

**Nonparametric Estimation and Model Combination in a
Bandit Problem with Covariates**

**A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY**

Wei Qian

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy**

Advisor: Dr. Yuhong Yang

July, 2014

© Wei Qian 2014
ALL RIGHTS RESERVED

Acknowledgements

I have received enormous support from a number of individuals over the past six years. I am very grateful that Professor Yuhong Yang has been my mentor, and guided me through several research projects including this dissertation research. His knowledge, generosity and patience have profoundly influenced me, and reshaped my life forever in many important aspects. The extremely rewarding endeavor with him in the past few years has ignited my ever-lasting interests and passion for statistics research and applications. What I have learned from him will certainly continue to be of great help for many years to come.

I would like to thank Professor Lan Wang for chairing my thesis committee, and being an excellent instructor for multiple fundamentally important statistics courses I took. I would also like to thank Professor Adam Rothman and Professor Bert Fristedt for spending their time serving as my committee members, reviewing my thesis work, and providing insightful suggestions on the research and career development. I am also grateful to Professor Hui Zou, Professor Dennis Cook and Professor Tiefeng Jiang for helping to broaden my research horizons and further my research interests in various directions.

I have also been the beneficiary of the very supporting environment in the School of Statistics at the University of Minnesota. In addition to all faculty and staff members in our school, I would like to thank my fellow classmates for helping me in every step of my academic development. I truly appreciate their friendship and help from the very first day of my study here.

Finally, I want to thank my parents and Shanshan for their love and support. The happiness and trust they give me is always my source of strength that keeps me moving forward.

Dedication

To my parents Jianping Qian and Minhe Wei for their unwavering support and encouragement.

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. Under a flexible problem setup, we focus on a sequential randomized allocation strategy, under which, the “plug-in” regression methods for the estimation of mean reward functions play an important role in the algorithm performance. In the first part of the dissertation, we study the kernel estimation based randomized allocation strategy, and establish asymptotic strong consistency and finite-time regret analysis.

In addition, although many nonparametric and parametric estimation methods in supervised learning may be applied to the randomized allocation strategy in a convenient “plug-in” fashion, guidance on how to choose among these estimation methods is generally unavailable. In the second part of the dissertation, we study a model combining allocation strategy for adaptive performance, and establish its asymptotic strong consistency. Simulations and a real data evaluation are conducted to illustrate the performance of the proposed combining strategy.

In the existing literature of nonparametric bandit problem with covariates, it is generally assumed that the smoothness parameters of a Hölder condition for the reward functions are known. Also, the finite-time regret analysis in the first part of the dissertation remains minimax sub-optimal. In the third part of the dissertation, we address these two issues by proposing a multi-stage randomized allocation strategy with arm elimination. In particular, when the smoothness parameter is unknown, we equip the algorithm with a smoothness parameter selector based on Lepski’s method, and show that the regret minimax rate is achieved up to a logarithmic factor.

Contents

Acknowledgements	i
Dedication	ii
Abstract	iii
List of Tables	vi
List of Figures	vii
1 Introduction	1
1.1 Multi-Armed Bandit Problem	1
1.2 Bandit Problem with Covariates	4
2 Kernel Estimation in a Bandit Problem with Covariate	8
2.1 Problem Setup	8
2.2 Algorithm	9
2.3 Kernel Regression Procedures	12
2.3.1 Strong Consistency	12
2.3.2 Finite-Time Regret Analysis	14
2.3.3 Dimension Reduction	16
2.4 Proofs	17
2.4.1 Proof of Theorem 2.1	17
2.4.2 Proofs of Theorem 2.2 and Corollary 2.2	24

3	Model Combining Based Allocation	28
3.1	Strong Consistency	28
3.2	Simulations	30
3.2.1	Univariate Covariate	30
3.2.2	Multivariate Covariates with Dimension Reduction	35
3.3	Web-Based Personalized News Article Recommendation	35
3.4	Proof of Theorem 3.1	39
4	Adaptive Performance in Randomized Allocation with Arm Elimination	43
4.1	Algorithms	45
4.2	Smoothness Parameter Selector	48
4.3	Finite-Time Regret Analysis	49
4.4	Discussion	50
4.5	Proofs	53
4.5.1	Proof of Proposition 4.1	53
4.5.2	Proof of Theorem 4.1	57
5	Conclusion	67
	References	69

List of Tables

3.1	(Case 1) Weights of bandwidth choices for Nadaraya-Watson regressions from the last repeat	32
3.2	(Case 2) Weights for combining Nadaraya-Watson (NW) regression and linear regression	34
3.3	Comparing the estimated dimension reduction matrix $\hat{B}_{2,N}^*$ for the second arm between SIR and CIS-SIR.	36
3.4	Normalized CTRs of various algorithms on the news article recommendation dataset. CTRs are normalized with respect to the random algorithm.	38

List of Figures

3.1	(Case 1) Combining Nadaraya-Watson regressions with different bandwidth choices. Left panel: averaged per-round regret. Right panel: averaged inferior sampling rate.	32
3.2	(Case 1) Averaged per-round regret from combining different methods. Left panel: combining K -nearest neighbor methods with different choices of K . Right panel: combining different nonparametric methods.	33
3.3	Averaged per-round regret from combining different methods. Left panel: (Case 2) combining Nadaraya-Watson regression and linear regression. Right panel: (the multivariate covariate case) comparing SIR and CIS-SIR.	34
3.4	Boxplots of normalized CTRs of various algorithms on the news article recommendation dataset. 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.	39

Chapter 1

Introduction

1.1 Multi-Armed Bandit Problem

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” the gambler can pull to receive the reward. The distribution of reward for each arm is unknown to the gambler. At each round of the game, the gambler is allowed to play one and only one of these arms. 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). Recent overviews of Cesa-Bianchi and Lugosi (2006) and Bubeck and Cesa-Bianchi (2012) include many of its extensions.

An algorithm for bandit problem usually involves a trade-off between “exploration” and “exploitation”. On the one hand, we want to pull the arm of each machine as many times as possible so as to explore the true reward distribution of each arm; on the other hand, we want to exploit the information obtained from the previous arm pullings and play the “best” arm and realize the gain. Clearly, exploration alone or exploitation alone cannot result in an optimal strategy: excessive pulling of all arms would give results no

better than a completely randomized strategy, while constantly pulling the “best” arm is certainly sub-optimal if the exploitation decision is made based on the wrong initial information about the reward distributions.

Next, we review some main algorithms for the standard multi-armed bandit problem. Suppose in a classic l -armed bandit problem, the rewards associated with the bandit machines have expected values μ_1, \dots, μ_l and variance $\sigma_1^2, \dots, \sigma_l^2$. With a time horizon N , let $\hat{\mu}_1(n), \dots, \hat{\mu}_l(n)$, $n = 0, 1, \dots, N$, be the empirical estimate (typically calculated as the sample mean) of μ_1, \dots, μ_l at time n . Define the expected cumulative regret R_n of an algorithm to be the difference between the expected reward by always pulling the best arm and the expected reward of this algorithm from time 0 to time n . We can consider the following algorithms.

Pure Greedy. Pure greedy is the simple-minded exploitation-only algorithm. After initial random exploration, the player always pulls the arm with the highest empirical estimate of the expected reward. Clearly, this strategy can suffer from the insufficient exploration at the initial stage.

Upper Confidence Bound (UCB). UCB algorithm is first introduced by Lai and Robbins (1985). At time n , the upper bound index for the expected reward functions are calculated, and the arm with the highest upper bound index is pulled. In this classical paper, they show that for some specific family of reward distributions, any suboptimal arm i satisfies

$$E[N_i(n)] \leq \left(\frac{1}{D_i} + o(1) \right) \log n$$

where $N_i(n)$ is the number of times arm i is pulled during the first n plays, and D_i is the Kullback-Leibler divergence between the reward density of arm i and the reward density of the optimal arm. They also show that this upper bound is asymptotically optimal. Auer et al. (2002) proposes several computationally simpler UCB algorithms, and show that they achieve logarithmic regret uniformly instead of only asymptotically. In addition, the upper confidence bound used in their algorithms have the form very similar to the upper confidence bound of a regular sample mean for i.i.d observations.

Exponential Weighting. Let $p_i(n)$ denote the probability of pulling arm i , $1 \leq i \leq l$, at time n . For exponential weighting algorithms, $p_i(n+1)$ is updated based on the action and observation at previous time points. The SoftMax strategy (Luce, 1959) is a simple version of exponential weighting, which pulls the arm i at time n with probability

$$p_i(n) = \frac{\exp(\hat{\mu}_i(n)/\tau)}{\sum_{k=1}^l \exp(\hat{\mu}_k(n)/\tau)},$$

where τ is a tuning parameter. A more complicated version called “exponential-weight algorithm for exploration and exploitation” (Exp3) is first introduced by Auer (2002). With the implicit assumption that an infinite sequence of time-dependent reward has been assigned to each bandit machine, Exp3 starts with the weights $w_i(1) = 1$ for $i = 1, \dots, l$, and updates the probability by

$$p_i(n) = (1 - \gamma) \frac{w_i(n)}{\sum_{k=1}^l w_k(n)} + \frac{\gamma}{l}, \quad 1 \leq i \leq l.$$

If arm i is pulled at time n and gives the corresponding reward $y_i(n)$, update the weights by $w_i(n+1) = w_i(n) \exp(\gamma y_i(n)/p_i(n))$. Otherwise, update the weights by $w_i(n+1) = w_i(n)$. The tuning parameter $\gamma \in (0, 1]$ can be chosen by the user.

ϵ -Greedy. Different from pure greedy strategy, ϵ -greedy implements an enforced randomization strategy, pulling the arm of the highest empirical estimate with probability $1 - \epsilon$ while pulling the rest of the arms with equal probability $\frac{\epsilon}{l-1}$. With a constant ϵ , it is clear that this strategy can be inefficient since exploration of the sub-optimal arms continues even after the optimal arm is apparent. A closely related variant, which is sometimes called ϵ -decreasing, overcomes such inefficiency by letting $\epsilon \rightarrow 0$ as $n \rightarrow \infty$. By appropriately choosing a decreasing sequence of ϵ , Auer et al. (2002) shows that the expected regret has an optimal bound of $O(\log n)$. Extensive numerical study by Vermorel and Mohri (2005) also show that ϵ -decreasing strategy performs rather well compared with other sophisticated algorithms.

1.2 Bandit Problem with Covariates

Different variants of the bandit problem motivated by real applications have been studied extensively very recently. 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 $x \in \mathcal{R}^d$ at each round of the game, and the expected reward of each bandit arm given x can have a functional form $f_i(x)$, $i = 1 \cdots, l$. We call this variant **multi-armed bandit problem with covariates**, or MABC for its abbreviation. 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 to improve the overall well-being of the patients.

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 Woodroffe (1979) under a somewhat restrictive setting. With settings more flexible than that of Woodroffe (1979), a linear response bandit problem 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 are 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). Yang and Zhu (2002) 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 (2008) 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.

Another important line of development in 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 metric space. Like MABC, the setting with the arm space can be studied from both parametric (linear) framework and nonparametric framework. Examples of the 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 Lipschitz smoothness conditions are Auer et al. (2007), Kleinberg et al. (2007) and Bubeck et al. (2011). 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 follows the nonparametric framework of MABC in Yang and Zhu (2002) and Rigollet and Zeevi (2010) with finitely many arms. As one motivation of our work, previous nonparametric approaches to MABC are limited to simple averages of clusters of observed rewards, which gives only discontinuous estimated functions, while It is known that kernel methods can generate continuous estimates and potentially improve estimation efficiency for smooth targets. The kernel regression analysis results under i.i.d. or weak dependence settings are well-established in, e.g., Devroye, 1978; Härdle and Luckhaus, 1984; Hansen, 2008. One of our contributions is to show, under the MABC setting, that kernel methods enjoy estimation uniform strong consistency as well, which leads to strongly consistent allocation rules. In addition, with the help of the Hölder smoothness condition, we provide a finite-time regret analysis for the proposed randomized allocation strategy. 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. Moreover, with a model combining strategy along with the dimension reduction technique to be introduced later, the kernel method based allocation strategy can be quite flexible with potential practical use.

One natural and interesting issue in the randomized allocation strategy in MABC is how to choose the modeling methods among numerous nonparametric and parametric estimation approaches. The motivation of such question shares the flavor of model aggregation/combining in statistical learning, which targets to achieve prediction performance almost as well as the best of the prediction candidates (see, e.g., Audibert, 2009 and references therein). In the bandit problem literature, model combining is also quite relevant to the adversary bandit problem (Auer et al., 2003). As a recent example, Maillard and Munos (2011) study the history-dependent adversary bandit to target the best among a pool of history class mapping strategies.

As an empirical solution to our attempt to choose 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. As another contribution, 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.

Since the Hölder smoothness condition is usually assumed for MABC from nonparametric perspective, the last question this dissertation attempts to address is whether we can achieve a guaranteed regret upper bound without the prior knowledge of the smoothness parameter. Our solution to such question is closely related to the adaptive nonparametric estimation technique pioneered by Lepski (1990). The “Lepski-type” method has recently been studied to establish the adaptive confidence bands for density estimation and regression problems in Giné and Nickl (2010), Hoffmann and Nickl (2011) and Bull, 2012. Their “self-similarity” condition is employed here to study the adaptive performance of the proposed MABC algorithm. By imbedding the “Lepski-type” method and an arm-elimination subroutine (Even-Dar et al., 2006; Perchet and Rigollet,

2013) into the randomized allocation strategy, we show that the resulting cumulative regret adaptively achieves the minimax rate up to a logarithmic factor.

This dissertation is organized as follows. Chapter 2 introduces the problem setup for MABC and a randomized allocation strategy with model combination. In particular, we focus on the kernel estimation method, and study the strong consistency and the finite-time regret analysis for the proposed algorithm. In Chapter 3, we study the asymptotic property of the model combining allocation strategy, and evaluate its empirical performance by simulations and a web-based news article recommendation dataset. In Chapter 4, we propose a randomized allocation strategy with arm elimination to show the adaptive performance when combined with the “Lepski-type” method.

Chapter 2

Kernel Estimation in a Bandit Problem with Covariate

2.1 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, and (X_n, ε_n) are independent of the earlier observations. Different from Yang and Zhu (2002), we shall not assume that the error ε_n and the covariate X_n are independent. 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.

The errors ε_n are often assumed to have a bounded support in bandit problem literature. We will see that such an assumption can be avoided to allow distributions with other types of tails. When dealing with a continuous response, this weaker requirement substantially enhance applicability of the results in real problems.

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 $i^*(x) = \operatorname{argmax}_{1 \leq i \leq l} f_i(x)$ and $f^*(x) = f_{i^*(x)}(x)$. 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, \dots, 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 $\sum_{j=1}^n f_{I_j}(X_j)$. Thus we can measure the performance of an allocation rule η by the cumulative regret

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

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

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

To maintain the readability for the rest of this chapter, 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.

2.2 Algorithm

In this section, we present the randomized allocation strategy. For convenience, a model combining procedure is imbedded into the algorithm, which will be discussed in Chapter 3. 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 $Z^{n,i}$. 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 $\Delta = \{\delta_1, \dots, \delta_m\}$. Let $\hat{f}_{i,n,r}$ denote the regression estimate of procedure δ_r based on $Z^{n,i}$, 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 $I_1 = 1, I_2 = 2, \dots, I_l = l, I_{l+1} = 1, \dots, I_{2l} = l, \dots, I_{(n_0-1)l+1} = 1, \dots, I_{n_0l} = l$.

STEP 2. Initialize the weights and the error variance estimates. For $n = n_0l + 1$, 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 $n = n_0l + 1$, based on the current data $Z^{n,i}$, 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 $\hat{i}_n = \operatorname{argmax}_{1 \leq i \leq l} \hat{f}_{i,n}(X_n)$ (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

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

$$\hat{v}_{i,n+1,r} = \frac{\sum_{k=n_0l+1}^n (Y_{I_k,k} - \hat{f}_{I_k,k,r}(X_k))^2 I(I_k = i)}{\sum_{k=n_0l+1}^n I(I_k = i)}, \quad 1 \leq r \leq m,$$

and

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

where $I(\cdot)$ is the indicator function.

STEP 7. Repeat steps 3 - 6 for $n = n_0l + 2, n_0l + 3, \dots$, 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. 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.

2.3 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 sub-procedure is described in section 2.3.3.

2.3.1 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 $\|\hat{f}_{i,n,r} - f_i\|_\infty \rightarrow 0$ a.s. as $n \rightarrow \infty$. In the following, we do not assume the boundedness of the observed reward in our MABC setup.

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 $n \geq 1$,*

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

almost surely.

Assumption 0 means that the error distribution, which could depend on the covariates, satisfies 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 be used throughout this dissertation.

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 $u = (u_1, u_2, \dots, u_d) \in R^d$, define $\|u\| = \max\{|u_1|, |u_2|, \dots, |u_d|\}$. Consider two natural conditions on the mean reward functions and the covariate density as follows.

Assumption 2.1. *The mean reward 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 2.2. *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 $\underline{c} \leq p(x) \leq \bar{c}$ for some positive constants $\underline{c} \leq \bar{c}$.*

In addition, consider a multivariate nonnegative kernel function $K(u) : R^d \rightarrow R$ that satisfies both Lipschitz and bounded support conditions. We further assume $K(u)$ has bounded support, is uniformly upper bounded, and is bounded away from zero over a certain region around the origin.

Assumption 2.3. *For some constants $0 < \lambda < \infty$, we have $K(u) = 0$ for $\|u\| > L$, and*

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

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

Assumption 2.4. *There exist constants $L_1 \leq L$, $c_3 > 0$ and $c_4 \geq 1$ such that $K(u) = 0$ for $\|u\| > L$, $K(u) \geq c_3$ for $\|u\| \leq L_1$ and $K(u) \leq c_4$ for all $u \in R^d$.*

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

$$\hat{f}_{i,n+1}(x) = \frac{\sum_{j \in J_{i,n+1}} Y_{i,j} K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)}. \quad (2.1)$$

Theorem 2.1. *Suppose Assumptions 0-2.4 are satisfied. If h_n and π_n are chosen to satisfy $h_n \rightarrow 0$, $\pi_n \rightarrow 0$ and*

$$\frac{nh_n^{2d} \pi_n^4}{\log n} \rightarrow \infty,$$

then the Nadaraya-Watson estimators defined in (2.1) are strongly consistent in L_∞ norm for the functions f_i .

Together with Theorem 3.1 of the next chapter, it is an immediate consequence that including kernel methods in our allocation strategy can achieve strong consistency for MABC when they are properly combined with other candidate regression methods. 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 2.3 and Theorem 2.1 in the Appendix).

2.3.2 Finite-Time Regret Analysis

Next, we provide the 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

$$\omega(h; f_i) = \sup\{|f_i(x_1) - f_i(x_2)| : |x_{1k} - x_{2k}| \leq h \text{ for all } 1 \leq k \leq d\},$$

where x_{1k} and x_{2k} are the k th element of the vectors x_1 and x_2 , respectively. For technical convenience of guarding against the situation that the denominator of (2.1) is extremely small (which might occur with a non-negligible probability due to arm selection), in this subsection, we replace $K(\cdot)$ with the uniform kernel $I(\|u\| \leq L)$ when $\sum_{j \in J_{i,n+1}} K(\frac{x-X_j}{h_n}) < c_5 \sum_{j \in J_{i,n+1}} I(\|x - X_j\| \leq Lh_n)$ 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}_δ by

$$n_\delta = \min\left\{n > n_0 l : \sqrt{\frac{16v^2 \log(8lN^2/\delta)}{cn(2Lh_n)^d \pi_n}} \leq \frac{c_5 v^2}{c} \text{ and } \exp\left(-\frac{3cn(2Lh_n)^d \pi_n}{56}\right) \leq \frac{\delta}{4lN}\right\}. \quad (2.2)$$

Under the condition that $\lim_{n \rightarrow \infty} nh_n^d \pi_n / \log n = \infty$, we can see from (2.2) that $n_\delta / N \rightarrow 0$ as $N \rightarrow \infty$. As a result, if the total time horizon is long enough, we have $N > n_\delta$.

Theorem 2.2. *Suppose Assumptions 0-2.2 and 2.4 are satisfied, and assume $N > n_\delta$. then with probability larger than $1 - 2\delta$, the cumulative regret $R_N(\eta)$ satisfies*

$$R_N(\eta) < An_\delta + \sum_{n=n_\delta}^N \left(2 \max_{1 \leq i \leq l} \omega(Lh_n; f_i) + \frac{C_{N,\delta}}{\sqrt{nh_n^d \pi_n}} + (l-1)\pi_n\right) + A\sqrt{\frac{N}{2}} \log\left(\frac{1}{\delta}\right), \quad (2.3)$$

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

It is interesting to see from the right hand side of (2.3) 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: $\max_{1 \leq i \leq l} \omega(Lh_n; f_i)$ 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 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. Furthermore, we consider a smoothness assumption of the mean reward functions as follows.

Assumption 2.5. *There exist positive constants ρ and $\kappa \leq 1$ such that for each reward function f_i , the modulus of continuity satisfies*

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

Clearly, when $\kappa = 1$, Assumption 2.5 becomes Lipschitz continuity. As an immediate consequence of Theorem 2.2 and Assumption 2.5, 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 2.1. *Suppose Assumptions 0-2.2 and 2.4 are satisfied, and 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 $R_N(\eta)$ satisfies*

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

where $C_{N,\delta}^* = \sqrt{16c_4^2v^2(l-1)\log(8lN^2/\delta)/2^d c_5^2 \mathfrak{C}}$.

In Corollary 2.1, 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 2.2, which gives the upper bound for the mean of $R_N(\eta)$. Note that by definition of n_δ , the condition $N > n_{\delta^*}$ in Corollary 2.2 is satisfied if N is large enough.

Corollary 2.2. *Suppose Assumptions 0-2.2 and 2.4 are satisfied, and 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 $ER_N(\eta)$ satisfies*

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

The derived regret cumulative rate in Corollary 2.2 is suboptimal in the minimax sense (Perchet and Rigollet, 2013). Specifically, 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) (after

ignoring the margin condition). 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 strategy can be quite flexible with potential practical use.

2.3.3 Dimension Reduction

Recall that Z^n is the set of observations $\{(X_j, I_j, Y_{I_j, j}), 1 \leq j \leq n-1\}$, and $Z^{n,i}$ is the subset of Z^n where $I_j = i$. Then $M_{i,n}$ is the number of observations in $Z^{n,i}$. Let $X^{n,i}$ be the $M_{i,n} \times d$ design matrix consisting of all covariates in $Z^{n,i}$, and let $Y^{n,i} \in R^{M_{i,n}}$ be the observed reward vector corresponding to $X^{n,i}$. 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 $(X^{n,i}, Y^{n,i})$ first to obtain lower dimensional covariates before using kernel estimation.

Specifically, suppose for each arm i , there exists a reduction function $s_i : R^d \rightarrow R^{r_i}$ ($r_i < d$), such that $f_i(x) = g_i(s_i(x))$ for some function $g_i : R^{r_i} \rightarrow R$. 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 $\text{span}(B_i)$ 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 $(X^{n,i}, Y^{n,i})$ to obtain $\hat{B}_{i,n}$. In practice, it is convenient to have $X^{n,i}$ work on Z -scale (i.e., the sample mean is zero and the sample covariance matrix is the identity matrix). Suppose the Nadaraya-Watson estimation is used with $K_i(u) : R^{r_i} \rightarrow R$ 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 $X^{n,i}$ to the Z -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 $X^{n,i}$, respectively.

Step 2. Apply a dimension reduction method to $(X_*^{n,i}, Y^{n,i})$ to obtain the estimated $d \times r_i$ dimension reduction matrix $\hat{B}_{i,n}^*$, where $\hat{B}_{i,n}^{*T} \hat{B}_{i,n}^* = I_{r_i}$.

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 Z -scale. The Nadaraya-Watson estimator of $f_i(x)$ is

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

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. The numerical examples are given in the next chapter to show the performance of SIR and CIS-SIR.

2.4 Proofs

2.4.1 Proof of Theorem 2.1

Lemma 2.1. *Suppose $\{\mathcal{F}_j, j = 1, 2, \dots\}$ 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 $j \geq 1$, $E(|\varepsilon_j|^k | \mathcal{F}_j) \leq k! v^2 c^{k-2} / 2$, then for every $\epsilon > 0$ and every integer*

$n \geq 1$,

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

Proof of Lemma 2.1. Note that

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

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

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

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

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

where the last inequality is obtained by minimization over t . This completes the proof of Lemma 2.1. \square

Lemma 2.2. *Suppose $\{\mathcal{F}_j, j = 1, 2, \dots\}$ 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*

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

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

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

Lemma 2.2 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 2.3. 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}_j . By Bernstein's inequality,

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

Also, it is not hard to show that $\sum_{j=1}^n W_j$ is stochastically no smaller than $\sum_{j=1}^n \tilde{W}_j$, that is, for every t , $P(\sum_{j=1}^n W_j > t) \geq P(\sum_{j=1}^n \tilde{W}_j > t)$. Thus, Lemma 2.2 holds. \square

Lemma 2.3. *Under the settings of the kernel estimation in section 2.3.1, given arm i and a cube $A \subset [0, 1]^d$ with side width h , if Assumptions 0, 2.3 and 2.4 are satisfied, then for any $\epsilon > 0$,*

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

Proof of Lemma 2.3. 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 $G(x) = \sum_{j=1}^n \varepsilon_j W_j K\left(\frac{x - X_j}{h_n}\right)$. 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 $\|x - x'\| \leq \gamma_k/2$. Let $\tau_k(x) = \text{argmin}_{x' \in F_k} \|x - x'\|$ be the closest point to x in the net F_k . With the sequence F_0, F_1, F_2, \dots of $\gamma_0/2, \gamma_1/2, \gamma_2/2, \dots$ nets in A , it is easy to see that for every $x \in A$, $\|\tau_k(x) - \tau_{k-1}(x)\| \leq \gamma_k/2$ and $\lim_{k \rightarrow \infty} \tau_k(x) = x$.

Thus, by the continuity of the kernel function, we have $\lim_{k \rightarrow \infty} G(\tau_k(x)) = G(x)$. It follows that

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

Thus,

$$\begin{aligned} & P\left(\sup_{x \in A} G(x) > \frac{n\epsilon}{1 - 1/\sqrt{2}}\right) \\ &= P\left(\sup_{x \in A} \left\{G(\tau_0(x)) + \sum_{k=1}^{\infty} [G(\tau_k(x)) - G(\tau_{k-1}(x))]\right\} > \sum_{k=0}^{\infty} \frac{n\epsilon}{2^{k/2}}\right) \\ &\leq P\left(\sup_{x \in A} G(\tau_0(x)) > n\epsilon\right) + \sum_{k=1}^{\infty} P\left(\sup_{x \in A} [G(\tau_k(x)) - G(\tau_{k-1}(x))] > \frac{n\epsilon}{2^{k/2}}\right) \\ &\leq P\left(\sup_{x \in F_0} G(x) > n\epsilon\right) + \sum_{k=1}^{\infty} P\left(\sup_{\substack{x_2 \in F_k, x_1 \in F_{k-1} \\ \|x_2 - x_1\| \leq \gamma_k/2}} [G(x_2) - G(x_1)] > \frac{n\epsilon}{2^{k/2}}\right) \\ &\leq \text{card}(F_0) \max_{x \in F_0} P(G(x) > n\epsilon) + \sum_{k=1}^{\infty} 2^d \text{card}(F_{k-1}) \max_{\substack{x_2 \in F_k, x_1 \in F_{k-1} \\ \|x_2 - x_1\| \leq \gamma_k/2}} P\left(G(x_2) - G(x_1) > \frac{n\epsilon}{2^{k/2}}\right), \end{aligned} \tag{2.4}$$

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 $\|x_2 - x_1\| \leq \gamma_k/2$. Given $x \in F_0$, since $|W_j K(\frac{x - X_j}{h})| \leq c_4$ almost surely for all $j \geq 1$, it follows by Lemma 2.1 that

$$P(G(x) > n\epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2c_4^2(v^2 + c\epsilon/c_4)}\right). \tag{2.5}$$

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

$$\left|K\left(\frac{x_2 - X_j}{h}\right) - K\left(\frac{x_1 - X_j}{h}\right)\right| \leq \frac{\lambda \|x_2 - x_1\|}{h} \leq \frac{\lambda \gamma_k}{2h} = \frac{\lambda}{2^{k+1}}$$

almost surely, it follows by Lemma 2.1 that

$$\begin{aligned} P\left(G(x_2) - G(x_1) > \frac{n\epsilon}{2^{k/2}}\right) &= P\left(\sum_{j=1}^n \epsilon_j W_j \left[K\left(\frac{x_2 - X_j}{h}\right) - K\left(\frac{x_1 - X_j}{h}\right)\right] > \frac{n\epsilon}{2^{k/2}}\right) \\ &\leq \exp\left(-\frac{2^{k+2} n\epsilon^2}{2\lambda^2(v^2 + 2^{k/2+1} c\epsilon/\lambda)}\right). \end{aligned} \tag{2.6}$$

Thus, by (2.4), (2.5) and (2.6),

$$\begin{aligned}
& P\left(\sup_{x \in A} G(x) > \frac{n\epsilon}{1 - 1/\sqrt{2}}\right) \\
& \leq \exp\left(-\frac{n\epsilon^2}{2c_4^2(v^2 + c\epsilon/c_4)}\right) + \sum_{k=1}^{\infty} 2^{kd} \exp\left(-\frac{2^{k+2}n\epsilon^2}{2\lambda^2(v^2 + 2^{k/2+1}c\epsilon/\lambda)}\right) \\
& \leq \exp\left(-\frac{n\epsilon^2}{4c_4^2v^2}\right) + \exp\left(-\frac{n\epsilon}{4c_4c}\right) + \sum_{k=1}^{\infty} 2^{kd} \exp\left(-\frac{2^k n\epsilon^2}{\lambda^2 v^2}\right) + \sum_{k=1}^{\infty} 2^{kd} \exp\left(-\frac{2^{k/2}n\epsilon}{2\lambda c}\right).
\end{aligned}$$

This completes the proof of Lemma 2.3. \square

Proof of Theorem 2.1. Note that for each $x \in R^d$,

$$\begin{aligned}
|\hat{f}_{i,n+1}(x) - f_i(x)| &= \left| \frac{\sum_{j \in J_{i,n+1}} Y_{i,j} K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)} - f_i(x) \right| \\
&= \left| \frac{\sum_{j \in J_{i,n+1}} (f_i(X_j) + \varepsilon_j) K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)} - f_i(x) \right| \\
&= \left| \frac{\sum_{j \in J_{i,n+1}} (f_i(X_j) - f_i(x)) K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)} + \frac{\sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right)}{\sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)} \right| \\
&\leq \sup_{\{x,y:\|x-y\| \leq Lh_n\}} |f_i(x) - f_i(y)| + \left| \frac{\frac{1}{M_{i,n+1}h_n^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right)}{\frac{1}{M_{i,n+1}h_n^d} \sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right)} \right|, \tag{2.7}
\end{aligned}$$

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

$$\lim_{n \rightarrow \infty} \sup_{\{x,y:\|x-y\| \leq Lh_n\}} |f_i(x) - f_i(y)| = 0.$$

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

$$\sup_{x \in [0,1]^d} \left| \frac{\frac{1}{M_{i,n+1}h^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h}\right)}{\frac{1}{M_{i,n+1}h^d} \sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h}\right)} \right| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (2.8)$$

First, we want to show

$$\inf_{x \in [0,1]^d} \frac{1}{M_{i,n+1}h^d} \sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h}\right) > \frac{c_3 \underline{c} L_1^d \pi_n}{2}, \quad (2.9)$$

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

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

where the last inequality follows by Assumption 2.4. Consequently,

$$\begin{aligned} &P\left(\inf_{x \in \tilde{A}_k} \frac{1}{M_{i,n+1}h_n^d} \sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\right) \\ &\leq P\left(\inf_{x \in \tilde{A}_k} \frac{1}{nh_n^d} \sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\right) \\ &\leq P\left(\frac{c_3}{nh_n^d} \sum_{j=1}^n I(I_j = i, X_j \in \tilde{A}_k) \leq \frac{c_3 \underline{c} L_1^d \pi_n}{2}\right) \\ &= P\left(\sum_{j=1}^n I(I_j = i, X_j \in \tilde{A}_k) \leq \frac{cn(L_1 h_n)^d \pi_n}{2}\right). \end{aligned} \quad (2.10)$$

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 the

extended Bernstein inequality (Yang and Zhu, 2002, eq. 8) that

$$P\left(\sum_{j=1}^n I(I_j = i, X_j \in \tilde{A}_k) \leq \frac{cn(L_1 h_n)^d \pi_n}{2}\right) \leq \exp\left(-\frac{3cn(L_1 h_n)^d \pi_n}{28}\right). \quad (2.11)$$

Therefore,

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

where the last inequality follows by (2.10) and (2.11). With the condition $nh^{2d}\pi_n^4/\log n \rightarrow \infty$, we immediately obtain (2.9) by Borel-Cantelli lemma.

By (2.9), it follows that (2.8) holds if

$$\sup_{x \in [0,1]^d} \left| \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| = o(\pi_n). \quad (2.12)$$

In the rest of the proof, we want to show that (2.12) 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 $\epsilon > 0$,

$$\begin{aligned} & P\left(\sup_{x \in [0,1]^d} \left| \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| > \pi_n \epsilon\right) \\ & \leq B \max_{1 \leq k \leq B} P\left(\sup_{x \in A_k} \left| \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| > \pi_n \epsilon\right) \\ & \leq BP\left(\frac{M_{i,n+1}}{n} \leq \frac{\pi_n}{2}\right) + B \max_{1 \leq k \leq B} P\left(\sup_{x \in A_k} \left| \frac{1}{M_{i,n+1} h_n^d} \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| > \pi_n \epsilon, \frac{M_{i,n+1}}{n} > \frac{\pi_n}{2}\right) \\ & \leq BP\left(\frac{M_{i,n+1}}{n} \leq \frac{\pi_n}{2}\right) + B \max_{1 \leq k \leq B} P\left(\sup_{x \in A_k} \left| \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| > \frac{n\pi_n^2 h_n^d \epsilon}{2}\right) \\ & \leq B \exp\left(-\frac{3n\pi_n}{28}\right) + B \max_{1 \leq k \leq B} P\left(\sup_{x \in A_k} \left| \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| > \frac{n\pi_n^2 h_n^d \epsilon}{2}\right), \quad (2.13) \end{aligned}$$

where the last inequality follows by the extended Bernstein's inequality. Note that by Lemma 2.3,

$$\begin{aligned}
& P\left(\sup_{x \in A_k} \left| \sum_{j \in J_{i,n+1}} \varepsilon_j K\left(\frac{x - X_j}{h_n}\right) \right| > \frac{n\pi_n^2 h_n^d \epsilon}{2}\right) \\
& \leq 2 \exp\left(-\frac{(\sqrt{2}-1)^2 n \pi_n^4 h_n^{2d} \epsilon^2}{32c_4^2 v^2}\right) + 2 \exp\left(-\frac{(\sqrt{2}-1)n\pi_n^2 h_n^d \epsilon}{8\sqrt{2}c_4 c}\right) \\
& \quad + 2 \sum_{k=1}^{\infty} 2^{kd} \exp\left(-\frac{(\sqrt{2}-1)^2 2^k n \pi_n^4 h_n^{2d} \epsilon^2}{8\lambda^2 v^2}\right) + 2 \sum_{k=1}^{\infty} 2^{kd} \exp\left(-\frac{(\sqrt{2}-1)2^{k/2} n \pi_n^2 h_n^d \epsilon}{4\sqrt{2}\lambda c}\right).
\end{aligned} \tag{2.14}$$

Thus, by (2.13), (2.14) and the condition that $nh_n^{2d}\pi_n^4/\log n \rightarrow \infty$, (2.12) is an immediate consequence of Borel-Cantelli lemma. This completes the proof of Theorem 2.1. \square

2.4.2 Proofs of Theorem 2.2 and Corollary 2.2

Given $x \in [0, 1]^d$, $1 \leq i \leq l$ and $n \geq n_0 l + 1$, define $G_{n+1}(x) = \{j : 1 \leq j \leq n, \|x - X_j\| \leq Lh_n\}$ and $G_{i,n+1}(x) = \{j : 1 \leq j \leq n, I_j = i, \|x - X_j\| \leq Lh_n\}$. 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 2.4. *Suppose Assumptions 0, 2.1 and 2.4 are satisfied. Given $x \in [0, 1]^d$, $1 \leq i \leq l$ and $n \geq n_0 l + 1$, for every $\epsilon > \omega(Lh_n; f_i)$,*

$$P_{X^n}(|\hat{f}_{i,n+1}(x) - f_i(x)| \geq \epsilon) \leq \exp\left(-\frac{3M_{n+1}(x)\pi_n}{28}\right) + 4N \exp\left(-\frac{c_5^2 M_{n+1}(x)\pi_n(\epsilon - \omega(Lh_n; f_i))^2}{4c_4^2 v^2 + 4c_4 c(\epsilon - \omega(Lh_n; f_i))}\right), \tag{2.15}$$

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

Proof of Lemma 2.4. It is clear that if $M_{n+1}(x) = 0$, (2.15) trivially holds. Without loss of generality, assume $M_{n+1}(x) > 0$. Define the event $B_{i,n} = \left\{ \frac{1}{M_{i,n+1}(x)} \sum_{j \in J_{i,n+1}} K\left(\frac{x - X_j}{h_n}\right) \geq \right.$

$c_5\}$. Note that

$$\begin{aligned}
& P_{X^n}(|\hat{f}_{i,n+1}(x) - f_i(x)| \geq \epsilon) \\
& \leq P_{X^n}\left(\frac{M_{i,n+1}(x)}{M_{n+1}(x)} \leq \frac{\pi_n}{2}\right) + P_{X^n}\left(|\hat{f}_{i,n+1}(x) - f_i(x)| \geq \epsilon, \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}\right) \\
& \leq \exp\left(-\frac{3M_{n+1}(x)\pi_n}{28}\right) + P_{X^n}\left(|\hat{f}_{i,n+1}(x) - f_i(x)| \geq \epsilon, \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}, B_{i,n}\right) \\
& \quad + P_{X^n}\left(|\hat{f}_{i,n+1}(x) - f_i(x)| \geq \epsilon, \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}, B_{i,n}^c\right), \\
& =: \exp\left(-\frac{3M_{n+1}(x)\pi_n}{28}\right) + A_1 + A_2, \tag{2.16}
\end{aligned}$$

where the last inequality follows by the extended Bernstein inequality. Under $B_{i,n}$, by Assumption 2.4, the definition of the modulus continuity and the same argument as (2.7), we have

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

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

$$\begin{aligned}
A_1 &\leq P_{X^n}\left(\left| \sum_{j \in G_{i,n+1}(x)} \varepsilon_j K\left(\frac{x-X_j}{h_n}\right) \right| \geq c_5 M_{i,n+1}(x) (\epsilon - \omega(Lh_n; f_i)), \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}\right) \\
&\leq \sum_{\bar{n}=0}^N P_{X^n}\left(\left| \sum_{t=1}^{\bar{n}} \varepsilon_{\tilde{\sigma}_t} K\left(\frac{x-X_{\tilde{\sigma}_t}}{h_n}\right) \right| \geq c_5 \bar{n} (\epsilon - \omega(Lh_n; f_i)), \frac{M_{i,n+1}(x)}{M_{n+1}(x)} > \frac{\pi_n}{2}, M_{i,n+1}(x) = \bar{n}\right) \\
&\leq \sum_{\bar{n}=\lceil M_{n+1}(x)\pi_n/2 \rceil}^N P_{X^n}\left(\left| \sum_{t=1}^{\bar{n}} \varepsilon_{\tilde{\sigma}_t} K\left(\frac{x-X_{\tilde{\sigma}_t}}{h_n}\right) \right| \geq c_5 \bar{n} (\epsilon - \omega(Lh_n; f_i))\right) \\
&\leq \sum_{\bar{n}=\lceil M_{n+1}(x)\pi_n/2 \rceil}^N 2 \exp\left(-\frac{\bar{n}c_5^2(\epsilon - \omega(Lh_n; f_i))^2}{2c_4^2v^2 + 2c_4c(\epsilon - \omega(Lh_n; f_i))}\right) \\
&\leq 2N \exp\left(-\frac{c_5^2 M_{n+1}(x)\pi_n(\epsilon - \omega(Lh_n; f_i))^2}{4c_4^2v^2 + 4c_4c(\epsilon - \omega(Lh_n; f_i))}\right), \tag{2.17}
\end{aligned}$$

where the last to second inequality follows by Lemma 2.1 and the upper boundedness of the kernel function. By an argument similar to the previous two displays, it is not

hard to obtain that

$$A_2 \leq 2N \exp\left(-\frac{M_{n+1}(x)\pi_n(\epsilon - \omega(Lh_n; f_i))^2}{4v^2 + 4c(\epsilon - \omega(Lh_n; f_i))}\right). \quad (2.18)$$

Combining (2.16), (2.17), (2.18) and the fact that $0 < c_5 \leq 1 \leq c_4$, we complete the proof of Lemma 2.4. \square

Proof of Theorem 2.2. 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

$$\begin{aligned} & \sum_{n=n_0l+1}^N (f^*(X_n) - f_{I_n}(X_n)) \\ = & \sum_{n=n_0l+1}^N (f_{i^*(X_n)}(X_n) - \hat{f}_{i^*(X_n),n}(X_n) + \hat{f}_{i^*(X_n),n}(X_n) - \hat{f}_{\hat{i}_n}(X_n) + \hat{f}_{\hat{i}_n}(X_n) - f_{I_n}(X_n)) \\ \leq & \sum_{n=n_0l+1}^N (f_{i^*(X_n)}(X_n) - \hat{f}_{i^*(X_n),n}(X_n) + \hat{f}_{\hat{i}_n,n}(X_n) - \hat{f}_{\hat{i}_n}(X_n) + \hat{f}_{\hat{i}_n}(X_n) - f_{I_n}(X_n)) \\ \leq & \sum_{n=n_0l+1}^N \left(2 \sup_{1 \leq i \leq l} |\hat{f}_{i,n}(X_n) - f_i(X_n)| + AI(I_n \neq \hat{i}_n)\right). \end{aligned} \quad (2.19)$$

It can be seen from (2.19) 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 , $n \geq n_0l$ and $\epsilon > \omega(Lh_n; f_i)$,

$$\begin{aligned} & P\left(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \epsilon\right) \\ \leq & EP_{X_{n+1}}\left(M_{n+1}(X_{n+1}) \leq \frac{cn(2Lh_n)^d}{2}\right) \\ & + EP_{X_{n+1}}\left(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \epsilon, M_{n+1}(X_{n+1}) > \frac{cn(2Lh_n)^d}{2}\right). \end{aligned} \quad (2.20)$$

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

$$P_{X_{n+1}}\left(M_{n+1}(X_{n+1}) \leq \frac{cn(2Lh_n)^d}{2}\right) \leq \exp\left(-\frac{3cn(2Lh_n)^d}{28}\right). \quad (2.21)$$

By Lemma 2.4,

$$\begin{aligned} & P_{X_{n+1}} \left(|\hat{f}_{i,n+1}(X_{n+1}) - f_i(X_{n+1})| \geq \epsilon, M_{n+1}(X_{n+1}) > \frac{cn(2Lh_n)^d}{2} \right) \\ & \leq \exp\left(-\frac{3cn(2Lh_n)^d\pi_n}{56}\right) + 4N \exp\left(-\frac{c_5^2 cn(2Lh_n)^d\pi_n(\epsilon - \omega(Lh_n; f_i))^2}{8c_4^2 v^2 + 8c_4 c(\epsilon - \omega(Lh_n; f_i))}\right). \end{aligned} \quad (2.22)$$

Let

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

Then, by (2.20), (2.21), (2.22) and the definition of n_δ in (2.2), it follows that for every $n \geq n_\delta$,

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

which implies that

$$P\left(\sum_{n=n_\delta+1}^N 2 \sup_{1 \leq i \leq l} |\hat{f}_{i,n}(X_n) - f_i(X_n)| \geq \sum_{n=n_\delta+1}^N 2 \max_{1 \leq i \leq l} \tilde{\epsilon}_{i,n-1}\right) \leq \delta. \quad (2.23)$$

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

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

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

$$P\left(A \sum_{n=n_\delta+1}^N I(I_n \neq \hat{i}_n) \geq \sum_{n=n_\delta+1}^N (l-1)\pi_n + A\sqrt{\frac{N}{2}} \log\left(\frac{1}{\delta}\right)\right) \leq \delta. \quad (2.24)$$

Then, (2.19), (2.23) and (2.24) together complete the proof of Theorem 2.2. \square

Chapter 3

Model Combining Based Allocation

In this chapter, based on the allocation strategy described in section 4.1 of the previous chapter, we consider the general case that multiple candidate function estimation methods are used for model combining. First, we study the strong consistency of the algorithm. Then we evaluate the numerical performance with both simulation and a web-based real dataset.

3.1 Strong Consistency

We consider the general case that multiple function estimation methods are used for model combining. It is known from section 2.3 that strong consistency in L_∞ norm is desirable for the randomized allocation strategy. However, it is often 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 propose the combining algorithm to addresses 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.

Given an arm i , let $N_t^{(i)} = \inf\{n : \sum_{j=n_0l+1}^n I(I_j = i) \geq t\}$, $t \geq 1$, 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*

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

with probability one.

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

Assumption C. *$\|\hat{f}_{i,n,r} - f_i\|_\infty$ is upper bounded by a constant c_2 for all $1 \leq i \leq l$, $n \geq n_0l + 1$ and $1 \leq r \leq m$.*

Assumption D. *The variance estimates $\hat{v}_{i,n}$ are neither too close to zero nor too large: there exist constants $0 < p < q < \infty$ such that*

$$p \leq \hat{v}_{i,n} \leq q$$

with probability one for all $1 \leq i \leq l$ and $n \geq n_0l + 1$.

Assumption E. *The sequence $\{\pi_n, n \geq 1\}$ 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, $\sum_{t=1}^T (\hat{f}_{i,N_t,s}(X_{N_t}) - f_i(X_{N_t}))^2$ 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. If the mean reward functions f_i 's are continuous, Assumptions C can be satisfied by applying a truncation method on the function estimator. Similarly, Assumption D is satisfied by truncating the “combined” variance estimate $\hat{v}_{i,n}$ to be inside a positive interval. Assumption E ensures that N_t is finite as shown in Lemma 3.1 in the Appendix. As implied in Lemma 3.1, 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 3.1. *Under Assumption 0 and Assumptions A-E, the model combining allocation strategy is strongly consistent.*

With Theorem 3.1, 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, as long as the candidates are properly combined. The resulting per-round regret can be much improved if good methods (possibly different for different arms) are added in.

3.2 Simulations

For simulations, two examples with a univariate covariate are shown in section 3.2.1, and a more complicated example for multivariate covariates with the application of dimension reduction is given in section 3.2.2.

3.2.1 Univariate Covariate

The first case presented here has nonlinear reward functions with normal random errors, while the second case has binary responses with both linear and nonlinear reward functions.

Case 1

Consider a bandit problem with two arms. Suppose the true mean reward functions of the two arms on $[0, 1]$ are

$$\begin{aligned} f_1(x) &= 2e^{-200(x-0.2)^2} + 2e^{-200(x-0.8)^2}, \\ f_2(x) &= 0.5x^2 + e^{-(x-0.5)^2}. \end{aligned}$$

It is clear that no arm dominates the other over the entire domain $[0, 1]$, and the optimal decision should be made based on the value of the covariate. Given the time horizon $N = 800$, assume that for $1 \leq n \leq N$, the covariates are sampled from uniform(0,1), and the errors ε_n are normally distributed with mean 0 and variance $\sigma^2 = 0.5$. Let the first 20 rounds of the game be the initial forced sampling period.

We start with the Nadaraya-Watson regression described in section 2.3.1. Here we use the Epanechnikov quadratic kernel

$$K(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } |t| \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

For illustration, we consider three bandwidth choices: $h_1 = \frac{1}{(\log_2 n)^{0.25}}$, $h_2 = \frac{1}{\log_2 n}$ and $h_3 = \frac{1}{(\log_2 n)^2}$. Treating these three choices as different candidate modeling procedures, we run the model combining allocation strategy with the “inferior” arm sampling probability $\pi_n = \frac{1}{\log_2 n}$ to obtain the per-round regret r_n and the inferior sampling rate q_n . For comparison, each of the bandwidth choices is also run separately with the allocation strategy. We repeat this process 20 times to obtain the averaged per-round regret \bar{r}_n and the averaged inferior sampling rate \bar{q}_n , and plot the resulting \bar{r}_n and \bar{q}_n in Figure 3.1. We can see that in this case, the combining strategy performs even better than the winner of the three candidate procedures. It is also interesting to compare the weights of the three candidate procedures in the combining strategy at different rounds of the game (Table 3.1). As expected, the weights for a given arm can evolve as more and more rounds are played. We can also see that given n , the dominating candidate procedures are not necessarily the same for the two arms. Therefore, it supports that the model combining algorithm enables the allocation strategy to smartly evolve and appropriately prefer the better estimation methods for different arms at different time points to achieve an optimal performance.

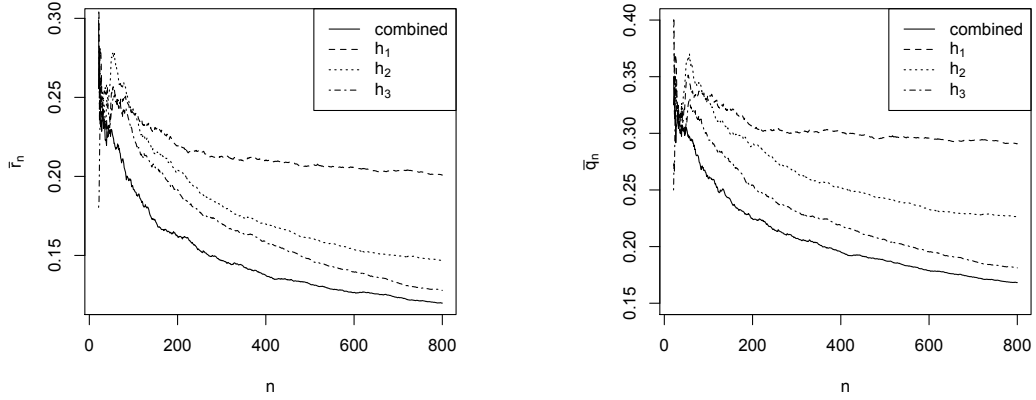


Figure 3.1: (Case 1) Combining Nadaraya-Watson regressions with different bandwidth choices. Left panel: averaged per-round regret. Right panel: averaged inferior sampling rate.

Table 3.1: (Case 1) Weights of bandwidth choices for Nadaraya-Watson regressions from the last repeat

bandwidth	arm 1			arm 2		
	h_1	h_2	h_3	h_1	h_2	h_3
weight ($n = 40$)	0.00	0.00	1.00	0.26	0.65	0.09
weight ($n = 800$)	0.00	0.00	1.00	0.03	0.97	0.00

As another example, we use the K -nearest neighbor method with $K_1 = \lfloor \frac{n}{(\log_2 n)^{0.25}} \rfloor$, $K_2 = \lfloor \frac{n}{\log_2 n} \rfloor$ and $K_3 = \lfloor \frac{n}{(\log_2 n)^2} \rfloor$, and repeat the simulation study described above. The averaged per-round regret in Figure 3.2 (left panel) shows that the performance of the combining strategy is satisfactorily close to the best of the three choices of K . Since the graphs for averaged inferior sampling rate look very similar to that of averaged per-round regret, we only present the graphs for per-round regret in the following numerical examples.

Next, we combine different nonparametric methods. Consider the following four nonparametric methods with the specified tuning parameters: histogram method with

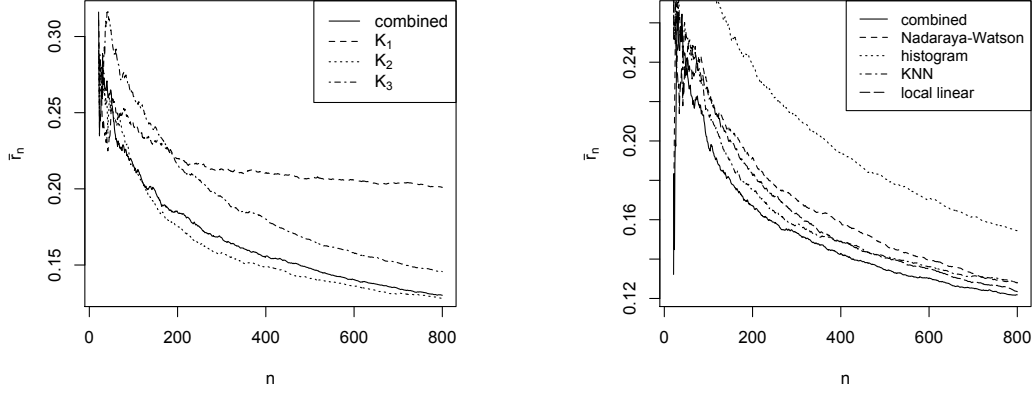


Figure 3.2: (Case 1) Averaged per-round regret from combining different methods. Left panel: combining K -nearest neighbor methods with different choices of K . Right panel: combining different nonparametric methods.

the bin width $\frac{1}{\lfloor (\log_2 n)^2 \rfloor}$, Nadaraya-Watson regression with the kernel bandwidth $\frac{1}{(\log_2 n)^2}$, local linear regression with the kernel bandwidth $\frac{1}{\log_2 n}$ and K -nearest neighbor method with $K = \lfloor \frac{n}{\log_2 n} \rfloor$. Repeat the simulation study under the same settings as described above, and the summary plot is shown in Figure 3.2 (right panel). Again, the combining strategy performs very well compared with the individual candidate procedures.

Case 2

Consider a two-armed bandit problem with 0-1 binary responses. Suppose the true reward functions (i.e., $P(Y = 1|X = x)$) of the two arms on $[0, 1]$ are

$$f_1(x) = 0.7e^{-30(x-0.2)^2} + 0.7e^{-30(x-0.8)^2},$$

$$f_2(x) = 0.65 - 0.3x.$$

Except for the mean reward functions and the error distribution, the settings of case 2 remain the same as case 1. Clearly, the error distribution here is dependent on the covariate. For combining modeling methods, we consider Nadaraya-Watson regression with bandwidth $h_1 = \frac{1}{\log_2 n}$ and $h_2 = \frac{1}{(\log_2 n)^2}$. In addition, we intentionally add linear regression as one candidate modeling method, which is not a strongly consistent method

for estimating arm 1, and is expected to perform poorly. The model combining strategy as well as the individual modeling candidates are repeated 50 times. By examining the weights of the combining strategy for the last run in Table 3.2, we can see that for arm 1, the linear regression is eventually assigned a very small weight while the Nadaraya-Watson regression with better bandwidth choice stands out. On the other hand, arm 2 seems to prefer the linear regression, which can be more efficient in estimating the linear mean reward function. The summary plot in Figure 3.3 (left panel) confirms that linear regression alone gives rather poor results, while the combining strategy once again performs very closely to the best individual modeling candidate.

Table 3.2: (Case 2) Weights for combining Nadaraya-Watson (NW) regression and linear regression

bandwidth	arm 1			arm 2		
	NW- h_1	NW- h_2	linear reg.	NW- h_1	NW- h_2	linear reg.
weight ($n = 40$)	0.10	0.78	0.12	0.20	0.00	0.80
weight ($n = 800$)	1.00	0.00	0.00	0.00	0.00	1.00

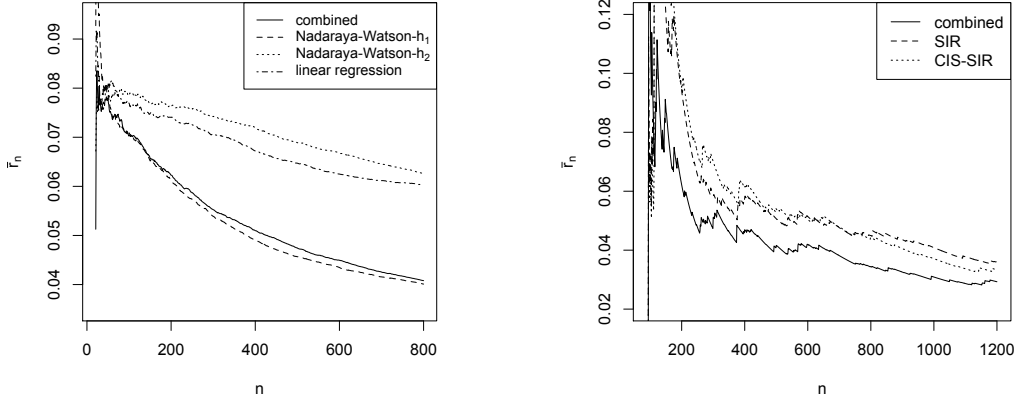


Figure 3.3: Averaged per-round regret from combining different methods. Left panel: (Case 2) combining Nadaraya-Watson regression and linear regression. Right panel: (the multivariate covariate case) comparing SIR and CIS-SIR.

3.2.2 Multivariate Covariates with Dimension Reduction

In this subsection, we use the dimension reduction function estimation procedures described in section 2.3.3 for bandit problem with multivariate covariates. Two 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

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

We consider a three-arm bandit model with $d = 10$. Assume that at each time n , the covariate is $X_n = (X_{n1}, X_{n2}, \dots, X_{nd})^T$, and X_{ni} 's ($i = 1, \dots, d$) 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 3.3 (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 3.3. As expected, CIS-SIR results in a sparse dimension reduction matrix with rows 1, 3 and 4 being non-zero.

3.3 Web-Based Personalized News Article Recommendation

In this section, we use the Yahoo! Front Page Today Module User Click Log dataset (Yahoo! Academic Relations, 2011) to evaluate the proposed allocation strategy. The

Table 3.3: Comparing the estimated dimension reduction matrix $\hat{B}_{2,N}^*$ for the second arm between SIR and CIS-SIR.

	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

complete dataset 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 dataset provides the necessary platform to test a MABC algorithm.

One remaining issue before algorithm evaluation is that the complete dataset 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 dataset is used in May et al. (2012) for MABC algorithm evaluation purposes. With the above, we obtain a reduced dataset 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” dataset (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” dataset (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” dataset (history) is unchanged. This process is run sequentially over all the interaction events to generate the final “valid” dataset, 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” dataset is like being obtained from running the MABC algorithm over a random sample of the underlying population.

With the reduced dataset 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 4.1 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 dataset 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 dataset. 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. 3.4. The means and standard deviations of the normalized CTRs are given in Table 3.4.

Table 3.4: Normalized CTRs of various algorithms on the news article recommendation dataset. CTRs are normalized with respect to the random algorithm.

	LinUCB	ϵ -greedy	SIR-kernel			model combining
			h_{n1}	h_{n2}	h_{n3}	
mean	1.239	1.189	1.235	1.236	1.236	1.238
std. dev.	0.041	0.005	0.015	0.017	0.016	0.018

Similar to what we have seen in the previous section, the choice of bandwidth sequences has limited effect on the performance of SIR-kernel method. The naive ϵ -greedy algorithm, however, clearly under-performs 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

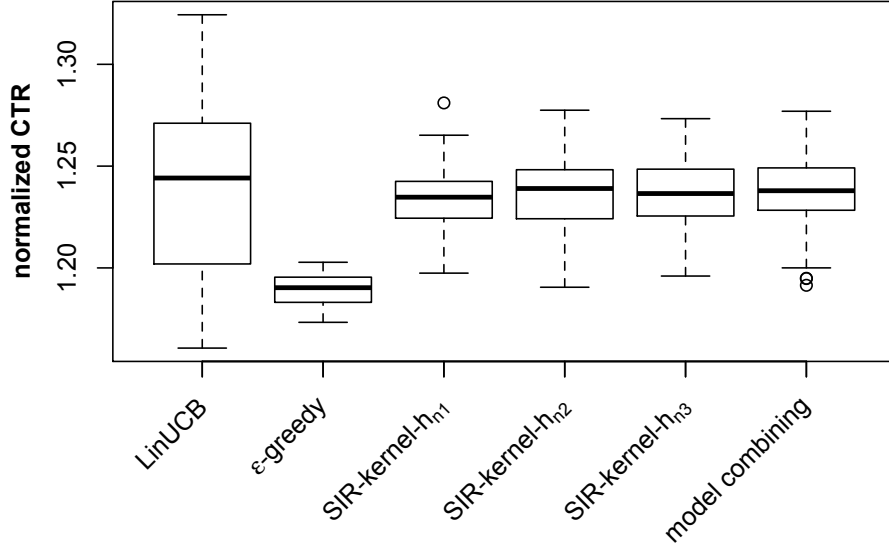


Figure 3.4: Boxplots of normalized CTRs of various algorithms on the news article recommendation dataset. Algorithms include (from left to right): LinUCB, ϵ -greedy, SIR-kernel (h_{n1}), SIR-kernel (h_{n2}), SIR-kernel (h_{n3}), model combining with SIR-kernel (h_{n3}) and ϵ -greedy. CTRs are normalized with respect to the random algorithm.

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 dataset are constructed with logistic regression related methods (Li et al., 2010), it is satisfactory to observe that SIR-kernel algorithm has similar performance with relatively small variation when compared with the LinUCB algorithm.

3.4 Proof of Theorem 3.1

Lemma 3.1. *Under Assumption E and the proposed allocation strategy, for each arm i*

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

Proof of Lemma 3.1. It suffices to check that

$$\sum_{j=n_0l+1}^{\infty} \mathbb{I}(I_j = i) = \infty \quad \text{a.s.} \quad (3.1)$$

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 $j \geq n_0l + 1$,

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

By Assumption E, $\sum_{j=n_0l+1}^{\infty} P(I_j = i | \mathcal{F}_{j-1}) = \infty$. Therefore, (3.1) is an immediate result of the Lévy's extension of the Borel-Cantelli lemmas (Williams, 1991, pp.124). \square

Proof of Theorem 3.1. The key to the proof is to show $\|\hat{f}_{i,n} - f_i\|_{\infty} \rightarrow 0$ 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 $\frac{W_{i,n,1}}{W_{i,n,2}} \rightarrow \infty$ almost surely as $n \rightarrow \infty$.

As defined before, $N_t = \inf\{n : \sum_{j=n_0l+1}^n \mathbb{I}(I_j = i) \geq t\}$, 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 3.1, $N_t < \infty$ 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 $\frac{W_{i,N_t+1}}{W_{i,N_t,2}} \rightarrow \infty$ almost surely as $t \rightarrow \infty$.

Note that for any $t \geq 1$,

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

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 $T \geq 1$,

$$\frac{W_{i,N_{T+1},1}}{W_{i,N_{T+1},2}} = \left(\prod_{t=1}^T \frac{\hat{v}_{i,N_t,2}^{1/2}}{\hat{v}_{i,N_t,1}^{1/2}}\right) \exp\left(\sum_{t=1}^T T_{1t} + \sum_{t=1}^T T_{2t}\right). \quad (3.2)$$

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 2.1 that for every $\tau > 0$ and every $T \geq 1$,

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

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

$$\hat{v}_{i,N_t,1} < \hat{v}_{i,N_t,2} \tag{3.3}$$

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 2.1 again to obtain that

$$\sum_{t=1}^T T_{2t} = o(\sqrt{T}(\log T)^b), \tag{3.4}$$

almost surely. By (3.3), (3.4) and Assumption A, we conclude from (3.2) that

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

This completes the proof of Theorem 3.1. □

Chapter 4

Adaptive Performance in Randomized Allocation with Arm Elimination

From Chapter 2, we see that the randomized allocation strategy is not minimax optimal, and the finite-time regret analysis have to rely on the knowledge of the smoothness parameter. In this chapter, we propose a new algorithm called randomized allocation with arm elimination (or RAAE for abbreviation) as an attempt to address these two issues. The RAAE algorithm has an explicit multi-stage structure, and can be viewed as a natural extension of the randomized allocation of Yang and Zhu (2002) and the adaptively binned successive elimination (ABSE) of Perchet and Rigollet (2013). In particular, new development is made in the following two aspects. First, a smoothness parameter selector modified from the Lepski's approach, a popular adaptive nonparametric estimation technique pioneered by Lepski (1990), is integrated into the RAAE algorithm to address the issue of unknown smoothness parameters. The resulting cumulative regret is shown to adaptively achieve the minimax rate up to a logarithmic factor. Second, due to the randomized allocation feature, the RAAE provides the flexibility to choose an appropriate reward function modeling method for each arm to exploit the response-covariate association while maintaining the near minimax optimality by imbedding the arm elimination procedure in the end of each stage to avoid over-exploration of the

lower rewarding arms.

Before introducing the algorithm, we revisit the Hölder smoothness condition, and a margin condition. 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 smoothness condition: for $f \in \Sigma(\kappa, \rho)$,

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

for every $x_1, x_2 \in [0, 1]^d$. To our knowledge, existing nonparametric MABC algorithms all require the knowledge of κ . Since $\Sigma(\tilde{\kappa}, \rho) \supseteq \Sigma(\kappa, \rho)$ for every $\tilde{\kappa} < \kappa$, the game player certainly wishes to choose the largest possible κ . However, such information is usually unavailable to the player. Efforts are made to provide a proper estimate for κ in section 4.2.

Besides the smoothness condition, a margin condition has 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 4.1. *There exist $\alpha \in (0, d/\kappa]$, $t_0 \in (0, 1)$ and $c_0 > 0$ such that*

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

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

Larger α in Assumption 4.1 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 $\alpha \leq d/\kappa$.

4.1 Algorithms

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 ($n_0 \geq 1$). 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 arm elimination mechanism. After the forced sampling step, the remaining time horizon is divided into $T + 1$ stages. Let $\tilde{N}_1 < \tilde{N}_2 < \dots < \tilde{N}_T$ 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 ($1 \leq t \leq T$) is denoted by $N_t = \tilde{N}_t - \tilde{N}_{t-1}$. Let $\{h_t, 1 \leq t \leq T\}$ be a sequence of bin width that satisfies $h_1 = 1$ and $h_{k+1} = h_k/2$, $1 \leq k \leq T-1$. At the end of stage t ($1 \leq t \leq T$), 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 $B \in \mathcal{B}_t$ ($1 \leq t \leq T$) and each stage s ($0 \leq s < t$), 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 $\{\pi_n, 1 \leq n \leq N\}$ be a sequence of positive numbers satisfying $(l-1)\pi_n < 1$ for every $1 \leq n \leq N$. 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 ($1 \leq i \leq l$), estimate κ by the smoothness parameter selector described in section 4.2. The resulting estimate for arm i is denoted by $\hat{\kappa}^{(i)}$. Define $\hat{\kappa}^* = \min_{1 \leq i \leq l} \hat{\kappa}^{(i)}$, 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 $\mathcal{S}_{\mathcal{X}} = \{1, 2, \dots, l\}$. Start stage $t = 1$ of the game. For $n = \tilde{N}_{t-1} + 1, \tilde{N}_{t-1} + 2, \dots, \tilde{N}_t$, 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 player-specified modeling method. 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 $H_{B,i} = \{n : \tilde{N}_{t-1} + 1 \leq n \leq \tilde{N}_t, X_n \in B, I_n = i\}$ 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 mean reward 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 mean reward 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 $t = 2, 3, \dots, T$.

Step 4. Repeat Step 1 for $n = \tilde{N}_T + 1, \tilde{N}_T + 2, \dots, N$ (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 ($1 \leq t \leq T + 1$), 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 ($1 \leq t \leq T$), 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 mean of each active arm during stage t . Subsequently, Step 2.3 eliminates the arms with low reward sample means compared to the highest. The remaining arms of each bin after elimination serve as the new active arms, and the next stage follows. Intuitively 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 4.3. 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 $n = N$.

4.2 Smoothness Parameter Selector

Suppose $f(x)$ is the mean reward function of a given arm, and a random sample $\{(X_i, Y_i), i = 1, \dots, n\}$ 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

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

and

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

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

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

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 $\hat{\kappa} = \kappa_{\hat{\tau}} - \frac{b_2 \log \log n}{\log n}$, 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

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

Then

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

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 (Giné and Nickl, 2010; Hoffmann and Nickl, 2011; Bull, 2012). We defer the discussion of this condition to section 4.4.

Assumption 4.2. *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 0, 2.2 and 4.2 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*

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

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 $\Sigma_0(\kappa, \rho)$.

4.3 Finite-Time Regret Analysis

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

$$T = \min\left\{t \in \mathbb{N} : \frac{h_1}{2^{t-1}} \leq 6 \left(\frac{l}{N}\right)^{\frac{1}{2\hat{\kappa}^* + d}}\right\}. \quad (4.2)$$

Given any stage t ($1 \leq t \leq T$), define $\tilde{\pi}_t = \min\{\pi_n : \tilde{N}_{t-1} + 1 \leq n \leq \tilde{N}_t\}$. Take $\alpha_t = 4\rho h_t^{\hat{\kappa}^*}$ and $N_t = \tilde{\gamma}_t h_t^{-(2\hat{\kappa}^*+d)}(1 \vee \log(Nh_t^{2\hat{\kappa}^*+d}))$, where $\tilde{\gamma}_t$ is a stage-dependent parameter chosen to make N_t a positive integer. In particular, it suffices to assume

$$\max\left\{\frac{16(v^2 + c\rho/2)}{c\rho^2\tilde{\pi}_t^2}, \frac{56}{3c\tilde{\pi}_t}\right\} \leq \tilde{\gamma}_t \leq \gamma < \infty, \quad (4.3)$$

where γ is a positive constant. Note that such γ exists if $\{\tilde{\pi}_t, t \geq 1\}$ is uniformly lower bounded by a positive constant.

Theorem 4.1. *Under Assumptions 0, 2.2, 4.1 and 4.2, the mean cumulative regret of the proposed algorithm satisfies*

$$ER_N(\eta) \leq C_\gamma N^{\frac{\kappa - \kappa\alpha + d}{2\kappa + d}} (\log N)^{c^*},$$

where C_γ is a positive constant (not depending on N) and $c^* = (2\kappa^* + d)^2(1 + 1/\kappa_*)/d$.

The cumulative regret rate in Theorem 4.1 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.

4.4 Discussion

In the context of 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 . In this work, we take the Lepski's approach to estimate κ while allowing a flexible modeling for estimating the reward functions.

Under 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 technically "complicated" case of the two in nonparametric estimation. Its counterpart in

the MABC problem (Lemma 4.1) turns out to be straightforward because the event probability can be bounded tightly by using the moment condition (Assumption 0) and a resulting exponential tail probability (Lemma 2.1) concerning the random errors. The observation that the former event has a tight probability is shared in, e.g., Lepski (1990) and Lepski et al. (1997) under a Gaussian white noise model. On the other hand, the latter event (i.e., over-estimation of κ) is usually considered 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, partly due to insufficiency in sample size.

The difficulty in the estimation of κ is shared in the adaptive confidence bound problems, and if we only consider the Hölder condition without further assumptions, it is known that the adaptive confidence bound does not exist (Low, 1997). To overcome such difficulty, Giné and Nickl (2010) propose a “self-similarity” condition, and show that the functions that do not satisfy the “self-similarity” condition is a negligible subset of Hölder class (see Giné and Nickl, 2010, Condition 3 and Proposition 4). It turns out that the function class $\Sigma_0(\kappa, \rho)$ defined in section 4.2 is closely related to the “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

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

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

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

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 $\rho_2 \leq \rho$ and a positive integer τ_0 such that for every integer $\tau \geq \tau_0$, $\rho_2 2^{-\tau\kappa} \leq \|\tilde{K}_\tau f - f\|_\infty \leq \rho 2^{-\tau\kappa}$) becomes essentially 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 a sense similar to Proposition 4 of Giné and Nickl (2010)). 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 $U_1, U_2 \neq 0$,

$$\lim_{v \rightarrow 0^+} \frac{f(x_0 + v) - f(x_0)}{|v|^\kappa} = U_1 \quad \text{or} \quad \lim_{v \rightarrow 0^-} \frac{f(x_0 + v) - f(x_0)}{|v|^\kappa} = U_2, \quad (4.4)$$

then $f \in \Sigma_0(\kappa, \rho)$. Because of this observation, the rate obtained in Theorem 4.1 remains to be the minimax rate for $\Sigma_0(\kappa, \rho)$ under Assumption 4.2. Indeed, the proof of such minimax result follows directly from Theorem 4.1 of Rigollet and Zeevi (2010) since the functions considered in their proof satisfies (4.4) and consequently belong to $\Sigma_0(\kappa, \rho)$.

Next, we state two straightforward observations.

Remark 1. The smoothness parameter selector proposed in section 4.2 is a “plug-in” type result, and therefore, we can also equip other nonparametric MABC algorithms such as ABSE algorithm (Perchet and Rigollet, 2013) to obtain the same regret rate.

Remark 2. Our algorithm implicitly assumes that ρ is known. However, it suffices to know the upper and lower bound of ρ in order to achieve the minimax regret rate.

4.5 Proofs

4.5.1 Proof of Proposition 4.1

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

Lemma 4.1. *Suppose $f(\cdot) \in \Sigma(\kappa, \rho)$ and Assumptions 0 and 2.2 hold. Then for $\hat{\kappa}$ obtained by procedures in section 4.2, there exists an integer n_* and a constant C_H such that*

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

for every $n > n_*$

Proof of Lemma 4.1. Define

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

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

$$\begin{aligned} & \{\check{\kappa} \leq \tilde{\kappa}\} \\ \Rightarrow & \bigcup_{\tilde{\tau}=1}^{\tau^*} \{\hat{\tau} = \tau\} \\ \Rightarrow & \bigcup_{\tau=\tilde{\tau}-1}^{\tau^*-1} \bigcup_{\tau_2=\tau+1}^{\tau^*} \{\|\hat{\theta}_\tau - \hat{\theta}_{\tau_2}\|_\infty > b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n\} \\ \Rightarrow & \bigcup_{\tau=\tilde{\tau}-1}^{\tau^*-1} \bigcup_{\tau_2=\tau+1}^{\tau^*} \left\{ \|\hat{\theta}_\tau - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2} \right\} \cup \left\{ \|\hat{\theta}_{\tau_2} - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2} \right\}. \end{aligned} \quad (4.5)$$

Given $\tau \in \mathbb{N}$, let \mathcal{M}_τ 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 $\tilde{\tau} - 1 \leq \tau \leq \tau_2 \leq \tau^*$, we have

$$P\left(\|\hat{\theta}_\tau - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\right) \leq \sum_{B \in \mathcal{M}_\tau} P\left(\sup_{x \in B} |\hat{\theta}_\tau(x) - f(x)| > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\right). \quad (4.6)$$

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

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

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

$$\sup_{x \in B} |\hat{\theta}_\tau(x) - f(x)| \leq \frac{\left| \sum_{i \in H_\tau(x_B^*)} \varepsilon_i \right|}{M_\tau(x_B^*)} + \rho u_\tau^\kappa. \quad (4.7)$$

Define

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

and

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

Then,

$$\begin{aligned} P(J_{\tau,B}) &\leq P(A_{\tau,B}^c) + P(J_{\tau,B}, A_{\tau,B}) \\ &\leq P(A_{\tau,B}^c) + P\left(\frac{\left| \sum_{i \in H_\tau(x_B^*)} \varepsilon_i \right|}{M_\tau(x_B^*)} > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2} - \rho u_\tau^\kappa, A_{\tau,B}\right) \\ &\leq P(A_{\tau,B}^c) + P\left(\frac{\left| \sum_{i \in H_\tau(x_B^*)} \varepsilon_i \right|}{M_\tau(x_B^*)} > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{4}, A_{\tau,B}\right), \end{aligned} \quad (4.8)$$

where the second inequality follows by (4.7) 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 2.1,

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

As a result,

$$\begin{aligned} &P\left(\frac{\left| \sum_{i \in H_\tau(x_B^*)} \varepsilon_i \right|}{M_\tau(x_B^*)} > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{4}, A_{\tau,B}\right) \\ &\leq \exp\left(-\frac{n \underline{c} u_\tau^d b_1^2 u_{\tau_2}^{2\kappa_{\tau_2}} \gamma_n^2}{64(v^2 + c b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n / 4)}\right) \\ &\leq \exp\left(-\frac{c b_1^2 \gamma_n^2}{128 v^2}\right) \\ &\leq n^{-\frac{d}{2\kappa_* + d} - \frac{1}{c_*}}, \end{aligned} \quad (4.9)$$

where the last two inequalities follow by the observation that $n u_\tau^d u_{\tau_2}^{2\kappa_{\tau_2}} \geq 1$, $\frac{c b_1 \gamma_n}{4 n^{c_*}} \leq v^2$ and $\frac{c b_1^2 \log n}{128 v^2} > \frac{d}{2\kappa_* + d} + \frac{1}{c_*}$ for large enough n . Also, since $P(I(X_i \in B)) \geq \underline{c} u_\tau^d$ for any $B \in \mathcal{D}_\tau$, by Lemma 2.2,

$$P(A_{\tau,B}^c) \leq \exp\left(-\frac{3c n u_\tau^d}{28}\right). \quad (4.10)$$

Thus, by (4.6), (4.8), (4.9), (4.10), and the fact that $u_\tau^{-d} \leq C_{H1} n^{\frac{d}{2\kappa_*+d}}$ for some constant $C_{H1} > 0$, we have

$$\begin{aligned} & P\left(\|\hat{\theta}_\tau - f\|_\infty > \frac{b_1 u_{\tau_2}^{\kappa_{\tau_2}} \gamma_n}{2}\right) \\ & \leq u_\tau^{-d} \exp\left(-\frac{3cn u_\tau^d}{28}\right) + u_\tau^{-d} n^{-\frac{d}{2\kappa_*+d} - \frac{1}{c_*}} \\ & \leq \frac{2C_{H1}}{n^{1/c_*}}. \end{aligned}$$

In together with (4.5) and $\tilde{\kappa} > \kappa - \frac{\Delta}{\log n}$, we know that there exists n_* and some constant C_H such that

$$P\left(\tilde{\kappa} \leq \kappa - \frac{\Delta}{\log n}\right) \leq P\left(\tilde{\kappa} \leq \tilde{\kappa}\right) \leq \frac{C_H (\log n)^2}{n^{1/c_*}}$$

for any $n > n_*$. This completes the proof of Lemma 4.1. \square

Lemma 4.2. *Suppose $f(\cdot) \in \Sigma_0(\kappa, \rho)$ and Assumptions 0 and 2.2 hold. Then for $\hat{\kappa}$ obtained by procedures in section 4.2, there exists an integer n^* and a constant $C_H^* > 0$ such that*

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

for every $n > n^*$.

Proof of Lemma 4.2. Let $\tilde{\tau}$, $\tilde{\kappa}$ and $\tilde{\kappa}$ be defined as in the proof of Lemma 4.1. Let $\kappa' = \kappa + \frac{b_2 \log \log n}{\log n}$. Define

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

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

$$\begin{aligned} & \{\tilde{\kappa} > \kappa'\} \\ & \Rightarrow \{\hat{\tau} \leq \tau'\} \\ & \Rightarrow \{\|\hat{\theta}_{\tau'} - \hat{\theta}_{\tilde{\tau}}\|_\infty \leq b_1 u_{\tilde{\tau}}^{\tilde{\kappa}}\} \\ & \Rightarrow \{\|\hat{\theta}_{\tau'} - f\|_\infty \leq \frac{3}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\} \cup \{\|\hat{\theta}_{\tilde{\tau}} - f\|_\infty > \frac{1}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\}. \end{aligned} \quad (4.11)$$

Recall from the proof of Lemma 4.1 that there is a constant C_{H1} such that

$$P\left(\|\hat{\theta}_{\tilde{\tau}} - f\|_\infty > \frac{1}{2} b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\right) \leq \frac{2C_{H1}}{n^{1/c_*}}. \quad (4.12)$$

It remains to find the upper bound for $P(\|\hat{\theta}_{\tau'} - f\|_\infty \leq \frac{3}{2}b_1u_{\tilde{\tau}}^{\tilde{\kappa}}\gamma_n)$. Note that by triangle inequalities,

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

The previous inequality implies that for large enough n ,

$$\begin{aligned} & \|\hat{\theta}_{\tau'} - f\|_\infty \\ &\geq \|K_{\tau'}f - f\|_\infty - \sup_x \left| \frac{\sum_{i \in H_{\tau'}(x)} f(X_i)}{M_{\tau'}(x)} - K_{\tau'}f(x) \right| - \sup_x \left| \frac{\sum_{i \in H_{\tau'}(x)} \varepsilon_i}{M_{\tau'}(x)} \right| \\ &=: \|K_{\tau'}f - f\|_\infty - \Gamma_1 - \Gamma_2 \\ &> \rho_1 u_{\tau'}^{\tilde{\kappa}} - \Gamma_1 - \Gamma_2 \\ &\geq 2b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n - \Gamma_1 - \Gamma_2 \end{aligned} \tag{4.13}$$

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

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

Also, by derivations similar to that of (4.9) and (4.10),

$$\begin{aligned} & P\left(\Gamma_2 \geq \frac{1}{4}b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\right) \\ &\leq u_{\tau'}^{-d} \left(\exp\left(-\frac{3cnu_{\tau'}^d}{28}\right) + \exp\left(-\frac{cb_1^2 \gamma_n^2}{256v^2}\right) \right) \\ &\leq \frac{2C_{H1}}{n^{1/c_*}}, \end{aligned} \tag{4.14}$$

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

$$\begin{aligned} & P\left(\Gamma_1 \geq \frac{1}{4}b_1 u_{\tilde{\tau}}^{\tilde{\kappa}} \gamma_n\right) \\ &\leq u_{\tau'}^{-d} \left(\exp\left(-\frac{3cnu_{\tau'}^d}{28}\right) + \exp\left(-\frac{(cnu_{\tau'}^d/2)b_1^2 u_{\tilde{\tau}}^{2\tilde{\kappa}} \gamma_n^2}{64\|f\|_\infty}\right) \right) \\ &\leq \frac{2C_{H1}}{n^{1/c_*}}, \end{aligned} \tag{4.15}$$

for all large enough n . The, by (4.13), (4.14) and (4.15),

$$\begin{aligned}
& P\left(\|\hat{\theta}_{\tau'} - f\|_\infty \leq \frac{3}{2}b_1u_{\tilde{\tau}}^{\tilde{\kappa}}\gamma_n\right) \\
& \leq P\left(2b_1u_{\tilde{\tau}}^{\tilde{\kappa}}\gamma_n - \Gamma_1 - \Gamma_2 \leq \frac{3}{2}b_1u_{\tilde{\tau}}^{\tilde{\kappa}}\gamma_n\right) \\
& \leq P\left(\Gamma_1 \geq \frac{1}{4}b_1u_{\tilde{\tau}}^{\tilde{\kappa}}\gamma_n\right) + P\left(\Gamma_2 \geq \frac{1}{4}b_1u_{\tilde{\tau}}^{\tilde{\kappa}}\gamma_n\right) \\
& \leq \frac{4C_{H1}}{n^{1/c_*}}.
\end{aligned}$$

Together with (4.11) and (4.12),

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

which completes the proof of Lemma 4.2. \square

Proof of Proposition 4.1. By Lemma 4.1 and Assumption 4.2,

$$P\left(\hat{\kappa}^* \leq \kappa - \frac{\Delta}{\log n} - \frac{b_2 \log \log n}{\log n}\right) \leq \sum_{i=1}^l P\left(\hat{\kappa}^{(*)} \leq \kappa - \frac{\Delta}{\log n} - \frac{b_2 \log \log n}{\log n}\right) \leq C_{Hl}(\log n)^2 n^{-1/c_*}.$$

Together with Lemma 4.2 and the fact that there exists $f_i \in \Sigma_0(\kappa, \rho)$, the proof of Proposition 4.1 is complete. \square

4.5.2 Proof of Theorem 4.1

Proof of Theorem 4.1. Let $A_0 = \{\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log \log n}{\log n} < \hat{\kappa}^* \leq \kappa\}$. Motivated 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 ($1 \leq t \leq T$), define the sets of arms

$$\mathcal{S}_{t,B,1} = \{1 \leq i \leq l : \text{there exists some } x \in p_t(B) \text{ such that } f^*(x) = f_i(x)\},$$

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

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 $8\rho h_t^{\hat{k}^*}$. Further define the sets

$$A_{t,B} = G_{t,B,1} \cap G_{t,B,2}, \quad (4.16)$$

$$F_{t,B} = \bigcap_{1 \leq k \leq t} A_{k,B}. \quad (4.17)$$

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

$$R_N(\eta) = R_N(\eta)I(A_0^c) + R_N(\eta)I(A_0) \quad (4.18)$$

and

$$\begin{aligned} R_N(\eta)I(A_0) &\leq Aln_0 + \sum_{n=\tilde{N}_0+1}^N (f^*(X_n) - f_{I_n}(X_n))I(A_0) \\ &\leq Aln_0 + \sum_{B \in \mathcal{B}_T} \sum_{n=\tilde{N}_0+1}^N (f^*(X_n) - f_{I_n}(X_n))I(A_0)I(X_n \in B) \\ &=: Aln_0 + \sum_{B \in \mathcal{B}_T} R_B. \end{aligned} \quad (4.19)$$

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

$$\begin{aligned} R_N^{(0)} &= \sum_{B \in \mathcal{B}_T} R_B I(A_{1,B}^c) + \sum_{B \in \mathcal{B}_T} R_B I(F_{1,B} \cap A_{2,B}^c) + \cdots \\ &\quad + \sum_{B \in \mathcal{B}_T} R_B I(F_{T-1,B} \cap A_{T,B}^c) + \sum_{B \in \mathcal{B}_T} R_B I(F_{T,B}) \\ &=: R_1 + R_2 + \cdots + R_T + R_{T+1}. \end{aligned} \quad (4.20)$$

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

$$\begin{aligned} R_1 &= \sum_{n=\tilde{N}_0+1}^N \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n))I(X_n \in B)I(A_0 \cap A_{1,B}^c) \\ &\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_0 \cap A_{1,B}^c) + \sum_{n=\tilde{N}_1+1}^N \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_0 \cap A_{1,B}^c). \end{aligned}$$

Let $E^{(0)}(\cdot)$ and $P^{(0)}(\cdot)$ denote the conditional expectation and conditional probability given $\hat{\kappa}^* = \kappa_0$ ($\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log \log n}{\log n} < \hat{\kappa}^* \leq \kappa$), 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 $\hat{\kappa}^* = \kappa_0$,

$$\begin{aligned}
E^{(0)}(R_1) &\leq E^{(0)}\left(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_{1,B}^c)\right) + \sum_{n=\tilde{N}_1+1}^N \sum_{B \in \mathcal{B}_T} AP(X_n \in B)P^{(0)}(A_{1,B}^c) \\
&\leq E^{(0)}\left(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_{1,B}^c)\right) + \sum_{n=\tilde{N}_1+1}^N A \max_{B \in \mathcal{B}_T} P^{(0)}(A_{1,B}^c) \\
&\leq E^{(0)}\left(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_{1,B}^c)\right) + 4Alh_1^{-(2\kappa_0+d)}, \tag{4.21}
\end{aligned}$$

where the last inequality follows by Lemma 4.3. Similarly, by definition, for $2 \leq t \leq T$,

$$\begin{aligned}
R_t &= \sum_{n=\tilde{N}_0+1}^N \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n))I(X_n \in B)I(A_0 \cap F_{t-1,B} \cap A_{t,B}^c) \\
&\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_0 \cap F_{t-1,B} \cap A_{t,B}^c) \\
&\quad + \sum_{k=1}^{t-1} \left(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n)) \right. \\
&\quad \quad \cdot I(X_n \in B, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\hat{\kappa}^*})I(A_0 \cap F_{t-1,B} \cap A_{t,B}^c) \Big) \\
&\quad + \sum_{n=\tilde{N}_t+1}^N \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n)) \\
&\quad \quad \cdot I(X_n \in B, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\hat{\kappa}^*})I(A_0 \cap F_{t-1,B} \cap A_{t,B}^c) \\
&\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(A_0 \cap F_{t-1,B} \cap A_{t,B}^c) \\
&\quad + \sum_{k=1}^{t-1} \left(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} 8\rho h_k^{\hat{\kappa}^*} I(0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\hat{\kappa}^*}) \right) \\
&\quad + \sum_{n=\tilde{N}_t+1}^N \sum_{B \in \mathcal{B}_T} 8\rho h_{t-1}^{\hat{\kappa}^*} I(X_n \in B, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\hat{\kappa}^*})I(A_0 \cap F_{t-1,B} \cap A_{t,B}^c)
\end{aligned}$$

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

$$\begin{aligned}
E^{(0)}(R_t) &\leq E^{(0)}\left(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(F_{t-1,B} \cap A_{t,B}^c)\right) + \sum_{k=1}^{t-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha} N_{k+1} \\
&\quad + \sum_{n=\tilde{N}_t+1}^N \sum_{B \in \mathcal{B}_T} 8\rho h_{t-1}^{\kappa_0} P(X_n \in B, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_{t-1}^{\kappa_0}) P^{(0)}(F_{t-1,B} \cap A_{t,B}^c) \\
&\leq E^{(0)}\left(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B)I(F_{t-1,B} \cap A_{t,B}^c)\right) \\
&\quad + \sum_{k=1}^{t-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha} \gamma h_{k+1}^{-(2\kappa_0+d)} \log(Nh_{k+1}^{2\kappa_0+d}) + 4lc_0(8\rho h_{t-1}^{\kappa_0})^{1+\alpha} h_t^{-(2\kappa_0+d)},
\end{aligned} \tag{4.22}$$

where the first inequality follows by Assumption 4.1, and the second inequality follows by Assumption 4.1, Lemma 4.3 and the choice of $\{N_k, 1 \leq k \leq t\}$. Similarly, by

definition,

$$\begin{aligned}
R_{T+1} &= \sum_{n=\tilde{N}_0+1}^N \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n)) I(X_n \in B) I(A_0 \cap F_{T,B}) \\
&\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B) I(A_0 \cap F_{T,B}) \\
&\quad + \sum_{k=1}^{T-1} \left(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n)) \right. \\
&\quad \quad \cdot I(X_n \in B, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_k^{\hat{\kappa}^*}) I(A_0 \cap F_{T,B}) \Big) \\
&\quad + \sum_{n=\tilde{N}_T+1}^N \sum_{B \in \mathcal{B}_T} (f^*(X_n) - f_{I_n}(X_n)) I(X_n \in B, 0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_T^{\hat{\kappa}^*}) I(A_0 \cap F_{T,B}) \\
&\leq \sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B) I(A_0 \cap F_{T,B}) \\
&\quad + \sum_{k=1}^{T-1} \left(\sum_{n=\tilde{N}_k+1}^{\tilde{N}_{k+1}} 8\rho h_k^{\hat{\kappa}^*} I(0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_k^{\hat{\kappa}^*}) \right) \\
&\quad + \sum_{n=\tilde{N}_T+1}^N 8\rho h_T^{\hat{\kappa}^*} I(0 < f^*(X_n) - f^\sharp(X_n) \leq 8\rho h_T^{\hat{\kappa}^*}).
\end{aligned}$$

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

$$\begin{aligned}
E^{(0)}(R_{T+1}) &\leq E^{(0)} \left(\sum_{n=\tilde{N}_0+1}^{\tilde{N}_1} \sum_{B \in \mathcal{B}_T} AI(X_n \in B) I(F_{T,B}) \right) \\
&\quad + \sum_{k=1}^{T-1} c_0 (8\rho h_k^{\kappa_0})^{1+\alpha} \gamma h_{k+1}^{-(2\kappa_0+d)} \log(N h_{k+1}^{2\kappa_0+d}) + N (8\rho h_T^{\kappa_0})^{1+\alpha}.
\end{aligned} \tag{4.23}$$

Combining (4.20)-(4.23), we have

$$\begin{aligned}
E^{(0)}(R_N^{(0)}) &\leq A\gamma h_1^{-(2\kappa_0+d)} \log(Nh_1^{2\kappa_0+d}) + 4Alh_1^{-(2\kappa_0+d)} \\
&\quad + \sum_{t=2}^T \sum_{k=1}^{t-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha} \gamma h_{k+1}^{-(2\kappa_0+d)} \log(Nh_{k+1}^{2\kappa_0+d}) + \sum_{t=2}^T 4lc_0(8\rho h_{t-1}^{\kappa_0})^{1+\alpha} h_t^{-(2\kappa_0+d)} \\
&\quad + \sum_{k=1}^{T-1} c_0(8\rho h_k^{\kappa_0})^{1+\alpha} \gamma h_{k+1}^{-(2\kappa_0+d)} \log(Nh_{k+1}^{2\kappa_0+d}) + N(8\rho h_T^{\kappa_0})^{1+\alpha} \\
&\leq A\gamma \log N + 4Al + C_{\gamma 1} h_T^{-(\kappa_0 - \kappa_0 \alpha + d)} (1 + \log(Nh_T^{2\kappa_0+d})) + C_{\gamma 2} N h_T^{\kappa_0 + \kappa_0 \alpha} \\
&\leq C_{\gamma 3} N^{\frac{\kappa_0 - \kappa_0 \alpha + d}{2\kappa_0 + d}} \\
&\leq C_{\gamma 4} N^{\frac{\kappa - \kappa \alpha + d}{2\kappa + d}} (\log N)^{c^*}, \tag{4.24}
\end{aligned}$$

where $C_{\gamma 1}, \dots, C_{\gamma 4}$ are some positive constants, and the last inequality follows by $\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log \log n}{\log n} < \hat{\kappa}^* \leq \kappa$. Then, by (4.18), (4.19), (4.24) and Proposition 4.1, there exists some constant $C_\gamma > 0$ such that

$$ER_N(\eta) \leq ANP(A_0^c) + Aln_0 + ER_N^{(0)} \leq C_\gamma N^{\frac{\kappa - \kappa \alpha + d}{2\kappa + d}} (\log N)^{c^*}$$

This completes the proof of Theorem 4.1. \square

The proof of Theorem 4.1 above needs the following lemma.

Lemma 4.3. *Suppose the conditions of Theorem 4.1 are satisfied. If the events $A_{t,B}$ and $F_{t,B}$ ($1 \leq t \leq T$) are defined as in (4.16) and (4.17), respectively, then given any κ_0 satisfying $\kappa - \frac{\Delta}{\log n} - \frac{b_2 \log \log n}{\log n} < \hat{\kappa}^* \leq \kappa$,*

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

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

Proof of Lemma 4.3. Given $2 \leq t \leq T - 1$ and $B \in \mathcal{B}_T$, to find $P(F_{t-1,B} \cap A_{t,B}^c)$, 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 $N_{\tilde{B},i} \neq 0$, $\bar{Y}_{\tilde{B},i} = \sum_{n \in H_{\tilde{B},i}} Y_{i,n} / N_{\tilde{B},i}$. Then, by the arm elimination mechanism, there exists an arm $i_2 \in \mathcal{S}_{\tilde{B}}$ such that

$$\bar{Y}_{\tilde{B},i_2} - \bar{Y}_{\tilde{B},i_1} > \alpha_t = 4\rho h_t^{\kappa_0}. \quad (4.25)$$

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

$$\begin{aligned} \bar{f}_{\tilde{B},i_2} - \bar{f}_{\tilde{B},i_1} &= \frac{\sum_{n \in H_{\tilde{B},i_2}} f_{i_2}(X_n)}{N_{\tilde{B},i_2}} - \frac{\sum_{n \in H_{\tilde{B},i_1}} f_{i_1}(X_n)}{N_{\tilde{B},i_1}} \\ &\leq \max_{x \in \tilde{B}} f^*(x) - \frac{\sum_{n \in H_{\tilde{B},i_1}} f_{i_1}(X_n)}{N_{\tilde{B},i_1}} \\ &= \frac{\sum_{n \in H_{\tilde{B},i_1}} (\max_{x \in \tilde{B}} f^*(x) - f_{i_1}(X_n))}{N_{\tilde{B},i_1}}. \end{aligned} \quad (4.26)$$

Since $i_1 \in \mathcal{S}_{t,B,1}$, by Assumption 4.2, for every $x' \in \tilde{B}$, $\max_{x \in \tilde{B}} f^*(x) - f_{i_1}(x') \leq 2\rho h_t^\kappa$. Therefore, we have by (4.26) that

$$\bar{f}_{\tilde{B},i_2} - \bar{f}_{\tilde{B},i_1} \leq 2\rho h_t^\kappa. \quad (4.27)$$

By (4.25), (4.27) and the fact that both arms i_1 and i_2 are in $\mathcal{S}_{p_{t-1}(B)}$, we conclude that under $F_{t-1,B} \cap G_{t,B,1}^c$, there exists an arm $i \in \mathcal{S}_{p_{t-1}(B)}$ such that $N_{\tilde{B},i} \neq 0$ and

$$|\bar{Y}_{\tilde{B},i} - \bar{f}_{\tilde{B},i}| = \left| \frac{\sum_{n \in H_{\tilde{B},i}} \varepsilon_n}{N_{\tilde{B},i}} \right| > \rho h_t^{\kappa_0}. \quad (4.28)$$

Next, we assume that the event $F_{t-1,B} \cap G_{t,B,1} \cap G_{t,B,2}^c$ happens. Since $G_{t,B,2}^c = \{\mathcal{S}_{\tilde{B}} \subseteq \mathcal{S}_{t,B,2}\}^c$, 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,

$$f_{i_4}(\tilde{x}) - f_{i_3}(\tilde{x}) > 8\rho h_t^{\kappa_0}. \quad (4.29)$$

Then, by Assumption 4.2, if $N_{\tilde{B},i_3} \neq 0$ and $N_{\tilde{B},i_4} \neq 0$,

$$\begin{aligned}
\bar{f}_{\tilde{B},i_4} - \bar{f}_{\tilde{B},i_3} &= \frac{\sum_{n \in H_{\tilde{B},i_4}} f_{i_4}(X_n)}{N_{\tilde{B},i_4}} - \frac{\sum_{n \in H_{\tilde{B},i_3}} f_{i_3}(X_n)}{N_{\tilde{B},i_3}} \\
&\geq \frac{\sum_{n \in H_{\tilde{B},i_4}} (f_{i_4}(\tilde{x}) - \rho h_t^\kappa)}{N_{\tilde{B},i_4}} - \frac{\sum_{n \in H_{\tilde{B},i_3}} (f_{i_3}(\tilde{x}) + \rho h_t^\kappa)}{N_{\tilde{B},i_3}} \\
&= f_{i_4}(\tilde{x}) - f_{i_3}(\tilde{x}) - 2\rho h_t^\kappa \\
&> 6\rho h_t^{\kappa_0},
\end{aligned} \tag{4.30}$$

where the last inequality follows by (4.29). 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 $N_{\tilde{B},i_4} \neq 0$,

$$|\bar{Y}_{\tilde{B},i_4} - \bar{Y}_{\tilde{B},i_3}| \leq \alpha_t = 4\rho h_t^{\kappa_0}. \tag{4.31}$$

By (4.30) and (4.31), we conclude that under $F_{t-1,B} \cap G_{t,B,1} \cap G_{t,B,2}^c$, 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

$$|\bar{Y}_{\tilde{B},i} - \bar{f}_{\tilde{B},i}| = \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}. \tag{4.32}$$

Combining (4.28) and (4.32), we know that under event $F_{t-1,B} \cap A_{t,B}^c$, 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

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

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

$$\begin{aligned}
&P(F_{t-1,B} \cap A_{t,B}^c) \\
&\leq P(\exists \text{ arm } i \in \mathcal{S}_{p_{t-1}(B)} \text{ such that } N_{\tilde{B},i} = 0) \\
&\quad + P\left(\exists \text{ arm } i \in \mathcal{S}_{p_{t-1}(B)} \text{ such that } N_{\tilde{B},i} \neq 0 \text{ and } \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}\right) \\
&\leq l \max_{1 \leq i \leq l} P\left(N_{\tilde{B},i} = 0 \mid \text{arm } i \in \mathcal{S}_{p_{t-1}(B)}\right) \\
&\quad + l \max_{1 \leq i \leq l} P\left(N_{\tilde{B},i} \neq 0, \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0} \mid \text{arm } i \in \mathcal{S}_{p_{t-1}(B)}\right).
\end{aligned} \tag{4.33}$$

Given $1 \leq i \leq l$, 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 (4.33), note that

$$P\left(N_{\tilde{B},i} = 0 \mid C_{t-1}^{(i)}\right) \leq P\left(\frac{N_{\tilde{B},i}}{N_t} \leq \frac{\underline{c}h_t^d \tilde{\pi}_t}{2} \mid C_{t-1}^{(i)}\right) \leq \exp\left(-\frac{3\underline{c}N_t h_t^d \tilde{\pi}_t}{28}\right), \quad (4.34)$$

where the last inequality follows by Lemma 2.2 and the fact that $P(X_n \in \tilde{B}, I_n = i \mid C_{t-1}^{(i)}) \geq \underline{c}h_t^d \tilde{\pi}_t$ for all $\tilde{N}_{t-1} + 1 \leq n \leq \tilde{N}_t$. To provide the upper bound for the second term in (4.33), define $H_{\tilde{B}} = \{n : \tilde{N}_{t-1} + 1 \leq n \leq \tilde{N}_t, X_n \in \tilde{B}\}$ 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,

$$\begin{aligned} & P\left(N_{\tilde{B},i} \neq 0, \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0} \mid C_{t-1}^{(i)}\right) \\ & \leq P\left(\frac{N_{\tilde{B}}}{N_t} \leq \frac{\underline{c}h_t^d}{2}\right) + P\left(N_{\tilde{B},i} \neq 0, \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}, \frac{N_{\tilde{B}}}{N_t} > \frac{\underline{c}h_t^d}{2} \mid C_{t-1}^{(i)}\right) \\ & \leq P\left(\frac{N_{\tilde{B}}}{N_t} \leq \frac{\underline{c}h_t^d}{2}\right) + E_c P_{X^t}\left(N_{\tilde{B},i} \neq 0, \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}, \frac{N_{\tilde{B}}}{N_t} > \frac{\underline{c}h_t^d}{2}\right), \end{aligned} \quad (4.35)$$

where $P_{X^t}(\cdot)$ denotes the conditional probability given $(X_{N_{t-1}+1}, X_{N_{t-1}+2}, \dots, X_{N_t})$, $C_{t-1}^{(i)}$ and $\{\hat{\kappa}^* = \kappa_0\}$, and $E_c(\cdot)$ denotes the conditional expectation given $C_{t-1}^{(i)}$ and $\{\hat{\kappa}^* = \kappa_0\}$. Since $P(X_n \in \tilde{B}) \geq \underline{c}h_t^d$, by the extended Bernstein's inequality,

$$P\left(\frac{N_{\tilde{B}}}{N_t} \leq \frac{\underline{c}h_t^d}{2}\right) \leq \exp\left(-\frac{3\underline{c}N_t h_t^d}{28}\right). \quad (4.36)$$

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

$$\begin{aligned} & P_{X^t}\left(N_{\tilde{B},i} \neq 0, \frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}\right) \\ & \leq P_{X^t}\left(\frac{N_{\tilde{B},i}}{N_{\tilde{B}}} \leq \frac{\tilde{\pi}_t}{2}\right) + P_{X^t}\left(\frac{|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n|}{N_{\tilde{B},i}} > \rho h_t^{\kappa_0}, \frac{N_{\tilde{B},i}}{N_{\tilde{B}}} > \frac{\tilde{\pi}_t}{2}\right) \\ & \leq \exp\left(-\frac{3N_{\tilde{B}} \tilde{\pi}_t}{28}\right) + P_{X^t}\left(\left|\sum_{n \in H_{\tilde{B},i}} \varepsilon_n\right| > \frac{N_{\tilde{B}} \tilde{\pi}_t \rho h_t^{\kappa_0}}{2}\right), \end{aligned} \quad (4.37)$$

where the last inequality follows by Lemma 2.2 and the fact that $P(I_n = i \mid X_n \in \tilde{B}) \geq \tilde{\pi}_t$

for all $\tilde{N}_{t-1} + 1 \leq n \leq \tilde{N}_t$. Define $W_{n,i} = I(I_n = i)$. Then

$$\begin{aligned}
& P_{X^t} \left(\left| \sum_{n \in H_{\tilde{B},i}} \varepsilon_n \right| > \frac{N_{\tilde{B}} \tilde{\pi}_t \rho h_t^{\kappa_0}}{2} \right) \\
&= P_{X^t} \left(\left| \sum_{n \in H_{\tilde{B}}} W_{n,i} \varepsilon_n \right| > \frac{N_{\tilde{B}} \tilde{\pi}_t \rho h_t^{\kappa_0}}{2} \right) \\
&\leq \exp \left(-\frac{N_{\tilde{B}} \tilde{\pi}_t^2 \rho^2 h_t^{2\kappa_0}}{8(v^2 + c\tilde{\pi}_t \rho h_t^{\kappa_0}/2)} \right), \tag{4.38}
\end{aligned}$$

where the last inequality follows by Lemma 2.1 and Assumption 0. Thus, by (4.37) and (4.38),

$$\begin{aligned}
& P_{X^t} \left(N_{\tilde{B},i} \neq 0, \frac{\left| \sum_{n \in H_{\tilde{B},i}} \varepsilon_n \right|}{N_{\tilde{B},i}} > \rho h_t^\kappa, \frac{N_{\tilde{B}}}{N_t} > \frac{c h_t^d}{2} \right) \\
&\leq \begin{cases} 0 & \text{if } \frac{N_{\tilde{B}}}{N_t} \leq \frac{c h_t^d}{2}, \\ \exp \left(-\frac{3N_{\tilde{B}} \tilde{\pi}_t}{28} \right) + \exp \left(-\frac{N_{\tilde{B}} \tilde{\pi}_t^2 \rho^2 h_t^{2\kappa_0}}{8(v^2 + c\tilde{\pi}_t \rho h_t^{\kappa_0}/2)} \right) & \text{if } \frac{N_{\tilde{B}}}{N_t} > \frac{c h_t^d}{2}. \end{cases} \tag{4.39}
\end{aligned}$$

Combining (4.33)-(4.36) and (4.39), we have

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

It follows immediately by (4.3) that $P(F_{t-1,B} \cap A_{t,B}^c) \leq 4l/Nh_t^{2\kappa_0+d}$.

Lastly, noting that $P(A_{1,B}^c) \leq 4l/Nh_1^{2\kappa_0+d}$ can be derived by the same argument as that of $P(F_{t-1,B} \cap A_{t,B}^c) \leq 4l/Nh_t^{2\kappa_0+d}$, we complete the proof of Lemma 4.3. \square

Chapter 5

Conclusion

In this dissertation, under a general framework that allows for both binary and continuous responses, we focus our attention on a randomized allocation strategy that has the flexibility to incorporate different regression methods. In particular, we study the application of kernel regression method under a nonparametric framework, and evaluate the algorithm performance by studying the cumulative regret by both asymptotic and finite-time analysis. For asymptotic analysis, the Nadaraya-Watson estimation is shown to satisfy a uniform strong consistency, which implies the asymptotic optimality of the proposed algorithm. For the finite-time analysis, although the derived upper bound is sub-optimal in the minimax sense, 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. Moreover, by integrating a model combination strategy in together with the dimension reduction technique, the kernel estimation based randomized allocation strategy is shown to be very flexible in our simulation and real data evaluation studies.

As another main contribution of this dissertation, we attempt to design an algorithm that can be adaptive to the Hölder smoothness parameter. To achieve this goal, we investigate a smoothness parameter selection algorithm modified from the Lepski's method, and show that the cumulative regret of the randomized allocation with arm elimination strategy can achieve a minimax-optimal rate up to a logarithmic factor when the smoothness parameter is unknown.

It is generally assumed in the nonparametric MABC literature that the smoothness

parameter κ under the Hölder condition is no more than 1. In the future, one interesting but challenging direction is to see if a better finite-time result can be achieved with a more general smoothness condition. It is also of interest to see if efficient algorithm can be designed when covariates and arm features are simultaneously considered and their corresponding dimensions are both high.

References

- Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, 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 20th Annual Conference on Learning Theory*, 2007.
- 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 non stochastic 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.
- A. D. Bull. Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics*, 6:1490–1516, 2012.
- 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, 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 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 27th Annual Conference on Uncertainty in Artificial Intelligence*, 2011.

- E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.
- E. Giné and R. Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38:1122–1170, 2010.
- J. C. Gittins. *Multi-Armed Bandit Allocation Indices*. Wiley, New York, 1989.
- A. Goldenshluger and A. Zeevi. Woodrooffe’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.
- W. Härdle and S. Luckhaus. Uniform consistency of a class of regression function estimators. *The Annals of Statistics*, 12:612–623, 1984.
- B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748, 2008.
- W. Härdle, G. Kerkycharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics*. Springer, New York, 1998.
- M. Hoffmann and R. Nickl. On adaptive inference and confidence bands. *The Annals of Statistics*, 39:2383–2409, 2011.
- R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of 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 *Advances in Neural Information Processing Systems 20*, 2008.
- O. V. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, 35:454–466, 1990.

- O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25:929–947, 1997.
- 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.
- M. G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25:2547–2554, 1997.
- T. Lu, D. Pál, and M. Pál. Showing relevant ads via lipschitz context multi-armed bandits. In *Proceedings of 14th International Conference on Artificial Intelligence and Statistics*, 2010.
- R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, Wiley, 1959.
- O.-A. Maillard and R. Munos. Adaptive bandits: Towards the best history-dependent strategy. In *Proceedings of 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.
- 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 24th Annual Conference on Learning Theory*, pages 679–702, 2011.
- J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proceedings of the 16th European Conference on Machine Learning*, 2005.
- 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. Woodroffe. 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.
- H. Zou and Y. Yang. Combining time series models for forecasting. *International Journal of Forecasting*, 20:69–84, 2004.