Electronic Journal of Statistics
Vol. 10 (2016) 242–270
ISSN: 1935-7524
DOI: 10.1214/15-EJS1104

Randomized allocation with arm elimination in a bandit problem with covariates

Wei Qian

School of Mathematical Sciences, Rochester Institute of Technology
Rochester, NY 14623
e-mail: wxqsma@rit.edu

and

Yuhong Yang

School of Statistics, University of Minnesota
Minneapolis, MN 55455
e-mail: yyang@stat.umn.edu

Abstract: Motivated by applications in personalized web services and clinical research, we consider a multi-armed bandit problem in a setting where the mean reward of each arm is associated with some covariates. A multi-stage randomized allocation with arm elimination algorithm is proposed to combine the flexibility in reward function modeling and a theoretical guarantee of a cumulative regret minimax rate. When the function smoothness parameter is unknown, the algorithm is equipped with a histogram estimation based smoothness parameter selector using Lepski's method, and is shown to maintain the regret minimax rate up to a logarithmic factor under a "self-similarity" condition.

MSC 2010 subject classifications: Primary 62G08; secondary 62L05.
Keywords and phrases: Contextual bandit problem, MABC, nonparametric bandit, adaptive estimation, regret bound.

Received October 2014.

1. Introduction

The multi-armed bandit problem is an optimization game with promising applications in, e.g., web services and clinical research. Under a prototypical framework, a bandit problem consists of several gambling machines, and the underlying reward distribution of each machine is unknown to the game player. Each time, the player can pull only one of the machine arms to receive reward. Given a finite number of times to play the machines, the goal is to devise a sequential arm allocation algorithm to maximize the cumulative reward, and equivalently, to minimize the cumulative regret (the shortfall of the reward of the algorithm compared to an oracle). A balance between exploration and exploitation is usually required for a bandit problem algorithm.

The standard setting of a bandit problem assumes that the reward response of each arm is "homogeneous" with no available covariates. Since the seminal work of Robbins (1954), the standard bandit problem is studied extensively, the representative early work of which includes Lai and Robbins (1985), Berry and Fristedt (1985), Gittins (1989) and Auer, Cesa-Bianchi and Fischer (2002). See also Cesa-Bianchi and Lugosi (2006) and Bubeck and Cesa-Bianchi (2012) for recent reviews of its various extensions. The "homogeneity" assumption of the standard setting, however, can be too restrictive in real applications. An increasingly popular but a much less studied setting is to assume that the mean reward is associated with some covariates, that is, the game player is given a d-dimensional covariate $x \in R^d$ as additional information before deciding which arm to pull, and the expected reward of a bandit arm given covariate x takes a functional form $f(x)$. Such variant of bandit problem is called **m**ulti-**a**rmed **b**andit problem with **c**ovariates, or MABC for its abbreviation.

The MABC problem first appears under a parametric framework in Woodroofe (1979). Attracted by promising applications in personalized web and medical services, more and more attentions are directed to the MABC problem in recent years. For example, with settings more flexible than that of Woodroofe (1979), a linear response bandit problem is recently studied under a minimax framework with margin conditions (Goldenshluger and Zeevi, 2009; Goldenshluger and Zeevi, 2013 and references therein). The well-known upper confidence bound (UCB) type algorithms are also extended to linear parametric settings, and are studied empirically in, e.g., Li et al. (2010).

The MABC problem from a nonparametric perspective is initiated by Yang and Zhu (2002). They propose a randomized allocation algorithm with histogram and K-nearest neighbor methods, the cumulative reward of which is shown to be asymptotically equivalent to that of an oracle. Although it is a very flexible and often effective algorithm, a finite-time regret analysis by Qian and Yang (2016) suggests that it may converge sub-optimally in terms of the minimax rate of the regret established by Rigollet and Zeevi (2010) due to its over-exploration in the randomization process. Perchet and Rigollet (2013) propose algorithms with an important step of arm elimination that originally appeared in a standard bandit problem setting (Even-Dar, Mannor and Mansour, 2006). They provide more rigorous and fine-tuned arguments for the standard setting, and further obtain performance bounds for their arm elimination algorithms devised to deal with the MABC problem. In particular, by a dyadic binning process, their adaptively binned successive elimination (ABSE) algorithm achieves the regret minimax rate, and is adaptive to a margin condition. The aforementioned nonparametric MABC algorithms, however, all assume a known Hölder smoothness condition on the mean reward functions. It is of interest to find algorithms that are adaptive to both the smoothness and the margin conditions.

Other settings of MABC problems have been studied in, e.g., Langford and Zhang (2007) and Dudik et al. (2011), where algorithms are designed to target the performance of the best arm-pulling policy among a class of finitely many candidate policies. May et al. (2012) study MABC from a Bayesian perspective.

In addition, differently from the MABC problem, a related setting considers the arm space (with possibly infinitely many arms) instead of the covariate space (see, e.g., Dani, Hayes and Kakade, 2008; Rusmevichientong and Tsitsiklis, 2010; Auer, Ortner and Szepesvári, 2007; Kleinberg, Slivkins and Upfal, 2007). The bandit problem that considers the joint covariate and arm space is studied in Lu, Pál and Pál (2010) and Slivkins (2011).

In this article, we follow the line of nonparametric MABC problem. The primary task is to address the question of whether we can achieve a near minimax optimal regret upper bound without the prior knowledge of the smoothness parameter. Our solution to this question is closely related to the adaptive nonparametric estimation technique pioneered by Lepski (1990). The "Lepski-type" method is recently studied in Giné and Nickl (2010), Hoffmann and Nickl (2011) and Bull (2012), and a "self-similarity" condition is used for establishing the adaptive confidence bands in both density estimation and regression problems. As the most important contribution of this work, we propose the strategy of integrating the Lepski's method with a nonparametric MABC algorithm, and show that under a "self-similarity" condition, the resulting cumulative regret can adaptively achieve the minimax rate up to a logarithmic factor. In particular, the ABSE algorithm (Perchet and Rigollet, 2013) can be used for adaptively achieving a near minimax rate when equipped with the Lepski-type smoothness parameter selector (see Remark 5.1).

It is noted that the regret minimization in the MABC problem differs from the usual purpose of nonparametric function estimation, but shares the difficulties involved in establishing adaptive confidence bands. A more detailed discussion regarding the connection of the adaptive nonparametric estimation with the MABC problem is deferred to section 6.

We present the proposed strategy using a nonparametric MABC algorithm called randomized allocation with arm elimination (or RAAE for abbreviation). Motivated by the observation in Qian and Yang (2016) that using randomized allocation strategy alone may give sub-optimal rate for the cumulative regret, the RAAE algorithm is proposed to embed the *key* arm-elimination technique developed in Perchet and Rigollet (2013) into the randomized allocation and can be shown to achieve the same minimax rate as the ABSE (with known smoothness). In our view, the feature of randomized allocation procedure (in addition to arm elimination) is practically useful because it provides a user with additional flexibility of applying a regression modeling method (e.g., kernel regression) for each arm to further exploit the response-covariate association. The practical implications of the randomized allocation step in RAAE are discussed in Remark 3.1 and are numerically illustrated in Appendix B with simulation examples.

The remainder of this article is organized as follows. The MABC problem setup is introduced in section 2. The RAAE algorithm and the integrated smoothness parameter selector are described in sections 3 and 4, respectively. The finite-time regret analysis is done in section 5. A final discussion is given in section 6. The technical lemmas and proofs are given in Appendix A and a simulation experiment regarding the randomized allocation in RAAE is shown in Appendix B.

2. Problem setup

Consider an l-armed bandit problem ($l \geq 2$) and suppose the covariates take values in the hypercube $[0,1]^d$. Let $f_i(x)$ denote the (conditional) mean reward function of an arm i ($1 \leq i \leq l$) given a covariate x. We model the observed reward as $f_i(x)+\varepsilon$, where ε is the random error with mean 0. The mean reward functions and the random error distributions are unknown.

Let $\{X_n, n \geq 1\}$ be a sequence of independent covariates with an unknown probability distribution P_X supported in $[0,1]^d$. Given any time point n ($n \geq 1$), let $Y_{i,n}$ denote the observed reward from pulling arm i ($1 \leq i \leq l$), and let I_n denote the arm chosen by a sequential allocation rule η. The MABC problem works as follows at each time point n. First, the covariate X_n is observed. Based on X_n and the previous observations $(X_j, I_j, Y_{I_j,j})$, $1 \leq j \leq n-1$, the allocation rule η is subsequently applied to decide which arm to pull. Then, the game player pulls the chosen arm I_n and receives the corresponding reward $Y_{I_n,n}$. The received reward is generated by $Y_{I_n,n} = f_{I_n}(X_n) + \varepsilon_n$, where ε_n is the random error, and (X_n, ε_n) is independent of the previous observations. We assume the covariate and the random error satisfy the following conditions.

Assumption 2.1. *The design distribution of the covariate 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 $\underline{c} \leq \overline{c}$.*

Assumption 2.2. *The errors satisfy a (conditional) moment condition that there exist positive constants v and c such that for all integers $k \geq 2$ and $n \geq 1$,*

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

almost surely.

Assumption 2.1 is used by the smoothness parameter selector to ensure that the histogram estimation is close to the true reward function uniformly. Assumption 2.2 is a (conditional) moment assumption known as refined Bernstein condition (e.g., Birgé and Massart, 1998). Note that under Assumption 2.2, the random error can be dependent on the covariate, and is not necessarily bounded. When the response is bounded (e.g., binary), Assumption 2.2 trivially holds. In general, it is satisfied if the error has a finite exponential moment, and thus allows error distributions with tails heavier than normal distribution.

Define, at given x, $i^*(x) = \operatorname{argmax}_{1 \leq i \leq l} f_i(x)$ to be the best arm, $f^*(x) = f_{i^*(x)}(x)$ to be the best mean reward, and let $w = \sup_{1 \leq i \leq l} \sup_{x \in [0,1]^d} (f^*(x) - f_i(x))$. We measure the performance of an allocation rule η using cumulative regret $R_n(\eta)$, per-round regret $r_n(\eta)$ and inferior sampling rate $q_n(\eta)$, which are defined by

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

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

and

$$q_n(\eta) = \frac{1}{n}\sum_{j=1}^{n} I\big(I_j \neq i^*(X_j)\big),$$

respectively.

Next, we introduce a Hölder smoothness condition and a margin condition, both of which have been studied in the context of nonparametric estimation (Audibert and Tsybakov, 2005; Audibert and Tsybakov, 2007) and classification (Mammen and Tsybakov, 1999; Tsybakov, 2004). Let $\|\cdot\|$ be the sup-norm on a d-dimensional vector. Suppose κ_* and κ^* are two known constants satisfying $0 < \kappa_* < \kappa^* \leq 1$. Given $\kappa \in [\kappa_*, \kappa^*]$ and $\rho > 0$, define $\Sigma(\kappa, \rho)$ to be the class of functions that satisfies the following Hölder smoothness condition: for $f \in \Sigma(\kappa, \rho)$,

$$|f(x_1) - f(x_2)| \leq \rho\|x_1 - x_2\|^\kappa,$$

for every $x_1, x_2 \in [0,1]^d$. As mentioned in the introduction, to our knowledge, existing nonparametric MABC algorithms all require the knowledge of κ for optimal properties. However, such information is typically not available to the game player. Efforts are made to provide a proper estimate for κ in section 4.

The margin condition has also been used in the MABC problem to control the game complexity (Goldenshluger and Zeevi, 2009; Perchet and Rigollet, 2013). Given $x \in [0,1]^d$, define $f^\sharp(x)$ to be

$$f^\sharp(x) = \begin{cases} \max_{1\leq i\leq l}\{f_i(x) : f_i(x) < f^*(x)\} & \text{if } \min_{1\leq i\leq l} f_i(x) < f^*(x), \\ f^*(x) & \text{otherwise.} \end{cases}$$

Assumption 2.3. *There exist $\alpha \in (0, d/\kappa]$, $t_0 \in (0,1)$ and $c_0 > 0$ such that*

$$P_X\big(0 < f^*(X) - f^\sharp(X) \leq t\big) \leq c_0 t^\alpha$$

for all $t \in [0, t_0]$.

Larger α in Assumption 2.3 indicates an easier MABC game in the sense that except on a subset of the domain with a small P_X-probability, it happens that either all the mean rewards are the same for all arms, or the optimal mean reward is well-separated from the sub-optimal ones. In particular, when $\alpha > d/\kappa$, one arm dominates over the entire domain (Perchet and Rigollet, 2013, Proposition 3.1) and the standard bandit problem algorithms will suffice in this case. Since this simple situation is not the interest of this article, we assume that $\alpha \leq d/\kappa$.

Next, we want to devise an algorithm that does not rely on the knowledge of either κ or α, but still achieves the (nearly) optimal regret cumulative rate as if we knew them in advance.

3. Algorithm

The algorithm consists of a forced sampling step followed by a randomized allocation with arm elimination mechanism. Suppose N is the total time horizon. The algorithm starts with a forced sampling step, in which every arm is pulled n_0 times ($1 \le n_0 \ll N$). The random sample of each arm thus obtained feeds into a smoothness parameter selector, which can be subsequently used to choose related parameters of the remaining steps. After the forced sampling step, the remaining time horizon is divided into $T+1$ stages. Let $\tilde{N}_1 < \tilde{N}_2 < \cdots < \tilde{N}_T$ be the end time points of the first T stages, and define $\tilde{N}_0 = n_0 l$. The number of time points in stage t ($1 \le t \le T$) is denoted by $N_t = \tilde{N}_t - \tilde{N}_{t-1}$. Let $\{h_t,\, 1 \le t \le T\}$ be a sequence of bin width that satisfies $h_1 = 1$ and $h_{k+1} = h_k/2$, $1 \le k \le T-1$. At the end of stage t ($1 \le t \le T$), for arm elimination, we partition the domain $[0,1]^d$ into $1/h_t^d$ bins with bin width h_t. Let $\mathcal{B}_t$ denote the set of these bins, and let $\mathcal{B}_t(x)$ denote the bin in $\mathcal{B}_t$ that contains the covariate $x \in [0,1]^d$. For notational convenience, define $h_0 = 1$ and bin $\mathcal{X} = [0,1]^d$. Also define $\mathcal{B}_0 = \{\mathcal{X}\}$ and $\mathcal{B}_0(x) = \mathcal{X}$ for every $x \in [0,1]^d$. By the choice of bin width sequence, we can see that for each bin $B \in \mathcal{B}_t$ ($1 \le t \le T$) and each stage s ($0 \le s < t$), there is a unique (larger) bin $B' \in \mathcal{B}_s$ that contains B. We denote B' by $p_s(B)$ and call it the "parent" bin of B at stage s. Let $\{\pi_n, 1 \le n \le N\}$ be a sequence of positive numbers satisfying $(l-1)\pi_n < 1$ for every $1 \le n \le N$. The algorithm for MABC works as follows.

Step 0. Initialize the game with the forced sampling step.

Step 0.1. Obtain a random sample of each arm by pulling each arm n_0 times.

Step 0.2. If the smoothness parameter κ is unknown, for every given arm i ($1 \le i \le l$), estimate κ by the smoothness parameter selector described in section 4. The resulting estimate for arm i is denoted by $\hat{\kappa}^{(i)}$. Define $\hat{\kappa}^* = \min_{1 \le i \le l} \hat{\kappa}^{(i)}$, which is used to determine parameters of the following steps. If κ is known, simply set $\hat{\kappa}^* = \kappa$.

Step 1. Define the initial set of active arms in bin $\mathcal{X}$ to be $\mathcal{S}_{\mathcal{X}} = \{1, 2, \cdots, l\}$. Start stage $t = 1$ of the game. For $n = \tilde{N}_{t-1}+1, \tilde{N}_{t-1}+2, \cdots, \tilde{N}_t$, perform the following substeps.

Step 1.1. Observe covariate X_n and locate the bin with bin width h_{t-1} that contains X_n by $B = \mathcal{B}_{t-1}(X_n)$. Find $\mathcal{S}_B$, the set of active arms in bin B. Denote the number of arms in $\mathcal{S}_B$ by l_B.

Step 1.2. For each arm $i \in \mathcal{S}_B$, based on the previously obtained sample of covariates and rewards, estimate the mean reward $f_i(X_n)$ by some user-specified regression modeling method (e.g., kernel regression). The estimator is denoted by $\hat{f}_{i,n}(X_n)$.

Step 1.3. Estimate the best arm, select and pull. Define $\hat{i}_n = \operatorname{argmax}_{i \in \mathcal{S}_B} \hat{f}_{i,n}(X_n)$ (If there is a tie, any tie-breaking rule may apply). Choose an arm, with probability $1-(l_B-1)\pi_n$ for arm $\hat{i}_n$ (the currently most

promising choice) and with probability π_n for each of the remaining arms in $\mathcal{S}_B$. That is,

$$I_n = \begin{cases} \hat{i}_n, & \text{with probability } 1-(l_B-1)\pi_n, \\ i, & \text{with probability } \pi_n,\ i \neq \hat{i}_n,\ i \in \mathcal{S}_B. \end{cases}$$

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

Step 2. At the end of stage t, perform arm elimination for the bins in $\mathcal{B}_t$ (with bin width h_t). For each bin $B \in \mathcal{B}_t$, do the following substeps.

Step 2.1. Identify the parent bin $B' = p_{t-1}(B)$ and the set of active arms $\mathcal{S}_{B'}$ for bin B'.

Step 2.2. For each arm $i \in \mathcal{S}_{B'}$, let $H_{B,i} = \{n : \tilde{N}_{t-1}+1 \le n \le \tilde{N}_t,\ X_n \in B,\ I_n = i\}$ be the set of time points during stage t at which the covariate falls in bin B and arm i is pulled. Let $N_{B,i}$ be the size of $H_{B,i}$. Find the arms in $\mathcal{S}_{B'}$ with $N_{B,i} \neq 0$ and define

$$\mathcal{S}_B^{(0)} = \{i \in \mathcal{S}_{B'} : N_{B,i} \neq 0\}.$$

Calculate the sample average of each arm $i \in \mathcal{S}_B^{(0)}$ during stage t inside bin B by $\bar{Y}_{B,i} = \sum_{n \in H_{B,i}} Y_{i,n}/N_{B,i}$. Calculate the maximum sample average by $\bar{Y}_B^* = \max_{i \in \mathcal{S}_B^{(0)}} \bar{Y}_{B,i}$.

Step 2.3. Identify the set of "bad" arms to be eliminated by

$$\mathcal{A}_B = \{i \in \mathcal{S}_B^{(0)} : \bar{Y}_B^* - \bar{Y}_{B,i} > \alpha_t\},$$

where α_t is a stage-dependent parameter. Obtain the set of active arms in bin B for the next stage by eliminating "bad" arms in $\mathcal{A}_B$ from $\mathcal{S}_{B'}$: $\mathcal{S}_B = \mathcal{S}_{B'} \setminus \mathcal{A}_B$.

Step 3. Repeat Step 1 and Step 2 for stage $t = 2, 3, \cdots, T$.

Step 4. Repeat Step 1 for $n = \tilde{N}_T + 1, \tilde{N}_T + 2, \cdots, N$ (it is stage $T+1$).

The forced sampling step obtains a random sample of each arm for the smoothness parameter selector. After the forced sampling step, $T+1$ stages of randomized allocation with arm elimination follow. For a given stage t ($1 \le t \le T+1$), Step 1 performs the randomized arm allocation. Specifically, Step 1.1 retrieves the set of active arms inherited from the previous stage. In particular, for stage $t = 1$, the set of active arms includes all the candidate arms. In Step 1.2, we have the flexibility to choose proper regression methods to estimate the mean reward functions of the active arms. Both parametric and nonparametric methods may apply. Step 1.3 is the randomized allocation that favors the arm with highest estimated reward and selects this arm with a high probability. At the end of a given stage t ($1 \le t \le T$), Step 2 follows to identify and eliminate the obvious bad-performing arms so that they do not get pulled in the next stage. For this purpose, the covariate domain is divided into $1/h_t^d$ bins with bin width h_t. For each of these bins, Step 2.2 calculates the reward sample average

of each active arm during stage t. Subsequently, Step 2.3 eliminates the arms with low sample average compared to the highest. The remaining arms of each bin after elimination serve as the new active arms, and the next stage follows. Heuristically speaking, Step 2 assists the randomized allocation mechanism of Step 1 to decrease the number of times the bad-performing arms get selected. The choice of algorithm parameters including n_0, T, $\tilde{N}_t$ and α_t depends on $\hat{\kappa}^*$, and is described in section 5. Note also that the algorithm above implicitly assume that $N > \tilde{N}_T$. If $\tilde{N}_T$ is chosen such that $N < \tilde{N}_T$, we simply stop the algorithm at $n = N$.

Remark 3.1. *Here, we provide some detailed discussion regarding the practical relevance of the randomized allocation procedure shown in Steps 1.2–1.3. From the perspectives of minimax optimality, under our settings and with the current technical tools available, if π_n's are uniformly lower bounded by a positive constant (and upper bounded by $1/l_B$ due to the natural requirement from randomized allocation), the RAAE algorithm can achieve the minimax regret rate of Rigollet and Zeevi (2010), irrespective of the regression modeling method chosen by the user. In particular, if we choose $\pi_n = 1/l_B$ (that is, each active arm has equal chance to be pulled), then the information from Step 1.2 is effectively ignored and the RAAE algorithm essentially becomes analogous to ABSE in the sense that both algorithms tend to pull each active arm an equal number of times. Practically, we advocate the use of smaller π_n to take advantage of the additional information gained from Step 1.2. For example, we may use kernel regression in Step 1.2 to estimate the reward function of each active arm. Then, in Step 1.3, if we choose $\pi_n = 0.05 \wedge \frac{1}{l_B}$, the arm with the highest estimated reward from Step 1.2 is pulled with larger probability than that of other active arms in the randomized allocation (assuming $l_B < 20$). Our empirical experience favors the latter choice of π_n. Simulation examples are given in Appendix B for comparison of the two different scenarios of π_n, with kernel regressions as the user-specified regression modeling method.*

4. Smoothness parameter selector

Suppose $f(x)$ is the mean reward function of a given arm, and a random sample $\{(X_i, Y_i), i = 1, \cdots, n\}$ of this arm is observed during the forced sampling step. Recall that κ_* and κ^* $(0 < \kappa_* < \kappa^* \leq 1)$ are the known lower and upper bound of κ, respective.

First, we make the following definitions. Define two integers

$$\tau^* = \max\left\{\tau + 1 : 2^\tau \leq n^{\frac{1}{2\kappa_* + d}}\right\}$$

and

$$\tau_* = \max\left\{\tau : 2^\tau \leq n^{\frac{1}{2\kappa^* + d}}\right\}.$$

For any $\tau \in \mathbb{N}$, define $u_\tau = 2^{-\tau}$, and let κ_τ be the real number that satisfies $u_\tau = n^{-\frac{1}{2\kappa_\tau + d}}$. Then, it is not hard to see that there exists a constant $\Delta > 0$

such that $\kappa_\tau - \kappa_{\tau+1} \leq \frac{\Delta}{\log n}$ for any $\tau \in [\tau_*, \tau^*]$. Given τ, we evenly partition the domain into $1/u_\tau^d$ bins with bin width u_τ, and let $\mathcal{D}_\tau(x)$ denote the bin that contains $x \in [0,1]^d$.

Next, with any given $x \in [0,1]^d$ and $\tau \in \mathbb{N}$, we can define a histogram estimator of $f(x)$ by

$$\hat{\theta}_\tau(x) = \frac{\sum_{i \in H_\tau(x)} Y_i}{M_\tau(x)},$$

where $H_\tau(x) = \{i : X_i \in \mathcal{D}_\tau(x), 1 \leq i \leq n\}$, and $M_\tau(x)$ is the size of $H_\tau(x)$. Define $\hat{\tau}$ to be

$$\min\{\tau \in [\tau_*, \tau^*] : \|\hat{\theta}_\tau - \hat{\theta}_{\tau_2}\|_\infty \leq b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n \text{ for every } \tau_2 \text{ satisfying } \tau < \tau_2 \leq \tau^*\}, \tag{4.1}$$

where $\|\cdot\|_\infty$ is the sup-norm, b_1 is a constant satisfying $b_1 > 4\rho$, and $\gamma_n = \log n$. Then the selected smoothness parameter for $f(x)$ is $\hat{\kappa} = \min\{\kappa_{\hat{\tau}} - \frac{b_2 \log\log n}{\log n}, \kappa^*\}$, where b_2 is a constant satisfying $b_2 > \frac{(2\kappa^*+d)^2}{2\kappa}$.

The smoothness parameter selector described above is essentially searching the largest possible u_τ such that its corresponding estimator for f does not differ too much from that of all smaller u_τ's under sup norm. The resulting $\kappa_{\hat{\tau}}$ after minor adjustment is used to approximate the smoothness parameter of the mean reward function.

To understand how well the method above performs when the knowledge of κ is absent, consider a sub-class $\Sigma_0(\kappa, \rho)$ of $\Sigma(\kappa, \rho)$ as follows. Given $\tau \in \mathbb{N}$ and $x \in [0,1]^d$, define

$$K_\tau f(x) =: E[f(X)|X \in \mathcal{D}_\tau(x)] = \frac{\int_{\mathcal{D}_\tau(x)} f(t) \mathrm{d}P_X(t)}{\int_{\mathcal{D}_\tau(x)} \mathrm{d}P_X(t)}.$$

Then

$$\Sigma_0(\kappa, \rho) =: \big\{f \in \Sigma(\kappa, \rho) : \text{there exists } 0 < \rho_1 < \rho \text{ and } \tau_0 > 0 \text{ such that } \|K_\tau f - f\|_\infty > \rho_1 u_\tau^\kappa \text{ for every } \tau \geq \tau_0\big\}.$$

It is not hard to see that for any $f \in \Sigma_0(\kappa, \rho)$, we have that $f \notin \Sigma(\tilde{\kappa}, \rho)$ for every $\tilde{\kappa} > \kappa$. It is worth emphasizing that $\Sigma_0(\kappa, \rho)$ is not an unnatural class of functions. Indeed, $\Sigma_0(\kappa, \rho)$ can be viewed as a class of functions that satisfies a "self-similarity" condition (Condition 3 in Giné and Nickl, 2010; see also Hoffmann and Nickl, 2011 and Bull, 2012). We defer the discussion of this condition to section 6.

Assumption 4.1. *The mean reward functions of all candidate arms are in $\Sigma(\kappa, \rho)$, and at least one reward function is in $\Sigma_0(\kappa, \rho)$.*

Proposition 4.1. *Suppose Assumptions 2.1, 2.2 and 4.1 hold. Then for $\hat{\kappa}^*$ obtained in Step 0 of the RAAE algorithm, there exist a constant $\tilde{C}_H$ and an integer n_H such that*

$$P\Big(\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n} < \hat{\kappa}^* \leq \kappa\Big) \geq 1 - \tilde{C}_H l (\log n)^2 n^{-1/c_*}$$

for every $n > n_H$, *where* $c_* = \frac{\kappa_*}{2\kappa_*+d}$.

Proposition 4.1 indicates that with high probability, the estimated smoothness parameter is no more than $O(\log\log n/\log n)$ smaller than κ, the largest possible smoothness parameter of the arm in $\Sigma_0(\kappa,\rho)$.

5. Finite-time regret analysis

The regret analysis of the RAAE algorithm relies on the appropriate choice of the corresponding parameters. Set the parameters as follows. Let $n_0 = \lceil N^{c_*} \rceil$ and $h_1 = 1$. Let stage number T be

$$T = \min\Big\{t \in \mathbb{N} : \frac{h_1}{2^{t-1}} \le 6\Big(\frac{l\log l}{N}\Big)^{\frac{1}{2\hat\kappa^*+d}}\Big\}. \tag{5.1}$$

Given any stage t $(1 \le t \le T)$, define $\tilde\pi_t = \min\{\pi_n : \tilde N_{t-1}+1 \le n \le \tilde N_t\}$. Take the threshold α_t in Step 2.3 to be $\alpha_t = 4\rho h_t^{\hat\kappa^*}$ (Alternatively, we may choose $\alpha_t = \sqrt{\frac{c(1\vee\log(Nh_t^{2\hat\kappa^*+d}))}{N_B^*\wedge N_{B,i}}}$, where N_B^* is the number of times the arm with the maximum sample average is pulled in bin B during stage t, and c is some constant. For brevity, we only show the proof under the former choice of α_t). Set $N_t = \tilde\gamma_t h_t^{-(2\hat\kappa^*+d)}(1\vee\log(Nh_t^{2\hat\kappa^*+d}))$, where $\tilde\gamma_t$ is a stage-dependent parameter chosen to make N_t a positive integer. In particular, it suffices to assume

$$\max\Big\{\frac{8(v^2+c\rho/2)}{\underline{c}\rho^2\tilde\pi_t}, \frac{56}{3\underline{c}\tilde\pi_t}\Big\} \le \tilde\gamma_t \le \gamma < \infty, \tag{5.2}$$

where γ is a positive constant. Note that such γ exists if $\{\tilde\pi_t, t \ge 1\}$ is uniformly lower bounded by a positive constant. For the proof of Theorem 5.1, we specifically set $\pi_n = 0.05 \wedge \frac{1}{l_B}$ in Step 1.3.

Theorem 5.1. *Under Assumptions 2.1, 2.2, 2.3 and 4.1, the mean cumulative regret of the proposed algorithm satisfies*

$$ER_N(\eta) \le \tilde C N\Big(\frac{l\log l}{N}\Big)^{\frac{\kappa(1+\alpha)}{2\kappa+d}}(\log N)^{c^*},$$

where $\tilde C$ *is a positive constant (not depending on* N *or* l*) and* $c^* = \frac{d(1+\alpha)(2\kappa^*+d)^2}{\kappa\kappa_*(2\kappa+d)}$.

The cumulative regret rate in Theorem 5.1 matches the minimax rate obtained by Perchet and Rigollet (2013) up to a logarithmic factor. The additional logarithmic term is the price we pay for not knowing κ. If the value of κ is available, we simply set $\hat\kappa^* = \kappa$ and the exact minimax rate can be achieved.

It is noted that the sample size n_0 used for the smoothness parameter selector in Step 0 of the RAAE algorithm has to be carefully chosen with the consideration of the subsequent steps. The sample size n_0 should be large enough so that the estimation of κ becomes accurate enough with a high probability before

its subsequent use. On the other hand, n_0 should be small enough so that the regret from Step 0 can be controlled within the desired range. It is also worth mentioning that although the proposed algorithm appears to assume a known value for ρ, it suffices to know the upper and lower bound of ρ to obtain the same rate.

Remark 5.1. *As is pointed out in section 1, the ABSE algorithm (Perchet and Rigollet, 2013) can also be used for adaptively achieving a near minimax rate when equipped with the Lepski-type smoothness parameter selector. Indeed, in the proof of Theorem 5.1, we can see that Step 0 essentially serves as a plugged-in estimator of the smoothness parameter κ, and, because of Proposition 4.1, the analysis can go through almost like we knew the true κ by using its estimator $\hat{\kappa}^*$ (in place of κ).*

6. Discussion

In the nonparametric MABC problem, as far as we know, no algorithms before this work have been shown to be minimax-rate optimal adaptively with respect to the unknown smoothness parameter κ. The Lepski's method is known to have successful applications in the context of adaptive nonparametric estimations. In the following, we discuss the connection of our proposed MABC algorithm with adaptive nonparametric estimation when the Lepski's method is applied.

In the context of the RAAE algorithm, heuristically speaking, under-estimation of κ results in overly small bin width so that the smoothness of the reward functions is not fully utilized. Over-estimation of κ leads to possible pre-mature elimination of good-performing arms, the probability of which cannot be properly bounded. Interestingly, in nonparametric estimation, the Lepski's approach also has to consider separately the events that its built-in selector generates too small or too large smoothness parameter estimates. The former event (i.e., under-estimation of κ) is usually considered the technically "complicated" case of the two in nonparametric estimation. Its counterpart in the MABC problem (see Lemma A.3) turns out to be straightforward because the event probability can be bounded tightly by using the moment condition (Assumption 2.2) and a Bernstein-type inequality. The observation that the former event has a tight probability is shared in, e.g., Lepski (1990) and Lepski, Mammen and Spokoiny (1997) under a Gaussian white noise model. On the other hand, the latter event (i.e., over-estimation of κ) is usually considered the technically "easy" case of the two in nonparametric estimation because of the straightforward use of the built-in selector's definition. But such "easy" results do not apply to the MABC problem since the over-estimation of κ will have adverse effects on subsequent procedures.

Indeed, the difficulty caused by the over-estimation of κ is shared in the adaptive confidence bound problems. If we only consider the Hölder condition without further assumptions, it is known that the adaptive confidence bound generally does not exist (Low, 1997). As one solution to overcome such difficulty,

Giné and Nickl (2010) propose a "self-similarity" condition, and show that the functions that do not satisfy this condition can be a negligible subset of Hölder class (see Condition 3 and Proposition 4 in Giné and Nickl, 2010). It turns out that the function class $\Sigma_0(\kappa, \rho)$ defined in section 4 takes the form of their "self-similarity" condition. To see such connection, we consider the special case in the rest of the discussion that the covariate is univariate and has the distribution $P_X \sim \text{Uniform}[0, 1]$.

Consider the wavelet kernel as follows (Härdle et al., 1998). Let ϕ and ψ be the father Harr wavelet and mother Harr wavelet, that is, $\phi(x) = I(x \in (0, 1])$ and $\psi(x) = I(x \in [0, \frac{1}{2}]) - I(x \in (\frac{1}{2}, 1])$. Let $\phi_{\tau k}(x) = 2^{\tau/2}\phi(2^\tau x - k)$. Define the wavelet kernel

$$K(x, x') = \sum_k \phi(x - k)\phi(x' - k),$$

and define $K_\tau(x, x') = 2^\tau K(2^\tau x, 2^\tau x')$. Then the projection of function $f \in \Sigma(\kappa, \rho)$ to the linear subspace with basis $V_\tau = \{\phi_{\tau k} : k \in \mathbb{Z}\}$ is

$$\tilde{K}_\tau f(x) =: \int_{[0,1]} K_\tau(x, z) f(z) \mathrm{d}z.$$

Note that if $x \in (\frac{k_0}{2^\tau}, \frac{k_0+1}{2^\tau}]$ for some $k_0 \in \{0, 1, \cdots, 2^\tau - 1\}$, then

$$\begin{aligned}
\tilde{K}_\tau f(x) &= \frac{1}{2^{-\tau}} \sum_k \int_{[0,1]} \phi(2^\tau x - k)\phi(2^\tau z - k) f(z) \mathrm{d}z \\
&= \frac{1}{2^{-\tau}} \int_{[0,1]} \phi(2^\tau z - k_0) f(z) \mathrm{d}z \\
&= \frac{\int_{(\frac{k_0}{2^\tau}, \frac{k_0+1}{2^\tau}]} f(z) \mathrm{d}z}{2^{-\tau}} \\
&= K_\tau f(x).
\end{aligned}$$

With the above, it is clear that if we only consider $f \in \Sigma(\kappa, \rho)$, then Condition 3 of Giné and Nickl (2010) (that is, there exist positive constants $\rho_2 \le \rho$ and a positive integer τ_0 such that for every integer $\tau \ge \tau_0$, $\rho_2 2^{-\tau\kappa} \le \|\tilde{K}_\tau f - f\|_\infty \le \rho 2^{-\tau\kappa}$) becomes largely equivalent to the definition of $\Sigma_0(\kappa, \rho)$. Inspired by such connection, it is conjectured that $\Sigma_0(\kappa, \rho)$ can be a "rich" sub-class in $\Sigma(\kappa, \rho)$. In fact, it is not hard to show that for any function $f \in \Sigma(\kappa, \rho)$, if for some $x_0 \in [0, 1]$ and some constants $U_1, U_2 \neq 0$,

$$\lim_{v \to 0+} \frac{f(x_0 + v) - f(x_0)}{|v|^\kappa} = U_1 \quad \text{or} \quad \lim_{v \to 0-} \frac{f(x_0 + v) - f(x_0)}{|v|^\kappa} = U_2, \tag{6.1}$$

then $f \in \Sigma_0(\kappa, \rho)$. Interestingly, since the functions constructed in Theorem 4.1 of Rigollet and Zeevi (2010) satisfy (6.1) and consequently belong to $\Sigma_0(\kappa, \rho)$, the rate obtained in Theorem 5.1 remains to be the near minimax rate for $\Sigma_0(\kappa, \rho)$ under Assumption 4.1.

Appendix A: Lemmas and proofs

The proofs of Proposition 4.1 and Theorem 5.1 are given in the sections A.1 and A.2, respectively. To keep this paper self-contained, we list the following two lemmas for convenience, and their proofs can be found in Qian and Yang (2016).

Lemma A.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* $n \geq 1$*,*

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

Lemma A.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* $n \geq 1$*,*

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

A.1. Proof of Proposition 4.1

Proposition 4.1 is a straightforward result of the following two lemmas.

Lemma A.3. *Suppose* $f(\cdot) \in \Sigma(\kappa, \rho)$ *and Assumptions 2.1 and 2.2 hold. Then for* $\hat{\kappa}$ *obtained by procedures in section 4, there exists an integer* n_* *and a constant* C_H *such that*

$$P\Big(\hat{\kappa} \leq \kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n}\Big) \leq \frac{C_H(\log n)^2}{n^{1/c_*}}$$

for every $n > n_*$

Proof of Lemma A.3. Define

$$\tilde{\tau} = \max\Big\{\tau + 1 : 2^{\tilde{\tau}} \leq n^{\frac{1}{2\kappa+d}}\Big\}.$$

Let $\tilde{\kappa} = \kappa_{\tilde{\tau}}$ and $\check{\kappa} = \kappa_{\hat{\tau}}$. Then by the definition in (4.1),

$$
\begin{aligned}
&\{\check{\kappa} \le \tilde{\kappa}\} \\
&\Rightarrow \bigcup_{\tilde{\tau}=1}^{\tau^*} \{\hat{\tau} = \tau\} \\
&\Rightarrow \bigcup_{\tau=\tilde{\tau}-1}^{\tau^*-1} \bigcup_{\tau_2=\tau+1}^{\tau^*} \{\|\hat{\theta}_\tau - \hat{\theta}_{\tau_2}\|_\infty > b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n\} \\
&\Rightarrow \bigcup_{\tau=\tilde{\tau}-1}^{\tau^*-1} \bigcup_{\tau_2=\tau+1}^{\tau^*} \Big\{\{\|\hat{\theta}_\tau - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\} \cup \{\|\hat{\theta}_{\tau_2} - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\}\Big\}.
\end{aligned} \tag{A.1}
$$

Given $\tau \in \mathbb{N}$, let $\mathcal{M}_\tau$ be the set of bins with bin width u_τ that partition the domain. Clearly, $|\mathcal{M}_\tau| = 1/u_\tau^d$.

Then, given any τ_2 and τ such that $\tilde{\tau} - 1 \le \tau \le \tau_2 \le \tau^*$, we have

$$
P\Big(\|\hat{\theta}_\tau - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\Big) \le \sum_{B \in \mathcal{M}_\tau} P\Big(\sup_{x \in B} |\hat{\theta}_\tau(x) - f(x)| > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\Big). \tag{A.2}
$$

To derive the upper bound for the inequality above, note that if $M_\tau(x) > 0$,

$$
\hat{\theta}_\tau(x) - f(x) = \frac{\sum_{i \in H_\tau(x)} \big(Y_i - f(x)\big)}{M_\tau(x)} = \frac{\sum_{i \in H_\tau(x)} \varepsilon_i}{M_\tau(x)} + \frac{\sum_{i \in H_\tau(x)} \big(f(X_i) - f(x)\big)}{M_\tau(x)}.
$$

Let x_B^* be a fix point in bin $B \in \mathcal{M}_\tau$, then the previous display implies that

$$
\sup_{x \in B} |\hat{\theta}_\tau(x) - f(x)| \le \frac{\Big|\sum_{i \in H_\tau(x_B^*)} \varepsilon_i\Big|}{M_\tau(x_B^*)} + \rho u_\tau^\kappa. \tag{A.3}
$$

Define

$$
A_{\tau,B} = \Big\{M_\tau(x_B^*) > \frac{n \underline{c} u_\tau^d}{2}\Big\}
$$

and

$$
J_{\tau,B} = \Big\{\sup_{x \in B} |\hat{\theta}_\tau(x) - f(x)| > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\Big\}.
$$

Then,

$$
\begin{aligned}
P(J_{\tau,B}) &\le P(A_{\tau,B}^c) + P(J_{\tau,B},\, A_{\tau,B}) \\
&\le P(A_{\tau,B}^c) + P\Big(\frac{\Big|\sum_{i \in H_\tau(x_B^*)} \varepsilon_i\Big|}{M_\tau(x_B^*)} > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2} - \rho u_\tau^\kappa,\, A_{\tau,B}\Big) \\
&\le P(A_{\tau,B}^c) + P\Big(\frac{\Big|\sum_{i \in H_\tau(x_B^*)} \varepsilon_i\Big|}{M_\tau(x_B^*)} > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{4},\, A_{\tau,B}\Big),
\end{aligned} \tag{A.4}
$$

where the second inequality follows by (A.3) and the last inequality follows by the fact that $\rho h_\tau^\kappa < b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n/4$ for large enough n. Note that by Lemma A.1,

$$P_{X^n}\Big(\frac{\Big|\sum_{i\in H_\tau(x_B^*)}\varepsilon_i\Big|}{M_\tau(x_B^*)} > \epsilon\Big) \le \exp\Big(-\frac{M_\tau(x_B^*)\epsilon^2}{2(v^2+c\epsilon)}\Big).$$

As a result,

$$\begin{aligned}
&P\Big(\frac{\Big|\sum_{i\in H_\tau(x_B^*)}\varepsilon_i\Big|}{M_\tau(x_B^*)} > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}}\gamma_n}{4},\, A_{\tau,B}\Big)\\
\le&\exp\Big(-\frac{n\underline{c}u_\tau^d b_1^2 u_{\tau_2}^{2\kappa_{\tau_2}}\gamma_n^2}{64(v^2+cb_1u_{\tau_2}^{\kappa_{\tau_2}}\gamma_n/4)}\Big)\\
\le&\exp\Big(-\frac{\underline{c}b_1^2\gamma_n^2}{128v^2}\Big)\\
\le&n^{-\frac{d}{2\kappa_*+d}-\frac{1}{c_*}},
\end{aligned} \tag{A.5}$$

where the last two inequalities follow by the observation that $nu_\tau^d u_{\tau_2}^{2\kappa_{\tau_2}} \ge 1$, $\frac{cb_1\gamma_n}{4n^{c_*}} \le v^2$ and $\frac{\underline{c}b_1^2\log n}{128v^2} > \frac{d}{2\kappa_*+d}+\frac{1}{c_*}$ for large enough n. Also, since $P(I(X_i\in B)) \ge \underline{c}u_\tau^d$ for any $B\in\mathcal{D}_\tau$, by Lemma A.2,

$$P(A_{\tau,B}^c) \le \exp\Big(-\frac{3\underline{c}nu_\tau^d}{28}\Big). \tag{A.6}$$

Thus, by (A.2), (A.4), (A.5), (A.6), and the fact that $u_\tau^{-d} \le C_{H1} n^{\frac{d}{2\kappa_*+d}}$ for some constant $C_{H1}>0$, we have

$$\begin{aligned}
&P\Big(\|\hat\theta_\tau - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}}\gamma_n}{2}\Big)\\
\le&u_\tau^{-d}\exp\Big(-\frac{3\underline{c}nu_\tau^d}{28}\Big) + u_\tau^{-d}n^{-\frac{d}{2\kappa_*+d}-\frac{1}{c_*}}\\
\le&\frac{2C_{H1}}{n^{1/c_*}}.
\end{aligned}$$

In together with (A.1) and $\tilde\kappa > \kappa - \frac{\Delta}{\log n}$, we know that there exists n_* and some constant C_H such that

$$P\Big(\check\kappa \le \kappa - \frac{\Delta}{\log n}\Big) \le P\Big(\check\kappa \le \tilde\kappa\Big) \le \frac{C_H(\log n)^2}{n^{1/c_*}}$$

for any $n > n_*$. This completes the proof of Lemma A.3. □

Lemma A.4. *Suppose $f(\cdot)\in\Sigma_0(\kappa,\rho)$ and Assumptions 2.1 and 2.2 hold. Then for $\hat\kappa$ obtained by procedures in section 4, there exists an integer n^* and a constant $C_H^*>0$ such that*

$$P(\hat\kappa > \kappa) \le \frac{C_H^*}{n^{1/c_*}}$$

for every $n > n^*$.

Proof of Lemma A.4. Let $\tilde{\tau}$, $\tilde{\kappa}$ and $\check{\kappa}$ be defined as in the proof of Lemma A.3. Let $\kappa' = \kappa + \frac{b_2 \log\log n}{\log n}$. Define the integer

$$\tau' = \max\{\tau : 2^\tau \le n^{\frac{1}{2\kappa'+d}}\}$$

Then by definition in (4.1) and the fact that $\tau' < \tilde{\tau}$,

$$\begin{aligned}
&\{\check{\kappa} > \kappa'\}\\
\Rightarrow&\{\hat{\tau} \le \tau'\}\\
\Rightarrow&\{\|\hat{\theta}_{\tau'} - \hat{\theta}_{\tilde{\tau}}\|_\infty \le b_1 u_{\tilde{\tau}}^{\tilde{\kappa}}\}\\
\Rightarrow&\{\|\hat{\theta}_{\tau'} - f\|_\infty \le \frac{3}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\} \cup \{\|\hat{\theta}_{\tilde{\tau}} - f\|_\infty > \frac{1}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\}. \qquad \text{(A.7)}
\end{aligned}$$

Recall from the proof of Lemma A.3 that there is a constant C_{H1} such that

$$P\Big(\|\hat{\theta}_{\tilde{\tau}} - f\|_\infty > \frac{1}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) \le \frac{2C_{H1}}{n^{1/c_*}}. \qquad \text{(A.8)}$$

It remains to find the upper bound for $P(\|\hat{\theta}_{\tau'} - f\|_\infty \le \frac{3}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n)$. Note that by triangle inequalities,

$$\begin{aligned}
&|\hat{\theta}_{\tau'}(x) - f(x)|\\
=&\Big|\frac{\sum_{i\in H_{\tau'}(x)} f(X_i)}{M_{\tau'}(x)} - K_{\tau'} f(x) + K_{\tau'} f(x) - f(x) + \frac{\sum_{i\in H_{\tau'}(x)} \varepsilon_i}{M_{\tau'}(x)}\Big|\\
\ge&|K_{\tau'} f(x) - f(x)| - \Big|\frac{\sum_{i\in H_{\tau'}(x)} f(X_i)}{M_{\tau'}(x)} - K_{\tau'} f(x)\Big| - \Big|\frac{\sum_{i\in H_{\tau'}(x)} \varepsilon_i}{M_{\tau'}(x)}\Big|.
\end{aligned}$$

The previous inequality implies that for large enough n,

$$\begin{aligned}
&\|\hat{\theta}_{\tau'} - f\|_\infty\\
\ge&\|K_{\tau'} f - f\|_\infty - \sup_x \Big|\frac{\sum_{i\in H_{\tau'}(x)} f(X_i)}{M_{\tau'}(x)} - K_{\tau'} f(x)\Big| - \sup_x \Big|\frac{\sum_{i\in H_{\tau'}(x)} \varepsilon_i}{M_{\tau'}(x)}\Big|\\
=:&\|K_{\tau'} f - f\|_\infty - \Gamma_1 - \Gamma_2\\
>&\rho_1 u_{\tau'}^{\kappa} - \Gamma_1 - \Gamma_2\\
\ge&2b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n - \Gamma_1 - \Gamma_2 \qquad \text{(A.9)}
\end{aligned}$$

where the second to last inequality follows by that $f \in \Sigma_0(\kappa, \rho)$, and the last inequality follows because

$$\frac{u_{\tau'}^{\kappa}}{u_{\tilde{\tau}}^{\tilde{\kappa}}} \ge \frac{n^{-\frac{\kappa}{2\kappa'+d}}}{n^{-\frac{\kappa+\Delta/\log n}{2\kappa+d}}} = e^{\frac{\Delta}{2\kappa+d}} n^{\frac{2\kappa(\kappa'-\kappa)}{(2\kappa+d)(2\kappa'+d)}} \ge e^{\frac{\Delta}{2\kappa+d}} (\log n)^{\frac{2\kappa b_2}{(2\kappa^*+d)^2}} > \frac{2b_1\gamma_n}{\rho_1}.$$

Also, by derivations similar to that of (A.5) and (A.6),

$$
\begin{aligned}
&P\Big(\Gamma_2 \geq \frac{1}{4} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) \\
&\leq u_{\tau'}^{-d}\Big(\exp\Big(-\frac{3\underline{c} n u_{\tau'}^d}{28}\Big) + \exp\Big(-\frac{\underline{c} b_1^2 \gamma_n^2}{256 v^2}\Big)\Big) \\
&\leq \frac{2C_{H1}}{n^{1/c_*}},
\end{aligned} \tag{A.10}
$$

for all large enough n. Similarly, we can apply Azuma's inequality to obtain that

$$
\begin{aligned}
&P\Big(\Gamma_1 \geq \frac{1}{4} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) \\
&\leq u_{\tau'}^{-d}\Big(\exp\Big(-\frac{3\underline{c} n u_{\tau'}^d}{28}\Big) + \exp\Big(-\frac{(\underline{c} n u_{\tau'}^d/2) b_1^2 u_{\tilde{\tau}}^{2\tilde{\kappa}} \gamma_n^2}{64\|f\|_\infty}\Big)\Big) \\
&\leq \frac{2C_{H1}}{n^{1/c_*}},
\end{aligned} \tag{A.11}
$$

for all large enough n. Then, by (A.9), (A.10) and (A.11),

$$
\begin{aligned}
&P\Big(\|\hat{\theta}_{\tau'} - f\|_\infty \leq \frac{3}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) \\
&\leq P\Big(2b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n - \Gamma_1 - \Gamma_2 \leq \frac{3}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) \\
&\leq P\Big(\Gamma_1 \geq \frac{1}{4} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) + P\Big(\Gamma_2 \geq \frac{1}{4} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\Big) \\
&\leq \frac{4C_{H1}}{n^{1/c_*}}.
\end{aligned}
$$

Together with (A.7) and (A.8),

$$
P(\check{\kappa} > \kappa') \leq \frac{6C_{H1}}{n^{1/c_*}},
$$

which completes the proof of Lemma A.4. □

Proof of Proposition 4.1. By Lemma A.3 and Assumption 4.1,

$$
\begin{aligned}
P\Big(\hat{\kappa}^* \leq \kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n}\Big) &\leq \sum_{i=1}^{l} P\Big(\hat{\kappa}^{(*)} \leq \kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n}\Big) \\
&\leq C_H l (\log n)^2 n^{-1/c_*}.
\end{aligned}
$$

Together with Lemma A.4 and the fact that there exists $f_i \in \Sigma_0(\kappa, \rho)$, the proof of Proposition 4.1 is complete. □

A.2. Proof of Theorem 5.1

Proof of Theorem 5.1. Let $V_0 = \{\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n} < \hat{\kappa}^* \leq \kappa\}$. Inspired by the technique employed in the proof of Theorem 5.1 in Perchet and Rigollet

(2013), we define some sets and events as follows. For every bin $B \in \mathcal{B}_T$ (at stage T), recall that $p_t(B)$ is the parent bin of set B at stage t, and $\mathcal{S}_{p_t(B)}$ is the set of arms in $p_t(B)$ that survive the stage t arm elimination. Then, for every bin $B \in \mathcal{B}_T$ and every t $(1 \le t \le T)$, define the sets of arms

$$\mathcal{S}_{t,B,1} = \{1 \le i \le l : \text{ there exists some } x \in p_t(B) \text{ such that } f^*(x) = f_i(x)\},$$
$$\mathcal{S}_{t,B,2} = \{1 \le i \le l : \text{ for every } x \in p_t(B),\ f^*(x) - f_i(x) \le 8\rho h_t^{\hat{\kappa}^*}\},$$

and define the events

$$G_{t,B,1} = \{\mathcal{S}_{t,B,1} \subseteq \mathcal{S}_{p_t(B)}\},$$
$$G_{t,B,2} = \{\mathcal{S}_{p_t(B)} \subseteq \mathcal{S}_{t,B,2}\}.$$

Here, we consider $G_{t,B,1}$ and $G_{t,B,2}$ as "good" events because $G_{t,B,1}$ means that all possible best arms in bin $p_t(B)$ survive the stage t arm elimination, and $G_{t,B,2}$ means that all survived arms in $\mathcal{S}_{p_t(B)}$ have regret no larger than $8\rho h_t^{\hat{\kappa}^*}$. Further define the sets

$$A_{t,B} = G_{t,B,1} \cap G_{t,B,2}, \tag{A.12}$$
$$F_{t,B} = \cap_{1 \le k \le t} A_{k,B}. \tag{A.13}$$

The set $A_{t,B}$ means that the "good" events happen at stage t, and $F_{t,B}$ means that such "good" events happen during all of the first t stages. Note that

$$R_N(\eta) = R_N(\eta) I(V_0^c) + R_N(\eta) I(V_0) \tag{A.14}$$

and

$$\begin{aligned} R_N(\eta) I(V_0) &\le w l n_0 + \sum_{n=\tilde{N}_0+1}^{N} \big(f^*(X_n) - f_{I_n}(X_n)\big) I(V_0) \\ &\le w l n_0 + \sum_{B \in \mathcal{B}_T} \sum_{n=\tilde{N}_0+1}^{N} \big(f^*(X_n) - f_{I_n}(X_n)\big) I(V_0) I(X_n \in B) \\ &=: w l n_0 + \sum_{B \in \mathcal{B}_T} R_B. \end{aligned} \tag{A.15}$$

Let $R_N^{(0)} = \sum_{B \in \mathcal{B}_T} R_B$. Then, by the tree diagram,

$$\begin{aligned} R_N^{(0)} &= \sum_{B \in \mathcal{B}_T} R_B I(A_{1,B}^c) + \sum_{B \in \mathcal{B}_T} R_B I(F_{1,B} \cap A_{2,B}^c) + \cdots \\ &\quad + \sum_{B \in \mathcal{B}_T} R_B I(F_{T-1,B} \cap A_{T,B}^c) + \sum_{B \in \mathcal{B}_T} R_B I(F_{T,B}) \\ &=: R_1 + R_2 + \cdots + R_T + R_{T+1}. \end{aligned} \tag{A.16}$$

Next, we provide upper bounds for $R_1, R_2, \cdots, R_{T+1}$. By definition,

$$R_1 = \sum_{n=\tilde{N}_0+1}^{N} \sum_{B \in \mathcal{B}_T} \big(f^*(X_n) - f_{I_n}(X_n)\big) I(X_n \in B) I(V_0 \cap A_{1,B}^c)$$

$$\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B\in\mathcal{B}_T} wI(X_n \in B)I(V_0 \cap A_{1,B}^c) + \sum_{n=\tilde{N}_1+1}^{N} \sum_{B\in\mathcal{B}_T} wI(X_n \in B)I(V_0 \cap A_{1,B}^c).$$

Let $E^{(0)}(\cdot)$ and $P^{(0)}(\cdot)$ denote the conditional expectation and conditional probability given $\hat{\kappa}^* = \kappa_0$ ($\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n} < \kappa_0 \leq \kappa$), respectively. Then, by independence of the event $\{X_n \in B\}$ with $A_{1,B}^c$ ($\tilde{N}_1+1 \leq n \leq N$) given $\hat{\kappa}^* = \kappa_0$,

$$\begin{aligned} E^{(0)}(R_1) &\leq E^{(0)}\Big(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B\in\mathcal{B}_T} wI(X_n \in B)I(A_{1,B}^c)\Big) + \sum_{n=\tilde{N}_1+1}^{N} \sum_{B\in\mathcal{B}_T} wP(X_n \in B)P^{(0)}(A_{1,B}^c) \\ &\leq E^{(0)}\Big(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B\in\mathcal{B}_T} wI(X_n \in B)I(A_{1,B}^c)\Big) + \sum_{n=\tilde{N}_1+1}^{N} w \max_{B\in\mathcal{B}_T} P^{(0)}(A_{1,B}^c) \\ &\leq E^{(0)}\Big(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B\in\mathcal{B}_T} wI(X_n \in B)I(A_{1,B}^c)\Big) + 4wlh_1^{-(2\kappa_0+d)}, \quad \text{(A.17)} \end{aligned}$$

where the last inequality follows by Lemma A.5. Similarly, by definition, for $2 \leq t \leq T$,

$$\begin{aligned} R_t &= \sum_{n=\tilde{N}_0+1}^{N} \sum_{B\in\mathcal{B}_T} \big(f^*(X_n) - f_{I_n}(X_n)\big) I(X_n \in B) I(V_0 \cap F_{t-1,B} \cap A_{t,B}^c) \\ &\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B\in\mathcal{B}_T} wI(X_n \in B) I(V_0 \cap F_{t-1,B} \cap A_{t,B}^c) \\ &\quad + \sum_{k=1}^{t-1} \Big(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} \sum_{B\in\mathcal{B}_T} \big(f^*(X_n) - f_{I_n}(X_n)\big) \times \\ &\qquad I(X_n \in B,\, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\hat{\kappa}^*}) I(V_0 \cap F_{t-1,B} \cap A_{t,B}^c) \Big) \\ &\quad + \sum_{n=\tilde{N}_t+1}^{N} \sum_{B\in\mathcal{B}_T} \big(f^*(X_n) - f_{I_n}(X_n)\big) \times \\ &\qquad I(X_n \in B,\, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\hat{\kappa}^*}) I(V_0 \cap F_{t-1,B} \cap A_{t,B}^c) \end{aligned}$$

$$
\begin{aligned}
&\le \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B\in\mathcal{B}_T} wI(X_n\in B)I(V_0\cap F_{t-1,B}\cap A_{t,B}^c)\\
&\quad+\sum_{k=1}^{t-1}\Big(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} 8\rho h_k^{\hat{\kappa}^*} I(0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_{t-1}^{\hat{\kappa}^*})\Big)\\
&\quad+\sum_{n=\tilde{N}_t+1}^{N}\sum_{B\in\mathcal{B}_T} 8\rho h_{t-1}^{\hat{\kappa}^*} I(X_n\in B,\,0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_{t-1}^{\hat{\kappa}^*})\times\\
&\quad I(V_0\cap F_{t-1,B}\cap A_{t,B}^c)
\end{aligned}
$$

where the second to last inequality follows by the definition of event $F_{t-1,B}$. Then, by conditional independence of the event $\{X_n\in B,\,0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_{t-1}^{\hat{\kappa}^*}\}$ with $F_{t-1,B}\cap A_{t,B}^c$ $(\tilde{N}_t+1\le n\le N)$, given $\hat{\kappa}^*=\kappa_0$,

$$
\begin{aligned}
&E^{(0)}(R_t)\\
&\le E^{(0)}\Big(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1}\sum_{B\in\mathcal{B}_T} wI(X_n\in B)I(F_{t-1,B}\cap A_{t,B}^c)\Big)\\
&\quad+\sum_{k=1}^{t-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha}N_{k+1}\\
&\quad+\sum_{n=\tilde{N}_t+1}^{N}\sum_{B\in\mathcal{B}_T} 8\rho h_{t-1}^{\kappa_0}P(X_n\in B,\,0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_{t-1}^{\kappa_0})\\
&\quad\times P^{(0)}(F_{t-1,B}\cap A_{t,B}^c)\\
&\le E^{(0)}\Big(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1}\sum_{B\in\mathcal{B}_T} wI(X_n\in B)I(F_{t-1,B}\cap A_{t,B}^c)\Big)\\
&\quad+\sum_{k=1}^{t-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha}\gamma h_{k+1}^{-(2\kappa_0+d)}\log(Nh_{k+1}^{2\kappa_0+d})+4lc_0(8\rho h_{t-1}^{\kappa_0})^{1+\alpha}h_t^{-(2\kappa_0+d)},
\end{aligned}
\tag{A.18}
$$

where the first inequality follows by Assumption 2.3, and the second inequality follows by Assumption 2.3, Lemma A.5 and the choice of $\{N_k,1\le k\le t\}$. Similarly, by definition,

$$
\begin{aligned}
R_{T+1}&=\sum_{n=\tilde{N}_0+1}^{N}\sum_{B\in\mathcal{B}_T}\big(f^*(X_n)-f_{I_n}(X_n)\big)I(X_n\in B)I(V_0\cap F_{T,B})\\
&\le\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1}\sum_{B\in\mathcal{B}_T} wI(X_n\in B)I(V_0\cap F_{T,B})
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{k=1}^{T-1}\Big(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}}\sum_{B\in\mathcal{B}_T}\big(f^*(X_n)-f_{I_n}(X_n)\big)\times\\
&\quad I(X_n\in B,\,0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_k^{\hat{\kappa}^*})I(V_0\cap F_{T,B})\Big)\\
&+\sum_{n=\tilde{N}_T+1}^{N}\sum_{B\in\mathcal{B}_T}\big(f^*(X_n)-f_{I_n}(X_n)\big)\times\\
&\quad I(X_n\in B,\,0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_T^{\hat{\kappa}^*})I(V_0\cap F_{T,B})\\
&\le\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1}\sum_{B\in\mathcal{B}_T} wI(X_n\in B)I(V_0\cap F_{T,B})\\
&\quad+\sum_{k=1}^{T-1}\Big(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} 8\rho h_k^{\hat{\kappa}^*}I(0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_k^{\hat{\kappa}^*})\Big)\\
&\quad+\sum_{n=\tilde{N}_T+1}^{N} 8\rho h_T^{\hat{\kappa}^*}I(0<f^*(X_n)-f^\sharp(X_n)\le 8\rho h_T^{\hat{\kappa}^*}).
\end{aligned}
$$

Then, given $\hat{\kappa}^*=\kappa_0$,

$$
\begin{aligned}
E^{(0)}(R_{T+1})&\le E^{(0)}\Big(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1}\sum_{B\in\mathcal{B}_T} wI(X_n\in B)I(F_{T,B})\Big)\\
&\quad+\sum_{k=1}^{T-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha}\gamma h_{k+1}^{-(2\kappa_0+d)}\log(Nh_{k+1}^{2\kappa_0+d})+N(8\rho h_T^{\kappa_0})^{1+\alpha}.
\end{aligned}
\tag{A.19}
$$

Combining (A.16)–(A.19), we have

$$
\begin{aligned}
E^{(0)}(R_N^{(0)})&\le w\gamma h_1^{-(2\kappa_0+d)}\log(Nh_1^{2\kappa_0+d})+4wlh_1^{-(2\kappa_0+d)}\\
&\quad+\sum_{t=2}^{T}\sum_{k=1}^{t-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha}\gamma h_{k+1}^{-(2\kappa_0+d)}\log(Nh_{k+1}^{2\kappa_0+d})\\
&\quad+\sum_{t=2}^{T} 4lc_0(8\rho h_{t-1}^{\kappa_0})^{1+\alpha}h_t^{-(2\kappa_0+d)}\\
&\quad+\sum_{k=1}^{T-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha}\gamma h_{k+1}^{-(2\kappa_0+d)}\log(Nh_{k+1}^{2\kappa_0+d})+N(8\rho h_T^{\kappa_0})^{1+\alpha}\\
&\le w\gamma\log N+4wl+C_1lh_T^{-(\kappa_0-\kappa_0\alpha+d)}(1+\log(Nh_T^{2\kappa_0+d}))\\
&\quad+C_2Nh_T^{\kappa_0+\kappa_0\alpha}\\
&\le C_3N^{\frac{\kappa_0-\kappa_0\alpha+d}{2\kappa_0+d}}(l\log l)^{\frac{\kappa_0(1+\alpha)}{2\kappa_0+d}}\\
&\le C_4N\Big(\frac{l\log l}{N}\Big)^{\frac{\kappa(1+\alpha)}{2\kappa+d}}(\log N)^{c^*},
\end{aligned}
\tag{A.20}
$$

where $C_1, \cdots, C_4$ are some positive constants, and the last inequality follows by $\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n} < \kappa_0 \le \kappa$. Then, by (A.14), (A.15), (A.20) and Proposition 4.1, there exists some constant $\tilde{C} > 0$ such that

$$ER_N(\eta) \le wNP(V_0^c) + wln_0 + ER_N^{(0)} \le \tilde{C}N\Big(\frac{l\log l}{N}\Big)^{\frac{\kappa(1+\alpha)}{2\kappa+d}} (\log N)^{c^*}.$$

This completes the proof of Theorem 5.1. □

The proof of Theorem 5.1 above needs the following lemma.

Lemma A.5. *Suppose the conditions of Theorem 5.1 are satisfied. If the events $A_{t,B}$ and $F_{t,B}$ $(1 \le t \le T)$ are defined as in (A.12) and (A.13), respectively, then given any κ_0 satisfying $\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log\log n}{\log n} < \kappa_0 \le \kappa$,*

$$P^{(0)}(A_{1,B}^c) \le \frac{4l}{Nh_1^{2\kappa_0+d}} \quad \text{and} \quad P^{(0)}(F_{t-1,B} \cap A_{t,B}^c) \le \frac{4l}{Nh_t^{2\kappa_0+d}},\ 2 \le t \le T,$$

where $P^{(0)}(\cdot)$ is the conditional probability given $\hat{\kappa}^* = \kappa_0$.

Proof of Lemma A.5. Given $2 \le t \le T-1$ and $B \in \mathcal{B}_T$, to find $P^{(0)}(F_{t-1,B} \cap A_{t,B}^c)$, note that by definition, $A_{t,B}^c = G_{t,B,1}^c \cup (G_{t,B,1} \cap G_{t,B,2}^c)$. As a result, under $F_{t-1,B} \cap A_{t,B}^c$, either $F_{t-1,B} \cap G_{t,B,1}^c$ or $F_{t-1,B} \cap G_{t,B,1} \cap G_{t,B,2}^c$ happens.

First, we assume the event $F_{t-1,B} \cap G_{t,B,1}^c$ happens. Since $G_{t,B,1}^c = \{\mathcal{S}_{t,B,1} \subseteq \mathcal{S}_{p_t(B)}\}^c$, the event $F_{t-1,B} \cap G_{t,B,1}^c$ implies that there exists an arm $i_1 \in \mathcal{S}_{t,B,1}$ such that arm i_1 is eliminated at the end of stage t (within bin $p_t(B)$). For notation brevity, denote $p_t(B)$ by $\tilde{B}$. Recall that if $N_{\tilde{B},i} \ne 0$, we have $\bar{Y}_{\tilde{B},i} = \sum_{n \in H_{\tilde{B},i}} Y_{i,n}/N_{\tilde{B},i}$. Then, by the arm elimination mechanism, there exists an arm $i_2 \in \mathcal{S}_{\tilde{B}}$ such that

$$\bar{Y}_{\tilde{B},i_2} - \bar{Y}_{\tilde{B},i_1} > \alpha_t = 4\rho h_t^{\kappa_0}. \tag{A.21}$$

For every arm $1 \le i \le l$, define $\bar{f}_{\tilde{B},i} = \sum_{n \in H_{\tilde{B},i}} f_i(X_n)/N_{\tilde{B},i}$ if $N_{\tilde{B},i} \ne 0$. Then, since $N_{\tilde{B},i_1} \ne 0$ and $N_{\tilde{B},i_2} \ne 0$,

$$\begin{aligned}
\bar{f}_{\tilde{B},i_2} - \bar{f}_{\tilde{B},i_1} &= \frac{\sum_{n \in H_{\tilde{B},i_2}} f_{i_2}(X_n)}{N_{\tilde{B},i_2}} - \frac{\sum_{n \in H_{\tilde{B},i_1}} f_{i_1}(X_n)}{N_{\tilde{B},i_1}} \\
&\le \max_{x \in \tilde{B}} f^*(x) - \frac{\sum_{n \in H_{\tilde{B},i_1}} f_{i_1}(X_n)}{N_{\tilde{B},i_1}} \\
&= \frac{\sum_{n \in H_{\tilde{B},i_1}} \big(\max_{x \in \tilde{B}} f^*(x) - f_{i_1}(X_n)\big)}{N_{\tilde{B},i_1}}.
\end{aligned} \tag{A.22}$$

Since $i_1 \in \mathcal{S}_{t,B,1}$, by Assumption 4.1, for every $x' \in \tilde{B}$, $\max_{x \in \tilde{B}} f^*(x) - f_{i_1}(x') \le 2\rho h_t^{\kappa}$. Therefore, we have by (A.22) that

$$\bar{f}_{\tilde{B},i_2} - \bar{f}_{\tilde{B},i_1} \le 2\rho h_t^{\kappa}. \tag{A.23}$$

By (A.21), (A.23) and the fact that both arms i_1 and i_2 are in $\mathcal{S}_{p_{t-1}(B)}$, we conclude that under $F_{t-1,B} \cap G^c_{t,B,1}$, there exists an arm $i \in \mathcal{S}_{p_{t-1}(B)}$ such that $N_{\tilde{B},i} \neq 0$ and

$$|\bar{Y}_{\tilde{B},i} - \bar{f}_{\tilde{B},i}| = \left| \frac{\sum_{n \in H_{\tilde{B},i}} \varepsilon_n}{N_{\tilde{B},i}} \right| > \rho h_t^{\kappa_0}. \tag{A.24}$$

Next, we assume that the event $F_{t-1,B} \cap G_{t,B,1} \cap G^c_{t,B,2}$ happens. Since $G^c_{t,B,2} = \{\mathcal{S}_{\tilde{B}} \subseteq \mathcal{S}_{t,B,2}\}^c$, there exists an arm $i_3 \in \mathcal{S}_{\tilde{B}}$ and some $\tilde{x} \in \tilde{B}$ such that $f^*(\tilde{x}) - f_{i_3}(\tilde{x}) > 8\rho h_t^{\kappa_0}$. Also, by event $G_{t,B,1}$, there exists an arm $i_4 \in \mathcal{S}_{\tilde{B}}$ such that $f^*(\tilde{x}) = f_{i_4}(\tilde{x})$. Therefore,

$$f_{i_4}(\tilde{x}) - f_{i_3}(\tilde{x}) > 8\rho h_t^{\kappa_0}. \tag{A.25}$$

Then, by Assumption 4.1, if $N_{\tilde{B},i_3} \neq 0$ and $N_{\tilde{B},i_4} \neq 0$,

$$\begin{aligned}
\bar{f}_{\tilde{B},i_4} - \bar{f}_{\tilde{B},i_3} &= \frac{\sum_{n \in H_{\tilde{B},i_4}} f_{i_4}(X_n)}{N_{\tilde{B},i_4}} - \frac{\sum_{n \in H_{\tilde{B},i_3}} f_{i_3}(X_n)}{N_{\tilde{B},i_3}} \\
&\geq \frac{\sum_{n \in H_{\tilde{B},i_4}} \left(f_{i_4}(\tilde{x}) - \rho h_t^{\kappa}\right)}{N_{\tilde{B},i_4}} - \frac{\sum_{n \in H_{\tilde{B},i_3}} \left(f_{i_3}(\tilde{x}) + \rho h_t^{\kappa}\right)}{N_{\tilde{B},i_3}} \\
&= f_{i_4}(\tilde{x}) - f_{i_3}(\tilde{x}) - 2\rho h_t^{\kappa} \\
&> 6\rho h_t^{\kappa_0},
\end{aligned} \tag{A.26}$$

where the last inequality follows by (A.25). Also, since $i_3, i_4 \in \mathcal{S}_{\tilde{B}}$ implies that arms i_3 and i_4 are not eliminated at the end of stage t in bin $\tilde{B}$, if $N_{\tilde{B},i_3} \neq 0$ and $N_{\tilde{B},i_3} \neq 0$,

$$|\bar{Y}_{\tilde{B},i_4} - \bar{Y}_{\tilde{B},i_3}| \leq \alpha_t = 4\rho h_t^{\kappa_0}. \tag{A.27}$$

By (A.26) and (A.27), we conclude that under $F_{t-1,B} \cap G_{t,B,1} \cap G^c_{t,B,2}$, if $N_{\tilde{B},i} \neq 0$ for all $i \in \mathcal{S}_{p_{t-1}(B)}$, there exists an arm $i \in \mathcal{S}_{\tilde{B}}$ such that

$$|\bar{Y}_{\tilde{B},i} - \bar{f}_{\tilde{B},i}| = \frac{\left|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n\right|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}. \tag{A.28}$$

Combining (A.24) and (A.28), we know that under event $F_{t-1,B} \cap A^c_{t,B}$, if $N_{\tilde{B},i} \neq 0$ for all $i \in \mathcal{S}_{p_{t-1}(B)}$, there exists an arm $i \in \mathcal{S}_{p_{t-1}(B)}$ such that

$$\frac{\left|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n\right|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}.$$

Also, in the rest of this proof, we let $P(\cdot) = P^{(0)}(\cdot)$. Consequently,

$$\begin{aligned}
&P(F_{t-1,B} \cap A^c_{t,B}) \\
\leq\ &P(\exists \text{ arm } i \in \mathcal{S}_{p_{t-1}(B)} \text{ such that } N_{\tilde{B},i} = 0)
\end{aligned}$$

$$
\begin{aligned}
&+ P\Big(\exists \text{ arm } i \in \mathcal{S}_{p_{t-1}(B)} \text{ such that } N_{\tilde{B},i} \neq 0 \text{ and } \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}\Big) \\
&\le l \max_{1\le i\le l} P\Big(N_{\tilde{B},i} = 0 \,\Big|\, \text{arm } i \in \mathcal{S}_{p_{t-1}(B)}\Big) \qquad \text{(A.29)} \\
&\quad + l \max_{1\le i\le l} P\Big(N_{\tilde{B},i} \neq 0,\ \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0} \,\Big|\, \text{arm } i \in \mathcal{S}_{p_{t-1}(B)}\Big).
\end{aligned}
$$

Given $1 \le i \le l$, for notation brevity, define $C_{t-1}^{(i)} = \{\text{arm } i \in \mathcal{S}_{p_{t-1}(B)}\}$. For the upper bound of the first term in (A.29), note that

$$
P\Big(N_{\tilde{B},i} = 0 \,\Big|\, C_{t-1}^{(i)}\Big) \le P\Big(\frac{N_{\tilde{B},i}}{N_t} \le \frac{\underline{c} h_t^d \tilde{\pi}_t}{2} \,\Big|\, C_{t-1}^{(i)}\Big) \le \exp\Big(-\frac{3\underline{c} N_t h_t^d \tilde{\pi}_t}{28}\Big), \quad \text{(A.30)}
$$

where the last inequality follows by Lemma A.2 and the fact that $P(X_n \in \tilde{B},\, I_n = i \,|\, C_{t-1}^{(i)}) \ge \underline{c} h_t^d \tilde{\pi}_t$ for all $\tilde{N}_{t-1}+1 \le n \le \tilde{N}_t$. To provide the upper bound for the second term in (A.29), define $H_{\tilde{B}} = \{n : \tilde{N}_{t-1} + 1 \le n \le \tilde{N}_t,\, X_n \in \tilde{B}\}$ to be the set of time points during stage t at which the covariates fall into bin $\tilde{B}$. Let $N_{\tilde{B}}$ be the size of $H_{\tilde{B}}$. Then,

$$
\begin{aligned}
&P\Big(N_{\tilde{B},i} \neq 0,\ \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0} \,\Big|\, C_{t-1}^{(i)}\Big) \\
&\le P\Big(\frac{N_{\tilde{B}}}{N_t} \le \frac{\underline{c} h_t^d}{2}\Big) + P\Big(N_{\tilde{B},i} \neq 0,\ \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0},\ \frac{N_{\tilde{B}}}{N_t} > \frac{\underline{c} h_t^d}{2} \,\Big|\, C_{t-1}^{(i)}\Big) \\
&\le P\Big(\frac{N_{\tilde{B}}}{N_t} \le \frac{\underline{c} h_t^d}{2}\Big) + E_c P_{X^t}\Big(N_{\tilde{B},i} \neq 0,\ \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0},\ \frac{N_{\tilde{B}}}{N_t} > \frac{\underline{c} h_t^d}{2}\Big), \qquad \text{(A.31)}
\end{aligned}
$$

where $P_{X^t}(\cdot)$ denotes the conditional probability given $(X_{N_{t-1}} + 1, X_{N_{t-1}} + 2, \cdots, X_{N_t})$, $C_{t-1}^{(i)}$ and $\{\hat{\kappa}^* = \kappa_0\}$, and $E_c(\cdot)$ denotes the conditional expectation given $C_{t-1}^{(i)}$ and $\{\hat{\kappa}^* = \kappa_0\}$. Since $P(X_n \in \tilde{B}) \ge \underline{c} h_t^d$, by Lemma A.2,

$$
P\Big(\frac{N_{\tilde{B}}}{N_t} \le \frac{\underline{c} h_t^d}{2}\Big) \le \exp\Big(-\frac{3\underline{c} N_t h_t^d}{28}\Big). \qquad \text{(A.32)}
$$

Note that under the event $\{N_{\tilde{B}}/N_t > \underline{c} h_t^d/2\}$, we have

$$
\begin{aligned}
&P_{X^t}\Big(N_{\tilde{B},i} \neq 0,\ \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}\Big) \\
&\le P_{X^t}\Big(\frac{N_{\tilde{B},i}}{N_{\tilde{B}}} \le \frac{\tilde{\pi}_t}{2}\Big) + P_{X^t}\Big(\frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0},\ \frac{N_{\tilde{B},i}}{N_{\tilde{B}}} > \frac{\tilde{\pi}_t}{2}\Big) \\
&\le \exp\Big(-\frac{3N_{\tilde{B}}\tilde{\pi}_t}{28}\Big) + \exp\Big(-\frac{N_{\tilde{B}}\tilde{\pi}_t \rho^2 h_t^{2\kappa_0}}{4(v^2 + c\rho h_t^{\kappa_0})}\Big), \qquad \text{(A.33)}
\end{aligned}
$$

where the last inequality follows by Lemma A.1, Lemma A.2 and the fact that $P(I_n = i \,|\, X_n \in \tilde{B}) \geq \tilde{\pi}_t$ for all $\tilde{N}_{t-1} + 1 \leq n \leq \tilde{N}_t$. Thus, by (A.33),

$$\begin{aligned} &P_{X^t}\Big(N_{\tilde{B},i} \neq 0,\ \frac{\big|\sum_{n\in H_{\tilde{B},i}} \varepsilon_n\big|}{N_{\tilde{B},i}} > \rho h_t^{\kappa},\ \frac{N_{\tilde{B}}}{N_t} > \frac{\underline{c} h_t^d}{2}\Big) \\ &\leq \begin{cases} 0 & \text{if } \frac{N_{\tilde{B}}}{N_t} \leq \frac{\underline{c} h_t^d}{2}, \\ \exp\Big(-\frac{3N_{\tilde{B}}\tilde{\pi}_t}{28}\Big) + \exp\Big(-\frac{N_{\tilde{B}}\tilde{\pi}_t \rho^2 h_t^{2\kappa_0}}{4(v^2 + c\rho h_t^{\kappa_0})}\Big) & \text{if } \frac{N_{\tilde{B}}}{N_t} > \frac{\underline{c} h_t^d}{2}. \end{cases} \end{aligned} \tag{A.34}$$

Combining (A.29)–(A.32) and (A.34), we have

$$\begin{aligned} &P(F_{t-1,B} \cap A_{t,B}^c) \\ &\leq l\Big\{\exp\Big(-\frac{3\underline{c} N_t h_t^d \tilde{\pi}_t}{28}\Big) + \exp\Big(-\frac{3\underline{c} N_t h_t^d}{28}\Big) + \exp\Big(-\frac{3\underline{c} N_t h_t^d \tilde{\pi}_t}{56}\Big) \\ &\qquad + \exp\Big(-\frac{\underline{c}\rho^2 \tilde{\pi}_t^2 N_t h_t^{2\kappa_0+d}}{16(v^2 + c\rho\tilde{\pi}_t h_t^{\kappa_0}/2)}\Big)\Big\} \\ &\leq l\Big\{3\exp\Big(-\frac{3\underline{c}\tilde{\pi}_t \tilde{\gamma}_t h_t^{-2\kappa_0} \log(N h_t^{2\kappa_0+d})}{56}\Big) + \exp\Big(-\frac{\underline{c}\rho^2 \tilde{\pi}_t \tilde{\gamma}_t \log(N h_t^{2\kappa_0+d})}{8(v^2 + c\rho)}\Big)\Big\}. \end{aligned}$$

It follows immediately by (5.2) that $P(F_{t-1,B} \cap A_{t,B}^c) \leq 4l/N h_t^{2\kappa_0+d}$.

Lastly, noting that $P(A_{1,B}^c) \leq 4l/N h_1^{2\kappa_0+d}$ can be derived by the same argument as that of $P(F_{t-1,B} \cap A_{t,B}^c) \leq 4l/N h_t^{2\kappa_0+d}$, we complete the proof of Lemma A.5. □

Appendix B: Simulations

As is discussed in the Introduction and Remark 3.1, the RAAE algorithm includes a randomized allocation procedure to allow users to further explore the response-covariate association using a user-specified regression method. In the following, we use numerical examples to illustrate the impact of randomized allocation on the algorithm performance. Specifically, we compare two different choices for π_n. One choice is $\pi_n = 1/l_B$, under which RAAE becomes analogous to ABSE since it tends to select each active arm an equal number of times. Alternatively, we can choose $\pi_n = 0.05 \wedge \frac{1}{l_B}$, which implies that Step 1.2 takes effect and the arm with the highest reward estimate is more likely to be pulled than other active arms. Consider the following two cases.

Case 1. Suppose a three-armed bandit with $d = 1$ generates 0-1 binary responses using the following (conditional) mean reward functions with $\kappa = 1$:

$$\begin{aligned} f_1(x) &= 0.7\exp(-30(x-0.2)^2) + 0.7\exp(-30(x-0.8)^2), \\ f_2(x) &= 0.45 - 0.3x, \\ f_3(x) &= 0.1 + 2(x-0.5)^2. \end{aligned}$$

The covariates X_n's are i.i.d uniform(0,1) and the time horizon N is 40,000.

TABLE 1
Simulation results to compare the two different choices of π_n for RAAE (values shown in parenthesis are standard errors)

		$\pi_n = 1/l_B$		$\pi_n = 0.05$	
	ρ	$\bar{r}_n$	$\bar{q}_n$	$\bar{r}_n$	$\bar{q}_n$
Case 1	0.5	0.0402 (0.0005)	0.273 (0.003)	0.0215 (0.0013)	0.186 (0.005)
	1.0	0.0939 (0.0006)	0.489 (0.002)	0.0239 (0.0020)	0.197 (0.002)
Case 2	0.5	0.0553 (0.0005)	0.198 (0.002)	0.0069 (0.0002)	0.025 (0.001)
$(m = 4)$	1.0	0.1827 (0.0004)	0.489 (0.001)	0.0217 (0.0001)	0.060 (0.001)
Case 2	0.5	0.0786 (0.0005)	0.361 (0.002)	0.0134 (0.0003)	0.072 (0.002)
$(m = 10)$	1.0	0.1673 (0.0003)	0.602 (0.001)	0.0233 (0.0002)	0.095 (0.001)
Case 2	0.5	0.0840 (0.0003)	0.477 (0.002)	0.0258 (0.0006)	0.162 (0.004)
$(m = 20)$	1.0	0.1378 (0.0002)	0.644 (0.001)	0.0259 (0.0002)	0.142 (0.001)
Case 2	0.5	0.0786 (0.0003)	0.562 (0.001)	0.0383 (0.0007)	0.292 (0.006)
$(m = 40)$	1.0	0.1031 (0.0001)	0.660 (0.001)	0.0327 (0.0004)	0.241 (0.003)

Case 2. Suppose a three-armed bandit with $d = 1$ generates 0-1 binary responses using the following (conditional) mean reward functions with $\kappa = 0.5$:

$$f_1(x) = \begin{cases} (-1)^k \big(x - \frac{2k}{m}\big)^{0.5} + 0.5, & \text{if } \frac{2k}{m} \le x \le \frac{2k+1}{m},\ k = 0, 1, \cdots, \frac{m}{2} - 1, \\ (-1)^k \big(\frac{2k+2}{m} - x\big)^{0.5} + 0.5, & \text{if } \frac{2k+1}{m} \le x \le \frac{2k+2}{m},\ k = 0, 1, \cdots, \frac{m}{2} - 1, \end{cases}$$

$f_2(x) = -f_1(x)$ and $f_3(x) = 0.5$, where $m = 4, 10, 20,$ or 40. All the other settings of Case 2 remain the same as that of Case 1.

In this focused illustration with RAAE, κ is known to the user and set $n_0 = 20$, $\tilde{\gamma}_t = 1$, $\rho = 0.5$ or 1. The Nadaraya-Watson regression with Gaussian kernel is applied as the user-specified regression modeling method for each active arm in Step 1.2, and at each time point n, the bandwidth is $N_{i,n}^{-1/(2\kappa+d)}$, where $N_{i,n}$ is the total number of times arm i is pulled before the time point n. To compare the performance of using $\pi_n = 1/l_B$ versus $\pi_n = 0.05$, we run the algorithm 100 times for each choice of π_n. The averaged per-round regret $\bar{r}_N$ and the averaged inferior sampling rate $\bar{q}_N$ are computed over the 100 runs. All the numerical work was implemented in C++ and the code is available upon request.

Based on the results summarized in Table 1, we can see that in both cases, the choice of $\pi_n = 0.05$ (which uses the information obtained in Step 1.2 with Nadaraya-Watson regression) outperforms the choice of $\pi_n = 1/l_B$ (which ignores Step 1.2 and pulls each active arm with equal probability). Here, the RAAE algorithm shows its practical potential to improve algorithm performance by effectively employing user-specified regression modeling methods such as the Nadaraya-Watson regression to differentiate the active arms.

Acknowledgments

We would like to thank Richard Nickl, the Editor and two anonymous referees for their valuable suggestions. The research was supported by the NSF Grant DMS-1106576.

References

Audibert, J.-Y. and Tsybakov, A. B. (2005). Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*. MR2336861

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics* **35** 608–633. MR2336861

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

Auer, P., Ortner, R. and Szepesvári, C. (2007). Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of 20th Annual Conference on Learning Theory*. MR2397605

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, New York. MR0813698

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375. MR1653272

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

Bull, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics* **6** 1490–1516. MR2988456

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK. MR2409394

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

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

Even-Dar, E., Mannor, S. and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research* **7** 1079–1105. MR2274398

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics* **38** 1122–1170. MR2604707

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. Wiley, New York. MR0996417

Goldenshluger, a. and Zeevi, A. (2009). Woodrooofe's one-armed bandit problem revisited. *The Annals of Applied Probability* **19** 1603–1633. MR2538082

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems* **3** 230–261. MR3353472

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics*. Springer, New York. MR1618204

Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *The Annals of Statistics* **39** 2383–2409. MR2906872

Kleinberg, R., Slivkins, A. and Upfal, E. (2007). Multi-armed bandits in metric spaces. In *Proceedings of 40th Symposium on Theory of Computing*. MR2582691

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6** 4–22. MR0776826

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

Lepski, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications* **35** 454–466. MR1091202

Lepski, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics* **25** 929–947. MR1447734

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

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics* **25** 2547–2554. MR1604412

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

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics* **27** 1808–1829. MR1765618

May, B. C., Korda, N., A. Lee and Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* **13** 2069–2106. MR2956353

Perchet, V. and Rigollet, P. (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics* **41** 693–721. MR3099118

Qian, W. and Yang, Y. (2016). Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research* accepted.

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

Robbins, H. (1954). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58** 527–535. MR0050246

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research* **35** 395–411. MR2674726

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

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* **32** 135–166. MR2051002

Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association* **74** 799–806. MR0556471

Yang, Y. and Zhu, D. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics* **30** 100–121. MR1892657