Nonparametric Learning and Optimization with Covariates

Ningyuan Chen*[1] and Guillermo Gallego†[1]

[1]Department of Industrial Engineering & Decision Analytics
The Hong Kong University of Science and Technology

Abstract

Modern decision analytics frequently involves the optimization of an objective over a finite horizon where the functional form of the objective is unknown. The decision analyst observes covariates and tries to learn and optimize the objective by experimenting with the decision variables. We present a nonparametric learning and optimization policy with covariates. The policy is based on adaptively splitting the covariate space into smaller bins (hyper-rectangles) and learning the optimal decision in each bin. We show that the algorithm achieves a regret of order $O(\log(T)^2 T^{(2+d)/(4+d)})$, where T is the length of the horizon and d is the dimension of the covariates, and show that no policy can achieve a regret less than $O(T^{(2+d)/(4+d)})$ and thus demonstrate the near optimality of the proposed policy. The role of d in the regret is not seen in parametric learning problems: It highlights the complex interaction between the nonparametric formulation and the covariate dimension. It also suggests the decision analyst should incorporate contextual information selectively.

Keywords: multi-armed bandit, dynamic pricing, learning, regret analysis

*nychen@ust.hk
†ggallego@ust.hk

1. Introduction

Decision analysis in modern times is often extremely complex. Consider the following motivating example. A firm is pricing a product to customers with different profiles which can be based on their education, zip codes or other available data. The demand function is unknown and likely to be profile-based. The firm observes the profile of each arriving customer (e.g., through membership programs) and applies profile-based price discrimination. To maximize the long run revenue, what price should the firm charge for each customer?

This problem presents several challenges to the decision analyst (i.e., the firm). First, the functional form of the objective is unknown. Thus, the optimal decision cannot be obtained by solving a parametric optimization problem. There is usually a finite horizon that forces a trade-off between gathering more information (learning/exploration) and making sound decisions (earning/exploitation). Such problems, sometimes referred to as the exploration/exploitation dilemma, have attracted the attention of many researchers.

A second challenge is the increasing presence of contextual information, or covariates. In the example, the firm can access the profile information of each customer. In the context of retailing, the contextual information has been used extensively and the displayed set of product may depend on the customer's age, education background, and purchasing history. In medical decision making, doctors prescribe based on the health records or even genetic profiles of patients. On one hand, covariates provide extra information to help make better decisions. On the other hand, the objective function is peculiar to each instance and changes over time. This adds to the complexity of the unknown objective function. In the above example, the firm can only hope to learn the profile-based demand function by experimenting prices for many customers with similar profiles.

A third challenge is the selection of a predictive model. The choice of model reflects the decision analyst's belief about the unknown objective. In the above example, suppose the demand is based on the location of the customer and the firm postulates a linear model

$$\text{demand} = a - b \times \text{price} + \boldsymbol{c} \cdot (\text{latitude, longitude}).$$

Then the firm uses historical sales data to learn the parameters a, b and $\boldsymbol{c}$ and maximize revenue according to the estimation. However, whether the model is specified correctly

plays an important role in the performance of the policy. What if the actual demand is exponential in price? Or, the dependence on the coordinates is not linear but has a clustered structure, corresponding to community neighborhoods. By postulating a particular model, the decision analyst faces the risk of misspecification and not learning what is supposed to be learned.

In this paper we analyze a general learning problem that features the exploration/exploitation trade-off in the presence of covariates. We consider a decision analyst who tries to maximize the unknown expected reward function $f(\boldsymbol{x}, p)$, where p is a continuous decision variable and x represents the covariate which is an argument of the function and thus affects the optimal decision. We assume both are normalized so $\boldsymbol{x} \in [0, 1)^d$ and $p \in [0, 1]$. In period t, the decision analyst observes a random covariate $\boldsymbol{X}_t$, and s/he then relies on the past observations, i.e., the realized covariates, the made decisions, and the earned reward in periods $s < t$, to make a decision p at t. The earned reward is random and has expected value $f(\boldsymbol{X}_t, p)$. The primary motivation of the study is personalized dynamic pricing, in which p is the price, $\boldsymbol{X}_t$ is regarded as the profile of the tth customer, and $f(\boldsymbol{X}_t, p)$ represents the expected revenue.

We propose a nonparametric policy for the decision analyst. The policy achieves near-optimal performance compared to a clairvoyant decision analyst who knows $f(\boldsymbol{x}, p)$ and sets $p^*(\boldsymbol{x}) = \operatorname{argmax}_p f(\boldsymbol{x}, p)$ in each period. More precisely, the expected difference in total rewards between the proposed policy and the clairvoyant policy, which is referred to as the *regret* in the literature, grows at $O(\log(T)^2 T^{(2+d)/(4+d)})$ as $T \to \infty$. The rate is sublinear in T, implying that when the length of horizon tends to infinity, the average regret incurred per period becomes negligible. Moreover, we prove that no non-anticipating policies can achieve a lower regret than $O(T^{(2+d)/(4+d)})$ for a reasonable class of unknown reward functions f. Therefore, we successfully work out the exploration/exploitation dilemma with covariates.

There are two main contributions made in this paper. To the best of our knowledge, nonparametric policies have only been proposed for learning problems with discrete decision variables (multi-armed bandit problems) or without covariates in the previous literature. The design of our policy is substantially different from the UCB-type policies commonly used in learning problems. Instead, we incorporate the idea from tree-based methods in statistical learning (Hastie et al., 2001), which is a popular nonparametric predictive model but does not have an "exploitation" element. As a result, the regret

analysis of the policy does not follow the standard approach in the learning literature. We show that our policy achieves the best achievable rate of regret except for logarithm terms of T. Thus, the nonparametric solution in this paper cannot be further improved for the problem we consider.

Second, the best achievable rate of regret derived in this paper, $O(T^{(2+d)/(4+d)})$, sheds light on the complex nature of the interaction between the nonparametric formulation and the covariate dimension. Without covariates, it has been shown that parametric and nonparametric methods can achieve the same rate of regret $O(\sqrt{T})$ (e.g., compare Besbes and Zeevi 2009; Wang et al. 2014 and Keskin and Zeevi 2014; den Boer and Zwart 2014). Assuming a linear form of the covariate, Qiang and Bayati (2016); Javanmard and Nazerzadeh (2016); Ban and Keskin (2017) have shown that the best achievable rate of regret is still $O(\sqrt{T})$ or $O(\log(T))$, depending on specific assumptions. Our result demonstrates that lack of knowledge of the reward function's parametric form is extremely costly when the covariate is high-dimensional. In particular, the regret grows at $O(T^{(2+d)/(4+d)})$, which is almost linear in T when d is large.[1] This finding also suggests the decision analyst incorporate contextual information selectively when he/she is not confident in parametric models and decides to adopt a nonparametric formulation.

1.1. Literature Review

This paper is motivated by the recent literature that analyzes a firm's pricing problem when the demand function is unknown (e.g. Besbes and Zeevi, 2009; Araman and Caldentey, 2009; Farias and Van Roy, 2010; Broder and Rusmevichientong, 2012; den Boer and Zwart, 2014; Keskin and Zeevi, 2014; Cheung et al., 2017). den Boer (2015) provides a comprehensive survey for this area. Since the firm does not know the optimal price, it has to experiment different (suboptimal) prices and update its belief about the underlying demand function. Therefore, the firm has to balance the exploration/exploitation trade-off, which is usually referred to as the learning-and-earning problem in this line of literature. Our paper considers a generic version of the problem with contextual information and it does not consider the finite-inventory setting as in some of the papers mentioned above.

[1] This phenomenon also exists for multi-armed bandit problems (Goldenshluger and Zeevi, 2013; Rigollet and Zeevi, 2010).

More recently, several papers investigate the pricing problem with unknown demand in the presence of covariates (Nambiar et al., 2016; Qiang and Bayati, 2016; Javanmard and Nazerzadeh, 2016; Cohen et al., 2016; Ban and Keskin, 2017). In the pricing context, the covariate represents the personalized feature of a customer, such as his/her age, education background, marital status, etc., that are observed by the firm. Thus, on top of the unknown demand function, the firm has to learn the relationship between the customer features and the demand. The existing literature has adopted a parametric approach: in particular, the actual demand can be expressed in a linear form $\alpha^T \boldsymbol{x} + \beta^T \boldsymbol{x} p + \epsilon$, where $\boldsymbol{x}$ is the feature vector of a customer, α and β are vectorized coefficients, and ϵ is the random noise. Because of the parametric form, a key ingredient in the design of the algorithms in this line of literature is to plug in an estimator for the unknown parameters (α and β) in addition to some form of forced exploration. In contrast, we focus on a setting where the reward function (or equivalently, the demand function in the pricing problem) cannot be parametrized. Thus, the decision analyst cannot count on accurately estimating the reward function globally by estimating a few parameters. Instead, a localized optimal decision has to be made based on past covariates generated in the neighborhood. It highlights the different philosophies when designing algorithms for parametric/nonparametric learning problems with covariates. As a result, the best achievable regret deteriorates from $O(\sqrt{T})$ (parametric) to $O(T^{(2+d)/(4+d)})$ (nonparametric).

The dependence of the optimal rate of regret on the problem dimension d has been observed before. For example, Cohen et al. (2016) find a multi-dimensional binary search algorithm for feature-based dynamic pricing, which has regret $O(d^2 \log(T/d))$; Javanmard and Nazerzadeh (2016) propose a policy for a similar problem that achieves regret $O(s \log d \log T)$, where s represents the sparsity of the d features; in Ban and Keskin (2017), the near-optimal policy achieves regret $O(s\sqrt{T})$. In their parametric frameworks, the dependence of the regret on d is rather mild—it does not appear on the exponent of T. In contrast, in our nonparametric formulation, the optimal rate of regret $O(T^{(2+d)/(4+d)})$ increases dramatically in d, making the problem significantly harder to learn in high dimensions. This is similar to the nonparametric formulation in the network revenue management problem (Besbes and Zeevi, 2012), in which the dimension of the decision space is d and the optimal rate of regret is $O(T^{(2+d)/(3+d)})$. It seems such complexity only arises as a result of the interaction between the nonpara-

metric formulation and high dimensions: as shown in Besbes and Zeevi (2009); Wang et al. (2014); Lei et al. (2017), in the finite-inventory setting, nonparametric policies can achieve the same best achievable regret, $O(\sqrt{T})$, as parametric policies (den Boer and Zwart, 2014; Keskin and Zeevi, 2014).

Decision trees have been a very popular method in nonparametric statistical learning; see Breiman (2017); Hastie et al. (2001) for references. Based on decision trees, powerful statistical tools such as Random Forests (Breiman, 2001) and Gradient Boosting (Friedman, 2001) have been developed. The idea has been adapted to contextual learning in, e.g., Féraud et al. (2016); Elmachtoub et al. (2017); Lee and Chen (2017). The key component in the algorithm of this paper, adaptive binning, can also be viewed as a nonstandard form of decision tree. Another link of our paper to the statistics literature is the idea we use to prove the lower bound; see Györfi et al. (2006); Tsybakov (2009) for a more comprehensive description of these techniques in nonparametric estimation.

This paper is also related to the vast literature studying multi-armed bandit problems. See Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012) for a comprehensive survey. The classic multi-armed bandit problem involves finite arms, and the algorithms (Kuleshov and Precup, 2014; Agrawal and Goyal, 2012) cannot be applied directly to our setting. Recently, there is a stream of literature studying the so-called continuum-armed bandit problems (Agrawal, 1995; Kleinberg, 2005; Auer et al., 2007; Kleinberg et al., 2008; Bubeck et al., 2011), in which there are infinite number of arms (decisions). Although there is no contextual information in those papers, Kleinberg et al. (2008); Bubeck et al. (2011) have developed algorithms based on a similar idea to decision trees, because the potential arms form a high-dimensional space.

For multi-armed bandit problems with contextual information, parametric (regression based) algorithms have been proposed in, for example, Goldenshluger and Zeevi (2013); Bastani and Bayati (2015). Our paper is closely related to the literature studying contextual multi-armed bandit problems in a nonparametric framework Yang et al. (2002); Langford and Zhang (2008); Rigollet and Zeevi (2010); Perchet and Rigollet (2013); Slivkins (2014); Elmachtoub et al. (2017). Among them, Perchet and Rigollet (2013); Slivkins (2014) are the most relevant to this paper: they have used a similar idea to decision trees and derived dimension-dependent rate of regret. Our algorithm differs from those two papers in the following aspects: the algorithm design, the assumptions, and the achieved optimal rate of regret. For clarity, we defer the comparison and

discussion to Section 6 after the analysis of our algorithm.

2. Problem Formulation

Consider a function $f(\boldsymbol{x}, p)$, where $\boldsymbol{x} \in [0,1)^d$ and $p \in [0,1]$. Given $\boldsymbol{x}$, the function $f(\boldsymbol{x}, p)$ has a unique maximum $f^*(\boldsymbol{x})$, attained at $p^*(\boldsymbol{x})$. Here the scalar p represents the decision variable and $\boldsymbol{x}$ represents the covariate.

Initially, the form of f is unknown to the decision analyst. For $t = \{1, 2, \ldots, T\}$, the covariate $\boldsymbol{X}_t$ is generated and observed by the decision analyst sequentially. In period t, the decision analyst applies a non-anticipating policy π_t. Given $\boldsymbol{X}_t$, the reward in that period, Z_t, is a random variable with mean $f(\boldsymbol{X}_t, \pi_t)$ and is independent of everything else. The objective of the decision analyst is to design a policy π_t to maximize the total reward $\sum_{t=1}^{T} \mathsf{E}[f(\boldsymbol{X}_t, \pi_t)]$. The information structure of π_t requires that when making the decision in period t, π_t only relies on $\mathcal{F}_{t-1} \triangleq \sigma(\boldsymbol{X}_1, Z_1, \ldots, \boldsymbol{X}_{t-1}, Z_{t-1})$ as well as $\boldsymbol{X}_t$.

The main application of our model is personalized dynamic pricing. In this case $f(\boldsymbol{x}, p) = (p-c)d(\boldsymbol{x}, p)$ is the revenue/profit function and $d(\boldsymbol{x}, p)$ is the expected demand at $(\boldsymbol{x}, p)$, and c is the unit cost. Our method, however, can also be applied to the supply side where the price p is fixed and we are trying to maximize the expected profit $f(\boldsymbol{x}, q) = p\mathsf{E}[\min\{D(\boldsymbol{x}, p), q\}] - cq$ over the order quantity q.

2.1. Regret

A standard measure used in the literature for the performance of a policy is the regret incurred compared to the *clairvoyant* policy. Suppose $f(\boldsymbol{x}, p)$ is known to a clairvoyant decision analyst. Then the optimal policy is to set $p^*(\boldsymbol{X}_t)$ in period t and obtain a random reward with mean $f^*(\boldsymbol{X}_t)$. For the decision analyst we consider, the mean of the reward in period t is $f(\boldsymbol{X}_t, \pi_t)$. Clearly, its reward is less than the clairvoyant policy on average. Therefore, we define the regret of a policy π_t to be the expected reward gap

$$R_\pi(T) = \sum_{t=1}^{T} \mathsf{E}\left[(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t))\right].$$

In period t, the expectation is taken with respect to the distribution of $\boldsymbol{X}_t$ as well as π_t, which itself depends on $\mathcal{F}_{t-1}$ and $\boldsymbol{X}_t$. Our goal is to design a policy π_t that achieves

small $R_\pi(T)$ when $T \to \infty$.

However, because $R_\pi(T)$ also depends on the unknown function f, we require the designed policy to perform well for a wide class C of functions, i.e., we seek for optimal policies in terms of the minimax regret

$$\inf_{\pi_t} \sup_{f \in C} R_\pi(T).$$

Although it is usually impossible to find the exact policy that achieves the minimax regret, we focus on proposing a policy whose regret is at least comparable to (of the same order as) the minimax regret asymptotically when $T \to \infty$.

Evaluating policies by $\sup_{f \in C} R_\pi(T)$ is standard in the literature. The rationale for such measure can be illustrated by the following simple example. A naive policy $\pi_t \equiv 0$ may perform no worse than the clairvoyant policy for $f(\boldsymbol{x}, p) \equiv -p^2$; but is drastically outperformed by other well designed policies for a general reward function. Therefore, it is reasonable to focus on the worst-case regret in terms of possible reward functions. As a result, the class of functions C is particularly important in the regret analysis. In general, a larger class implies a worse minimax regret and vice versa.

2.2. Assumptions

In this section, we provide a set of assumptions that $f \in C$ has to satisfy and their justifications.

Assumption 1. The covariates $\boldsymbol{X}_t$ are i.i.d. for $t = 1, \ldots, T$. Given $\boldsymbol{x}$ and p, the reward Z satisfies $\mathsf{E}[Z|\boldsymbol{x}, p] = f(\boldsymbol{x}, p)$. It is independent of everything else and its distribution is sub-Gaussian, i.e., there exists a constant $\sigma > 0$ such that $\mathsf{E}[\exp(u(Z - f(\boldsymbol{x}, p)))] \leq \exp(\sigma u^2)$ for all $u \in \mathbb{R}$.

Both i.i.d. covariates and independent noise structure are standard in the literature. For the noise, there are usually two common forms: additive noise, i.e., $Z = f(\boldsymbol{x}, p) + \epsilon$ with i.i.d. noise ϵ, and binary outcomes, i.e., Zs are independent Bernoulli random variables with success rate $f(\boldsymbol{x}, p)$. Both forms are covered by this assumption. The sub-Gaussian assumption is merely a technical simplification: it implies that the noise does not have heavy tails, which is satisfied by any bounded reward (e.g., binary outcomes)

or normally distributed reward (e.g., additive noise with normal distribution). As shown in Bubeck et al. (2013) and discussed in Perchet and Rigollet (2013), it can be relaxed without affecting the regret bound.

Assumption 2. The functions $f(\cdot, p)$ and $f(\boldsymbol{x}, \cdot)$ are Lipschitz continuous given p and $\boldsymbol{x}$, i.e., there exists $M_1 > 0$ such that $|f(\boldsymbol{x}_1, p) - f(\boldsymbol{x}_2, p)| \leq M_1 \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ and $|f(\boldsymbol{x}, p_1) - f(\boldsymbol{x}, p_2)| \leq M_1 |p_1 - p_2|$ for all $\boldsymbol{x}_i$ and p_i $(i = 1, 2)$ in the domain.

This assumption is equivalent to $|f(\boldsymbol{x}_1, p_1) - f(\boldsymbol{x}_2, p_2)| \leq M_1(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + |p_1 - p_2|)$. Lipschitz continuity is a common assumption in the learning literature. If this assumption fails, past experiments are not informative even for a small neighborhood of their associated covariates and learning is virtually impossible. In revenue management applications, $f(\boldsymbol{x}, p) = pd(\boldsymbol{x}, p)$ is the revenue function and Assumption 2 only requires the demand function $d(\boldsymbol{x}, p)$ to be Lipschitz continuous in $\boldsymbol{x}$ and p.

Assumption 3. For any hyperrectangle $B \subset [0, 1)^d$, including a singleton $B = \{\boldsymbol{x}\}$, define the function $f_B(p) \triangleq \mathsf{E}[f(\boldsymbol{X}, p) | \boldsymbol{X} \in B]$. We assume that for any B,

1. The function $f_B(p)$ has a unique maximizer $p^*(B) \in [0, 1]$. Moreover, there exist uniform constants $M_2, M_3 > 0$ such that for all $p \in [0, 1]$, $M_2(p^*(B) - p)^2 \leq f_B(p^*(B)) - f_B(p) \leq M_3(p^*(B) - p)^2$.

2. The maximizer $p^*(B)$ is inside the interval $[\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}, \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}]$.

3. Let d_B be the diameter of B. Then there exists a uniform constant $M_4 > 0$ such that $\sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\} - \inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\} \leq M_4 d_B$.

This assumption is quite different from those in the no-covariates setting (Besbes and Zeevi, 2009; Wang et al., 2014; Lei et al., 2017) or the parametric setting (Ban and Keskin, 2017; Qiang and Bayati, 2016). To explain the intuition of the function $f_B(p)$, consider the following learning problem associated with B without covariates. If the decision analyst only observes $\mathbb{I}_{\{X \in B\}}$ but not the exact value of $\boldsymbol{X}$, then the randomness in $\boldsymbol{X}$ given $\boldsymbol{X} \in B$ becomes part of the noise. The learning objective is $f_B(p)$ and the clairvoyant policy that has the knowledge of $f(\boldsymbol{x}, p)$ is to set $p = p^*(B)$ in each period. This class of learning problems are important subroutines of the algorithm we propose and Assumption 3 guaranties that they can be effectively learned.

For part one of Assumption 3, we have the following result:

Proposition 1. *If $f_B(p)$ is continuous for $p \in [0, 1]$, and twice differentiable in an open interval containing the unique global maximizer $p^*(B)$ with $f_B''(p^*(B)) < 0$, then part one of Assumption 3 holds.*

Therefore, part one only requires smoothness and local concavity of $f_B(p)$ *at the global maximizer*. In particular, $f_B(p)$ does not even have to be unimodal. If B is a singleton, then it can be viewed as a weaker version of the concavity assumption in Wang et al. (2014); Lei et al. (2017), i.e., $0 < a < f''(p) < b$ for all p in their no-covariate setting, because Assumption 3 only requires local concavity. As a result, if $f_B(p)$ is the smooth revenue functions that are commonly used in the revenue management literature, e.g., $p \times (a - bp)$, $p \times \exp(-p/\theta)$, and $p \times ap^{-b}$, then part one is satisfied automatically. We also remark that the smoothness and local concavity at the maximizer imposed in part one *is not* a technical simplification. As we shall see in Section 6, the optimal rate of regret improves when the objective function is a little smoother than Lipschitz continuity at the maximum.

Part two of Assumption 3 prevents the following scenario: If $p^*(B)$ is far from $p^*(\boldsymbol{x})$ for $\boldsymbol{x} \in B$, even when d_B is relatively small, then collecting more information for $f_B(p)$ does not help to improve the decision making for any individual covariate $\boldsymbol{x} \in B$. Such obstacle may lead to failure to learn and is thus ruled out by the assumption. Similar types of assumptions have been imposed in revenue management. For example, one can show that the optimal price for the aggregated demand function lies in the convex hull of the optimal prices of individual demand functions under very mild conditions.

Part three imposes a continuity condition for maximizers. It is equivalent to, for example, some form of continuous differentiability of $f(\boldsymbol{x}, p)$, because $p^*(\boldsymbol{x})$ solves the implicit function from the first-order condition $f_p(\boldsymbol{x}, p) = 0$.

Remark 1. Assumptions 1 and 2 are variants of similar assumptions adopted in the literature. Assumption 3, although appearing nonstandard, is also satisfied by the parametric families studied by previous works. We give a few examples that satisfy Assumption 3.

- Dynamic pricing with linear covariate (Qiang and Bayati, 2016): if $f(\boldsymbol{x}, p) = p(\boldsymbol{\theta}^T \boldsymbol{x} - \alpha p)$, then $f_B(p) = p(\boldsymbol{\theta}^T \mathsf{E}[\boldsymbol{X} | \boldsymbol{X} \in B] - \alpha p)$ and $p^*(B) = \boldsymbol{\theta}^T \mathsf{E}[\boldsymbol{X} | \boldsymbol{X} \in B]/2\alpha$.

- Separable function: consider $f(\boldsymbol{x}, p) = \sum_{i=1}^{k} g_i(\boldsymbol{x}) h_i(p)$. Then $f_B(p) = \sum_{i=1}^{k} \mathsf{E}[g_i(\boldsymbol{X}) | \boldsymbol{X} \in B] h_i(p)$. If $h_i(p)$ are concave functions, then we may be able to solve the unique maximizer $p^*(B) = \mathsf{E}[g(\boldsymbol{X}) | \boldsymbol{X} \in B]$ for a continuous function g.
- Localized functions: the covariate only plays a role in a subset $B_0 \subset [0,1)^d$. See Section 5 for a concrete example.

On the other hand, from Section 5, the class of functions satisfying our assumptions includes highly localized functions that can hardly be represented by a parametric model.

We summarize the information available to the decision analyst. Before the learning begins, the length of the horizon T, the dimension d and the constants $\{M_i\}_{i=1}^4$ and σ are revealed to him/her[2]. In period t, the decision making can also depend on $\mathcal{F}_{t-1}$ and $\boldsymbol{X}_t$.

3. The ABE Algorithm

We next present a set of preliminary concepts related to the *bins* of the covariate space, and then introduce the Adaptive Binning and Exploration (ABE) algorithm.

3.1. Preliminary Concepts

Definition 1. A bin is a hyper-rectangle in the covariate space. More precisely, a bin is of the form

$$B = \{\boldsymbol{x} : a_i \le x_i < b_i,\ i = 1, \dots, d\}$$

for $0 \le a_i < b_i \le 1, i = 1, \dots, d$.

We can *split* a bin B by bisecting it in all the d dimensions to obtain 2^d *child* bins of B, all of equal size. For a bin B with boundaries a_i and b_i for $i = 1, \dots, d$, its children are indexed by $\boldsymbol{i} \in \{0,1\}^d$ and have the form

$$B_{\boldsymbol{i}} = \left\{ \boldsymbol{x} : a_j \le x_j < \frac{a_j + b_j}{2} \text{ if } \boldsymbol{i}_j = 0,\ \frac{a_j + b_j}{2} \le x_j < b_j \text{ if } \boldsymbol{i}_j = 1,\ j = 1, \dots, d \right\}.$$

[2] In fact, only σ and M_2 are needed in the algorithm.

Denote the set of child bins of B by $\boldsymbol{C}(B)$. Conversely, for any $B' \in \boldsymbol{C}(B)$, we refer to B as the *parent* bin of B', denoted by $P(B') = B$.

Our algorithm starts with a root bin $B_\emptyset \triangleq [0,1)^d$, which is the whole covariate space, and successively splits the bin as more data is collected. Therefore, any bin B produced during the process is the *offspring* of $B_\emptyset$, i.e., $P^{(k)}(B) = B_\emptyset$ for some $k > 0$, where $P^{(k)}$ is the kth composition of the parent function. Equivalently, $B_\emptyset$ is an *ancestor* of B. Therefore, one can use a sequence of indices $(\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_k)$ to represent a bin. As introduced above, the index $\boldsymbol{i}$ encodes the reference to a particular child bin when a parent bin is split. Likewise, $(\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_k)$ refers to a bin that is obtained by k split operations from $B_\emptyset$: when $B_\emptyset$ is split, we obtain its child $B_{\boldsymbol{i}_1}$; when $B_{\boldsymbol{i}_1}$ is split, we obtain its child $B_{\boldsymbol{i}_1\boldsymbol{i}_2}$; and so on. In the last operation, when $B_{\boldsymbol{i}_1\ldots\boldsymbol{i}_{k-1}}$ is split, we obtain its child $B_{\boldsymbol{i}_1\ldots\boldsymbol{i}_k}$. For such a bin, we define its *level* to be k, denoted by $l(B) = k$. Conventionally, let $l(B_\emptyset) = 0$.

In the algorithm, we keep a dynamic partition $\mathcal{P}_t$ of the covariate space consisting of offspring of $B_\emptyset$ in each period t. The partition is mutually exclusive and collectively exhaustive, so $B_i \cap B_j = \emptyset$ for $B_i, B_j \in \mathcal{P}_t$, and $\cup_{B_i \in \mathcal{P}_t} B_i = B_\emptyset$. Initially $\mathcal{P}_0 = \{B_\emptyset\}$. In the algorithm, we gradually refine the partition; that is, each bin in $\mathcal{P}_{t+1}$ has an ancestor (or itself) in $\mathcal{P}_t$.

An analogous, and probably more graphical, interpretation is to regard the sequential splitting as a *branching process* and relate it to decision trees in statistical learning. Consider $B_\emptyset$ as the *root* of a tree, or the initial *leaf* of the tree. When a split operation is performed, a leaf is branched into 2^d leaves. During the branching process, the set of all terminal leaves (those without offspring) form a partition of the covariate space. The algorithm involves gradually branching the tree as t increases and more data is collected.

3.2. Intuition

The intuitive idea behind the ABE algorithm is to use a partition $\mathcal{P}$ of the covariate space and try to find the optimal decision in each bin $B \in \mathcal{P}$, i.e., $p^*(B)$ defined in Section 2.2.

To do that, we keep a set of discrete decisions (referred to as the *decision set* hereafter) for each bin in the partition. The decision set consists of equally spaced grid points

of an interval associated with the bin. When a covariate $\boldsymbol{X}_t$ is generated inside a bin B, a decision is chosen successively in its decision set and applied to $\boldsymbol{X}_t$. The realized reward for this decision is recorded. When sufficient covariates are observed in B, the average reward for each decision p in the decision set is close to $f_B(p)$, which is defined as $\mathsf{E}[f(\boldsymbol{X}, p)|\boldsymbol{X} \in B]$ in Section 2.2. Therefore, the *empirically-optimal* decision in the decision set is close to $p^*(B)$, with high confidence.

There are two potential pitfalls of this approach. First, the number of decisions has an impact on the performance of the algorithm. If there are too many decisions in a set, then a given number of covariates generated in the associated bin need to be distributed among the decision set, with each getting relatively few observations. As a result, the confidence interval for the average reward is wide. On the other hand, if there are too few decisions, then inevitably the decision set has low resolution. That is, the *optimal* decision in the set could be far from the true maximizer $p^*(B)$. We have, therefore, to select a proper size for the decision set to balance this trade-off.

Second, even if the optimal decision $p^*(B)$ is correctly identified, it may not be a strong indicator for $p^*(\boldsymbol{x})$ for a particular $\boldsymbol{x} \in B$. Indeed, $f_B(p)$ averages out the randomness of $\boldsymbol{X} \in B$, and the optimal decision for individual $\boldsymbol{x}$ could be very different. This obstacle, however, can be overcome as the size B decreases, as implied by Assumption 3. In particular, part 2 and 3 of the assumption guarantee that when B is small, $p^*(\boldsymbol{x})$ is concentrated within a neighborhood of $p^*(B)$ as long as $\boldsymbol{x} \in B$. The cost of using a smaller bin, however, is the less frequency of observing a covariate inside it.

To remedy the second pitfall, the algorithm adaptively refines the partition (hence denoted $\mathcal{P}_t$) and decreases the size of the bins in $\mathcal{P}_t$ as t increases. When a bin $B \in \mathcal{P}_t$ is large, the optimal decision $p^*(B)$ is not a strong indicator for $p^*(\boldsymbol{x})$, $\boldsymbol{x} \in B$. As a result, we only need a rough estimate for it and split the bin when a relatively small number of covariates are observed in B. When a bin $B \in \mathcal{P}_t$ is small, its optimal decision $p^*(B)$ provides a strong indicator for $p^*(\boldsymbol{x})$, $\boldsymbol{x} \in B$. Therefore, we collect a large number of covariates $\boldsymbol{X} \in B$ to explore the decision set and accurately predict $p^*(B)$, before it splits.

A crucial step in the algorithm is to determine what information to inherit when a bin is split into child bins. The ABE algorithm records the empirically-optimal decision in the decision set of the parent bin. In the child bins, we use this information and set up their decision sets around it. As explained above, when the parent bin (and

thus the child bins) is large, its optimal decision does not predict those of the child bins well. Therefore, the algorithm sets up conservative decision sets for the child bins, i.e., they have wide intervals. On the other hand, when the parent bin is small, its optimal decision provides accurate indicator for those of the child bins. Thus, the algorithm constructs decision sets with narrow ranges for the child bins around the empirically-optimal decision inherited.

3.3. Description of the Algorithm

In this section, we elaborate on the detailed steps of the ABE algorithm, shown in Algorithm 1.

The parameters for the algorithm include

1. K, the maximal level of the bins. When a bin is at level K, the algorithm no longer splits it and simply applies the decision in its decision set whenever a covariate is generated in it.

2. Δ_k, the length of the interval that contains the decision set of level-k bins.

3. n_k, the maximal number of covariate observed in a level-k bin in the partition. When n_k covariates are observed, the bin splits.

4. N_k, the number of decisions to explore in the decision set of level-k bins. The decision set of bin B consists of equally spaced grid points of an interval $[p_l^B, p_u^B]$, to be adaptively specified by the algorithm.

We initialize the partition to include only the root bin $B_\emptyset$ in Step 4. Its decision set spans the whole interval $[0, 1]$ with N_0 equally spaced grid points. That is, the jth decision is $j\delta_{B_\emptyset} \triangleq j/(N_0 - 1)$ for $j = 0, \ldots, N_0 - 1$. The initial average reward and the number of explorations already applied to the jth decision are set to $\bar{Y}_{B_\emptyset,j} = N_{B_\emptyset,j} = 0$.

Suppose the partition is $\mathcal{P}_t$ at t and a covariate $\boldsymbol{X}_t$ is generated (Step 6). The algorithm determines the bin $B \in \mathcal{P}_t$ which the covariate falls into. The counter $N(B)$ records the number of covariates already observed in B up to t when B is in the partition (Step 8). If the level of B is $l(B) = k < K$ (i.e., B is not at the maximal level) and the number of covariates observed in B is not sufficient (Step 9 and Step 10), then the algorithm has assigned a decision set to the bin in previous steps, namely, $\{p_l^B + j\delta_B\}$

Algorithm 1 Adaptive Binning and Exploration (ABE)

Input: T, d
Constants: M_1, M_2, M_3, M_4, σ
Parameters: K; Δ_k, n_k, N_k for $k = 0, \ldots, K$
Initialize: partition $\mathcal{P} \leftarrow \{B_\emptyset\}$, $p_l^{B_\emptyset} \leftarrow 0$, $p_u^{B_\emptyset} \leftarrow 1$, $\delta_{B_\emptyset} \leftarrow 1/(N_0 - 1)$, $\bar{Y}_{B,j}, N_{B_\emptyset,j} \leftarrow 0$ for $j = 0, \ldots, N_0 - 1$
for $t = 1$ to T **do**
    Observe $\boldsymbol{X}_t$
    $B \leftarrow \{B \in \mathcal{P} : \boldsymbol{X}_t \in B\}$ ▷ The bin in the partition that $\boldsymbol{X}_t$ belongs to
    $k \leftarrow l(B)$, $N(B) \leftarrow N(B) + 1$ ▷ Determine the level and update the number of covariates observed in B
    if $k < K$ **then** ▷ If not reaching the maximal level K
        if $N(B) < n_k$ **then** ▷ If not sufficient data observed in B
            $j \leftarrow N(B) - 1 \pmod{N_k}$ ▷ Apply the jth grid point in the decision set
            $\pi_t \leftarrow p_l^B + j\delta_B$; apply π_t and observe Z_t
            $\bar{Y}_{B,j} \leftarrow \frac{1}{N_{B,j}+1}(N_{B,j}\bar{Y}_{B,j} + Z_t)$, $N_{B,j} \leftarrow N_{B,j} + 1$
        else ▷ When $N(B) = n_k$
            $j^* \in \operatorname{argmax}_{j \in \{0,1,\ldots,N_k-1\}} \{\bar{Y}_{B,j}\}$, $p^* \leftarrow p_l^B + j^*\delta_B$ ▷ Find the empirically-optimal decision; if there are multiple, choose any one of them
            $\mathcal{P} \leftarrow (\mathcal{P} \setminus B) \cup \mathcal{C}(B)$ ▷ Update the partition by removing B and adding its children
            for $B' \in \mathcal{C}(B)$ **do** ▷ Initialization for each child bin
                $N(B') \leftarrow 0$
                $p_l^{B'} \leftarrow \max\{0, p^* - \Delta_{k+1}/2\}$; $p_u^{B'} \leftarrow \min\{1, p^* + \Delta_{k+1}/2\}$ ▷ The range of the decision set
                $\delta_{B'} \leftarrow (p_u^{B'} - p_l^{B'})/(N_{k+1} - 1)$ ▷ The grid size of the decision set
                $N_{B',j}, \bar{Y}_{B',j} \leftarrow 0$, for $j = 0, \ldots, N_{k+1} - 1$ ▷ Initialize the average and number of explorations for each decision
            end for
        end if
    else ▷ If reaching the maximal level
        $\pi_t \leftarrow (p_l^B + p_u^B)/2$
    end if
end for

for $j = 0, \ldots, N_k - 1$. There are N_k decisions in the set and they are equally spaced in the interval $[p_l^B, p_u^B]$. They are explored sequentially as new covariates are observed in B (explore p_l^B for the first covariate observed in B, $p_l^B + \delta_B$ for the second covariate, $\ldots$, $p_l^B + (N_k - 1)\delta_B$ for the N_kth covariate, p_l^B again for the $(N_k + 1)$th covariate, etc.). Therefore, the algorithm applies decision $\pi_t = p_l^B + j\delta_B$ where $j = N(B) - 1 \pmod{N_k}$ to the $N(B)$th covariate observed in B (Step 11). Then, Step 13 updates the average reward and the number of explorations for the jth decision.

If the level of B is $l(B) = k < K$ and we have observed sufficient covariates in B (Step 9 and Step 14), then the algorithm splits B and replaces it by its 2^d child bins in the partition (Step 16). For each child bin, Step 18 to Step 21 initialize the counter, the interval that encloses the decision set, the grid size of the decision set, and the average reward/number of explorations that have been conducted for each decision in the decision set, respectively. In particular, to construct the decision set of a child bin, the algorithm first computes the empirically-optimal decision in the decision set of the parent bin B; that is, $j^* \in \operatorname{argmax}_{j \in \{0,1,\ldots,N_k-1\}} \{\bar{Y}_{B,j}\}$ in Step 15. Then, the algorithm creates an interval centered at this optimal decision with width Δ_{k+1}, properly cut off by the boundaries $[0, 1]$. The decision set is then an equally spaced grid of the above interval (Step 19 and Step 20).

If the level of B is already K, then the algorithm simply applies a single decision inherited from its parent (Step 25) repeatedly without further exploration. For such a bin, its size is sufficiently small and the algorithm has narrowed the range of the decision set K times. The applied decision, which is the middle point of the interval, is close enough to all $p^*(\boldsymbol{x})$, $\boldsymbol{x} \in B$, with high probability.

3.4. Choice of Parameters

We set $K = \lfloor \frac{\log(T)}{(d+4)\log(2)} \rfloor$, $\Delta_k = 2^{-k}\log(T)$, $N_k = \lceil \log(T) \rceil$, and

$$n_k = \max\left\{0, \left\lceil \frac{2^{4k+18}\sigma}{M_2^2 \log^3(T)} (\log(T) + \log(\log(T)) - (d+2)k\log(2)) \right\rceil \right\}.$$

To give a sense of their magnitudes, the maximal level of bins is $K \approx \log(T)/(d+4)$. The range of the decision set (Δ_k) is proportional to the edge length of the bin (2^{-k}). The number of decisions in a decision set is approximately $\log(T)$. Therefore, the grid

size $\delta_B \approx 2^{-k}$ for a level-k bin B. The number of covariates to collect in a level-k bin B is roughly $n_k \approx 2^{4k}/\log(T)^2$. When k is small, n_k can be zero according to the expression. In this case, the algorithm immediately splits the bin without collecting any covariate in it.

3.5. A Schematic Illustration

We illustrate the key steps of the algorithm by an example with $d = 2$. Figure 1 illustrates a possible outcome of the algorithm in period $t_1 < t_2 < t_3$ (top panel, mid panel, and bottom panel respectively). Up until period t_1, there is a single bin and the observed values $\boldsymbol{X}_t$ for $t \leq t_1$ are illustrated in the top left panel. The algorithm has explored the objective in the decision set, in this case $p \in \{0.1, 0.2, \ldots, 0.9\}$, and recorded the average reward $\bar{Y}_{B,j}$, illustrated by the top right panel. At $t_1 + 1$, sufficient observations are collected and Step 14 is triggered in the algorithm. Therefore, the bin is split into four child bins.

From period $t_1 + 1$ to t_2, new covariates are observed in each child bin (mid left panel). Note that the covariates generated before t_1 in the parent bin are no longer used and colored in gray. For each child bin (the bottom-left bin is abbreviated as BL, etc.), the average reward for the decision set is demonstrated in the mid right panel. The decision sets are centered at the empirically-optimal decision of their parent bin, in this case $p^* = 0.6$ from the top right panel. They have narrower ranges and finer grids than that of the parent bin. At $t_2 + 1$, sufficient observations are collected for BL, and it is split into four child bins.

From period $t_2 + 1$ to t_3, the partition consists of seven bins, as shown in the bottom left panel. The BR, TL and TR bins keep collecting covariates and updating the average reward, because they have not collected sufficient data. Their status at t_3 is shown in the bottom panels. In the four newly created child bins of BL (the bottom-left bin of BL is abbreviated as BL-BL, etc.), the decisions in the decision sets are applied successively and their average rewards are illustrated in the bottom right panel.

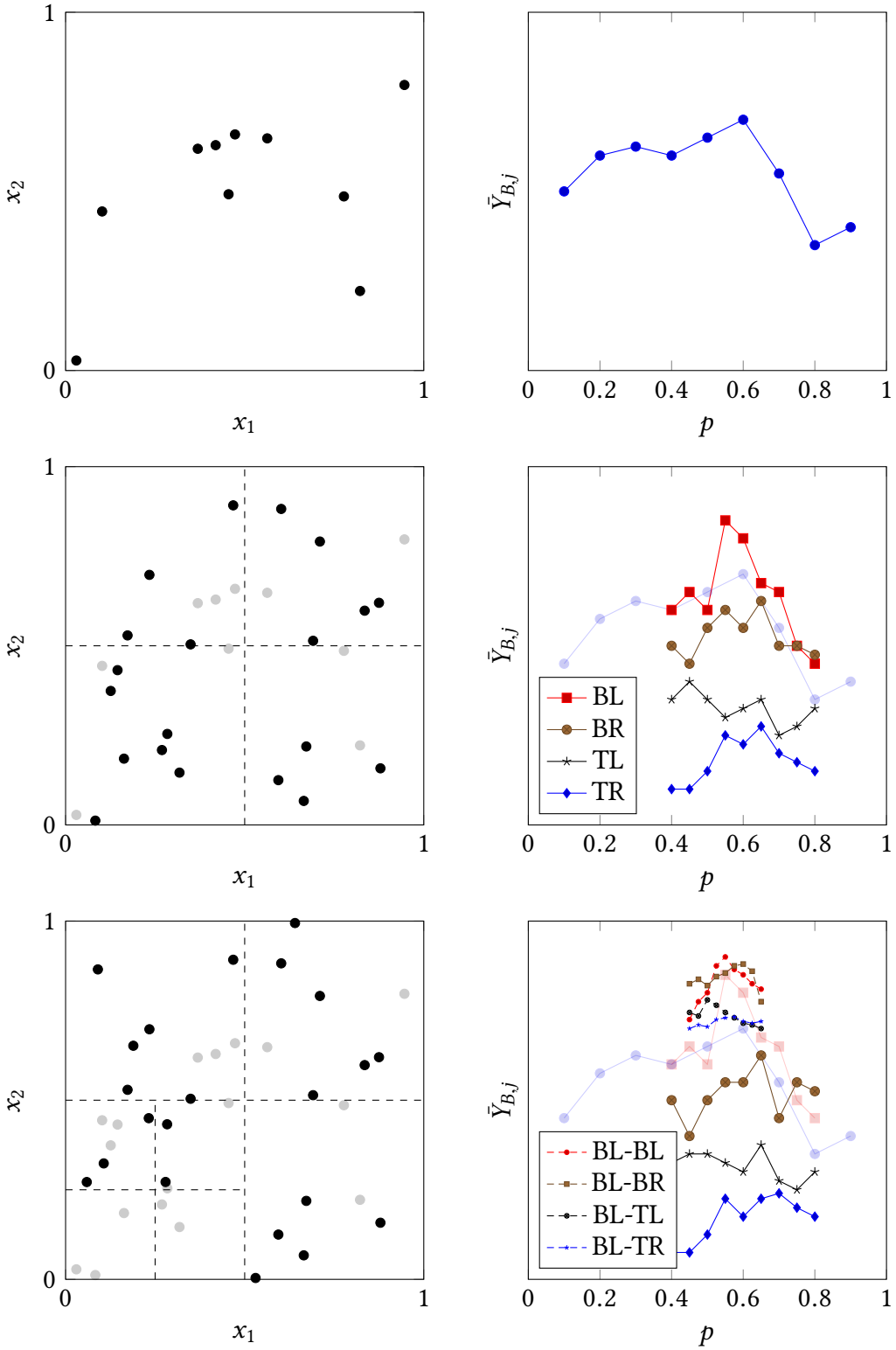

Figure 1: A schematic illustration of the ABE algorithm.

4. Regret Analysis: Upper Bound

To measure the performance of the ABE algorithm, we provide an upper bound for its regret.

Theorem 1. *For any function f satisfying Assumption 1 to 3, the regret incurred by the ABE algorithm is bounded by*

$$R_{\pi_{ABE}}(T) \leq CT^{\frac{2+d}{4+d}} \log(T)^2$$

for a constant $C > 0$ that is independent of T.

We provide a sketch of the proof here and present details in the Appendix. In period t, if $\boldsymbol{X}_t \in B$ for a bin in the partition $B \in \mathcal{P}_t$, then the expected regret incurred by the ABE algorithm is $\mathsf{E}[(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t))\mathbb{I}_{\{\boldsymbol{X}_t \in B, B \in \mathcal{P}_t\}}]$. Since the total regret simply sums up the above quantity over $t = 1, \ldots, T$ and all possible Bs, it suffices to focus on the regret for given t and B. Two possible scenarios can arise: (1) the optimal decision of B, i.e., $p^*(B)$, is inside the range of the decision set, i.e., $p^*(B) \in [p_l^B, p_u^B]$ (Step 19); (2) the optimal decision $p^*(B)$ is outside the range of the decision set.

Scenario one represents the regime where the algorithm is working "normally": up until t, the algorithm has successfully narrowed the optimal decision $p^*(B)$ (which provides a useful indicator for all $p^*(\boldsymbol{x})$, $\boldsymbol{x} \in B$ when B is small) down to $[p_l^B, p_u^B]$. By Assumption 3 part one, the regret in this scenario can be decomposed into two terms

$$f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t) \leq M_3(|\pi_t - p^*(B)| + |p^*(B) - p^*(\boldsymbol{X}_t)|)^2.$$

The first term can be bounded by the length of the interval $p_u^B - p_l^B$. The second term can be bounded by the size of B given $\boldsymbol{X}_t \in B$ by Assumption 3 part two and three. By the choice of parameters in Section 3.4, the length of the interval decreases as the bin size decreases. Therefore, both terms can be well controlled when the size of B is sufficiently small, or equivalently, when the level $l(B)$ is sufficiently large. This is why a properly chosen n_k can guarantee that the algorithm spends little time for large bins, when each period incurs substantial regret, and collect a large number of covariate observations for small bins. When the bin level reaches K, the above two terms are small enough and no more exploration is needed.

Scenario two represents the regime where the algorithm works "abnormally". In scenario two, the difference $f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t)$ can no longer be controlled as in scenario one because π_t and $p^*(\boldsymbol{X}_t)$ can be far apart. To make things worse, $p^*(B) \notin [p_l^B, p_u^B]$ usually implies $p^*(B') \notin [p_l^{B'}, p_u^{B'}]$ with high probability, where B' is a child of B. This is because (1) $p^*(B)$ is close to $p^*(B')$ for small B, and (2) $[p_l^{B'}, p_u^{B'}]$ is created around the empirically-optimal decision for B, and thus largely overlapping with $[p_l^B, p_u^B]$. Therefore, for all period s following t, the worst-case regret is $O(1)$ in that period if $\boldsymbol{X}_s \in B$ or its offspring.

To bound the regret in scenario two, we have to bound the probability, which requires delicate analysis of the events. If scenario two occurs for B, then during the process that we sequentially split $B_\emptyset$ to obtain B, we can find an ancestor bin of B (which can be B itself) that scenario two happens for the first time along the "branch" from $B_\emptyset$ all the way down to B. More precisely, denoting the ancestor bin by B_a and its parent by $P(B_a)$, we have (1) $p^*(P(B_a))$ is inside $[p_l^{P(B_a)}, p_u^{P(B_a)}]$ (scenario one); (2) after $P(B_a)$ is split, $p^*(B_a)$ is outside $[p_l^{B_a}, p_u^{B_a}]$ (scenario two). Denote the empirically-optimal decision in the decision set of $P(B_a)$ by p^*. For such an event to occur, the center of the decision set of B_a, which is p^*, has to be at least $\Delta_{l(B_a)}/2$ away from $p^*(B_a)$.[3] Because of Assumption 3 and the choice of Δ_k, the distance between $p^*(P(B_a))$ and $p^*(B_a)$ is relatively small compared to $\Delta_{l(B_a)}$. Therefore, the empirically-optimal decision p^* must be far away from $p^*(P(B_a))$. The probability of such event can be bounded using classic concentration inequalities for sub-Gaussian random variables: the decisions that are closer to $p^*(P(B_a))$ and thus have higher means turn out to have lower average reward than p^*; this event is extremely unlikely to happen when we have collected a large amount of reward for each decision in the decision set.

The total regret aggregates those in scenario one and scenario two for all possible combinations of B and B_a. It matches the quantity $O(T^{\log(T)^2(2+d)/(4+d)})$ presented in the theorem.

[3]Recall that $p_u^{B_a} - p_l^{B_a} = \Delta_{l(B_a)}$.

5. Regret Analysis: Lower Bound

In the last section, we have shown that for any function satisfying the assumptions, the ABE algorithm achieves regret of order $\log(T)^2 T^{(2+d)/(4+d)}$. To complete the argument, we will show in this section that the minimax regret is no lower than $cT^{(2+d)/(4+d)}$ for some constant c. Combining the two quantities, we shall conclude that no non-anticipating policy does better than the ABE algorithm in terms of the order of magnitude of the regret in T (neglecting logarithmic terms).

We first identify a class C of functions that satisfy Assumption 1 to 3. The functions in the class are selected to be "difficult" to distinguish. By doing so, we will prove that any policy has to spend a substantial amount of time exploring decisions that generate low reward but help to differentiate the functions. Otherwise, the incapability to correctly identify the underlying function is costly in the long run. Therefore, unable to contain both sources of regret at the same time, no policy can achieve lower regret than the quantity stated in Theorem 2.

Before introducing the class of functions, we define ∂B to be the boundary of a convex set in $[0,1)^d$. Let $d(B_1, B_2)$ be the Euclidean distance between two sets B_1 and B_2. That is $d(B_1, B_2) \triangleq \inf\{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 : \boldsymbol{x}_1 \in B_1, \boldsymbol{x}_2 \in B_2\}$. We allow B_1 or B_2 to be a singleton. To define C, partition the covariate space $[0,1)^d$ into M^d equally sized bins. That is, each bin has the following form: for $(k_1, \ldots, k_d) \in \{1, \ldots, M\}^d$,

$$\left\{\boldsymbol{x} : \frac{k_i - 1}{M} \le x_i < \frac{k_i}{M}, \quad \forall\, i = 1, \ldots, d\right\}.$$

We number those bins by $1, \ldots, M^d$ in an arbitrary order, i.e., $B_1, \ldots, B_{M^d}$. Each function $f(\boldsymbol{x}, p) \in C$ is represented by a tuple $w \in \{0,1\}^{M^d}$, whose jth index determines the behavior of $f_w(\boldsymbol{x}, p)$ in B_j. More precisely, for $\boldsymbol{x} \in B_j$,

$$f_w(\boldsymbol{x}, p) = \begin{cases} -p^2 & w_j = 0 \\ -p^2 + 2pd(\boldsymbol{x}, \partial B_j) & w_j = 1 \end{cases}$$

By the form of f_w, the optimal decision satisfies $p^*(\boldsymbol{x}) = 0$ if $w_j = 0$ and $p^*(\boldsymbol{x}) = d(\boldsymbol{x}, \partial B_j)$ if $w_j = 1$. In other words, when $w_j = 1$, the optimal decision $p^*(\boldsymbol{x})$ is zero when $\boldsymbol{x}$ is at the boundary of B_j; as $\boldsymbol{x}$ moves away from the boundary, $p^*(\boldsymbol{x})$ increases and reaches

its maximum $1/2M$ when $\boldsymbol{x}$ is at the center of B_j.

The construction of C follows a similar idea to Rigollet and Zeevi (2010). For a given $f_w \in C$, we can always find another function $f_{w'} \in C$ that only differs from f in a single bin B_j by setting w' to be equal to w except for the jth index. The decision analyst can only rely on the covariates generated in B_j to distinguish between f_w and $f_{w'}$. For small bins (i.e., large M), this is particularly costly because there are only a tiny fraction of covariates generated in a particular bin and the difference $|f_w - f_{w'}| = pd(\boldsymbol{x}, \partial B_j) \leq 1/2M$ becomes tenuous. Now a policy has to carry out the task for M^d bins, i.e., distinguishing the underlying function f_w with M^d tuples that only differ from w in one index. The cost cannot be avoided and adds to the lower bound of the regret.

For the distribution of the covariate $\boldsymbol{X}$ and the reward Z, let $\boldsymbol{X}$ be uniformly distributed in $[0, 1)^d$ and $Z = f(\boldsymbol{x}, p) + \epsilon$ where ϵ are i.i.d. normal random variables with mean 0 and variance 1.

Next we show that Assumption 1 to 3 are satisfied by the above setup.

Proposition 2. *The choice of $f \in C$, $\boldsymbol{X}$ and Z satisfies Assumption 1 to 3 with $\sigma = 1/2$, $M_1 = 4$, $M_2 = 1/2$, $M_3 = 2$, and $M_4 = 1$.*

The detailed proof of Proposition 2 is provided in the appendix. To give some intuitions, note that Assumption 1 is satisfied because the reward Z is a standard normal random variable, and thus $\mathsf{E}[\exp(t(Z - \mathsf{E}[Z]))] = \exp(t^2/2)$. For Assumption 2 and 3, note that by the construction of f, both $f_w(\boldsymbol{x}, p)$ and $p^*(\boldsymbol{x})$ are Lipschitz continuous in $[0, 1)^d$. Such continuity guarantees the desired properties.

The next theorem shows the lower bound for the regret.

Theorem 2. *For all non-anticipating polices, we have*

$$\inf_{\pi} \sup_{f \in C} R_\pi \geq cT^{\frac{2+d}{4+d}}$$

for a constant $c > 0$.

The proof uses Kullback-Leibler (KL) divergence to measure the "distinguishability" of the underlying functions. Such information-theoretical approach has been a standard technique in the learning literature. The proof is outlined in the following.

Among all functions $f \in C$, we focus on each pair of f_w and $f_{w'}$ that only differ in a single bin. For example, consider $w = (w_1, w_2, \ldots, w_{j-1}, 0, w_{j+1}, \ldots, w_{M^d})$ and

$w' = (w_1, w_2, \ldots, w_{j-1}, 1, w_{j+1}, \ldots, w_{M^d})$ for some j. Because the indices of w and w' are identical except for the jth, f_w and $f_{w'}$ only differ in bin B_j. Denote $w = (w_{-j}, 0)$ and $w' = (w_{-j}, 1)$ to highlight this fact. Distinguishing between $f_{w_{-j},0}$ and $f_{w_{-j},1}$ poses a challenge to any policy. In particular, for $\boldsymbol{x} \in B_j$, the difference of the two functions $|f_{w_{-j},0}(\boldsymbol{x}, p) - f_{w_{-j},1}(\boldsymbol{x}, p)| = pd(\boldsymbol{x}, \partial B_j)$ is increasing in p. Thus, applying a large decision $p = \pi_t$ makes the difference more visible and helps to distinguish between $f_{w_{-j},0}$ and $f_{w_{-j},1}$. However, if p deviates too much from the optimal decision $p^*(\boldsymbol{x}) = 0$ or $p^*(\boldsymbol{x}) = d(\boldsymbol{x}, \partial B_j)$, then significant regret is incurred in that period.

To capture this trade-off, for a given $j = 1, \ldots, M^d$ and $w_{-j} \in \{0, 1\}^{M^d - 1}$, define the following quantity

$$z_{w_{-j}} = \sum_{t=1}^{T} \frac{1}{8M^2} \mathsf{E}^{\pi}_{f_{w_{-j},0}} \left[\pi_t^2 \mathbb{I}_{\{X_t \in B_j\}} \right] \tag{1}$$

where the expectation is taken with respect to a policy π and the underlying function $f_{w_{-j},0}$. This quantity is crucial in analyzing the regret. More precisely, if $z_{w_{-j}}$ is large (which implies that π_t is large), then $f_{w_{-j},0}$ and $f_{w_{-j},1}$ are easy to distinguish but the regret becomes uncontrollable.

Lemma 1.

$$\sup_{f \in C} R_\pi \geq \frac{8M_3}{2^{M^d}} M^2 \sum_{j=1}^{M^d} \sum_{w_{-j}} z_{w_{-j}}.$$

On the other hand, if $z_{w_{-j}}$ is small, then the KL divergence of the measures associated with $f_{w_{-j},0}$ and $f_{w_{-j},1}$ is also small. In other words, the decision analyst cannot easily distinguish between $f_{w_{-j},0}$ and $f_{w_{-j},1}$ which impedes learning and incurs substantial regret.

Lemma 2.

$$\sup_{f \in C} R_\pi \geq \frac{M_3 T}{2^{M^d+9} M^{d+2}} \sum_{j=1}^{M^d} \sum_{w_{-j}} \exp(-z_{w_{-j}}).$$

Since the effects of $z_{w_{-j}}$ are opposite in Lemma 1 and Lemma 2, combining the two

bounds, we have that for $c_1 = M_3/512$ and $c_2 = 8M_3$,

$$\begin{aligned}\sup_{f\in C} R_\pi &\geq \frac{1}{2^{M^d}+1}\sum_{j=1}^{M^d}\sum_{w_{-j}}\left(\frac{c_1 T}{M^{d+2}}\exp\left(-z_{w_{-j}}\right)+c_2 M^2 z_{w_{-j}}\right)\\ &\geq \frac{1}{2^{M^d}+1}\sum_{j=1}^{M^d}\sum_{w_{-j}} c_2 M^2\left(1+\log\left(\frac{c_1 T}{c_2 M^{d+4}}\right)\right)\\ &\geq \frac{c_2 M^{d+2}}{4}\left(1+\log\left(\frac{c_1 T}{c_2 M^{d+4}}\right)\right).\end{aligned}$$

In the second inequality above, we minimize the expression over positive $z_{w_{-j}}$. Since M can be an arbitrary positive integer, we let $M = \lceil T^{1/(d+4)}\rceil$ in the last quantity. Calculation shows that it is lower bounded by $cT^{(2+d)/(4+d)}$ for a constant $c > 0$.

6. Discussions

In this section, we discuss some features and potential extensions of our algorithm and compare it to those presented in the literature.

6.1. Comparison with the Literature

As mentioned in Section 1.1, this paper is closely related to the literature of nonparametric formulations of contextual bandit problems, in particular, Perchet and Rigollet (2013); Slivkins (2014). Perchet and Rigollet (2013) study a problem with finite arms while the covariates are generated in a hypercube like in our setting. Their algorithm applies adaptive binning to the covariate space and eliminates suboptimal arms for the leaf bins in the process. In Slivkins (2014), the arms are continuous and high-dimensional. Their algorithm adaptively constructs balls (instead of hyperrectangles) in the product space of the covariate and the arm; it picks an arbitrary arm belonging to a ball which is selected according to a UCB-type criterion. We compare our algorithm to theirs in several respects.

The algorithm design. The algorithms of those two papers and this paper, roughly speaking, have two components: decision (arm) selection and binning. For decision selection, Perchet and Rigollet (2013) successively eliminates arms as bins are split.

This applies to their setting of finite arms. Slivkins (2014) adopts a unified approach that bins the decision and covariate space simultaneously. It is worth mentioning that unlike the other two, our algorithm does not track the confidence intervals. Instead, we choose the range of the decision set carefully to avoid removing optimal decisions. For binning, Perchet and Rigollet (2013) and this paper use a similar approach. Slivkins (2014) uses balls instead of hyperrectangles, and the parent balls are not removed. Thus, the "bins" do not form a partition of the covariate space.

The assumptions. Since the setting of Perchet and Rigollet (2013) is different, we compare our assumptions to those in Slivkins (2014). The key additional assumption in our setting is part one of Assumption 3. Roughly speaking, it states that the objective function is smooth and locally concave at the global maximum. Similar assumptions have been imposed in the literature (Assumption 2 in Auer et al. 2007, the Margin Condition in Perchet and Rigollet 2013). The intuition of why smoothness and local concavity matter can be illustrated by a simplified example. Consider $f_1(x) = -|x|$ and $f_2 = -x^2$. At the maximizer $x^* = 0$, smoothness makes $f_2(x)$ converge at a higher rate as $x \to 0$, compared to $f_1(x)$. As a result, extra care needs to be taken when the candidate decisions are close to the maximizer. In our algorithm, this is done implicitly by the choice of Δ_k, n_k and N_k. Note that the intricacy of different converging rates is not new in the literature. In the context of learning and dynamic pricing with inventory constraints, Wang et al. (2014); Lei et al. (2017) handle sufficient capacity (learning the unconstrained maximizer) and insufficient capacity (learning the market-clearing price) differently, exactly because of the smoothness of the former and the resulting different converging rates. For network revenue management, Besbes and Zeevi (2012) find that the smoothness of the demand functions may change the rate of regret.

The rate of regret. Because of the assumption mentioned above, our algorithm can achieve an improved rate of regret, $T^{(2+d)/(4+d)}$, compared to $T^{(2+d)/(3+d)}$ in Equation (3) of Slivkins (2014) with $d_Y = 1$. In other words, if the objective function is indeed smooth and locally concave at the maximizer, then it is beneficial to treat it specially (our algorithm) and obtain lower regret than the procedure in Slivkins (2014) designed for Lipschitz continuous functions. In fact, even with Assumption 3, the algorithm in Slivkins (2014) seems unable to be adapted to achieve the improved rate of regret. Since smooth objective functions are common in optimization problems in practice, this paper provides the first algorithm that achieves the optimal rate of regret in this case.

We plan to investigate how the rate of regret depends on the degree of smoothness at the maximizer, similar to the result in Auer et al. (2007), in a future study.

6.2. Discretizing Decisions

The major distinction between multi-armed bandit problems and the problem we study is the form of the decision space. Namely, the decision analyst faces a continuum of decisions in this paper. If the decision analyst discretizes the decision space in advance, then one might argue that the algorithms for multi-armed bandit problems could be applied to our setting as well. To examine this intuition, we consider the algorithm in Perchet and Rigollet (2013). To apply it to our setting, we set $K = T^{(2+d)/(4+d)}$ equally spaced arms so that the discretization error is $O(T^{(2+d)/(4+d)})$ by Lipschitz continuity. In addition, we let $\alpha = 2$ in their margin condition (by part one of our Assumption 3) and $\beta = 1$ in their smoothness condition (by our Assumption 2) to match our assumptions. However, the rate of regret derived in Perchet and Rigollet (2013) does not match that in this paper. Although we do not have a definitive explanation to why discretization (in the simplest form) fails in our setting, one reason might be the smoothness and local concavity of the objective function at the maximum. As discussed in Section 6.1, to utilize the different converging rate at the maximum, the decision analyst would have discretized at a different resolution when the decision is close to the maximizer, or sampled the nearby arms at a different rate. This, however, is infeasible if the decision space is discretized in advance without any information of the objective function. Therefore, the problem studied in this paper cannot be regarded as a trivial extension to the multi-armed bandit literature.

6.3. Adaptive Binning

In the ABE algorithm, the covariate space is refined adaptively: a bin is split only when we have collected sufficient observations in it. An alternative idea of binning, similar to the algorithm designed in Rigollet and Zeevi (2010), is to pre-define a set of static bins that are sufficiently small. The algorithm then performs parallel learning in each bin whenever a covariate is generated in it. In our algorithm, it is equivalent to setting $n_k = 0$ for all $k < K$, i.e., we start to collect observations only for bins of level K. Based on the results in Rigollet and Zeevi (2010) and Perchet and Rigollet

(2013), the benefit of adaptive binning might not be reflected in the asymptotic rate of regret. That is, if designed carefully, static binning may achieve the same rate of regret. However, adaptive binning arguably outperforms static binning in practice. When binning adaptively, a covariate and its reward observed in a parent bin provide some information for all offspring bins. Such pooling effect makes the exploration more effective. While in static binning, the information learned from an observation is only restricted in its own bin, which is usually very small by design in order to achieve acceptable regret. In Section 7, we conduct numerical experiments and demonstrate the performance of adaptive/static binning in practice.

A second reason to use adaptive rather than static binning is its extensibility and potential to incorporate other machine learning algorithms. As explained in Section 6.5, the combination with regression/classification trees, or even boosting (combining many trees to achieve better prediction; see Friedman 2001 for more details), may allow the algorithm to identify the sparsity structure of the covariate and thus improve the incurred regret. Static binning, however, does not easily accommodate such extensions: when pre-defining a set of bins, the decision analyst has no knowledge about the sparsity structure.

6.4. Complexity

We first analyze the time complexity of the ABE algorithm. At each t, the decision analyst observes the covariate $\boldsymbol{X}_t$. It first determines which bin in the partition it belongs to, and then explores the decisions in the decision set sequentially and update its average reward. The second step takes $O(1)$ computations. For the first step, since the maximum level of the bins is K, such determination takes $O(dK) = O(\log(T))$ computations (at most K binary searches along each dimension) by the choice of K in Section 3.4. Therefore, the total time complexity of the ABE algorithm is $O(T\log(T))$.

For the space complexity, note that the algorithm records the information for at most $O(2^{dK}) = O(T^{d/(d+4)})$ bins. Each bin can be assigned an identifier (the binary vector $(\boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_k)$) with at most $dK = O(\log(T))$ 0s and 1s. For each bin, the empirical average and number of trials of each decision in the decision set are stored. Since the decision set has $O(\log(T))$ grid points, the total space complexity is $O(\log(T)T^{d/(d+4)})$.

6.5. The Order of the Regret and Sparsity

As shown in Theorem 1 and Theorem 2, the best achievable regret of the formulated problem is of order $T^{(2+d)/(4+d)}$. To understand the intuition behind the exponent $(2+d)/(4+d)$, note that when $d = 0$, i.e., there is no covariate, we recover the $\sqrt{T}$ regret shown in the literature (Wang et al., 2014). For $d > 0$, consider a static version of the ABE algorithm: the covariate space is binned in advance into $T^{d/(d+4)}$ identical hypercubes (each hypercube has edges of length $T^{-1/(d+4)}$). There are roughly $T^{4/(d+4)}$ covariates generated in each bin over the horizon. If we conduct parallel learning without covariates for each bin B, then the regret would be $\sqrt{T^{4/(d+4)}}$, compared to the clairvoyant policy that knows $p^*(B)$. Moreover, the regret generated from the gap between $p^*(B)$ and $p^*(\boldsymbol{x})$ for $\boldsymbol{x} \in B$, is of order $f^*(\boldsymbol{x}) - f(\boldsymbol{x}, p^*(B)) \sim (p^*(B) - p^*(\boldsymbol{x}))^2 \sim d_B^2 \sim T^{2/(d+4)}$. Therefore, the total regret grows roughly in the order $T^{2/(d+4)} \times T^{d/(d+4)} = T^{(2+d)/(4+d)}$.

Note that the parametric version of the problem (e.g., Ban and Keskin 2017) can achieve regret of order $\sqrt{T}$ or even $\log(T)$ under certain conditions (e.g., Qiang and Bayati 2016; Javanmard and Nazerzadeh 2016), which is the same order as the problem without covariates (e.g., den Boer and Zwart 2014). In other words, when the interaction of the covariate and the reward is parametric (linear), it does not complicates the learning problem, regardless of the dimension of the covariate. In comparison, when the reward function does not have a parametric form as in our problem, the dimension of the covariate significantly affects how well the decision analyst can do at best. For large d, no algorithms can achieve regret of order less than $T^{(2+d)/(4+d)}$ which is almost linear in T. It implies that when the dimension of the covariate grows, the learning problem becomes quite intractable and the incurred regret is almost linear in T, which is considered unsatisfactory because an arbitrary policy can achieve linear regret.

As a result, knowledge of the sparsity structure of the covariate is essential in designing algorithms. More precisely, the provided covariate is of dimension d, while $f(\boldsymbol{X}, p)$ may only depend on d' entries of the covariate where $d' \ll d$. In this case, unable to identify the d' entries out of d significantly increases the incurred regret from $T^{(2+d')/(4+d')}$ to $T^{(2+d)/(4+d)}$. Indeed, in the ABE algorithm, if the sparsity structure is known, then a bin is split into $2^{d'}$ instead of 2^d child bins. It pools the covariate observations that only differ in the "useless" dimensions so that more observations are

available in a bin, and thus substantially reduces the exploration cost.

To design algorithms when the covariate is sparse, it is worth reflecting on the parametric (linear) setting. When $f(\boldsymbol{X}, p) = f(\boldsymbol{\theta}^T \boldsymbol{X}, p)$ is linear in $\boldsymbol{X}$ with coefficient $\boldsymbol{\theta} \in \mathbb{R}^n$, LASSO regression serves as a powerful tool to identify the value of $\boldsymbol{\theta}$ as well as the positions of its zero entries. The asymptotic statistical properties of LASSO have been studied extensively in the statistics and computer science community. Moreover, the sparsity structure is not as important in the parametric setting: in the best achievable regret, d does not appear in the exponent of T (Ban and Keskin, 2017); so even the least square estimator for $\boldsymbol{\theta}$ that ignores its sparsity achieves the same regret asymptotically in T.

In the nonparametric setting, however, the sparsity structure has a significant impact on the incurred regret as explained above. Recently, statistical tools to handle nonparametric sparsity have been developed (Lafferty et al., 2008; Rosasco et al., 2013). To circumvent the sparsity problem, we plan to develop an improved version of the ABE algorithm that is based on regression/classification trees (Hastie et al., 2001). More precisely, when a bin is split, instead of bisecting all d dimensions in the middle and thus obtaining 2^d child bins, we may carefully select one dimension and the position of the cutoff. The criterion based on which the dimension is selected may depend on the observed rewards for past covariates. For example, we may select a dimension and the associated cutoff so that the sum of the sample standard deviations in the two resulting child bins is minimized. Similar ideas are widely adopted for regression/classification trees. The intuition for such criterion is that for the entries of the covariate that do not affect the reward function, splitting along their dimensions does not actually refine the covariate space and thus does not significantly reduce the total in-bin standard deviation. The new algorithm can potentially improve the regret in sparse situations: In Step 16 of Algorithm 1, the split only increases the size of the partition by one instead of $2^d - 1$. The pooling of past observations can lead to more effective explorations. The detail of this algorithm and its regret analysis remain a topic for future research.

In practice, data preprocessing procedures may help improve the performance when the covariate has a sparse structure. For example, based on the experience or insights of the decision analyst, some contextual information may be discarded manually to reduce the dimension of the covariate and thus the magnitude of the regret. Principal component analysis provides a more systematic approach: the decision analyst may

collect the contextual information for a fraction of T and extracts a more compact representation of the covariate via principal components. The raw covariate is then replaced by the first few principal components for the rest of the horizon.

7. Numerical Experiment

In this section, we apply the algorithm to the example mentioned in the introduction. In the area $[0, 1)^2$, there are three community centers: $(0.2, 0.2)$, $(0.2, 0.8)$, and $(0.8, 0.2)$, corresponding to three demand functions $1-p$, $1-2p$, and $1-p/2$. For a customer from location (x_1, x_2), her ℓ_1 distance to the three centers are denoted $d_1 = |x_1-0.2|+|x_2-0.2|$, $d_2 = |x_1 - 0.2| + |x_2 - 0.8|$ and $d_3 = |x_1 - 0.8| + |x_2 - 0.2|$. The customer makes a binary decision, buying or not buying, based on the purchase probability

$$d(\boldsymbol{x}, p) = \max\left\{\frac{1}{1/d_1 + 1/d_2 + 1/d_3}\left(\frac{1}{d_1}(1-p) + \frac{1}{d_2}(1-2p) + \frac{1}{d_3}\left(1-\frac{p}{2}\right)\right), 0\right\}.$$

That is, the similarity between her purchase probability and the demand of community center i is proportional to the reciprocal of their ℓ_1 distance. Therefore, the reward (revenue) associated with location (x_1, x_2) and decision p is a random variable valued p with probability $d(\boldsymbol{x}, p)$ and 0 with probability $1 - d(\boldsymbol{x}, p)$, whose mean is $f(\boldsymbol{x}, p) = pd(\boldsymbol{x}, p)$.

The firm applies the ABE algorithm for $T = 1\,000\,000$ sequentially arriving customers, whose locations are uniformly distributed on $[0, 1)^2$. At the end of the horizon, the firm obtains a partition and an optimal price in each bin in the partition. Figure 2 compares the actual optimal price and that learned from the algorithm. Although the optimal price output by the algorithm is piecewise constant by nature, it captures the basic structure of the actual optimal price.

To obtain quantitative insights into the performance of the algorithm and the effects of various assumptions, we choose different T and conduct the following experiments in the same context.

1. Apply the ABE algorithm as described above.

2. The covariates in each period are no longer stationary. More precisely, for every $T/10$ periods, we generate a quadruplet $(a_1, a_2, b_1, b_2) \in [0, 1]^4$ randomly. All

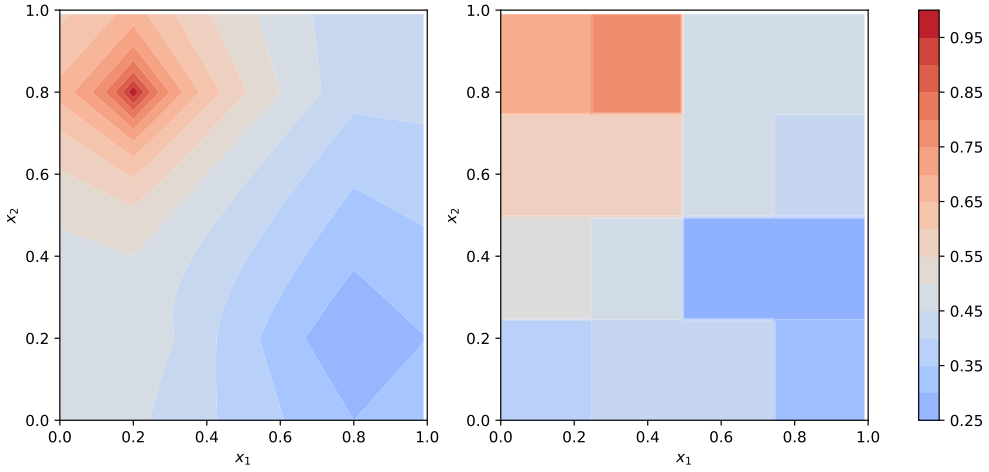

Figure 2: The actual optimal price (left) and the optimal price in each bin output by the ABE algorithm (right) for customers from different locations.

covariates $\boldsymbol{X}_t$ in the next $T/10$ periods are i.i.d. and drawn uniformly from

$$(\min\{a_1, a_2\}, \max\{a_1, a_2\}) \times (\min\{b_1, b_2\}, \max\{b_1, b_2\}).$$

The goal is to test the performance of the algorithm when the assumption of i.i.d. covariates fails.

3. Replace the demand function by

$$d(\boldsymbol{x}, p) = \max\Big\{(1-p)\mathbb{I}_{\{d_1 \le d_2, d_1 \le d_3\}} + (1-2p)\mathbb{I}_{\{d_2 < d_1, d_2 \le d_3\}} + \left(1 - \frac{p}{2}\right)\mathbb{I}_{\{d_3 < d_1, d_3 < d_2\}}, 0\Big\}.$$

Therefore, the demand function of a customer located at $\boldsymbol{x}$ is no longer a weighted average of those of the community centers, but identical to the closest community center. Consequently, the objective function is not continuous any more.

4. Replace the demand functions of the community centers by $1-p$, 0.5, and 0. The demand of location $\boldsymbol{x}$ is averaged in the same way. Since we restrict $p \in [0, 1]$, the maximizer may be located at the boundary $p^*(\boldsymbol{x}) = 1$ and not have zero derivatives. As a result, Assumption 3 fails. Note that the unlike the last scenario, the objective function is still continuous.

5. We test the static version of the ABE algorithm. More precisely, the covariate space is binned in advance into $T^{2/(d+4)}$ identical hyperrectangles (each hyperrectangle is of size $T^{-1/(d+4)} \times T^{-1/(d+4)}$). The choice of the size is consistent with the level-K bins which the ABE algorithm would stop splitting. Then the algorithm carries out parallel learning inside each bin, by first discretizing $p \in [0, 1]$ and then applying the UCB algorithm. The decision interval $[0, 1]$ is discretized into $\log(T)T^{1/(d+4)}$ equally spaced arms, matching the grid size of the decision set of level-K bins in the ABE algorithm. The goal is to demonstrate the performance improvement of adaptive binning.

6. The decision analyst considers a misspecified linear model: $d(\boldsymbol{x}, p) = a - bp + c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2$. In each period, he/she applies the least square estimator to the historical price and demand data to estimate the coefficients; the estimated coefficients are then used to compute the optimal price based on the misspecified model. This procedure is referred to as the greedy iterated least square policy (Keskin and Zeevi, 2014).

For each of the above scenarios, we execute the algorithm for selected $T \in [10^4, 10^7]$. Figure 3 illustrates the recorded regret against T in a log-log plot. The slope of the curves indicates the exponent of the rate of regret in T. For scenario 1 and 2, the slope roughly matches the theoretical prediction $(2 + d)/(4 + d) \approx 0.66$. The non-stationary covariates do not seem to affect the performance of the ABE algorithm. For scenario 3, 4 and 5, although the growth of the regret is still sublinear (below the dotted curve), its rate deteriorates significantly. This implies that in practice, a discontinuous objective function and a function without smoothness at the maximizers are much harder to learn for our algorithm; static binning with prespecified discretized decisions does not perform well. For the misspecified linear model, the greedy policy performs quite well for small T; as T increases, the regret grows almost linearly.

8. Conclusion

In this paper, we propose an algorithm that learns an unknown objective function and optimizes simultaneously, with contextual information. The algorithm achieves the optimal rate of regret, $O(T^{(2+d)/(4+d)})$, within a logarithmic term. Its dramatic increase in

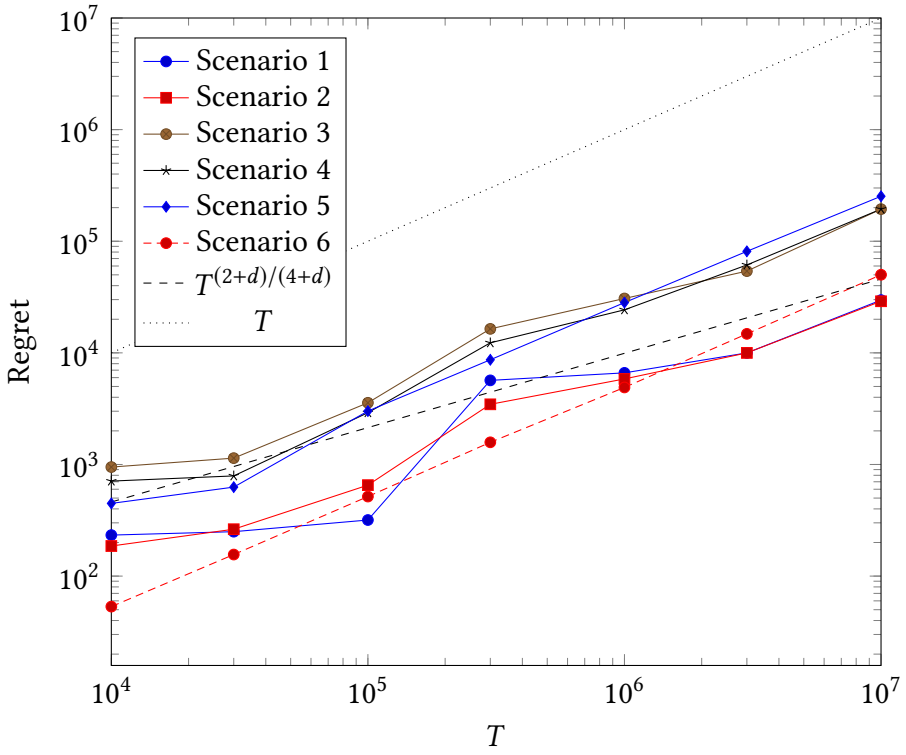

Figure 3: The incurred regret for different T in the four scenarios.

the covariate dimension, d, demonstrates the complex nature of nonparametric learning in high dimensions. It also calls for nonparametric learning algorithms that handle sparse covariates. This remains a topic for our future research.

References

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM journal on control and optimization 33*(6), 1926–1951.

Agrawal, S. and N. Goyal (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 39–1.

Araman, V. F. and R. Caldentey (2009). Dynamic pricing for nonperishable products with demand learning. *Operations research 57*(5), 1169–1188.

Auer, P., R. Ortner, and C. Szepesvári (2007). *Improved Rates for the Stochastic Continuum-Armed Bandit Problem*, pp. 454–468. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ban, G. and N. B. Keskin (2017). Personalized dynamic pricing with machine learning. *Working paper*.

Bastani, H. and M. Bayati (2015). Online decision-making with high-dimensional covariates. *Working paper*.

Besbes, O. and A. Zeevi (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research 57*(6), 1407–1420.

Besbes, O. and A. Zeevi (2012). Blind network revenue management. *Operations research 60*(6), 1537–1550.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Broder, J. and P. Rusmevichientong (2012). Dynamic pricing under a general parametric choice model. *Operations Research 60*(4), 965–980.

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

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

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

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge university press.

Cheung, W. C., D. Simchi-Levi, and H. Wang (2017). Dynamic pricing and demand learning with limited price experimentation. *Operations Research 65*(6), 1722–1731.

Cohen, M. C., I. Lobel, and R. Paes Leme (2016). Feature-based dynamic pricing. *Working paper*.

den Boer, A. V. (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science 20*(1), 1–18.

den Boer, A. V. and B. Zwart (2014). Simultaneously learning and optimizing using controlled variance pricing. *Management Science 60*(3), 770–783.

Elmachtoub, A. N., R. McNellis, S. Oh, and M. Petrik (2017). A practical method for solving contextual bandit problems using decision trees. *Working paper*.

Farias, V. F. and B. Van Roy (2010). Dynamic pricing with a prior on market response. *Operations Research 58*(1), 16–29.

Féraud, R., R. Allesiardo, T. Urvoy, and F. Clérot (2016). Random forest for the contextual bandit problem. In *Artificial Intelligence and Statistics*, pp. 93–101.

Foucart, S. and H. Rauhut (2013). *A mathematical introduction to compressive sensing*, Volume 1. Birkhäuser Basel.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Goldenshluger, A. and A. Zeevi (2013). A linear response bandit problem. *Stochastic Systems 3*(1), 230–261.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Javanmard, A. and H. Nazerzadeh (2016). Dynamic pricing in high-dimensions. *Working paper*.

Keskin, N. B. and A. Zeevi (2014). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research 62*(5), 1142–1167.

Kleinberg, R., A. Slivkins, and E. Upfal (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 681–690. ACM.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pp. 697–704.

Kuleshov, V. and D. Precup (2014). Algorithms for multi-armed bandit problems. *Working paper*.

Lafferty, J., L. Wasserman, et al. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics 36*(1), 28–63.

Langford, J. and T. Zhang (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pp. 817–824.

Lee, D. K. and N. Chen (2017). Boosting hazard regression with time-varying covariates. *Working paper*.

Lei, Y., S. Jasin, and A. Sinha (2017). Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. *Working paper*.

Nambiar, M., D. Simchi-Levi, and H. Wang (2016). Dynamic learning and price optimization with endogeneity effect. *Working paper*.

Perchet, V. and P. Rigollet (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics 41*(2), 693–721.

Qiang, S. and M. Bayati (2016). Dynamic pricing with demand covariates. *Working paper*.

Rigollet, P. and A. Zeevi (2010). Nonparametric bandits with covariates. *arXiv:1003.1630*.

Rosasco, L., S. Villa, S. Mosci, M. Santoro, and A. Verri (2013). Nonparametric sparsity and regularization. *The Journal of Machine Learning Research 14*(1), 1665–1714.

Slivkins, A. (2014). Contextual bandits with similarity information. *The Journal of Machine Learning Research 15*(1), 2533–2568.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation* (1 ed.). Springer-Verlag New York.

Wang, Z., S. Deng, and Y. Ye (2014). Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research 62*(2), 318–331.

Yang, Y., D. Zhu, et al. (2002). Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics 30*(1), 100–121.

A. Table of Notations

$\lceil x \rceil$	The smallest integer that does not exceed x
$(x)^+$	The positive part of x
$\#\{\}$	The cardinality of a set
$\mathcal{F}_t$	The σ-algebra generated by $(X_1, \pi_1, Z_1, \ldots, X_{t-1}, \pi_{t-1}, Z_{t-1})$
μ_X	The distribution of the covariate $\boldsymbol{X}$ over $[0,1)^d$
$P^*(B)$	$\text{argmax}_{p\in[0,1]} \{\mathsf{E}[f(\boldsymbol{X}, p)\|\boldsymbol{X} \in B]\}$
P^*_B	The empirically-optimal decision in the decision set for B
∂B	The boundary of B

Table 1: A table of notations used in the paper.

B. Proofs

Proof of Proposition 1: Define for $p \in [0,1]$

$$g(p) = \begin{cases} \frac{f_B(p^*(B)) - f_B(p)}{(p^*(B)-p)^2} & p \neq p^*(B) \\ -\frac{f_B''(p^*(B))}{2} & p = p^*(B). \end{cases}$$

By L'Hopital's rule, $g(p)$ is continuous at $p^*(B)$. In addition, because $f_B(p)$ is continuous, $g(p)$ is continuous for all $p \in [0,1]$. By Weierstrass's extreme value theorem, we have $g(p) \in [M_2, M_3]$ and both M_2 and M_3 are attained. By the definition and the uniqueness of the maximizer, $g(p) > 0$ for $p \in [0,1]$. Therefore, we must have $M_2, M_3 > 0$. This establishes the result. ■

To prove Theorem 1, we first introduce the following lemma.

Lemma 3. *Suppose for given $\boldsymbol{x}$ and p, the random variable $Z(\boldsymbol{x}, p)$ is sub-Gaussian with parameter σ, i.e.,*

$$\mathsf{E}[\exp(t(Z - \mathsf{E}[Z])] \leq \exp(\sigma t^2)$$

for all $t \in \mathbb{R}$. Then the distribution of $Z(\boldsymbol{X}, p)$ conditional on $\boldsymbol{X} \in B$ for a set B is still sub-Gaussian with the same parameter.

Proof. Let μ_X denote the distribution of $\boldsymbol{X}$. We have that for all $t \in \mathbb{R}$

$$\begin{aligned}
&\mathsf{E}[\exp(t(Z(\boldsymbol{X},p) - \mathsf{E}[Z(\boldsymbol{X},p)|\boldsymbol{X} \in B])|\boldsymbol{X} \in B] \\
&= \frac{\int_B \mathsf{E}[\exp(t(Z(\boldsymbol{x},p) - \mathsf{E}[Z(\boldsymbol{X},p)|\boldsymbol{X} \in B] d\mu_X(\boldsymbol{x})}{\int_B d\mu_X(\boldsymbol{x})} \\
&= \frac{\int_B \mathsf{E}[\exp(t(Z(\boldsymbol{x},p) - \mathsf{E}[Z(\boldsymbol{x},p)])] d\mu_X(\boldsymbol{x})}{\int_B d\mu_X(\boldsymbol{x})} \times \frac{\int_B \exp(t\mathsf{E}[Z(\boldsymbol{x},p)]) d\mu_X(\boldsymbol{x})}{\exp(t\mathsf{E}[Z(\boldsymbol{X},p)|\boldsymbol{X} \in B]) \int_B d\mu_X(\boldsymbol{x})} \\
&\leq \frac{\int_B \exp(\sigma t^2) d\mu_X(\boldsymbol{x})}{\int_B d\mu_X(\boldsymbol{x})} \times 1 = \exp(\sigma t^2),
\end{aligned}$$

where the last inequality is by the definition of conditional expectations. Hence the result is proved. ■

Proof of Theorem 1: According to the algorithm (Step 7), let $\mathcal{P}_t$ denote the partition formed by the bins at time t when $\boldsymbol{X}_t$ is generated. The regret associated with $\boldsymbol{X}_t$ can be counted by bins $B \in \mathcal{P}_t$ into which $\boldsymbol{X}_t$ falls. Meanwhile, the level of B is at most K. Therefore,

$$\begin{aligned}
R_{\pi_{ABE}}(T) &= \mathsf{E}\left[\sum_{t=1}^{T}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t))\right] = \mathsf{E}\left[\sum_{t=1}^{T}\sum_{B\in\mathcal{P}_t}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t))\mathbb{I}_{\{X_t\in B\}}\right] \\
&= \mathsf{E}\left[\sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t))\mathbb{I}_{\{X_t\in B, B\in\mathcal{P}_t\}}\right]
\end{aligned}$$

We will define the following random event for each bin B:

$$E_B = \left\{p^*(B) \in \left[p_l^B, p_u^B\right]\right\}.$$

Recall that $p^*(B)$ is the unique maximizer for $f_B(p) = \mathsf{E}[f(\boldsymbol{X},p)|\boldsymbol{X} \in B]$ by Assumption 3; $[p_l^B, p_u^B]$ is the range of the decision set to explore for B. According to Step 19 of the ABE algorithm, the interval $[p_l^B, p_u^B]$ is constructed around $p^*_{P(B)}$, the empirically-optimal decision of the parent bin $P(B)$ that maximizes the empirical average $\bar{Y}_{P(B),j}$.

We will decompose the regret depending on whether E_B occurs.

$$\mathsf{E}[R_{\pi_{ABE}}] = \underbrace{\mathsf{E}\left[\sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t,\pi_t))\mathbb{I}_{\{\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, E_B^c\}}\right]}_{\text{term 1}}$$
$$+ \underbrace{\mathsf{E}\left[\sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t,\pi_t))\mathbb{I}_{\{\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, E_B\}}\right]}_{\text{term 2}} \tag{2}$$

We first analyze term 1. Because $E_{B_\emptyset}$ is always true ($[p_l^{B_\emptyset}, p_u^{B_\emptyset}] = [0,1]$ (Step 4) always encloses $p^*(B_\emptyset)$), we can find an ancestor of B, say B_a (which can be B itself), such that $E_{B_a}^c \cap E_{P(B_a)} \cap E_{P(P(B_a))} \cap \ldots \cap E_{B_\emptyset}$ occurs. In other words, up until B_a, the algorithm always correctly encloses the optimal decision $p^*(P^{(k)}(B_a))$ of the ancestor bin of B_a in their decision intervals $\left[p_l^{P^{(k)}(B_a)}, p_u^{P^{(k)}(B_a)}\right]$. Therefore, when E_B^c occurs, we can rearrange the event by such B_a. Term 1 in (2) can be bounded by

$$\begin{aligned}
&\mathsf{E}\left[\sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t,\pi_t))\mathbb{I}_{\{\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, E_B^c\}}\right] \\
&\le \sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}} M_1\mathsf{P}(\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, E_B^c) \\
&\le \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\{B:l(B)=k\}} M_1\mathsf{P}\left(\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, E_{B_a}^c \cap E_{P(B_a)} \cap E_{P(P(B_a))} \cap \ldots \cap E_{B_\emptyset}\right) \\
&\le \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\{B:l(B)=k\}} M_1\mathsf{P}(\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, E_{B_a}^c \cap E_{P(B_a)}) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\{B:l(B)=k\}}\sum_{k'=1}^{K}\sum_{B':l(B')=k'} M_1\mathsf{P}(\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, B_a=B', E_{B'}^c \cap E_{P(B')})
\end{aligned} \tag{3}$$

The first inequality is due to Assumption 2. In the second inequality, we start enumerating from $k=1$ instead of $k=0$ because $E_{B_\emptyset}^c$ never occurs. In the last equality, we rearrange the probabilities by counting the deterministic bins B' instead of the random bins B_a.

Now note that $\left\{\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, B_a=B', E_{B'}^c \cap E_{P(B')}\right\}$ are exclusive for different Bs because $\mathcal{P}$ is a partition and $\boldsymbol{X}_t$ can only fall into one bin. Moreover, $\left\{\boldsymbol{X}_t\in B, B\in\mathcal{P}_t, B_a=B', E_{B'}^c \cap E_{P(B')}\right\}$ c

$\left\{\boldsymbol{X}_t \in B', E^c_{B'} \cap E_{P(B')}\right\}$ because $B \subset B_a$. Therefore,

$$\sum_{k=1}^{K} \sum_{\{B:l(B)=k\}} \mathsf{P}\left(\boldsymbol{X}_t \in B, B \in \mathcal{P}_t, B_a = B', E^c_{B'} \cap E_{P(B')}\right) \leq \mathsf{P}\left(\boldsymbol{X}_t \in B', E^c_{B'} \cap E_{P(B')}\right).$$

Thus, we can further simplify (3):

$$\begin{aligned}(3) &\leq \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{\{B':l(B')=k\}} M_1 \mathsf{P}\left(\boldsymbol{X}_t \in B', E^c_{B'} \cap E_{P(B')}\right) \\ &= \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{\{B:l(B)=k\}} M_1 \mathsf{P}(\boldsymbol{X}_t \in B) \mathsf{P}(E^c_B \cap E_{P(B)}). \end{aligned} \tag{4}$$

The last equality is because of the fact that for given t and B, the event $\{\boldsymbol{X}_t \in B\} \in \sigma(X_t)$ and $E^c_B \cap E_{P(B)} \in \mathcal{F}_{t-1}$. Therefore, the two events are independent.

Next we analyze the event $E^c_B \cap E_{P(B)}$ in more detail given $l(B) = k$ in order to bound (4). This event implies that when the parent bin $P(B)$ is created, its optimal decision $p^*(P(B))$ is inside the decision interval $[p_l^{P(B)}, p_u^{P(B)}]$. At the end of Step 14, when n_{k-1} covariate data points have been observed in $P(B)$, it is split into 2^d children. The optimal decision of one of the children of $P(B)$, that is $p^*(B)$, is no longer inside $[p_l^B, p_u^B]$. By Step 19, $p_l^B = \max\{p^*_{P(B)} - \Delta_k/2, 0\}$ and $p_u^B = \min\{p^*_{P(B)} + \Delta_k/2, 1\}$, where $p^*_{P(B)}$ is the empirically-optimal decision for $P(B)$. Therefore, E^c_B implies that $p^*(B) \notin [p^*_{P(B)} - \Delta_k/2, p^*_{P(B)} + \Delta_k/2]$. Combined with Assumption 3 part 2, which states that $p^*(B) \in [\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}, \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}] \subset \inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\}, \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\}$, we have

$$[\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\}, \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\}] \not\subset [p^*_{P(B)} - \Delta_k/2, p^*_{P(B)} + \Delta_k/2].$$

That is, either $\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\} < p^*_{P(B)} - \Delta_k/2$ or $\sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\} > p^*_{P(B)} + \Delta_k/2$. By Assumption 3 part 3, $\sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\} - \inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\} \leq M_4 d_{P(B)} = M_4\sqrt{d}2^{-(k-1)}$ because the level of $P(B)$ is $k-1$. Hence by Assumption 3 part 2, $\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\} \geq p^*(P(B)) - M_4\sqrt{d}2^{-(k-1)}$ and $\sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in P(B)\} \leq p^*(P(B)) + M_4\sqrt{d}2^{-(k-1)}$. Combining the above observations, E^c_B could only happen when $|p^*(P(B)) - p^*_{P(B)}| > \Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}$. On the other hand, $E_{P(B)}$ implies that $p^*(P(B)) \in [p_l^{P(B)}, p_u^{P(B)}]$. Therefore, $E^c_B \cap E_{P(B)}$ could occur only if there exist two grid points $0 \leq j_1, j_2 \leq N_{k-1} - 1$ in Step 14 for bin $P(B)$, such that

1. The j_2th grid point is the closest to the optimal decision for the bin $p^*(P(B))$. That

is, $|p_l^{P(B)} + j_2\delta_{P(B)} - p^*(P(B))| \le \delta_{P(B)}/2$.

2. The j_1th grid point maximizes $\bar{Y}_{P(B),j}$. That is $p_l^{P(B)} + j_1\delta_{P(B)} = P^*_{P(B)}$. It implies that $\bar{Y}_{P(B),j_1} \ge \bar{Y}_{P(B),j_2}$ in Step 15.

3. $|p^*(P(B)) - p^*_{P(B)}| > \Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}$.

In other words, the empirically-optimal decision is the j_1th grid point, while the j_2th grid point is closest to the true reward maximizer in bin $P(B)$, i.e., $p^*(P(B))$. Given that the two grid points are far apart (by part 3 above), the probability of this event should be small.

To further bound the probability, consider $\bar{Y}_{P(B),j}$. In Step 15, it is the sum of $\lfloor n_{k-1}/N_{k-1}\rfloor$ or $\lceil n_{k-1}/N_{k-1}\rceil$ independent random variables with mean $\mathsf{E}[f(\boldsymbol{X}, p_l^{P(B)} + j\delta_{P(B)})|\boldsymbol{X} \in P(B)]$. By Lemma 3, they are still sub-Gaussian with parameter σ. This gives the following probabilistic bound (recall the definition of $f_B(p)$ in Section 2.2):

$$
\begin{aligned}
\mathsf{P}(E_B^c \cap E_{P(B)}) &\le \mathsf{P}(\bar{Y}_{P(B),j_1} \ge \bar{Y}_{P(B),j_2}) \\
&= \mathsf{P}\Big(\frac{1}{t_1}\sum_{i=1}^{t_1} X_i^{(1)} - \frac{1}{t_2}\sum_{i=1}^{t_2} X_i^{(2)} \ge f_{P(B)}(p_l^{P(B)} + j_2\delta_{P(B)}) \\
&\qquad - f_{P(B)}(p_l^{P(B)} + j_1\delta_{P(B)})\Big) \\
&\le \mathsf{P}\Big(\frac{1}{t_1}\sum_{i=1}^{t_1} X_i^{(1)} - \frac{1}{t_2}\sum_{i=1}^{t_2} X_i^{(2)} \ge M_2\left(\left(\Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}\right)^+\right)^2 \\
&\qquad - M_3\delta^2_{P(B)}/4\Big).
\end{aligned}
$$

Here t_1 and t_2 can be either $\lfloor n_{k-1}/N_{k-1}\rfloor$ or $\lceil n_{k-1}/N_{k-1}\rceil$; $X_i^{(1)}$ and $X_i^{(2)}$ are independent mean-zero sub-Gaussian random variables with parameter σ. Their averages are the centered version of $\bar{Y}_{P(B),j_1}$ and $\bar{Y}_{P(B),j_2}$, and thus their means are moved to the right-hand side. In the last inequality, $(\cdot)^+$ represents the positive part. The inequality follows from Assumption 3 part 1 and the previously derived facts that $|p_l^{P(B)} + j_2\delta_{P(B)} - p^*(P(B))| \le \delta_{P(B)}/2$ and $|p_l^{P(B)} + j_1\delta_{P(B)} - p^*(P(B))| \ge \Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}$. By the property of sub-Gaussian random variables (for example, see Theorem 7.27 in Foucart and Rauhut,

2013), the above probability is bounded by

$$\mathrm{P}(E_B^c \cap E_{P(B)}) \leq \exp\left(-\frac{\left(\left(M_2\left(\left(\Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}\right)^+\right)^2 - M_3\delta_{P(B)}^2/4\right)^+\right)^2}{4\sigma(1/t_1 + 1/t_2)}\right)$$

$$\leq \exp\left(-\frac{n_{k-1}\left(\left(M_2\left(\left(\Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}\right)^+\right)^2 - M_3\delta_{P(B)}^2/4\right)^+\right)^2}{8\sigma(N_{k-1}+1)}\right)$$

By our choice of parameters, $\Delta_k = 2^{-k}\log(T)$, $N_k \equiv \lceil \log(T) \rceil$, $\delta_{P(B)} \leq \Delta_{k-1}/N_{k-1} \leq 2^{-(k-1)}$. Therefore, when $T \geq \max\{\exp(8M_4\sqrt{d}), \exp(4\sqrt{2M_3/M_2})\}$, we have:

$$\frac{\Delta_k}{4} - M_4\sqrt{d}2^{-(k-1)} = 2^{-(k+2)}\log(T) - M_4\sqrt{d}2^{-(k-1)} \geq 0$$

$$\Rightarrow \frac{\Delta_k}{2} - M_4\sqrt{d}2^{-(k-1)} \geq \frac{\Delta_k}{4}$$

$$\frac{M_2\Delta_k^2}{32} - \frac{M_3\delta_{P(B)}^2}{4} \geq \frac{M_2 2^{-2k}\log^2(T)}{32} - M_3 2^{-2k} \geq 0$$

$$\Rightarrow M_2\left(\Delta_k/2 - M_4\sqrt{d}2^{-(k-1)}\right)^2 - M_3\delta_{P(B)}^2/4 \geq \frac{M_2\Delta_k^2}{16} - \frac{M_3\delta_{P(B)}^2}{4} \geq \frac{M_2\Delta_k^2}{32}.$$

Therefore, there exists a constant $c_1 = M_2^2/(8192\sigma)$ such that

$$\mathrm{P}(E_B^c \cap E_{P(B)}) \leq \exp\left(-c_1\frac{\Delta_k^4 n_{k-1}}{\log(T)+1}\right).$$

With this bound, we can proceed to provide an upper bound for (4). Because $\sum_{\{B:l(B)=k\}} \mathrm{P}(\boldsymbol{X}_t \in B) = 1$, we have

$$\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\{B:l(B)=k\}} M_1\mathrm{P}(\boldsymbol{X}_t \in B)\mathrm{P}(E_B^c \cap E_{P(B)}) \leq M_1 T\sum_{k=1}^{K}\exp\left(-c_1\frac{\Delta_k^4 n_{k-1}}{\log(T)+1}\right). \quad (5)$$

We next analyze term 2 of (2). By Assumption 3 part 1, $f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t) \leq M_3(p^*(\boldsymbol{X}_t) - \pi_t)^2 \leq M_3(|p^*(B) - \pi_t| + |p^*(\boldsymbol{X}_t) - p^*(B)|)^2$. By the design of the algorithm (Step 12 and 25), $\pi_t \in [p_l^B, p_u^B]$; conditional on the event $E_B = \{p^*(B) \in [p_l^B, p_u^B]\}$, we have $|p^*(B) - \pi_t| \leq p_u^B - p_l^B \leq \Delta_k$ for $l(B) = k$. On the other hand, by Assumption 3 part

2 and 3, $|p^*(\boldsymbol{X}_t) - p^*(B)| \leq \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\} - \inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\} \leq M_4 d_B \leq M_4\sqrt{d}2^{-k}$ for $l(B) = k$. Therefore, term 2 can be bounded by

$$\begin{aligned}
&\mathsf{E}\left[\sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}}(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t))\mathbb{I}_{\{\boldsymbol{X}_t \in B, B \in \mathcal{P}_t, E_B\}}\right] \\
&\leq \mathsf{E}\left[\sum_{t=1}^{T}\sum_{k=0}^{K}\sum_{\{B:l(B)=k\}} M_3^2(\Delta_k + M_4\sqrt{d}2^{-k})^2\mathbb{I}_{\{\boldsymbol{X}_t \in B, B \in \mathcal{P}_t, E_B\}}\right] \\
&\leq \mathsf{E}\left[\sum_{k=0}^{K-1}\sum_{\{B:l(B)=k\}}\sum_{t=1}^{T} M_3^2(\Delta_k + M_4\sqrt{d}2^{-k})^2\mathbb{I}_{\{\boldsymbol{X}_t \in B, B \in \mathcal{P}_t\}}\right] \\
&\quad + \sum_{t=1}^{T} M_3^2(\Delta_K + M_4\sqrt{d}2^{-K})^2 \sum_{\{B:l(B)=K\}} \mathsf{P}(\boldsymbol{X}_t \in B)
\end{aligned} \tag{6}$$

For the first term in (6), note that $\{X_t \in B, B \in \mathcal{P}_t\}$ occurs for at most n_k times for given B with $l(B) = k$. Moreover, there are 2^{dk} bins with level k, i.e., $\#\{B : l(B) = k\} = 2^{dk}$. Therefore, substituting $\Delta_k = 2^{-k}\log(T)$ into the first term yields an upper bound $\sum_{k=0}^{K-1} M_3^2 n_k 2^{(d-2)k}(\log(T) + M_4\sqrt{d})^2$. For the second term in (6), $\sum_{\{B:l(B)=K\}} \mathsf{P}(\boldsymbol{X}_t \in B) = 1$ because $\{B : l(B) = K\}$ form a partition of the covariate space and $\boldsymbol{X}$ always falls into one of the bins. Therefore, (6) is bounded by

$$M_3^2\left(\sum_{k=0}^{K-1} n_k 2^{(d-2)k} + T2^{-2K}\right)\left(\log(T) + M_4\sqrt{d}\right)^2 \leq c_3 M_3^2\left(\sum_{k=0}^{K-1} n_k 2^{(d-2)k} + T2^{-2K}\right) \tag{7}$$

for $c_3 = \left(\log(2) + M_4\sqrt{d}\right)^2/\log(2)^2$ and $T \geq 2$.

Combining (5) and (7), we can find constants $c_2 = c_1/2^5 = M_2^2/(2^{18}\sigma)$ such that

$$\begin{aligned}
\mathsf{E}[R_{\pi_{ABE}}] &\leq \sum_{k=0}^{K-1}\left(c_3\log(T)^2 n_k 2^{(d-2)k} + M_1 T\exp(-c_1 2^{-4k-4}\log^3(T)n_k/2)\right) + c_3\log(T)^2 T2^{-2K} \\
&\leq \sum_{k=0}^{K-1}\left(c_3\log(T)^2 n_k 2^{(d-2)k} + M_1 T\exp(-c_2 2^{-4k}\log^3(T)n_k)\right) + c_3\log(T)^2 T2^{-2K}
\end{aligned} \tag{8}$$

We choose n_k

$$n_k = \max\left\{0, \left\lceil \frac{2^{4k+18}\sigma}{M_2^2 \log^3(T)}(\log(T) + \log(\log(T)) - (d+2)k\log(2)) \right\rceil\right\}$$

to minimize $(c_3 \log(T))^2 n_k 2^{(d-2)k} + M_1 T \exp(-c_2 2^{-4k} \log^3(T) n_k)$ in (8). More precisely,

$$\begin{aligned} c_3(\log(T))^2 n_k 2^{(d-2)k} &\leq c_4 \frac{\log(T)^2}{\log^3(T)} 2^{(d+2)k}(\log(T) + \log(\log(T))) \\ &\leq c_4 2^{(d+2)k} \end{aligned}$$

for some constants $c_4 > 0$, and

$$\begin{aligned} M_1 T \exp(-c_2 2^{-4k} \log^3(T) n_k) &\leq M_1 T \exp(-\log(T) - \log(\log(T)) + (d+2)k\log(2)) \\ &\leq c_5 2^{(d+2)k} \end{aligned}$$

for a constant $c_5 > 0$. Therefore, (8) implies that we can find a constant $c_6 = c_4 + c_5$ such that

$$\begin{aligned} \mathsf{E}[R_{\pi_{ABE}}] &\leq \sum_{k=0}^{K-1} c_6 2^{(d+2)k} + c_3 \log(T)^2 T 2^{-2K} \\ &\leq c_6 2^{(d+2)K} + c_3 \log(T)^2 T 2^{-2K}. \end{aligned}$$

Therefore, by our choice of $K = \lfloor \frac{\log(T)}{(d+4)\log(2)} \rfloor$, the regret is bounded by

$$c_7 \log^2(T) T^{\frac{d+2}{d+4}}.$$

for some constant c_7. Hence we have completed the proof. ■

Proof of Proposition 2: For normal random variables with distribution $N(0,1)$, their moment generating function is $\mathsf{E}[\exp(tZ)] = \exp(t^2/2)$. Therefore, Assumption 1 is satisfied with $\sigma = 1/2$.

For Assumption 2, we discuss two cases. The first case is $\boldsymbol{x}_1, \boldsymbol{x}_2 \in B_j$, i.e., the two covariates are in the same bin. In this case,

$$|f_w(\boldsymbol{x}_1, p_1) - f_w(\boldsymbol{x}_2, p_2)| \leq \begin{cases} |p_1^2 - p_2^2| \leq 2|p_1 - p_2| & w_j = 0 \\ |p_1^2 - p_2^2| + 2|p_1 d(\boldsymbol{x}_1, \partial B_j) - p_2 d(\boldsymbol{x}_2, \partial B_j)| & w_j = 1 \end{cases}$$

When $w_j = 0$, the assumption is already satisfied. When $w_j = 1$, by the triangle

inequality we have

$$\begin{aligned}|p_1 d(\boldsymbol{x}_1, \partial B_j) - p_2 d(\boldsymbol{x}_2, \partial B_j)| &\leq |p_1 - p_2| d(\boldsymbol{x}_1, \partial B_j) + p_2 |d(\boldsymbol{x}_1, \partial B_j) - d(\boldsymbol{x}_2, \partial B_j)| \\ &\leq \frac{1}{2M} |p_1 - p_2| + |d(\boldsymbol{x}_1, \partial B_j) - d(\boldsymbol{x}_2, \partial B_j)| \\ &\leq \frac{1}{2} |p_1 - p_2| + \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2\end{aligned}$$

The first inequality is because $p_2 \leq 1$ and $d(\boldsymbol{x}_1, \partial B_j) \leq 1/2M \leq 1/2$ when $\boldsymbol{x}_1 \in B_j$. The second inequality is because

$$\begin{aligned}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + d(\boldsymbol{x}_2, \partial B_j) &= \min_{a \in \partial B_j} \{\|a - \boldsymbol{x}_2\|_2 + \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2\} \geq \min_{\boldsymbol{a} \in \partial B_j} \{\|a - \boldsymbol{x}_1\|_2\} \\ &= d(\boldsymbol{x}_1, \partial B_j)\end{aligned}$$

and similarly $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + d(\boldsymbol{x}_1, \partial B_j) \geq d(\boldsymbol{x}_2, \partial B_j)$. Therefore, we have shown that $|f_w(\boldsymbol{x}_1, p_1) - f_w(\boldsymbol{x}_2, p_2)| \leq 3|p_1 - p_2| + 2\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ for case one. The second case is $\boldsymbol{x}_1 \in B_{j_1}$ and $\boldsymbol{x}_2 \in B_{j_2}$ for $j_1 \neq j_2$. If $j_1 = j_2 = 0$, then by the previous analysis, we already have $|f_w(\boldsymbol{x}_1, p_1) - f_w(\boldsymbol{x}_2, p_2)| \leq 2|p_1 - p_2|$. If $j_1 = 1$ and $j_2 = 0$, then

$$|f_w(\boldsymbol{x}_1, p_1) - f_w(\boldsymbol{x}_2, p_2)| \leq 2|p_1 - p_2| + 2p_1 d(\boldsymbol{x}_1, \partial B_{j_1}) \leq 2|p_1 - p_2| + 2\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2.$$

The last inequality is because the straight line connecting $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ must intersects B_{j_1}. The distance from $\boldsymbol{x}_1$ to the intersection is no less than $d(\boldsymbol{x}_1, \partial B_{j_1})$. Therefore, $d(\boldsymbol{x}_1, \partial B_{j_1}) \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$. If $j_1 = 0$ and $j_2 = 1$, then the result follows similarly. If $j_1 = j_2 = 1$, then by the same argument,

$$|p_1 d(\boldsymbol{x}_1, \partial B_{j_1}) - p_2 d(\boldsymbol{x}_2, \partial B_{j_2})| \leq d(\boldsymbol{x}_1, \partial B_{j_1}) + d(\boldsymbol{x}_2, \partial B_{j_2}) \leq 2\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2.$$

Therefore, combining both cases, we always have

$$|f_w(\boldsymbol{x}_1, p_1) - f_w(\boldsymbol{x}_2, p_2)| \leq 4|p_1 - p_2| + 4\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2.$$

For Assumption 3, note that

$$\begin{aligned}f_B(p) &= -p^2 + 2p\mathsf{E}\left[\sum_{j=1}^{M^d} w_j \mathbb{I}_{\{X \in B_j\}} d(X, \partial B_j) | X \in B\right] \\ &= -p^2 + 2p \sum_{j=1}^{M^d} w_j \mathsf{P}(B_j \cap B) \mathsf{E}[d(X, \partial B_j) | X \in B_j \cap B].\end{aligned}$$

For part one, because the second-order derivative of $f_B(p)$ is always one, we have

$$f_B(p^*(B)) - f_B(p) = (p - p^*(B))^2$$

and part one holds for $M_2 = /2$, $M_3 = 2$. For part two, note that the maximizer

$$p^*(B) = \sum_{j=1}^{M^d} w_j \mathsf{P}(B_j \cap B)\mathsf{E}[d(\boldsymbol{X}, \partial B_j)|\boldsymbol{X} \in B_j \cap B].$$

If we regard $p^*(\boldsymbol{X}') = \sum_{j=1}^{M^d} w_j \mathbb{I}_{\{X' \in B_j\}} d(\boldsymbol{X}', \partial B_j)$ as a function of the random variable $\boldsymbol{X}'$ that is uniformly distributed on B, then $p^*(B) = \mathsf{E}[p^*(\boldsymbol{X}')]$. Therefore, part two of Assumption 3 holds as $\mathsf{E}[p^*(\boldsymbol{X}')] \in [\inf\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}, \sup\{p^*(\boldsymbol{x}) : \boldsymbol{x} \in B\}]$. For part three, we have that for $\boldsymbol{x}_1 \in B \cap B_{j_1}$ and $\boldsymbol{x}_2 \in B \cap B_{j_2}$

$$p^*(\boldsymbol{x}_1) - p^*(\boldsymbol{x}_2) = w_{j_1} d(\boldsymbol{x}_1, \partial B_{j_1}) - w_{j_2} d(\boldsymbol{x}_2, \partial B_{j_2}).$$

If $w_{j_1} = 0$ and $w_{j_2} = 0$, then $p^*(\boldsymbol{x}_1) - p^*(\boldsymbol{x}_2) = 0$. If either $w_{j_1} = 0$ or $w_{j_2} = 0$, then $p^*(\boldsymbol{x}_1) - p^*(\boldsymbol{x}_2) \leq \max\left\{d(\boldsymbol{x}_1, \partial B_{j_1}), d(\boldsymbol{x}_2, \partial B_{j_2})\right\} \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leq d_B$ by the previous analysis. If $w_{j_1} = 1$ and $w_{j_2} = 1$, then again by the previous analysis $p^*(\boldsymbol{x}_1) - p^*(\boldsymbol{x}_2) \leq |d(\boldsymbol{x}_1, \partial B_{j_1}) - d(\boldsymbol{x}_2, \partial B_{j_2})| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \leq d_B$. Therefore, part four holds with $M_4 = 1$. ■

To prove Theorem 2, i.e., Lemma 1 and Lemma 2, we introduce the following lemmas.

Lemma 4 (KL divergence for normal random variables). *For $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$, we have*

$$\mathcal{K}(\mu_{X_1}, \mu_{X_2}) = \frac{(\theta_1 - \theta_2)^2}{2}.$$

Proof. The result from the following calculation

$$\begin{aligned}
\mathcal{K}(\mu_{X_1}, \mu_{X_2}) &= \int \log\left(\frac{\exp(-(x-\theta_1)^2/2)}{\exp(-(x-\theta_2)^2/2)}\right) \frac{1}{\sqrt{2\pi}} \exp(-(x-\theta_1)^2/2) dx \\
&= \int \left((\theta_1 - \theta_2)x - (\theta_1^2 - \theta_2^2)/2\right) \frac{1}{\sqrt{2\pi}} \exp(-(x-\theta_1)^2/2) dx \\
&= \frac{(\theta_1 - \theta_2)^2}{2}.
\end{aligned}$$

■

Lemma 5 (The chain rule of the KL divergence). *Given joint distributions $p(x, y)$ and $q(x, y)$, we have*

$$\mathcal{K}(p(x, y), q(x, y)) = \mathcal{K}(p(x), q(x)) + \mathsf{E}_{p(x)}[\mathcal{K}(p(y|x), q(y|x)],$$

where $p(\cdot)$ and $q(\cdot)$ represent the marginal distribution, $p(\cdot|x)$ and $q(\cdot|x)$ represent the conditional distribution.

Proof. The proof can be found from standard textbooks and is thus omitted. ■

Proof of Lemma 1: We use E_f^π to highlight the dependence of the expectation on the policy π and the underlying function f. Note that

$$\begin{aligned}\sup_{f\in C} R_\pi = \sup_{f\in C}\sum_{t=1}^{T}\mathsf{E}\left[f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t,\pi_t)\right] &= \sup_{f\in C}\sum_{t=1}^{T}\sum_{j=1}^{M^d}\mathsf{E}\left[(f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t,\pi_t))\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right]\\ &\geq M_2\sup_{f\in C}\sum_{t=1}^{T}\sum_{j=1}^{M^d}\mathsf{E}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right]\\ &\geq \frac{M_2}{2^{M^d}}\sum_{w}\sum_{t=1}^{T}\sum_{j=1}^{M^d}\mathsf{E}_{f_w}^\pi\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right].\end{aligned}$$

In the last inequality, we have used the fact that $\#\{C\} = 2^{M^d}$ and the supremum is always no less than the average.

For a given bin B_j, we focus on $f_{w_{-j},0}$ and $f_{w_{-j},1}$, which only differ for $\boldsymbol{x}\in B_j$. Therefore, we can rearrange $\sum_w$ to $\sum_{w_{-j}\in\{0,1\}^{M^d-1}}\sum_{w_j\in\{0,1\}}$. We have the following lower bound for the regret

$$\begin{aligned}\sup_{f\in C} R_\pi &\geq \frac{M_2}{2^{M^d}}\sum_{j=1}^{M^d}\sum_{w_{-j}\in\{0,1\}^{M^d-1}}\sum_{w_j\in\{0,1\}}\sum_{t=1}^{T}\mathsf{E}_{f_{w_{-j},w_j}}^\pi\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right]\\ &\geq \frac{M_2}{2^{M^d}}\sum_{j=1}^{M^d}\sum_{w_{-j}}\sum_{t=1}^{T}\mathsf{E}_{f_{w_{-j},0}}^\pi\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right]\\ &= \frac{M_2}{2^{M^d}}\sum_{j=1}^{M^d}\sum_{w_{-j}}\sum_{t=1}^{T}\mathsf{E}_{f_{w_{-j},0}}^\pi\left[\pi_t^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right]\\ &= \frac{8M_2}{2^{M^d}}M^2\sum_{j=1}^{M^d}\sum_{w_{-j}} z_{w_{-j}}.\end{aligned}$$

In the second inequality, we have neglected the regret for $f_{w_{-j},1}$. The last equality is by the definition of $z_{w_{-j}}$ in (1). Hence we have proved the result. ■

Proof of Lemma 2: By the same argument as in the proof of Lemma 1, we have

$$\sup_{f\in C} R_\pi \geq \frac{M_3}{2M^d}\sum_{j=1}^{M^d}\sum_{t=1}^{T}\sum_{w_{-j}\in\{0,1\}^{M^d-1}}\sum_{w_j\in\{0,1\}}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right].$$

Because $\boldsymbol{X}_t$ is uniformly distributed in $[0,1)^d$, $\mathsf{P}(\boldsymbol{X}_t\in B_j)=M^{-d}$. By conditioning on the event $\boldsymbol{X}_t\in B_j$, we have

$$\begin{aligned}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right] &= \mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2|\boldsymbol{X}_t\in B_j\right]\mathsf{P}(\boldsymbol{X}_t\in B_j)\\ &= \frac{1}{M^d}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2|\boldsymbol{X}_t\in B_j\right].\end{aligned}$$

Since $(p^*(\boldsymbol{X}_t)-\pi_t)^2$ is measurable with respect to the σ-algebra generated by $\mathcal{F}_{t-1}$ and $\boldsymbol{X}_t$, by the tower property, we have

$$\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right] = \frac{1}{M^d}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[\mathsf{E}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2|\mathcal{F}_{t-1},\boldsymbol{X}_t\in B_j\right]\right]$$

Let $\mathsf{E}^{\pi,t-1}_{f_{w_{-j},w_j}}[\cdot]$ denote $\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}[\mathsf{E}[\cdot|\mathcal{F}_{t-1}]]$ and let $\mathsf{P}^{B,t-1}_{\boldsymbol{X}_t}(\cdot)$ denote the conditional probability $\mathsf{P}(\cdot|\mathcal{F}_{t-1},\boldsymbol{X}_t\in B)$. By Markov's inequality, for any constant $s>0$ we have

$$\begin{aligned}&\sum_{w_j\in\{0,1\}}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2\mathbb{I}_{\{\boldsymbol{X}_t\in B_j\}}\right] \qquad (9)\\ &= \frac{1}{M^d}\sum_{w_j\in\{0,1\}}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[\mathsf{E}\left[(p^*(\boldsymbol{X}_t)-\pi_t)^2|\boldsymbol{X}_t\in B_j,\mathcal{F}_{t-1}\right]\right]\\ &\geq \frac{1}{M^d}\sum_{w_j\in\{0,1\}}\frac{s^2}{M^2}\mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(\left|p^*(\boldsymbol{X}_t)-\pi_t\right|\geq\frac{s}{M}\right)\right]\\ &= \frac{s^2}{M^{d+2}}\left(\mathsf{E}^{\pi}_{f_{w_{-j},0}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(|\pi_t|\geq\frac{s}{M}\right)\right]+\mathsf{E}^{\pi}_{f_{w_{-j},1}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(|d(\boldsymbol{X}_t,\partial B_j)-\pi_t|\geq\frac{s}{M}\right)\right]\right)\\ &\geq \frac{s^2}{M^{d+2}}\left(\mathsf{E}^{\pi}_{f_{w_{-j},0}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(|\pi_t|\geq\frac{s}{M},A\right)\right]+\mathsf{E}^{\pi}_{f_{w_{-j},1}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(|d(\boldsymbol{X}_t,\partial B_j)-\pi_t|\geq\frac{s}{M},A\right)\right]\right)\end{aligned}$$

where we define event $A=B_j\cap\{d(\boldsymbol{X}_t,\partial B_j)>2s/M\}$. In the last equality, we have used the fact that for $\boldsymbol{X}_t\in B_j$, when $w_j=0$, $p^*(\boldsymbol{X}_t)=0$; when $w_j=1$, $p^*(\boldsymbol{X}_t)=d(\boldsymbol{X}_t,\partial B_j)$. The motivation of introducing A is as follows: consider the classification rule $\Pi_t\mapsto\{0,1\}$

associated with π_t tries to distinguish between $w_j = 0$ and $w_j = 1$. It is defined as

$$\Pi_t = \begin{cases} 0 & |\pi_t| \le |d(\boldsymbol{X}_t, \partial B_j) - \pi_t| \\ 1 & \text{otherwise.} \end{cases}$$

In other words, Π_t classifies the underlying function as $f_{w_{-j},0}$ if π_t is closer to the optimal decision $p^*(\boldsymbol{X}_t)$ of $f_{w_{-j},0}$, and as $f_{w_{-j},1}$ vice versa. For $f_{w_{-j},0}$, a misclassification on the event A, $A \cap \{\Pi_t = 1\}$, implies $A \cap \{|\pi_t| \ge s/M\}$. This is because $|\pi_t| + |d(\boldsymbol{X}_t, \partial B_j) - \pi_t| \ge d(\boldsymbol{X}_t, \partial B_j) \ge 2s/M$ on A and $|\pi_t| \ge |d(\boldsymbol{X}_t, \partial B_j) - \pi_t|$ due to misclassification. Similarly, $A \cap \{\Pi_t = 0\} \subset A \cap \{|d(\boldsymbol{X}_t, \partial B_j) - \pi_t| \ge s/M\}$. Therefore, by the fact that $\mathsf{P}(\boldsymbol{X} \in A) = (1 - 4s)^d / M^d$, we have

$$\begin{aligned} &\mathsf{E}^{\pi}_{f_{w_{-j},0}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(|\pi_t| \ge \frac{s}{M}, A\right)\right] + \mathsf{E}^{\pi}_{f_{w_{-j},1}}\left[\mathsf{P}^{B_j,t-1}_{\boldsymbol{X}_t}\left(|d(\boldsymbol{X}_t, \partial B_j) - \pi_t| \ge \frac{s}{M}, A\right)\right] \\ &\ge \mathsf{E}^{\pi}_{f_{w_{-j},0}}\left[\mathsf{P}^{B_j}_{\boldsymbol{X}_t,t-1}\left(A \cap \{\Pi_t = 1\}\right)\right] + \mathsf{E}^{\pi}_{f_{w_{-j},1}}\left[\mathsf{P}^{B_j}_{\boldsymbol{X}_t,t-1}\left(A \cap \{\Pi_t = 0\}\right)\right] \\ &= (1 - 4s)^d \left(\mathsf{P}^{\pi}_{f_{w_{-j},0}}(\Pi_t = 1 | \boldsymbol{X}_t \in A) + \mathsf{P}^{\pi}_{f_{w_{-j},1}}(\Pi_t = 0 | \boldsymbol{X}_t \in A)\right). \end{aligned} \tag{10}$$

Next we lower bound the misclassification error (10) by the Kullback-Leibler (KL) divergence between the two probability measures associated with $f_{w_{-j},0}$ and $f_{w_{-j},1}$. Intuitively, if the two probability measures are close, then no classification (including Π_t) can incur very small misclassification error. Formally, introduce the KL divergence between two probability measures P and Q as

$$\mathcal{K}(P, Q) = \begin{cases} \int \log \frac{dP}{dQ} dP & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases},$$

where $P \ll Q$ indicates that P is absolute continuous w.r.t. Q. By the independence of $\mathcal{F}_{t-1}$ and $\boldsymbol{X}_t$, the two measures we want to distinguish in (10), $\mu^{\pi}_{f_{w_{-j},0}}(\cdot | \boldsymbol{X}_t \in A)$ and $\mu^{\pi}_{f_{w_{-j},1}}(\cdot | \boldsymbol{X}_t \in A)$, can be expressed as product measures

$$\begin{aligned} \mu^{\pi}_{f_{w_{-j},0}}(\cdot | \boldsymbol{X}_t \in A) &= \mu^{\pi,t-1}_{f_{w_{-j},0}}(\cdot) \times \mu^{A}_{\boldsymbol{X}_t}(\cdot) \\ \mu^{\pi}_{f_{w_{-j},1}}(\cdot | \boldsymbol{X}_t \in A) &= \mu^{\pi,t-1}_{f_{w_{-j},1}}(\cdot) \times \mu^{A}_{\boldsymbol{X}_t}(\cdot), \end{aligned}$$

where $\mu^{\pi,t-1}_{f_{w_{-j},0}}(\cdot)$ is a measure of $(\boldsymbol{X}_1, Z_1, \ldots, \boldsymbol{X}_{t-1}, Z_{t-1})$ depending on π and $f_{w_{-j},0}$ and

$\mu^A_{X_t}(\cdot)$ is a measure of X_t conditional on $X_t \in A$. By Theorem 2.2 (iii) in Tsybakov (2009),

$$
\begin{aligned}
(10) &\geq \frac{(1-4s)^d}{2} \exp\left(-\mathcal{K}\left(\mu^{\pi,t-1}_{f_{w_{-j},0}} \times \mu^A_{X_t}, \mu^{\pi,t-1}_{f_{w_{-j},1}} \times \mu^A_{X_t}\right)\right) \\
&= \frac{(1-4s)^d}{2} \exp\left(-\mathcal{K}\left(\mu^{\pi,t-1}_{f_{w_{-j},0}}, \mu^{\pi,t-1}_{f_{w_{-j},1}}\right) - \mathsf{E}^{\pi,t-1}_{f_{w_{-j},0}}\left[\mathcal{K}\left(\mu^A_{X_t}, \mu^A_{X_t}\right)\right]\right) \\
&= \frac{(1-4s)^d}{2} \exp\left(-\mathcal{K}\left(\mu^{\pi,t-1}_{f_{w_{-j},0}}, \mu^{\pi,t-1}_{f_{w_{-j},1}}\right)\right). \qquad (11)
\end{aligned}
$$

The second line follows from Lemma 5; the third line follows from the fact that $\mu^A_{X_t}$ is the same distribution for $f_{w_{-j},0}$ and $f_{w_{-j},1}$, independent of $\mathcal{F}_{t-1}$.

To further simplify the expression, note that $\mu^{\pi,t}_{f_{w_{-j},0}}(\cdot)$ can be decomposed as

$$
\mu^{\pi,t}_{f_{w_{-j},0}}(\cdot) = \mu^{\pi,t-1}_{f_{w_{-j},0}}(\cdot) \times \mu_X(\cdot) \times \mu^{Z_t}_{f_{w_{-j},0}}(\cdot|\mathcal{F}_{t-1}, X_t),
$$

where μ_X is the measure (uniform distribution) of X_t and $\mu^{Z_t}_{f_{w_{-j},0}}(\cdot|\mathcal{F}_{t-1}, X_t)$ is the measure of Z_t conditional on $\mathcal{F}_{t-1}$ and X_t. We apply Lemma 5 again:

$$
\begin{aligned}
\mathcal{K}\left(\mu^{\pi,t}_{f_{w_{-j},0}}, \mu^{\pi,t}_{f_{w_{-j},1}}\right) = \mathcal{K}\left(\mu^{\pi,t-1}_{f_{w_{-j},0}}, \mu^{\pi,t-1}_{f_{w_{-j},1}}\right) &+ \mathsf{E}^{\pi,t-1}_{f_{w_{-j},0}}\left[\mathcal{K}(\mu_{X_t}, \mu_{X_t})\right] \\
&+ \mathsf{E}^{\pi,t-1}_{f_{w_{-j},0}}\left[\mathsf{E}_X\left[\mathcal{K}\left(\mu^{Z_t}_{f_{w_{-j},0}}(\cdot|\mathcal{F}_{t-1}, X_t), \mu^{Z_t}_{f_{w_{-j},1}}(\cdot|\mathcal{F}_{t-1}, X_t)\right)\right]\right].
\end{aligned}
$$

It is easy to see that the second term is zero. For the third term, we first conditional on $\mathcal{F}_{t-1}$ and then on the covariate X_t. Because π_t depends only on $\mathcal{F}_{t-1}$ and X_t, $\pi_t(X_t)$ is the same for $f_{w_{-j},0}$ and $f_{w_{-j},1}$ conditional on $\mathcal{F}_{t-1}$ and X_t. Therefore, $\mu^{Z_t}_{f_{w_{-j},0}}(\cdot|\mathcal{F}_{t-1}, X_t)$ and $\mu^{Z_t}_{f_{w_{-j},1}}(\cdot|\mathcal{F}_{t-1}, X_t)$ are two normal distributions with variance one and means $f_{w_{-j},0}(X_t, \pi_t)$ and $f_{w_{-j},1}(X_t, \pi_t)$, respectively. By Lemma 4, we have

$$
\begin{aligned}
\mathcal{K}\left(\mu^{Z_t}_{f_{w_{-j},0}}(\cdot|\mathcal{F}_{t-1}, X_t), \mu^{Z_t}_{f_{w_{-j},1}}(\cdot|\mathcal{F}_{t-1}, X_t)\right) &\leq \frac{1}{2}\left(f_{w_{-j},0}(X_t, \pi_t) - f_{w_{-j},1}(X_t, \pi_t)\right)^2 \\
&= \frac{1}{2}\left(\pi_t d(X_t, \partial B_j)\mathbb{I}_{\{X_t \in B_j\}}\right)^2 \\
&\leq \frac{1}{8M^2}\pi_t^2 \mathbb{I}_{\{X_t \in B_j\}},
\end{aligned}
$$

where in the last inequality, we have used the fact that the distance of a vector inside B_j to the boundary of B_j is at most $1/2M$. Therefore, we can obtain an upper bound for

$$\mathcal{K}\left(\mu^{\pi,t}_{f_{w_{-j},0}}, \mu^{\pi,t}_{f_{w_{-j},1}}\right)$$

$$\begin{aligned}
\mathcal{K}\left(\mu^{\pi,t}_{f_{w_{-j},0}}, \mu^{\pi,t}_{f_{w_{-j},1}}\right) &\leq \sum_{i=1}^{t} \mathsf{E}^{\pi,i-1}_{f_{w_{-j},0}}\left[\mathsf{E}_X\left[\mathcal{K}\left(\mu^{Z_i}_{f_{w_{-j},0}}(\cdot|\mathcal{F}_{i-1}, \boldsymbol{X}_i), \mu^{Z_i}_{f_{w_{-j},1}}(\cdot|\mathcal{F}_{i-1}, \boldsymbol{X}_i)\right)\right]\right] \\
&\leq \sum_{i=1}^{t} \mathsf{E}^{\pi,i-1}_{f_{w_{-j},0}}\left[\mathsf{E}_X\left[\frac{1}{8M^2}\pi_i^2 \mathbb{I}_{\{\boldsymbol{X}_i \in B_j\}}\right]\right] \\
&\leq \sum_{t=1}^{T} \mathsf{E}^{\pi}_{f_{w_{-j},0}}\left[\frac{1}{8M^2}\pi_t^2 \mathbb{I}_{\{\boldsymbol{X}_t \in B_j\}}\right] = z_{w_{-j}}.
\end{aligned}$$

Therefore, combining it with (9), (10) and (11), we have shown the lemma:

$$\begin{aligned}
&\sup_{f\in C} \sum_{t=1}^{T} \mathsf{E}\left[f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, \pi_t)\right] \\
&\geq \frac{M_3}{2^{M^d}} \sum_{t=1}^{T} \sum_{j=1}^{M^d} \sum_{w_{-j}} \sum_{w_j \in \{0,1\}} \mathsf{E}^{\pi}_{f_{w_{-j},w_j}}\left[(p^*(\boldsymbol{X}_t) - \pi_t)^2 \mathbb{I}_{\{\boldsymbol{X}_t \in B_j\}}\right] \\
&\geq \frac{TM_3 s^2 (1-4s)^2}{2^{M^d+1} M^{d+2}} \sum_{j=1}^{M^d} \sum_{w_{-j}} \exp\left(-z_j\right) \\
&= \frac{M_3 T}{2^{M^d+9} M^{d+2}} \sum_{j=1}^{M^d} \sum_{w_{-j}} \exp\left(-z_j\right)
\end{aligned}$$

where in the last step, we have set $s = 1/8$. ■