 \leq i \leq l$ and $n \geq n_0 l + 1$, define $G_{n+1}(x) = \{j : 1 \leq j \leq n, \|x - X_j\|_\infty \leq Lh_n\}$ and $G_{i,n+1}(x) = \{j : 1 \leq j \leq n, I_j = i, \|x - X_j\|_\infty \leq Lh_n\}$. Let $M_{n+1}(x)$ and $M_{i,n+1}(x)$ be the size of the sets $G_{n+1}(x)$ and $G_{i,n+1}(x)$, respectively. Then, the kernel method estimator $\hat{f}_{i,n+1}(x)$ satisfies the following lemma.

Lemma 4. *Suppose Assumptions 0, 1 and 4 are satisfied, and $\{\pi_n\}$ is a decreasing sequence. Given $x \in [0,1]^d$, $1 \le i \le l$ and $n \ge n_0 l + 1$, for every $\epsilon > \omega(Lh_n; f_i)$,*

$$\begin{aligned} P_{X^n}\big(|\hat{f}_{i,n+1}(x) - f_i(x)| \ge \epsilon\big) \le \exp\Big(-\frac{3M_{n+1}(x)\pi_n}{28}\Big) \\ + 4N\exp\Big(-\frac{c_5^2 M_{n+1}(x)\pi_n\big(\epsilon - \omega(Lh_n; f_i)\big)^2}{4c_4^2 v^2 + 4c_4 c\big(\epsilon - \omega(Lh_n; f_i)\big)}\Big), \end{aligned} \tag{19}$$

where $P_{X^n}(\cdot)$ denotes the conditional probability given design points $X^n = (X_1, X_2, \cdots, X_n)$.

Proof of Lemma 4. It is clear that if $M_{n+1}(x) = 0$, (19) trivially holds. Without loss of generality, assume $M_{n+1}(x) > 0$. Define the event $B_{i,n} = \{\frac{1}{M_{i,n+1}(x)}\sum_{j\in J_{i,n+1}} K(\frac{x-X_j}{h_n}) \ge c_5\}$. Note that

$$\begin{aligned} & P_{X^n}\big(|\hat{f}_{i,n+1}(x) - f_i(x)| \ge \epsilon\big) \\ \le & P_{X^n}\Big(\frac{M_{i,n+1}(x)}{M_{n+1}(x)} \le \frac{\pi_n}{2}\Big) + P_{X^n}\Big(|\hat{f}_{i,n+1}(x) - f_i(x)| \ge \epsilon,\ \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}\Big) \\ \le & \exp\Big(-\frac{3M_{n+1}(x)\pi_n}{28}\Big) + P_{X^n}\Big(|\hat{f}_{i,n+1}(x) - f_i(x)| \ge \epsilon,\ \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2},\ B_{i,n}\Big) \\ & \quad + P_{X^n}\Big(|\hat{f}_{i,n+1}(x) - f_i(x)| \ge \epsilon,\ \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2},\ B_{i,n}^c\Big), \\ =: & \exp\Big(-\frac{3M_{n+1}(x)\pi_n}{28}\Big) + A_1 + A_2, \end{aligned} \tag{20}$$

where the last inequality follows by Lemma 2. Under $B_{i,n}$, by Assumption 4, the definition of the modulus continuity and the same argument as (11), we have

$$\begin{aligned} |\hat{f}_{i,n+1}(x) - f_i(x)| &= \left| \frac{\sum_{j\in J_{i,n+1}} Y_{i,j} K\Big(\frac{x-X_j}{h_n}\Big)}{\sum_{j\in J_{i,n+1}} K\Big(\frac{x-X_j}{h_n}\Big)} \right| \\ &\le \omega(Lh_n; f_i) + \frac{1}{c_5 M_{i,n+1}(x)} \Big| \sum_{j\in G_{i,n+1}(x)} \varepsilon_j K\Big(\frac{x-X_j}{h_n}\Big)\Big|. \end{aligned}$$

Define $\tilde{\sigma}_t = \inf\{\tilde{n} : \sum_{j=1}^{\tilde{n}} I(I_j = i \text{ and } \|x - X_j\|_\infty \leq Lh_n) \geq t\}$, $t \geq 1$. Then, by the previous display, for every $\epsilon > \omega(Lh_n; f_i)$,

$$\begin{aligned}
A_1 \leq\ & P_{X^n}\Big(\Big|\sum_{j \in G_{i,n+1}(x)} \varepsilon_j K\Big(\frac{x - X_j}{h_n}\Big)\Big| \geq c_5 M_{i,n+1}(x)\big(\epsilon - \omega(Lh_n; f_i)\big),\ \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}\Big) \\
\leq\ & \sum_{\bar{n}=0}^{n} P_{X^n}\Big(\Big|\sum_{t=1}^{\bar{n}} \varepsilon_{\tilde{\sigma}_t} K\Big(\frac{x - X_{\tilde{\sigma}_t}}{h_n}\Big)\Big| \geq c_5 \bar{n}\big(\epsilon - \omega(Lh_n; f_i)\big),\ \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}, M_{i,n+1}(x) = \bar{n}\Big) \\
\leq\ & \sum_{\bar{n}=\lceil M_{n+1}(x)\pi_n/2\rceil}^{n} P_{X^n}\Big(\Big|\sum_{t=1}^{\bar{n}} \varepsilon_{\tilde{\sigma}_t} K\Big(\frac{x - X_{\tilde{\sigma}_t}}{h_n}\Big)\Big| \geq c_5 \bar{n}\big(\epsilon - \omega(Lh_n; f_i)\big)\Big) \\
\leq\ & \sum_{\bar{n}=\lceil M_{n+1}(x)\pi_n/2\rceil}^{n} 2\exp\Big(-\frac{\bar{n} c_5^2\big(\epsilon - \omega(Lh_n; f_i)\big)^2}{2c_4^2 v^2 + 2c_4 c\big(\epsilon - \omega(Lh_n; f_i)\big)}\Big) \\
\leq\ & 2N \exp\Big(-\frac{c_5^2 M_{n+1}(x)\pi_n\big(\epsilon - \omega(Lh_n; f_i)\big)^2}{4c_4^2 v^2 + 4c_4 c\big(\epsilon - \omega(Lh_n; f_i)\big)}\Big), \qquad (21)
\end{aligned}$$

where the last to second inequality follows by Lemma 1 and the upper boundedness of the kernel function. By an argument similar to the previous two displays (using the uniform kernel), it is not hard to obtain that

$$A_2 \leq 2N \exp\Big(-\frac{M_{n+1}(x)\pi_n\big(\epsilon - \omega(Lh_n; f_i)\big)^2}{4v^2 + 4c\big(\epsilon - \omega(Lh_n; f_i)\big)}\Big). \qquad (22)$$

Combining (20), (21), (22) and the fact that $0 < c_5 \leq 1 \leq c_4$, we complete the proof of Lemma 4. □

Proof of Theorem 2. Since $\hat{f}_{i^*(X_n),n}(X_n) \leq \hat{f}_{\hat{i}_n,n}(X_n)$, the regret accumulated after the initial forced sampling period satisfies that

$$\begin{aligned}
& \sum_{n=n_0 l+1}^{N} \big(f^*(X_n) - f_{I_n}(X_n)\big) \\
=\ & \sum_{n=n_0 l+1}^{N} \big(f_{i^*(X_n)}(X_n) - \hat{f}_{i^*(X_n),n}(X_n) + \hat{f}_{i^*(X_n),n}(X_n) - f_{\hat{i}_n}(X_n) + f_{\hat{i}_n}(X_n) - f_{I_n}(X_n)\big) \\
\leq\ & \sum_{n=n_0 l+1}^{N} \big(f_{i^*(X_n)}(X_n) - \hat{f}_{i^*(X_n),n}(X_n) + \hat{f}_{\hat{i}_n,n}(X_n) - f_{\hat{i}_n}(X_n) + f_{\hat{i}_n}(X_n) - f_{I_n}(X_n)\big) \\
\leq\ & \sum_{n=n_0 l+1}^{N} \Big(2 \sup_{1 \leq i \leq l} |\hat{f}_{i,n}(X_n) - f_i(X_n)| + AI(I_n \neq \hat{i}_n)\Big). \qquad (23)
\end{aligned}$$

It can be seen from (23) that the error upper bound consists of the estimation error regret and randomization error regret.

First, we find the upper bound of the estimation error regret. Given arm i, $n \geq n_0 l$ and $\epsilon > \omega(Lh_n; f_i)$,

$$\begin{aligned} & P\Big(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \epsilon\Big) \\ \leq & EP_{X^{n+1}}\Big(M_{n+1}(X_{n+1}) \leq \frac{\underline{c}n(2Lh_n)^d}{2}\Big) \\ & + EP_{X^{n+1}}\Big(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \epsilon, M_{n+1}(X_{n+1}) > \frac{\underline{c}n(2Lh_n)^d}{2}\Big). \end{aligned} \tag{24}$$

Since for every $x \in [0,1]^d$, $P(\|x - X_j\|_\infty \leq Lh_n) \geq \underline{c}(2Lh_n)^d$, $1 \leq j \leq n$, we have by the extended Bernstein's inequality that

$$P_{X^{n+1}}\Big(M_{n+1}(X_{n+1}) \leq \frac{\underline{c}n(2Lh_n)^d}{2}\Big) \leq \exp\Big(-\frac{3\underline{c}n(2Lh_n)^d}{28}\Big). \tag{25}$$

By Lemma 4,

$$\begin{aligned} & P_{X^{n+1}}\Big(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \epsilon, M_{n+1}(X_{n+1}) > \frac{\underline{c}n(2Lh_n)^d}{2}\Big) \\ \leq & \exp\Big(-\frac{3\underline{c}n(2Lh_n)^d\pi_n}{56}\Big) + 4N\exp\Big(-\frac{c_5^2\underline{c}n(2Lh_n)^d\pi_n\big(\epsilon - \omega(Lh_n; f_i)\big)^2}{8c_4^2v^2 + 8c_4c\big(\epsilon - \omega(Lh_n; f_i)\big)}\Big). \end{aligned} \tag{26}$$

Let

$$\tilde{\epsilon}_{i,n} = \omega(Lh_n; f_i) + \sqrt{\frac{16c_4^2v^2\log(8lN^2/\delta)}{c_5^2\underline{c}(2L)^d nh_n^d\pi_n}}.$$

Then, by (24), (25), (26) and the definition of n_δ in (3), it follows that for every $n \geq n_\delta$,

$$P\Big(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \tilde{\epsilon}_{i,n}\Big) \leq \frac{\delta}{4lN} + \frac{\delta}{4lN} + \frac{\delta}{2lN} = \frac{\delta}{lN},$$

which implies that

$$P\Big(\sum_{n=n_\delta+1}^{N} 2\sup_{1\leq i\leq l}|\hat{f}_{i,n}(X_n) - f_i(X_n)| \geq \sum_{n=n_\delta+1}^{N} 2\max_{1\leq i\leq l}\tilde{\epsilon}_{i,n-1}\Big) \leq \delta. \tag{27}$$

Next, we want to bound the randomization error regret. Given $\epsilon > 0$, since $P(I_n \neq \hat{i}_n) = (l-1)\pi_n$, we have by Hoeffding's inequality that

$$P\Big(A(\sum_{n=n_\delta+1}^{N} I(I_n \neq \hat{i}_n) - \sum_{n=n_\delta+1}^{N}(l-1)\pi_n) \geq \epsilon\Big) \leq \exp\Big(-\frac{2\epsilon^2}{NA^2}\Big).$$

Taking $\epsilon = A\sqrt{N/2}\log(1/\delta)$, we immediately get

$$P\Big(A\sum_{n=n_\delta+1}^{N} I(I_n \neq \hat{i}_n) \geq A\sum_{n=n_\delta+1}^{N}(l-1)\pi_n + A\sqrt{\frac{N}{2}}\log(\frac{1}{\delta})\Big) \leq \delta. \tag{28}$$

Then, (23), (27) and (28) together complete the proof of Theorem 2. □

B.3 Proof of Theorem 3

Lemma 5. *Under Assumption E and the proposed allocation strategy, for each arm i*

$$N_t < \infty \quad \textit{a.s. for all } t \geq 1.$$

Proof of Lemma 5. It suffices to check that

$$\sum_{j=n_0 l+1}^{\infty} I(I_j = i) = \infty \quad \text{a.s..} \tag{29}$$

Indeed, let $\mathcal{F}_n$, $n \geq 1$ be the σ-field generated by (Z^n, X_n, I_n). By the proposed allocation strategy, for all $j \geq n_0 l + 1$,

$$P(I_j = i | \mathcal{F}_{j-1}) \geq \pi_j.$$

By Assumption E, $\sum_{j=n_0 l+1}^{\infty} P(I_j = i|\mathcal{F}_{j-1}) = \infty$. Therefore, (29) is an immediate result of the Lévy's extension of the Borel-Cantelli lemmas (Williams, 1991, pp.124). □

Proof of Theorem 3. The key to the proof is to show $\|\hat{f}_{i,n} - f_i\|_\infty \to 0$ almost surely for $1 \leq i \leq l$ (Yang and Zhu, 2002, Theorem 1). Without loss of generality, assume Δ includes only two candidate procedures ($m = 2$). Given $1 \leq i \leq l$, assume that procedure $\delta_1 \in \Delta_{i1}$ and procedure $\delta_2 \in \Delta_{i2}$ (the case of $\delta_1, \delta_2 \in \Delta_{i1}$ is trivial). Since

$$\begin{aligned}\|\hat{f}_{i,n} - f_i\|_\infty &= \|W_{i,n,1}(\hat{f}_{i,n,1} - f_i) + W_{i,n,2}(\hat{f}_{i,n,2} - f_i)\|_\infty \\ &\leq W_{i,n,1}\|\hat{f}_{i,n,1} - f_i\|_\infty + W_{i,n,2}\|\hat{f}_{i,n,2} - f_i\|_\infty,\end{aligned}$$

it suffices to prove that $\frac{W_{i,n,1}}{W_{i,n,2}} \to \infty$ almost surely as $n \to \infty$.

As defined before, $N_t = \inf\{n : \sum_{j=n_0 l+1}^{n} \mathrm{I}(I_j = i) \geq t\}$, and let $\mathcal{F}_n$ be the σ-field generated by (Z^n, X_n, I_n). Then for any $t \geq 1$, N_t is a stopping time relative to $\{\mathcal{F}_n, n \geq 1\}$. By Lemma 5, $N_t < \infty$ a.s. for all $t \geq 1$. Therefore, the weights $W_{i,N_t,1}$, $W_{i,N_t,2}$ and the variance estimates $\hat{v}_{i,N_t,1}$, $\hat{v}_{i,N_t,2}$ and $\hat{v}_{i,N_t}$ for $t \geq 1$ are well-defined. By the allocation strategy, the weight associated with arm i is updated only after this arm is pulled. Consequently, we only need to show $\frac{W_{i,N_t,1}}{W_{i,N_t,2}} \to \infty$ almost surely as $t \to \infty$.

Note that for any $t \geq 1$,

$$\begin{aligned}
\frac{W_{i,N_{t+1},1}}{W_{i,N_{t+1},2}} =& \frac{W_{i,N_t,1}}{W_{i,N_t,2}} \times \frac{\hat{v}_{i,N_t,2}^{1/2}}{\hat{v}_{i,N_t,1}^{1/2}} \exp\left(-\frac{(\hat{f}_{i,N_t,1}(X_{N_t}) - Y_{i,N_t})^2 - (\hat{f}_{i,N_t,2}(X_{N_t}) - Y_{i,N_t})^2}{2\hat{v}_{i,N_t}}\right) \\
=& \frac{W_{i,N_t,1}}{W_{i,N_t,2}} \times \frac{\hat{v}_{i,N_t,2}^{1/2}}{\hat{v}_{i,N_t,1}^{1/2}} \\
& \times \exp\left(-\frac{(\hat{f}_{i,N_t,1}(X_{N_t}) - f_i(X_{N_t}) - \varepsilon_{N_t})^2 - (\hat{f}_{i,N_t,2}(X_{N_t}) - f_i(X_{N_t}) - \varepsilon_{N_t})^2}{2\hat{v}_{i,N_t}}\right) \\
=& \frac{W_{i,N_t,1}}{W_{i,N_t,2}} \times \frac{\hat{v}_{i,N_t,2}^{1/2}}{\hat{v}_{i,N_t,1}^{1/2}} \exp\left(\frac{(\hat{f}_{i,N_t,2}(X_{N_t}) - f_i(X_{N_t}))^2 - (\hat{f}_{i,N_t,1}(X_{N_t}) - f_i(X_{N_t}))^2}{2\hat{v}_{i,N_t}}\right) \\
& \times \exp\left(\frac{\varepsilon_{N_t}(\hat{f}_{i,N_t,1}(X_{N_t}) - \hat{f}_{i,N_t,2}(X_{N_t}))}{\hat{v}_{i,N_t}}\right) \\
=& \frac{W_{i,N_t,1}}{W_{i,N_t,2}} \times \frac{\hat{v}_{i,N_t,2}^{1/2}}{\hat{v}_{i,N_t,1}^{1/2}} \exp(T_{1t} + T_{2t}),
\end{aligned}$$

where

$$T_{1t} = \frac{(\hat{f}_{i,N_t,2}(X_{N_t}) - f_i(X_{N_t}))^2 - (\hat{f}_{i,N_t,1}(X_{N_t}) - f_i(X_{N_t}))^2}{2\hat{v}_{i,N_t}},$$

and

$$T_{2t} = \frac{\varepsilon_{N_t}(\hat{f}_{i,N_t,1}(X_{N_t}) - \hat{f}_{i,N_t,2}(X_{N_t}))}{\hat{v}_{i,N_t}}.$$

Thus, for each $T \geq 1$,

$$\frac{W_{i,N_{T+1},1}}{W_{i,N_{T+1},2}} = \left(\prod_{t=1}^{T} \frac{\hat{v}_{i,N_t,2}^{1/2}}{\hat{v}_{i,N_t,1}^{1/2}}\right) \exp\left(\sum_{t=1}^{T} T_{1t} + \sum_{t=1}^{T} T_{2t}\right). \tag{30}$$

Then define $\xi_t = \varepsilon_{N_t}(\hat{f}_{i,N_t,1}(X_{N_t}) - f_i(X_{N_t}))$ and $\xi'_t = \varepsilon_{N_t}(\hat{f}_{i,N_t,2}(X_{N_t}) - f_i(X_{N_t}))$. Since $E(\varepsilon_{N_t}|\mathcal{F}_{N_t}) = 0$, it follows by Assumption C, Assumption 0 and Lemma 1 that for every $\tau > 0$ and every $T \geq 1$,

$$P(\sum_{t=1}^{T} \xi_t > T\tau) < \exp\Big(-\frac{T\tau^2}{2c_2^2(v^2 + c\tau/c_2)}\Big).$$

Replacing τ by $\frac{(\log T)^b}{\sqrt{T}}\tau'$, we obtain

$$\sum_{t=1}^{T} \xi_t = o(\sqrt{T}(\log T)^b)$$

almost surely by Borel-Cantelli lemma. By the same argument, $\sum_{t=1}^{T} \xi'_t = o(\sqrt{T}(\log T)^b)$ almost surely. Note that for each $T \geq 1$,

$$\begin{aligned}
\hat{v}_{i,N_{T+1},1} =& \frac{\sum_{t=1}^{T}(\hat{f}_{i,N_t,1}(X_{N_t}) - Y_{i,N_t})^2}{T} \\
=& \frac{\sum_{t=1}^{T}(\hat{f}_{i,N_t,1}(X_{N_t}) - f_i(X_{N_t}) - \varepsilon_{N_t})^2}{T} \\
=& \frac{\sum_{t=1}^{T}(\hat{f}_{i,N_t,1}(X_{N_t}) - f_i(X_{N_t}))^2}{T} + \frac{\sum_{t=1}^{T}\varepsilon_{N_t}^2}{T} - \frac{2\sum_{t=1}^{T}\xi_t}{T}.
\end{aligned}$$

Similarly, for each $T \geq 1$

$$\hat{v}_{i,N_{T+1},2} = \frac{\sum_{t=1}^{T}(\hat{f}_{i,N_t,2}(X_{N_t}) - f_i(X_{N_t}))^2}{T} + \frac{\sum_{t=1}^{T}\varepsilon_{N_t}^2}{T} - \frac{2\sum_{t=1}^{T}\xi'_t}{T}.$$

By Assumption A and the previous two equations, we obtain that

$$\hat{v}_{i,N_t,1} < \hat{v}_{i,N_t,2} \tag{31}$$

almost surely for large enough t.

The boundedness of $\{\hat{v}_{i,N_t},\ t \geq 1\}$ as implied in Assumption D enables us to apply Lemma 1 again to obtain that

$$\sum_{t=1}^{T} T_{2t} = o(\sqrt{T}(\log T)^b), \tag{32}$$

almost surely. By (31), (32) and Assumption A, we conclude from (30) that

$$\frac{W_{i,N_{T+1},1}}{W_{i,N_{T+1},2}} \to \infty \quad \text{a.s. as } T \to \infty.$$

This completes the proof of Theorem 3. □

B.4 Proof of Theorem 4

Proof of Theorem 4. First, note that

$$\begin{aligned}
R_N(\eta) =& \sum_{n=1}^{N}\Big(f^*(X_n) - f_{I_n}(X_n)\Big)I(I_n = \tilde{i}_n) + \sum_{n=1}^{N}\Big(f^*(X_n) - f_{I_n}(X_n)\Big)I(I_n \neq \tilde{i}_n) \\
\leq& \sum_{n=1}^{N}\big(f^*(X_n) - f_{I_n}(X_n)\big)I(I_n = \tilde{i}_n,\, \tilde{i}_n \neq i^*(X_n)) + \sum_{n=1}^{N} AI(I_n \neq \tilde{i}_n) \\
=& \sum_{i=1}^{l}\sum_{n=1}^{N}\big(f^*(X_n) - f_{I_n}(X_n)\big)I\big(I_n = i,\, \tilde{i}_n \neq i^*(X_n), \tilde{i}_n = i\big) + \sum_{n=1}^{N} AI(I_n \neq \tilde{i}_n) \\
=:& \sum_{i=1}^{l}\sum_{n=1}^{N} R_{i,n} + R_{N,2}.
\end{aligned} \tag{33}$$

Then we partition the domain into $1/h^d$ bins with bin width h and denote the set of these bins by $\mathcal{B}$.

Given bin $B \in \mathcal{B}$, define $R_{i,N,B} = \sum_{n=1}^{N} R_{i,n} I(X_n \in B)$. Let x_B be the center point in B. Define the (nearly) optimal arm $\bar{i} = \bar{i}_B = \operatorname{argmax}_{1 \leq i \leq l} f_i(x_B)$ and $\Delta_{i,B} = f_{\bar{i}}(x_B) - f_i(x_B)$. Let $S_{i,B} = \{n : ln_0 + 1 \leq n \leq N, I_n = i, \tilde{\bar{i}}_n \neq i^*(X_n), \tilde{i}_n = i, X_n \in B\}$. Define $N_{i,B} = \max S_{i,B}$ if $S_{i,B} \neq \varnothing$ and $N_{i,B} = 0$ if $S_{i,B} = \varnothing$. Given $N_{i,B} = \tilde{n}$, let $\sigma_t = \sigma_{i,t,\tilde{n}} = \min\{n : \sum_{j=1}^{n} I(I_j = i, K(\frac{X_j - X_{\tilde{n}}}{h}) \neq 0) \geq t\}$ be the earliest time point where $\{I_j = i, K(\frac{X_j - X_{\tilde{n}}}{h}) \neq 0\}$ happens t times. Define $\tau_i = \tau_{i,\tilde{n}} = \max\{t : \sigma_{i,t,\tilde{n}} < \tilde{n}\}$, which is the number of times $\{I_j = i, K(\frac{X_j - X_{\tilde{n}}}{h}) \neq 0\}$ happens before the time point $\tilde{n}$. Similarly, for the (nearly) optimal arm $\bar{i}$, define $\eta_t = \min\{n : \sum_{j=1}^{n} I(I_j = \bar{i}, K(\frac{X_j - X_{\tilde{n}}}{h}) \neq 0) \geq t\}$ and $\bar{\tau} = \max\{t : \eta_t < \tilde{n}\}$. Then, if $N_{i,B} = \tilde{n} \neq 0$ and $\tau_i \geq 1$, by the kernel-UCB algorithm, we have $\hat{f}_{i,\tilde{n}}(X_{\tilde{n}}) + U_{i,\tilde{n}}(X_{\tilde{n}}) \geq \hat{f}_{\bar{i},\tilde{n}}(X_{\tilde{n}}) + U_{\bar{i},\tilde{n}}(X_{\tilde{n}})$, that is,

$$\frac{\sum_{t=1}^{\tau_i} Y_{i,\sigma_t}}{\tau_i} + \tilde{c}\sqrt{\frac{\log N}{\tau_i}} \geq \frac{\sum_{t=1}^{\bar{\tau}} Y_{\bar{i},\eta_t}}{\bar{\tau}} + \tilde{c}\sqrt{\frac{\log N}{\bar{\tau}}}, \tag{34}$$

Note that (34) implies at least one of the following three events occurs:

$$\begin{aligned}
G_B &=: \Big\{\frac{\sum_{t=1}^{\tau_i} Y_{i,\sigma_t}}{\tau_i} - \frac{\sum_{t=1}^{\tau_i} f_i(X_{\sigma_t})}{\tau_i} > \tilde{c}\sqrt{\frac{\log N}{\tau_i}}\Big\},\\
F_B &=: \Big\{\frac{\sum_{t=1}^{\bar{\tau}} Y_{\bar{i},\eta_t}}{\bar{\tau}} - \frac{\sum_{t=1}^{\bar{\tau}} f_{\bar{i}}(X_{\eta_t})}{\bar{\tau}} < -\tilde{c}\sqrt{\frac{\log N}{\bar{\tau}}}\Big\}, \text{ or}\\
H_B &=: \Big\{\frac{\sum_{t=1}^{\tau_i} f_i(X_{\sigma_t})}{\tau_i} + 2\tilde{c}\sqrt{\frac{\log N}{\tau_i}} > \frac{\sum_{t=1}^{\bar{\tau}} f_{\bar{i}}(X_{\eta_t})}{\bar{\tau}}\Big\}.
\end{aligned}$$

Since $\|f_i(X_{\sigma_t}) - f_i(x_B)\|_\infty \leq \rho h^\kappa$ and $\|f_{\bar{i}}(X_{\eta_t}) - f_{\bar{i}}(x_B)\|_\infty \leq \rho h^\kappa$,

$$\begin{aligned}
H_B &\Rightarrow f_i(x_B) + \rho h^\kappa + 2\tilde{c}\sqrt{\frac{\log N}{\tau_i}} > f_{\bar{i}}(x_B) - \rho h^\kappa\\
&\Rightarrow 2\tilde{c}\sqrt{\frac{\log N}{\tau_i}} > \Delta_{i,B} - 2\rho h^\kappa\\
&\Rightarrow \{\Delta_{i,B} \leq 4\rho h^\kappa\} \text{ or } \{\Delta_{i,B} > 4\rho h^\kappa,\, \tau_i < \frac{16\tilde{c}^2}{\Delta_{i,B}^2}\log N\}.
\end{aligned} \tag{35}$$

By Lemma 1,

$$\begin{aligned}
P(G_B,\, N_{i,B} \neq 0,\, \tau_i > \log N) &\leq N^2 \exp\Big(-\frac{\tilde{c}^2 \log N}{2(v^2 + c\tilde{c})}\Big)\\
&\leq \frac{1}{N},
\end{aligned} \tag{36}$$

where the last inequality holds since $\tilde{c} > \max\{2\sqrt{3}v, 12c\}$.

Similarly, we can show that

$$P(F_B,\, N_{i,B} \neq 0,\, \tau_i > \log N) \leq \frac{1}{N}. \tag{37}$$

Note that

$$\begin{aligned}
R_{i,N,B} \leq\ & R_{i,N,B}I(N_{i,B}=0) + R_{i,N,B}I(\tau_i \leq \log N,\ N_{i,B} \neq 0) + \\
& R_{i,N,B}I(\tau_i > \log N, N_{i,B} \neq 0, G_B) + R_{i,N,B}I(\tau_i > \log N, N_{i,B} \neq 0, F_B) + \\
& R_{i,N,B}I(\tau_i > \log N, N_{i,B} \neq 0, H_B) \\
\leq\ & R_{i,N,B}I(N_{i,B}=0) + A\log N + ANI(\tau_i > \log N, N_{i,B} \neq 0, G_B) + \\
& ANI(\tau_i > \log N, N_{i,B} \neq 0, F_B) + 6\rho h^{\kappa}\tau_{i,B} + (\frac{3}{2}\Delta_{i,B})(\frac{16\tilde{c}^2}{\Delta_{i,B}^2}\log N)I(\Delta_{i,B} > 4\rho h^{\kappa}),
\end{aligned}$$

where the last inequality follows by (35), and $\tau_{i,B} = \sum_{n=1}^{N} I(I_n = i, X_n \in B)$. Then by (36), (37) and the definition of $N_{i,B}$,

$$\begin{aligned}
E\Big(\sum_{i=1}^{l}\sum_{n=1}^{N} R_{i,n}\Big) &= \sum_{i=1}^{l}\sum_{B\in\mathcal{B}} E(R_{i,n,B}) \\
&\leq Aln_0 + Alh^{-d}\log N + 2A + 6\rho N h^{\kappa} + \frac{6\tilde{c}^2 l}{\rho h^{\kappa+d}}\log N. \qquad (38)
\end{aligned}$$

Also, taking $\delta = 1/N$, we have by (28) that

$$\begin{aligned}
E(R_{N,2}) &\leq AN\delta + Al\sum_{n=1}^{N}\pi_n + A\sqrt{\frac{N}{2}}\log(\frac{1}{\delta}) \\
&\leq A + Al\sum_{n=1}^{N}\pi_n + A\sqrt{\frac{N}{2}}\log N. \qquad (39)
\end{aligned}$$

By (33), (38), (39) and our choice of n_0, h and π_n, we obtain (7).

□

Appendix C. Additional Numerical Results

Under the same settings of Section 7 with the Yahoo! data set, we implement additional algorithms as follows.

Simple-SIR-kernel : This algorithm is the same as the SIR-kernel algorithm described in Section 7 except that the dimension reduction matrix is estimated using only the data collected during the forced sampling stage. That is, for every $n > ln_0$, the Nadaraya-Waston estimator of $f_i(x)$ shown in (5) is modified to be

$$\hat{f}_{i,n}(x) = \frac{\sum_{j\in J_{i,n}} Y_{i,j}K_i\left(\frac{\hat{B}^{*T}_{i,ln_0}x^* - \hat{B}^{*T}_{i,ln_0}X^*_j}{h_{n-1}}\right)}{\sum_{j\in J_{i,n}} K_i\left(\frac{\hat{B}^{*T}_{i,ln_0}x^* - \hat{B}^{*T}_{i,ln_0}X^*_j}{h_{n-1}}\right)}, \qquad (40)$$

where the forced sampling size for each arm is $n_0 = 1000$ and the bandwidth is $h_n = n^{-1/10}$.

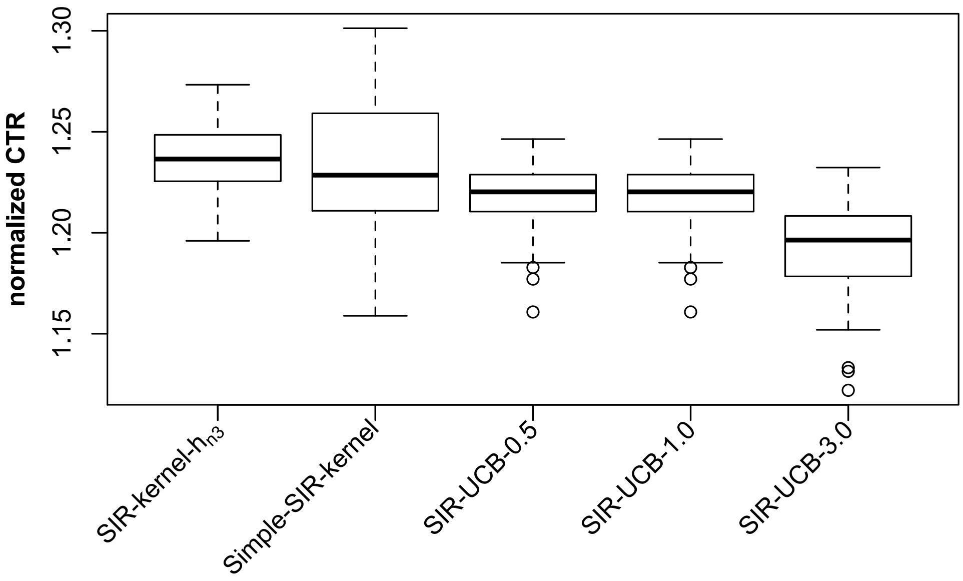

Figure 3: Boxplots of normalized CTRs of various algorithms on the news article recommendation data set. Algorithms include (from left to right): SIR-kernel (h_{n3}), Simple-SIR-kernel, SIR-UCB ($\tilde{c}_0 = 0.5$), SIR-UCB ($\tilde{c}_0 = 1.0$) and SIR-UCB ($\tilde{c}_0 = 3.0$). CTRs are normalized with respect to the random algorithm.

SIR-UCB : This algorithm modifies the kernel estimation based UCB algorithm described in Appendix A to handle covariates with high dimensions. Rather than using the original covariates, we apply the SIR method to estimate the dimension reduction matrices and use them to transform covariates to lower dimensions. These transformed covariates are subsequently applied to compute the kernel estimation based UCB index. That is, at each time point n after the forced sampling stage, we pull the arm with the highest UCB index $\hat{f}_{i,n}(X_n) + U^*_{i,n}(X_n)$, where $\hat{f}_{i,n}(X_n)$ is defined in (5) and

$$U^*_{i,n}(x) = \frac{\tilde{c}_0 \sqrt{\sum_{j \in J_{i,n}} K^2\Big(\frac{\hat{B}^{*T}_{i,n} x^* - \hat{B}^{*T}_{i,n} X^*_j}{h_{n-1}}\Big)}}{\sum_{j \in J_{i,n}} K\Big(\frac{\hat{B}^{*T}_{i,n} x^* - \hat{B}^{*T}_{i,n} X^*_j}{h_{n-1}}\Big)},$$

with $\hat{B}^*_{i,n}$, x^* and X^*_j defined in Section 4.3. We set $\tilde{c}_0 = 0.5, 1$ or 3 and $h_n = n^{-1/10}$.

The algorithms above are evaluated in the same manner as is described in Section 7, and the resulting normalized CTRs are summarized in the boxplots in Fig. 3. Although the averaged CTRs of the simple-SIR-kernel appears to be similar to SIR-kernel, the variation of the CTRs clearly enlarges as we use only the forced sampling stage to estimate the dimension reduction matrices. The SIR-UCB algorithm does not show significant improvement over the SIR-kernel algorithm either.

Remark 1. *Because of the curse of dimensionality, the kernel estimation in Section 4.1 cannot be directly applied to the Yahoo! data set. As described in Section 4.3, one way to address this issue is to assume that for each arm i* ($i = 1, 2, \cdots, l$)*, there exists a dimension reduction matrix* B_i *and a function* $g(\cdot)$ *such that* $\tilde{x}_i = B_i^T x$ *becomes lower dimensional covariate and* $f_i(x) = g_i(\tilde{x}_i)$*. If* B_i *(or more precisely,* $\texttt{span}(B_i)$*) is known, we can simply work with the lower dimensional covariates (which can be different for different arms), and the kernel estimation algorithm in Section 4.1 still applies. Indeed, we note that if* B_i *is known, the consistency and finite-time regret analysis (with rate in accordance with the lower dimension) can still be established in a way similar to that of Theorem 1 and Corollary 2.*

In practice, since the B_i *is unknown, it is natural to estimate it by introducing a dimension reduction method like SIR. The theoretical implications of the dimension reduction procedure is not yet clear to us. To provide some numerical guidance on how to apply SIR, we explored two different ways of estimating* B_i *using the Yahoo! data. One is the SIR-kernel algorithm, where the estimator for* B_i *gets updated as more and more data is collected throughout the total time horizon. Alternatively, we considered here the Simple-SIR-kernel algorithm, where only data from the initial forced sampling stage is used to generate a consistent estimator for* B_i *(Zhu et al., 1996); Subsequently, with the lower-dimensional covariates from a fixed dimension reduction matrix, the kernel estimation can be applied to the remaining time period. Our numerical result favors the former way of applying SIR.*

References

Y. Abbasi-Yadkori. Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, Department of Computing Science, University of Alberta, 2009.

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th Conference on Neural Information Processing Systems*, 2011.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37:1591–1646, 2009.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2003.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

G. Bartók and C. Szepesvári. Partial monitoring with side information. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, 2012.

D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall, New York, 1985.

L. Birgé and Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

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

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Thoery*, 59(11):7711–7717, 2013.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games.* Cambridge University Press, Cambridge, UK, 2006.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Proceedings of the 25th Conference on Neural Information Processing Systems*, pages 2249–2257, 2011.

X. Chen, C. Zou, and R. D. Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38:3696–3723, 2010.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

R. D. Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22:1–26, 2007.

R. D. Cook, L. M. Forzani, and D. R. Tomassi. LDR: A package for likelihood-based sufficient dimension reduction. *Journal of Statistical Software*, 39, 2011.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 355–366, 2008.

L. P. Devroye. The uniform convergence of the Nadaraya-Watson regression function estimate. *The Canadian Journal of Statistics*, 6:179–191, 1978.

M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence*, 2011.

J. C. Gittins. *Multi-Armed Bandit Allocation Indices.* Wiley, New York, 1989.

A. Goldenshluger and A. Zeevi. Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability*, 19:1603–1633, 2009.

A. Goldenshluger and A. Zeevi. A linear response bandit problem. *Stochastic Systems*, 3: 230–261, 2013.

B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748, 2008.

W. Härdle and S. Luckhaus. Uniform consistency of a class of regression function estimators. *The Annals of Statistics*, 12:612–623, 1984.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 18th Conference on Neural Information Processing Systems*, 2004.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Symposium on Theory of Computing*, 2007.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 21th Conference on Neural Information Processing Systems*, 2007.

K.-C. Li. Sliced inverse regression for dimension reduction, with discussions. *Journal of the American Statistical Association*, 86:316–342, 1991.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International World Wide Web Conference*, 2010.

T. Lu, D. Pál, and M. Pál. Showing relevant ads via Lipschitz context multi-armed bandits. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2010.

O.-A. Maillard and R. Munos. Adaptive bandits: Towards the best history-dependent strategy. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

B. C. May, N. Korda, A. Lee, and D. S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13:2069–2106, 2012.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41:693–721, 2013.

W. Qian and Y. Yang. Randomized allocation with dimension reduction in a bandit problem with covariates. In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 1537–1541, 2012.

P. Rigollet and A. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27:558–575, 2012.

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In *Proceedings of the 23rd International Conference on Learning Theory*, pages 54–66. Omnipress, 2010.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1954.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35:395–411, 2010.

A. Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 679–702, 2011.

Z. Wang, S. Paterlini, F. Gao, and Y. Yang. Adaptive minimax regression estimation over sparse l_q-hulls. *Journal of Machine Learning Research*, 15:1675–1711, 2014.

X. Wei and Y. Yang. Robust forecast combination. *Journal of Econometrics*, 22:1021–1040, 2012.

D. Williams. *Probability with Martingales.* Cambridge University Press, Cambridge, UK, 1991.

M. Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74:799–806, 1979.

Yahoo! Academic Relations. Yahoo! front page today module user click log dataset (version 1.0). 2011. Available from http://webscope.sandbox.yahoo.com.

Y. Yang. Combining forecasting procedures: Some theoretical results. *Econometric Theory*, 20:176–222, 2004.

Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30:100–121, 2002.

Li-Xing Zhu, Kai-Tai Fang, et al. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.

H. Zou and Y. Yang. Combining time series models for forecasting. *International Journal of Forecasting*, 20:69–84, 2004.