Dimension Reduction in Contextual Online Learning via Nonparametric Variable Selection

Preprint · September 2020

3 authors:

Wenhao Li
University of Technology Sydney

Ningyuan Chen
University of Toronto

L. Jeff Hong
Fudan University

Dimension Reduction in Contextual Online Learning via Nonparametric Variable Selection

Wenhao Li
Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong, wenhaoli6-c@my.cityu.edu.hk

Ningyuan Chen
Department of Management, University of Toronto Mississauga & Rotman School of Management, University of Toronto, Toronto, Cananda, ningyuan.chen@utoronto.ca

L. Jeff Hong
Department of Management Science, School of Management, Fudan University, Shanghai, China, hong_liu@fudan.edu.cn

Abstract

We consider a contextual online learning (multi-armed bandit) problem with high-dimensional covariate $\boldsymbol{x}$ and decision $\boldsymbol{y}$. The reward function to learn, $f(\boldsymbol{x}, \boldsymbol{y})$, does not have a particular parametric form. The literature has shown that the optimal regret is $\tilde{O}(T^{(d_x+d_y+1)/(d_x+d_y+2)})$, where d_x and d_y are the dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$, and thus it suffers from the curse of dimensionality. In many applications, only a small subset of variables in the covariate affect the value of f, which is referred to as *sparsity* in statistics. To take advantage of the sparsity structure of the covariate, we propose a variable selection algorithm called *BV-LASSO*, which incorporates novel ideas such as binning and voting to apply LASSO to nonparametric settings. Our algorithm achieves the regret $\tilde{O}(T^{(d_x^*+d_y+1)/(d_x^*+d_y+2)})$, where d_x^* is the effective covariate dimension. The regret matches the optimal regret when the covariate is d_x^*-dimensional and thus cannot be improved. Our algorithm may serve as a general recipe to achieve dimension reduction via variable selection in nonparametric settings.

Keywords: Contextual Bandits, Nonparametric Variable Selection, LASSO, Binning, Weighted Voting, High-dimensional Statistics

1 Introduction

Online learning is a popular paradigm to study dynamic decision making when new information can be collected actively to improve the quality of decisions simultaneously. It has seen numerous applications in the past decades in advertising, retailing, health care and so on. Among different frameworks for online learning, the multi-armed bandit (MAB) formulation is the most popular one. Interested readers may find a comprehensive review in Bubeck et al. [2012]. In the classic MAB framework, a decision maker faces a pool of candidate decisions (arms) in each period. The reward of each decision is random, whose mean is unknown a priori. The decision maker needs to learn the mean reward of each decision over time by active exploration and to choose the best one in the long term.

To accommodate the increasingly complex nature of many modern applications, the classic MAB framework has been extended in various directions, including

- A large (sometimes infinite) set of possible decisions. For instance, in dynamic pricing, a firm sets prices dynamically for a number of products over time, in order to learn the substitution patterns as well as the demand elasticity, to maximize revenues in the long run. The candidate decisions are the prices charged for various products, which are virtually infinite and high-dimensional. The discrete set of decisions used in MAB cannot properly capture the nature of dynamic pricing, and researchers have designed algorithms for continuous and high-dimensional decision variables.

- Contextual information or covariates. Covariates refer to the contextual information that is available for the decision maker to assess the current situation and make better decisions. In the example of dynamic pricing, when setting prices for a particular consumer, the personal information such as age, gender, and address can be used to infer the shopping habit of the consumer. It allows the firm to extract more revenues from consumers by price discrimination, but at the same time calls for more sophisticated decision rules to incorporate the covariates when learning the demand. The existence of covariates is ubiquitous in practice.

- Modeling the reward function. Learning and maximizing the reward function is the central goal of online learning. However, when little information is available, it is sometimes risky to even impose a model of *what to learn*. In dynamic pricing, it is tempting to assume that the demand is linear in the prices, and simplify the problem by learning only the linear coefficients. If the actual demand-price relationship is not linear, i.e. the model is misspecified, then the decision maker has little hope to find the optimal decision in the long run.

Next we informally describe the framework in the literature to incorporate those extensions. A formal introduction is included in Section 3. Consider a reward function $f(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ represents the covariate and $\boldsymbol{y}$ represents the decision. Both $\boldsymbol{x}$ and $\boldsymbol{y}$ can be vectors. The function is *non-parametric* and does not have a specific form except for a few general structures such as continuity and smoothness. In period t, a covariate $\boldsymbol{X}_t$ is generated and observed; the decision maker makes a decision $\boldsymbol{y}_t$ based on $\boldsymbol{X}_t$ as well as the historical information to maximize $f(\boldsymbol{X}_t, \boldsymbol{y}_t)$. The goal is to learn the optimal decision $\boldsymbol{y}(\boldsymbol{X}_t) = \arg\max_{\boldsymbol{y}} f(\boldsymbol{X}_t, \boldsymbol{y})$.

Unfortunately, it has been shown that the problem suffers from the *curse of dimensionality*. In particular, the optimal regret of the problem, a common metric in online learning, is[1] $\tilde{O}(T^{(d_x+d_y+1)/(d_x+d_y+2)})$ (see, e.g., Kleinberg et al. 2008, Slivkins 2014), where d_x and d_y are the

[1]We use $\tilde{O}$ to indicate asymptotic approximation neglecting logarithmic terms.

dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and T is the length of the learning horizon. In other words, the difficulty to learn the unknown reward function scales rapidly with d_x and d_y. No decision makers are able to break the fundamental limit without imposing additional assumptions on the reward function f.

On the other hand, in many applications, the information in the covariate $\boldsymbol{x}$ is likely to contain a great degree of redundancy. That is, out of d_x variables in $\boldsymbol{x}$, many may not affect the value of f at all. This is referred to as *sparsity* in statistics. In the example of dynamic pricing, for instance, the firm may have collected a rich set of personal information of a consumer (large d_x), while only a few key variables such as the income level actually affect the purchasing behavior. If we use d_x^* to denote the effective covariate dimension, or the number of *relevant* variables, then the question is, without knowing how many and which variables are redundant/relevant, can the decision maker achieve the regret $\tilde{O}(T^{(d_x^*+d_y+1)/(d_x^*+d_y+2)})$?

This paper provides an affirmative answer to the above question. Although such dimension reduction or variable selection has been one of the central topics in statistics for a few decades and has been well studied, the problem we consider is still very challenging because of the nonparametric nature of the reward function. In particular, statistical tools that are commonly used in variable selection such as LASSO [Hastie et al., 2015] are designed for certain parametric (linear) models. Applied to our nonparametric setting where any parametric family may be misspecified, it is unclear if they would work at all. Our paper addresses this challenge and contributes to the literature in the following aspects:

- Through the lens of online learning, we provide a nonparametric variable selection algorithm based on which the online learning can achieve regret $\tilde{O}(T^{(d_x^*+d_y+1)/(d_x^*+d_y+2)})$. In other words, the algorithm facilitates the learning of the decision maker as if s/he is informed of the sparsity structure of the covariate, i.e., how many and which variables are relevant, in advance. The regret matches the optimal regret when the covariate is only d_x^*-dimensional and thus cannot be improved. Therefore, we answer the fundamental question raised previously: when the covariate is sparse, we are able to identify the relevant variables and effectively lift the curse of dimensionality in online learning, even if the reward function is nonparametric.

- Our algorithm has two recipes that contribute to the successful variable selection in the nonparametric setting. Both may be of independent interest. The first one is *localized LASSO* (see Section 4). We partition the covariate space into small bins. Within each bin, we apply LASSO to the observations. Although LASSO only works for linear functions, we are able to show that the misspecification error incurred by approximating an arbitrary function f by linear functions can be controlled in a localized bin. That is, with properly chosen bin size and parameters, LASSO is able to identify relevant variables with high probability using the observations inside the bin despite the misspecification. This serves as the building block of our algorithm.

- Localized LASSO doesn't completely address the curse of dimensionality. To contain the approximation error of linear functions, the bin size needs to be small. As a result, the number of bins in a d_x-dimensional space grows exponentially in d_x and there are few observations in each bin. We resolve this issue by our second recipe, *weighted voting* (see Section 4). We aggregate the outcomes of variable selection in each bin and obtain a global set of selected variables. Each bin has a "vote" for whether a variable is relevant or not, and the weights of their votes depend on their "predictive power", which is calculated by our algorithm. For example, the localized LASSO applied to bin A predicts that x_1 is redundant, while bin B predicts the opposite. If A's vote carries more weight by our algorithm, possibly because it has more observations than B, then the algorithm makes a judgement that x_1 tends to be redundant. In this way, all the data in the covariate space are effectively utilized. The efficient use of data is reflected in our theoretical guarantee: the converence rate depends on the number of all observations as if the covariate space hadn't been partitioned.

We point out that although we introduce the algorithm in the online learning framework, it serves as a general recipe for variable selection in nonparametric settings. Therefore, it can be applied to other problems such as supervised learning. Next we review the related literature in the domain.

2 Related Literature

Our work is related to the literature studying nonparametric variable selection, contextual bandits and dynamic pricing with demand learning. We review the three streams below.

2.1 Nonparametric Variable Selection

In machine learning and statistics, the variable selection problem has been studied extensively. Suppose samples of $(Y, X_1, \ldots, X_{d_x})$ can be observed. Variable selection is concerned with the identification of relevant X_is that matter to the value of Y. Among the various methods proposed, LASSO is probably the most well-known. It combines computational efficiency and analytical tractability and is widely used in practice (see Bühlmann and Van De Geer 2011, Hastie et al. 2015 for a complete bibliography). However, LASSO assumes that Y depends on $(X_1, \ldots, X_{d_x})$ linearly. In general, variable selection is notoriously difficult in the nonparametric setting [Xu et al., 2016], when the dependence of Y on $(X_1, \ldots, X_{d_x})$ can be arbitrary. The difficulty lies in the potentially "local" behavior of a nonparametric function. Some variables may be irrelevant in some regions and affect the value of Y significantly elsewhere. One idea is to focus on the neighborhood of a given point and select relevant variables locally. For instance, Lafferty et al. [2008] identify the relevant variables by adjusting the bandwidth of a local linear regression. Bertin et al. [2008] apply LASSO to the observations locally near the given point. They provide consistency and finite sample bound when selecting variables in this way. Miller et al. [2010] discuss several local variable selection methods. It is not clear how to obtain a global sparsity structure from these methods, since locally the set of relevant variables may differ from region to region. The local methods also suffer from high dimensionality, as the observations in a neighborhood in a high-dimensional space are rather scarce. Although our algorithm builds on this idea, we provide an approach to aggregate the local predictions and create a global variable selector, which has a much better performance in high dimensions.

Other papers use the Reproducing Kernel Hilbert Space (RKHS) to represent nonparametric functions and conduct variable selection [Rosasco et al., 2013, Ye and Xie, 2012, Yang et al., 2016, He et al., 2018]. The choice of the kernel crucially determines the class of the functions. In a recent paper Xu et al. [2016] study the problem assuming the reward function $\mathbb{E}[Y] = f(X_1, \dots, X_{d_x})$ is convex and sparse. Different from these approaches, we do not impose kernel structures or shape constraints, and only assume more general structures such as continuity and smoothness.

Recently, there's a line of research using the "knockoff" framework. It's first proposed by Barber et al. [2015] for discovering relevant variables in a linear model. Candes et al. [2016], Barber et al. [2018, 2019], Ren and Candès [2020] extend it to more general, and in particular, nonparametric settings. Different from our setting, the knockoff method usually assumes the knowledge of the distribution of $(X_1, \dots, X_{d_x})$. Moreover, it focuses on the false positive rate (identifying a redundant variable as relevant) because of the bioinformatics applications such as finding the genetic expressions that affect the risk of a disease, in which the response only depends on a tiny fraction of variables. The setting and objective are different from the online learning problem that we are study.

Compared to the literature, the objective and method in this study are different. First, we do not allow d_x to scale with the number of observations, which is the focus of many studies in statistics. Moreover, besides selecting relevant variables, we do not want to recover the functional form $f(X_1, \dots, X_{d_x})$, which is the goal of sparse regression. They allow us to derive a strong theoretical guarantee and achieve near-optimal regret for online learning. Second, we provide a method called weighted voting, which effectively aggregates the information of local variable selections. It improves the localized methods in the literature and may be of independent interest.

2.2 Contextual Bandits

The literature on contextual bandits studies adaptive data collection and sequential decision-making (see Bubeck et al. 2012 for a complete bibliography). Many papers in this area consider linear reward in the covariates (see, e.g., Li et al. 2010). Among them, the sparsity structure of the contextual/covariate space has been studied by Carpentier and Munos [2012], Deshpande and Montanari [2012], Abbasi-Yadkori et al. [2012], Gilton and Willett [2017], Ren and Zhou [2020]. To our knowledge, Bastani and Bayati [2020] are the first to use the LASSO estimator to identify the sparsity. The proposed "LASSO bandit" algorithm obtains regret $O((d_x^*)^2(\log T + \log d_x)^2)$, which almost only depends on the effective dimension d_x^*, compared with the regret bound $O(d_x^3 \log T)$ of linear bandits without sparsity [Goldenshluger and Zeevi, 2013]. So the performance improves significantly if $d_x^* \ll d_x$. After that, Kim and Paik [2019], Wang et al. [2018] improve the regret by adopting doubly-robust and minimax concave penalized techniques. Recently, Oh et al. [2020] propose an algorithm solving the issue that the sparsity index d_x^* is not available in practice, which is required as prior knowledge in existing algorithms for sparse linear bandits. However, these methods are not applicable to the nonparametric setting that we consider in this paper. On one hand, there is no variable selection algorithm that is as powerful as LASSO in nonparametric settings. On the other hand, variable selection is particularly important for nonparametric online learning because the regret grows exponentially in the covariates dimension d_x. As a result, efficient nonparametric variable selection is both challenging and important. In this paper, we design new variable selection algorithms with a nonparametric setup and theoretically prove that the dependence of the regret on d_x can be reduced to d_x^* for online learning.

There are studies on nonparametric contextual bandits with finite arms and continuous reward functions [Yang et al., 2002, Rigollet and Zeevi, 2010, Perchet et al., 2013, Qian and Yang, 2016]. A similar stream of literature studies the continuum-armed bandits, where the arm/decision space is continuous just like the contextual space [Agrawal, 1995, Kleinberg, 2005, Auer et al., 2007, Kleinberg et al., 2008, Kleinberg and Slivkins, 2010, Bubeck et al., 2010, Magureanu et al., 2014]. A common result in the literature is that for continuous reward functions[2], the regret depends exponentially on d_x. For example, Lu et al. [2009], Slivkins [2014] present a uniformly partition and a zooming algorithm for reward functions that are Lipschitz continuous in both the decision and covariate. Both algorithms attain near-optimal regret $\tilde{O}(T^{1-1/(d_x+d_y+2)})$, where d_x, d_y are the dimensions of the covariate and decision space. Recently, Reeve et al. [2018], Guan and Jiang [2018] develop k-Nearest Neighbour (k-NN) based algorithms to address the dimensionality issue. Their algorithms automatically take advantage of the situations where the covariates are supported on a metric space of a lower effective dimension, such as a low-dimensional manifold embedded in a high dimensional space. However, they cannot be used to identify the sparsity structure. Our study attempts to lift the curse of dimensionality in the regret, particularly the exponential dependence on d_x. To the best of our knowledge, this is the first work to address the dimensionality issue in nonparametric contextual online learning by taking advantage of the sparsity structure. Although we formulate the problem for continuum-armed bandits, the approach can also be extended to discrete arms. Our work contributes to the contextual bandits literature by providing a general recipe to mitigate the curse of dimensionality for online learning.

There are papers focusing on the dimension reduction of the decision/arm space. Kwon et al. [2017], Kwon and Perchet [2016] consider bandit problem with finite arms, where only a fraction of arms have a positive (non-zero) expected reward. Considering contextual continuum-armed bandits, it turns out that if f is globally concave in $\boldsymbol{y}$, then algorithms can be developed to achieve regret $\tilde{O}(d_y T^{(d_x+1)/(d_x+2)})$ [Li et al., 2019, Cesa-Bianchi et al., 2017]. The regret grows linearly with d_y instead of exponentially. Applying the variable selection algorithm in our paper as a subroutine,

[2] For reward functions with a higher order of smoothness, the regret may be lower. See Hu et al. [2019], Gur et al. [2019].

their algorithms can achieve a smaller regret $\tilde{O}(T^{(d_x^*+1)/(d_x^*+2)})$ under sparsity.

2.3 Dynamic Pricing with Demand Learning

Our paper is also related to the literature on personalized dynamic pricing with demand learning [Besbes and Zeevi, 2009, Keskin and Zeevi, 2014, den Boer and Zwart, 2014, den Boer, 2015]. In this stream of literature, demand functions are typically assumed to be linear in prices and consumer features (covariates). Qiang and Bayati [2016] show a myopic pricing policy can exhibit near-optimal revenue performance with regret $O(d_x \log T)$. Cohen et al. [2020] find a multi-dimensional binary search algorithm for adversarial features, which has regret $O(d_x^2 \log(T/d_x))$. Javanmard and Nazerzadeh [2019] consider the sparsity structure of features and propose a pricing policy achieving regret $O(d_x^* \log d_x \log T)$, where d_x^* represents the effective dimension of the d_x features. Ban and Keskin [2020], Keskin and Zeevi [2014] also provide methods to deal with sparse covariates. In particular, Ban and Keskin [2020] take into account the feature-dependent price sensitivity and show that when not all prices are informative, the regret is at least $d_x\sqrt{T}$ under any admissible policy. They design a near-optimal pricing policy achieving regret $O(d_x^*\sqrt{T}(\log d_x + \log T))$. Javanmard et al. [2019] extend the result of Ban and Keskin [2020] to multi-product setting, considering the interaction between different products. They use the multinomial logit choice model and propose a pricing policy with regret $O(\log(d_x T)(\sqrt{T} + d_x \log T))$. Nambiar et al. [2019] study the effect of model misspecification in which the true demand model is quasi-linear, but the seller incorrectly assumes linearity. To address the "price endogeneity" issue caused by model misspecification, they propose a "random price shock" algorithm incurring regret $O(d_x\sqrt{T})$. In the studies above, the dependence of regret on d_x or d_x^* is not exponential as the demand is assumed to have a parametric (linear) form.

Going beyond the parametric setting, Hong et al. [2020] consider multi-product pricing problem without covariates. The proposed algorithm achieves the regret lower bound in parametric setting. They show that under some curvature conditions, such as concavity/convexity, the non-parametric pricing problems are not necessarily more difficult to solve than the parametric ones. Considering the contextual information, Chen and Gallego [Accepted] propose a nonparametric pricing policy that adaptively splits the covariate space into smaller bins and offers similar prices for customers with features belonging to the same bins. The policy achieves a near-optimal regret of $O((\log T)^2 T^{(2+d_x)/(4+d_x)})$, which indeed depends on d_x exponentially. A similar dependence is found in network revenue management [Besbes and Zeevi, 2012] in which the dimension of the decision space d_y appears in the regret $O(T^{(2+d_y)/(3+d_y)})$. Therefore, the dimension of the covariate significantly complicates the learning problem in the nonparametric formulation. Our work proposes a dimension reduction method that significantly mitigates the dimensionality problem. Although we formulate the problem for online learning in general, our approach is applicable to dynamic pricing with consumer features.

3 Problem Formulation

We now formulate the online learning problem. We define the decision and covariate space as $\mathcal{X} := [0,1]^{d_x}$ and $\mathcal{Y} := [0,1]^{d_y}$. Let $\mathcal{T} = \{1, 2, \ldots, T\}$ denote the sequence of decision periods faced by the decision maker. At the beginning of each period $t \in \mathcal{T}$, the covariate $\boldsymbol{X}_t \in \mathcal{X}$, drawn independently from some unknown distribution[3], is revealed to the decision maker. Then the decision maker chooses a decision $\boldsymbol{Y}_t$ in $\mathcal{Y}$. The reward in period t is a random variable Z_t:

$$Z_t = f(\boldsymbol{X}_t, \boldsymbol{Y}_t) + \epsilon_t,$$

where $f(\boldsymbol{X}_t, \boldsymbol{Y}_t)$ is the mean reward function which is unknown. The noises ϵ_t satisfy the following standard assumption.

Assumption 1 (Sub-Gaussian Noise)**.** *The noises $\{\epsilon_t\}_{t=1}^T$ are independent σ sub-Gaussian, i.e., for any $\xi \geq 0$,*

$$\mathbb{P}(\epsilon_t \geq \xi) \leq \exp\left(-\frac{\xi^2}{2\sigma^2}\right).$$

Assumption 1 is widely used in statistics and many classical distributions are sub-Gaussian, such as any bounded and centered distribution or the normal distribution.

Now we formally define *policy* and *regret* which are critical concepts in designing online learning algorithms.

Policy. Before making decisions in period t, the information revealed to the decision-maker includes observed covariates $\{\boldsymbol{X}_s\}_{s=1}^t$, the adopted decisions $\{\boldsymbol{Y}_s\}_{s=1}^{t-1}$ and the realized rewards $\{Z_s\}_{s=1}^{t-1}$. A policy π_t is defined as a function mapping the past history to the decision space:

$$\boldsymbol{Y}_t = \pi_t\left(\boldsymbol{X}_t, Z_{t-1}, \boldsymbol{Y}_{t-1}, \boldsymbol{X}_{t-1}, Z_{t-2}, \boldsymbol{Y}_{t-2}, \boldsymbol{X}_{t-2}, \ldots, Z_1, \boldsymbol{Y}_1, \boldsymbol{X}_1\right).$$

Regret. If the reward function is known, then the optimal decision and reward given covariate $\boldsymbol{x}$ are

$$\boldsymbol{y}^*(\boldsymbol{x}) := \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}), \quad f^*(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}),$$

Since the decision maker does not have access to the unknown reward function, the total expected reward of any policy π is always lower than $\sum_{t=1}^T \mathbb{E}[f^*(\boldsymbol{X}_t)]$. A standard performance measure of a policy is defined as the expected gap between the reward with known f and the reward under policy π, aggregated over the entire time horizon, i.e.,

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

For the decision maker, the objective is thus to design a policy that achieves small regret for a class

[3] Slivkins [2014] assumes that the covariate arrivals x_t are fixed before the first round. We follow Perchet et al. [2013] and assume that $\boldsymbol{X}_t$ are i.i.d.

of functions f.

Remark 1. *To motivate the formulation, consider the following example of personalized dynamic pricing. An online retailer sets personalized prices for an assortment of products to consumers with observable features such as education backgrounds, incomes, occupations, etc. The demand for the products depends not only on the prices, but also on the personal information. The retailer observes the information of each arriving customer ($\boldsymbol{X}_t$), decides personalized prices ($\boldsymbol{Y}_t$) accordingly, and observes the revenue (Z_t). If the relationship (f) between customers' information, prices and revenue is unknown to the retailer, then it has to be learned from historical observations and the goal is to maximize the long-run revenue.*

A standard assumption in online learning of nonparametric functions is that $f(\boldsymbol{x}, \boldsymbol{y})$ is continuous, as it is virtually impossible to learn f if it can be arbitrarily discontinuous. Therefore, we assume that

Assumption 2 (Continuously Differentiable)**.** *The function $f(\boldsymbol{x}, \boldsymbol{y})$ is continuously differentiable.*

Under a slightly weaker assumption that $f(\boldsymbol{x}, \boldsymbol{y})$ is Lipschitz continuous in both $\boldsymbol{x}$ and $\boldsymbol{y}$, the optimal rate of regret is (see, e.g., Slivkins 2014)

$$\min_{\pi} \sup_{f} R_\pi(T) \geq \Omega(T^{1-1/(2+d_x+d_y)}). \tag{1}$$

The lower bound here reflects the curse of dimensionality in nonparametric online learning. The regret grows almost linearly in T for large d_x and d_y. For example, if $d_x = d_y = 5$, then $R_\pi(T) \geq \Omega(T^{\frac{11}{12}})$, which is much worse than $\Omega(\sqrt{T})$, the typical lower bound in the parametric setting. Since the regret in (1) cannot be further improved under the assumption that f is Lipschitz continuous, the dependence on dimensionality looks dire. We next introduce a sparsity structure on the covariate space that may remedy the high dimension d_x. In this paper, as we focus on the dimension reduction in the covariate space, we set $d_y = 1$ in the rest of the paper for the ease of exposition. All the results can be generalized to the cases where $d_y > 1$.

3.1 Assumptions on the Sparsity Structure

In many practical cases, not all the variables in the covariate have an impact on the value of f. In other words, out of d_x variables in the covariate, many are redundant. Such sparsity has been one of the central topics in statistics. More precisely, we consider

Assumption 3 (Sparse Covariate)**.** *There exists $d_x^* \leq d_x$, a subset $J = \{i_1, \ldots, i_{d_x^*}\} \subset \{1, \ldots, d_x\}$, and a function $g : [0,1]^{d_x^*} \mapsto \mathbb{R}$ such that for all $\boldsymbol{x} = (x_1, \ldots, x_{d_x}) \in \mathcal{X}$ and any $y \in \mathcal{Y}$ [4], we have*

$$f(x_1, \ldots, x_{d_x}, y) = g(x_{i_1}, \ldots, x_{i_{d^*}}, y).$$

[4] Since $d_y = 1$, we use a scalar y instead of a vector $\boldsymbol{y}$ from here.

Assumption 3 gives a rigorous definition of the sparsity. We refer to the variables in J as *relevant* variables and those in $J^c := \{1, \ldots, d_x\} \setminus J$ as *redundant* variables. With a slight abuse of notations, we denote $J^{(i)} = 1$ if $i \in J$ and $J^{(i)} = 0$ otherwise. Since the redundant variables do not affect f, their partial derivatives are always zero:

$$J^c = \left\{ i \in \{1, 2, \ldots, d_x\} : \frac{\partial f(\boldsymbol{x}, y)}{\partial x_i} = 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, \forall y \in \mathcal{Y} \right\}.$$

However, in the nonparametric setting, Assumption 3 alone is not sufficient to characterize the sparsity structure. Suppose f changes slightly along the direction of x_1, only when y is in a small region. For example,

$$f(\boldsymbol{x}, y) = g(x_2, \ldots, x_{d_x}, y) + \mathbb{I}(0 \leq y \leq \epsilon/2)(-4y^3 + 3\epsilon y^2)x_1 + \mathbb{I}(\epsilon/2 < y \leq \epsilon)(-4(\epsilon - y)^3 + 3\epsilon(\epsilon - y)^2)x_1,$$

for an arbitrarily small $\epsilon > 0$. The function f satisfies Assumption 2 if g is continuously differentiable. We see that x_1 plays a role when $y \leq \epsilon$, and technically speaking, it is a relevant variable. However, it is almost impossible for any methods to detect the relevance of x_1, since the partial derivatives $\partial f(\boldsymbol{x}, y) \partial x_1$ diminish for infinitesimal ϵ. To resolve this issue, we impose a stronger assumption that $\partial f(\boldsymbol{x}, y) \partial x_i$ is non-vanishing for all $y \in \mathcal{Y}$ and all $i \in J$.

Assumption 4 (Global Relevance)**.** *For all $i \in J$, we have*

$$\frac{\partial f(\boldsymbol{x}, y)}{\partial x_i} \neq 0, \quad \forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}.$$

Assumption 4 states that the relevant variables must play a *global* role, not only for all $y \in \mathcal{Y}$, but also for all $\boldsymbol{x} \in \mathcal{X}$. Their partial derivatives are non-vanishing everywhere. Note that Assumption 4 includes functions that do not belong to any parametric family. For example, the variables are allowed to have complex interactions.

For certain applications, Assumption 4 may be too strong, especially when some relevant variables are relevant locally but not globally in $\mathcal{X}$.

Remark 2. *Considering the dynamic pricing example, even the variables that strongly predict consumer behavior are not always relevant. For instance, the demand for a product may be significantly increased when the income ranges from "low" to "medium", while the income level becomes almost irrelevant when it is above a certain threshold. Technically speaking, the partial derivatives are not always bounded away from zero, in which case Assumption 4 may fail.*

To make our approach more practical, we relax Assumption 4 below.

Assumption 4' (Local Relevance)**.** *For all $i \in J$, there exists $\boldsymbol{x}_{(i)} \in \mathcal{X}$ such that*

$$\frac{\partial f(\boldsymbol{x}_{(i)}, y)}{\partial x_i} \neq 0, \quad \forall y \in \mathcal{Y}$$

Assumption 4' is much weaker than Assumption 4. For $i \in J$, it assumes non-vanishing partial derivatives at one point $\boldsymbol{x}_{(i)}$ in the domain, for all y. By Assumption 2, Assumptions 4 and 4' have the following implications.

Lemma 1. *1. Suppose Assumptions 2 and 4 hold. There exists a constant $C > 0$ such that*

$$\left|\frac{\partial f(\boldsymbol{x}, y)}{\partial x_i}\right| \geq C, \quad \forall i \in J, \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}. \tag{2}$$

2. Suppose Assumptions 2 and 4' hold and in addition, f is twice-differentiable with respect to $\boldsymbol{x}$ (e.g., see Assumption 5). There exists a constant $C > 0$ such that

$$\left|\frac{\partial f(\boldsymbol{x}, y)}{\partial x_i}\right| \geq C, \quad \forall i \in J, \boldsymbol{x} \in \mathcal{H}_i, y \in \mathcal{Y}, \tag{3}$$

where $\mathcal{H}_i \subset \mathcal{X}$ is a hypercube centred at $\boldsymbol{x}_{(i)}$.

Lemma 1 states that, under some smoothness conditions, the partial derivatives are not only non-vanishing but also bounded away from a constant C.

For exposition, we first introduce our algorithm that works for Assumption 4 in Section 4 and Section 5. Then we show that with some adjustment, the algorithm has the same theoretical guarantee under Assumption 4' in Section 6.

3.2 Online Learning with Nonparametric Variable Selection

If the set of relevant variables J were known a priori, then the decision maker would discard the redundant variables and apply online learning algorithms only for the effective variables with dimension d_x^*. For example, existing algorithms for contextual bandits in nonparametric settings [Kleinberg, 2005, Lu et al., 2009, Slivkins, 2014] can achieve the near-optimal regret of the order $\tilde{O}(T^{1-1/(d_x^*+3)})$. (Recall that we set $d_y = 1$.)

We propose a two-phase approach to handle the problem. In particular, we design a subroutine to select variables before applying the online learning algorithms. We hope to collect data to provide an estimated set of relevant variables, $\hat{J}$, within the first $n < T$ periods. If $\hat{J} = J$ with high probability and $n \ll T$, then the online learning algorithms can be executed as if J were known and the regret does not deteriorate significantly. We elaborate this idea below.

Variable Selection Phase. We refer to the first n periods devoted to variable selection as the variable selection phase. In this phase, the main goal of the algorithm is to correctly identify the set of relevant variables J with high probability. By Assumption 3, the sparsity structure remains identical for all y. Therefore, in this phase, the decision maker may simply use a fixed decision $y \in \mathcal{Y}$.

Therefore, the observed reward is generated by

$$Z_t = f(\boldsymbol{X}_t, y) + \epsilon_t, \quad t = 1, \dots, n.$$

Our goal is to use $\{(\boldsymbol{X}_t, Z_t)\}_{t=1}^n$ to select relevant variables. We describe the variable selection algorithm in details in Section 4.

Online Learning Phase. We refer to the remaining $T-n$ periods as the online learning phase. Given that the relevant variables in the covariate have been correctly identified, we may apply the existing algorithms [Kleinberg, 2005, Lu et al., 2009, Slivkins, 2014] for contextual bandits. Denote the expected cumulative regret in the remaining $T-n$ periods as $R_2(T-n)$. On the correctly selected covariate space, the Uniform algorithm in Kleinberg [2005], Lu et al. [2009] and the Contextual Zooming algorithm in Slivkins [2014] can achieve regret

$$R_2(T-n) = O\left((T-n)^{1-1/(d_x^*+3)} \log(T-n)\right). \tag{4}$$

We may use either as a subroutine in the online learning phase.

Combined Regret. The cumulative regret of the two phases depends on the probability of successful variable selection in the first phase and the regret of the subroutine in the second phase. More precisely, the expected cumulative regret of our algorithm over T periods is

$$R_\pi(T) \leq 2n \max_{\boldsymbol{x}\in\mathcal{X}, y\in\mathcal{Y}} |f(\boldsymbol{x}, y)| + \mathbb{P}(\hat{J} = J) R_2(T-n) + 2 \max_{\boldsymbol{x}\in\mathcal{X}, y\in\mathcal{Y}} |f(\boldsymbol{x}, y)| \mathbb{P}(\hat{J} \neq J)(T-n).$$

The first term reflects the regret incurred in the variable selection phase, because $f^*(\boldsymbol{X}_t) - f(\boldsymbol{X}_t, y) \leq 2\max_{\boldsymbol{x},y} |f(\boldsymbol{x}, y)|$ in a single period. The regret in the online learning phase combines two scenarios: a "good" event that the variable selection phase correctly identifies the relevant variables and a "bad" event, where incorrect variable selection leads to linearly growing regret. The following proposition shows a sufficient condition for the total regret of both phases to achieve the optimal rate of regret.

Proposition 1. *If $n \leq T^{1-1/(d_x^*+3)}$ and $\mathbb{P}(\hat{J} \neq J) \leq n^{-1/(d_x^*+2)}$, then $R_\pi(T) = O\left(T^{1-1/(d_x^*+3)} \log(T)\right)$.*

The proposition provides a guideline for the algorithmic design of the variable selection phase. In the next two sections, we elaborate on the details.

Remark 3. *If the dimension of decision space is d_y, then Proposition 1 is modified as: If $n \leq T^{1-1/(d_x^*+d_y+2)}$ and $\mathbb{P}(\hat{J} \neq J) \leq n^{-1/(d_x^*+d_y+1)}$, we have $R_\pi(T) = O\left(T^{1-1/(d_x^*+d_y+2)} \log(T)\right)$.*

Remark 4. *Arguably, a more integrated variable selection method performs better when T is not known a priori, in which case n cannot be determined. For example, online learning may be intertwined with variable selection to achieve the optimal regret. This is left for future research. We point out that a fixed decision in the initial phase of the horizon is a common practice. For instance, a firm may stick to an incumbent pricing decision, which has proved to perform reasonably well in the market, before committing to risky new decisions that are required for online learning. This pattern is consistent with our design.*

4 Variable Selection for Global Relevance

In this section, we propose a new variable selection algorithm, which is referred to as "Binning and Voting LASSO" (BV-LASSO). The algorithm utilizes the idea of LASSO, a well-known method in statistics and machine learning, to achieve nonparametric variable selection and thus dimension reduction.

For linear models, LASSO has proved to be extremely successful in practice with strong theoretical guarantees and computational efficiency [Zhao and Yu, 2005, 2006]. If applied to our data, the standard LASSO estimator solves the following problem:

$$(\theta_0, \boldsymbol{\theta}^{lasso}) = \underset{\theta_0, \boldsymbol{\theta}}{\arg\min} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left(Z_t - \theta_0 - \boldsymbol{X}_t^T \boldsymbol{\theta} \right)^2 + 2\lambda \|\boldsymbol{\theta}\|_1 \right\}, \tag{5}$$

where the hyper-parameter λ penalizes the ℓ_1-norm of the parameter $\boldsymbol{\theta}$. The basic intuition of LASSO is that the ℓ_1 loss function creates sparsity. If f is a linear function, then with properly chosen λ, the estimators $\boldsymbol{\theta}_i^{lasso}$ of redundant variables x_i tend to be zero, while the estimators of relevant variables remain non-zero with high probability. As a result, the set of relevant variables can be identified from the sign of $\boldsymbol{\theta}^{lasso}$.

However, in our setting f is not necessarily linear and LASSO may fail. For example, consider $d_x = d_y = 1$ and $f(x_1, y) = (x_1 - 0.5)^2$ where X_1 has a uniform distribution in $[0, 1]$. The LASSO estimator θ_1 returns zero, because it is the best linear estimator for the quadratic function, thus falsely ruling out the relevant variable x_1. On the flip side, LASSO may also return false positives for nonlinear functions, identifying redundant variables as relevant. For instance, consider $d_x = 3$, $d_y = 1$ and let $X_1 \sim U[0, 1]$, $X_2 \sim U[0, 1] \perp X_1$, $X_3 = 0.5X_1 + 0.5X_2$. LASSO can correctly identify the relevant variables for the linear function $f(x_1, x_2, y) = -x_1 + 2x_2$ and rule out X_3 as redundant. However, for a nonlinear function $f(x_1, x_2, y) = -x_1 + e^{2x_2}$, X_3 would be falsely identified as relevant.

Having highlighted the technical difficulties, we introduce two mild technical assumptions required for our algorithm.

Assumption 5. *(Second-order Smoothness) The function f is twice-differentiable with respect to $\boldsymbol{x}$, i.e., there exists $L > 0$ such that*

$$|f(\boldsymbol{x_1}, y) - f(\boldsymbol{x_2}, y) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x_2}, y)^T (\boldsymbol{x_1} - \boldsymbol{x_2})| \leq L \|\boldsymbol{x_1} - \boldsymbol{x_2}\|_\infty^2,$$

for all $\boldsymbol{x_1}, \boldsymbol{x_2} \in \mathcal{X}$, $y \in \mathcal{Y}$.

Assumption 5 imposes the smoothness condition of f and is widely adopted in many problems in statistics and optimization. It allows a second-order approximation for f in a small area. An implication of Assumption 5 is that the infinity norm of the Hessian matrix, $\|\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}, y)\|_\infty$, is bounded by $2L$.

Next, we impose an assumption on the distribution of the covariate.

Assumption 6 (Regular Covariate)**.** *The covariate $\boldsymbol{X} \in \mathcal{X}$ has a probability density function $\mu(\boldsymbol{x})$ and there exist $\mu_m, \mu_M, L_\mu > 0$ such that*

1. $\mu_m \leq \mu(\boldsymbol{x}) \leq \mu_M$ *for all* $\boldsymbol{x} \in \mathcal{X}$*,*

2. *The density function μ is L_μ-Lipschitz, i.e., $\mu(\boldsymbol{x}) - \mu(\boldsymbol{x}') \leq L_\mu \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.*

Assumption 6 imposes bounded and continuous density functions and is easy to satisfy in many cases. There is a more general and less interpretable version of Assumption 6, which we defer to Appendix A.2 for the exposition.

Now that we have introduced all the assumptions, next we propose the BV-LASSO algorithm. Before describing our algorithms in detail, we remark on the information available to the decision maker initially: the decision maker knows T, d_x, σ, μ_m, μ_M, and L but doesn't know d_x^*, J, L_μ or C.

4.1 Binning and Local Linear Approximation

We first partition the covariate space regularly into k^{d_x} hypercubes (bins), each with side length $h = 1/k$, denoted by

$$\mathcal{B}_h = \{B_j |\ j = 1, 2, \ldots, h^{-d_x}\}.$$

The intuition is that, although f is nonlinear, it can be approximated by a linear function in a small bin by the Taylor series expansion. The approximation error can be controlled by the size of the bins. More importantly, the approximation error becomes small *relative to* the statistical error of LASSO when the side length h is small enough.

To formalize the intuition, for a given bin B, we project the function f to the functional vector space spanned by linear functions of the variables for a fixed y (we omit the dependence on y if it doesn't cause confusions):

$$\theta_0 = \int_{\boldsymbol{x} \in B} f(\boldsymbol{x}, y)\, d\boldsymbol{x},\ \ \theta_i = \frac{\int_{\boldsymbol{x} \in B} [f(x_1, \ldots, x_{d_x}, y) - \theta_0] x_i \, d\boldsymbol{x}}{\int_{\boldsymbol{x} \in B} x_i^2 \, d\boldsymbol{x}}, \quad \text{for } i = 1, 2, \ldots, d_x \tag{6}$$

The projection $\theta_0 + \sum_{i=1}^{d_x} \theta_i x_i$ is the "best" linear approximation of $f(\boldsymbol{x}, y)$ in the sense of minimizing integrated squared error, i.e.,

$$(\theta_0, \theta_1, \ldots, \theta_{d_x}) = \underset{\theta_0, \theta_1, \ldots, \theta_{d_x}}{\arg\min} \int_{\boldsymbol{x} \in B} \left(f(x_1, \ldots, x_{d_x}, y) - \theta_0 - \sum_{i=1}^{d_x} \theta_i x_i \right)^2 \, dx_1 \ldots dx_{d_x}.$$

If the sparsity structure of the projection maintains that of the original function f, then we may attempt to run LASSO on the projection and recover the sparsity of f. To do so, we need to calibrate the approximation error, in order to compare it with the statistical properties of LASSO later. The following lemma provides such calibration.

Lemma 2. *Suppose $(\theta_0, \theta_1, \ldots, \theta_{d_x})$ are the coefficients of the linear projection of f in B shown in* (6)*. Under Assumptions 2, 3, 4 and 5, we have*

1. $|\theta_i| \geq C$ *for any* $i \in J$ *and* $|\theta_i| = 0$ *for any* $i \notin J$*, where* C *is a constant satisfying* (2)*.*

2. $|f(\boldsymbol{x}, y) - \theta_0 - \sum_{i=1}^{d_x} \theta_i x_i| \leq (4\sqrt{3}+1) L d_x h^2$ *for all* $\boldsymbol{x} = (x_1, \ldots, x_{d_x}) \in B$*, where the constant* L *is presented in Assumption 5.*

The first point of the lemma shows that the linear projection maintains the sparsity structure of f. More importantly, it doesn't diminish the partial derivatives. The second point shows that the approximation error of the linear approximation is $O(h^2)$. This is crucial in the subsequent analysis, as we would like to control the bias or the approximation error of LASSO by the bin size.

4.2 Localized LASSO

Next, we apply LASSO to a given bin B_j. Suppose there are n_j periods in which the generated covariate falls in B_j. With a slight abuse of notation, let $\boldsymbol{X}_t \in B_j$ for $t = 1, 2, \ldots, n_j$. We first *normalize* the data by defining

$$\boldsymbol{U}_t \coloneqq (\boldsymbol{X}_t - C_{B_j})/h \tag{7}$$

where C_{B_j} is the geometric centre of B_j. The LASSO selector for B_j solves the penalized least square problem and identifies the non-zero coefficients:

$$\hat{J}_j = \operatorname{supp}\left\{ \underset{\theta_0, \boldsymbol{\theta}}{\arg\min} \left\{ \frac{1}{n_j} \sum_{t=1}^{n_j} \left(Z_t - \theta_0 - \boldsymbol{U}_t^T \boldsymbol{\theta} \right)^2 + 2\lambda \|\boldsymbol{\theta}\|_1 \right\} \right\}, \tag{8}$$

where the operator supp selects the subset of $\boldsymbol{\theta}$ that are non-zero. Note that the normalization is an affine mapping and thus doesn't change $\hat{J}_j$ as long as λ is properly scaled. Indeed, we normalize in order to keep a constant λ that does not scale with h in the analysis.

Our hope is that $\hat{J}_j$ would be identical to J for small h. As shown in the second point of Lemma 2, the approximation error is $O(h^2)$. If LASSO selects the relevant variables for the linear projection when the approximation error is small, then $\hat{J}_j = J$ because of the first point of Lemma 2. This intuition is formalized below.

Proposition 2 (Variable Selection of Localized LASSO)**.** *For a given bin B_j of side length h, under Assumptions 1, 2, 3, 4, 5 and 6, choosing $h \leq b_3$ and $\lambda = b_2 h^2$ in* (8)*, we have*

$$\mathbb{P}\left(\hat{J}_j = J\right) \geq 1 - p_j, \tag{9}$$

where $p_j \coloneqq b_0 \exp(-b_1 n_j h^4)$, for constants b_0, b_1, b_2, and b_3 presented in Section 5.

To streamline the proposition, we postpone the expressions of the constants to Section 5. Proposition 2 provides an accurate characterization of the probability of $\hat{J}_j = J$. In particular, h needs to be less than b_3, which itself depends on other constants. For example, it is understandable that if C is large, then J is easier to identify and the requirement b_3 can be larger. Once h is sufficiently small, the probability of $\hat{J}_j \neq J$ diminishes exponentially in $n_j h^4$. Proposition 2 serves as the backbone of the analysis of our algorithm.

Now that we have applied localized LASSO to a single bin, the next question is how to combine them to identify J. Because of the sheer number of bins ($1/h^{d_x}$), it is very unlikely that the sets of selected variables $\hat{J}_j$ are identical for all j despite the probability guarantee in Proposition 2. Next we introduce a scheme to aggregate $\hat{J}_j$ referred to as *weighted voting*.

4.3 Weighted Voting

After applying LASSO to all the bins, we have h^{-d_x} selectors $\{\hat{J}_j,\ j = 1, 2, \ldots, h^{-d_x}\}$, each representing a set of relevant variables. A straightforward idea would be to only trust the bin with most observations and use the outcome in that bin as the global selector. As n increases with T, the bin contains at least nh^{d_x} observations and Proposition 2 guarantees the correct selection with high probability. However, this method performs terribly in terms of efficient data utilization. For small h, any single bin would contain only a tiny fraction of all the observations $\{\boldsymbol{X}_t\}_{t=1}^n$. Such waste of data limits its practical use despite the asymptotic properties.

To fully exploit all the observations, we propose the idea of "weighted voting". For variable x_i, the outcome of LASSO in bin B_j, $\hat{J}_j^{(i)}$, is binary. If $i \in \hat{J}_j$, then bin B_j votes "yes" for x_i and $\hat{J}_j^{(i)} = 1$. Otherwise, the vote is "no" and $\hat{J}_j^{(i)} = 0$. If a majority of bins vote "yes", then x_i is likely to be relevant. Moreover, if B_j contains more observations, then we would expect $\hat{J}_j$ to be more reliable. This intuition is supported by Proposition 2, as the probability of false selection diminishes in n_j. Therefore, we assign more weights to the votes from the bins with more observations. In this way, all the observations are exploited as votes from all the bins are aggregated.

Next we describe the details of the procedure. For x_i, consider the linear combination of $\hat{J}_j^{(i)}$ over j:

$$\hat{J}^{(i)} = \sum_{j=1}^{h^{-d_x}} w_j \hat{J}_j^{(i)}, \tag{10}$$

where the weights $\{w_j\}$ satisfy

$$\sum_{j=1}^{h^{-d_x}} w_j = 1,\ w_j \geq 0.$$

If $\hat{J}^{(i)}$ is greater than $1/2$, implying that x_i has a weighted majority of "yes" votes, then we classify it as "relevant". Otherwise, we classify it as "redundant". The key questions to address are (1) how to properly choose the weights, and (2) how to control the errors, i.e., $\mathbb{P}\left(\hat{J}^{(i)} < 1/2 \middle| J^{(i)} = 1\right)$ and $\mathbb{P}\left(\hat{J}^{(i)} \geq 1/2 \middle| J^{(i)} = 0\right)$. Proposition 3 answers both questions.

Proposition 3. *Suppose $n \geq \log(2b_0)/(b_1 h^{d_x+4})$, $h \leq b_3$ and the weights are set to*

$$w_j = \begin{cases} \dfrac{\log 2 + \log p_j}{\sum_{k:p_k \leq 0.5}(\log 2 + \log p_k)} & \text{if } p_j \leq 0.5 \\ 0 & \text{if } p_j > 0.5 \end{cases},$$

where p_j is defined in Proposition 2. Then under Assumptions 1, 2, 3, 4, 5, and 6, we have

$$\mathbb{P}\left(\left|\hat{J}^{(i)} - J^{(i)}\right| \geq \frac{1}{2}\right) \leq \exp\left\{\frac{1}{2}\left(h^{-d_x}(1 + \log b_0 + \log 2) - b_1 n h^4\right)\right\}.$$

Moreover, the union bound implies

$$\mathbb{P}(\hat{J} = J) \geq 1 - d_x \exp\left\{\frac{1}{2}\left(h^{-d_x}(1 + \log b_0 + \log 2) - b_1 n h^4\right)\right\}.$$

Compared to Proposition 2, the probability bound improves from $\exp(-n_j h^4)$ to $\exp(-nh^4)$ omitting terms independent of n. This is a significant improvement when h is small and $n_j \ll n$. It demonstrates the power of weighted voting as it aggregates all the available data.

Remark 5. *We provide some intuitions for the convergence rate $O(\exp(h^{-d_x} - nh^4))$. It is well known that the false selection probability of LASSO for linear functions is $O(\exp(-n))$ (Theorem 11.3 in Hastie et al. 2015). Our bound has an additional term $\exp(h^{-d_x})$, because we have to discretize the covariate space into h^{-d_x} bins for the nonparametric setting. Also, there is another term h^4 in the convergence rate, which comes from approximating f by a linear function. There are two inferior alternatives to weighted voting: (1) If we just focus on a single bin, then roughly nh^{d_x} observations are used. So the convergence rate $O(\exp(h^{-d_x} - nh^{4+d_x}))$ is much worse than weighted voting. (2) If we assign the same weight to all the bins, then the votes from bins with fewer observations may tilt the outcome disproportionately, leading to noisy estimates.*

4.4 BV-LASSO and the Regret Analysis

After binning the observations, applying localized LASSO and weighted voting, the algorithm proceeds to the online learning phase and only focuses on the relevant variables in $\hat{J}$. Algorithm 1 demonstrates the complete algorithm combining the two phases, which we refer to as "BV-LASSO and Learning".

The regret analysis of the algorithm follows from Proposition 3. Since the false selection probability decreases exponentially with nh^4, BV-LASSO easily meets the rate required in Proposition 1. For properly chosen parameters, we have

Theorem 1. *If*

$$T \geq \exp\left\{\max\left\{(3 + \log 2 + \log b_0)/b_1, (b_3)^{-d_x}\right\}\right\},$$

Algorithm 1 BV-LASSO and Learning

Input: $T, d_x, \mu_m, \mu_M, L, \sigma$
Tunable parameters: n, h, λ
for $t = 1, 2, \ldots, n$ **do**
  Observe covariate $\boldsymbol{X}_t$
  Choose a fixed decision $Y_t = y$
  Observe Z_t {colleting observations in the variable selection phase}
end for
Partition the covariate space into $\mathcal{B}_h$
for $j = 1, 2, \ldots, h^{-d_x}$ **do**
  $\hat{J}_j = \text{supp}\left\{\arg\min_{\theta_0, \boldsymbol{\theta}} \left\{\frac{1}{n_j}\sum_{t=1}^{n_j}\left(Z_t - \theta_0 - \boldsymbol{U}_t^T\boldsymbol{\theta}\right)^2 + 2\lambda\|\boldsymbol{\theta}\|_1\right\}\right\}$ {applying LASSO to bin B_j}
end for
for $i = 1, 2, \ldots, d_x$ **do**
  $\hat{J}^{(i)} = \sum_{j=1}^{h^{-d_x}} w_j \hat{J}_j^{(i)}$ {w_j defined in Proposition 3}
end for
Let $\hat{J} = \{i : \hat{J}^{(i)} \geq 0.5\}$ {the set of selected coordinates}
for $t = n+1, n+2, \ldots, T$ **do**
  Apply contextual bandits algorithm to the variables in $\hat{J}$
end for

then taking $n = (\log T)^{2+\frac{4}{d_x}}$, $h = (\log T)^{-\frac{1}{d_x}}$ *and* $\lambda = b_2 h^2$, *we have*

$$R_\pi(T) = O\left(T^{1-1/(d_x^*+d_y+2)}\log(T)\right).$$

Note that the constants b_0, b_1, b_2, b_3, similar to Proposition 2, are given in Section 5. We do point out that to set the values of n, h, and λ, we need to be able to access some model parameters $(\sigma, \mu_m, \mu_M, L)$ and compute those constants. We discuss this point in Remark 6.

We have shown that BV-LASSO doesn't significantly increase the regret relative to the regret incurred in the online learning phase, demonstrated by the optimal rate of regret. As a general tool, we believe it has potential to be implemented for other nonparametric variable selection problems outside online learning.

5 Theoretical Analysis

In this section, we provide the detailed analysis of the theoretical results required for Theorem 1.

5.1 Analysis of Localized LASSO

In this section, we provide the major steps of the proof for Proposition 2. The proof is related to the variable selection consistency of LASSO [Zhao and Yu, 2006, Meinshausen et al., 2006, Wainwright, 2009]. We use some of the core ideas in proving the theoretical properties of LASSO and adapt them to the case when f is not necessarily linear.

Notations and Characterizations of LASSO. We rewrite the observations in bin B_j in the following form:

$$Z_t = f(\boldsymbol{X}_t, y) + \epsilon_t = \bar{\boldsymbol{U}}_t^T \boldsymbol{\theta}^* + \Delta_t + \epsilon_t =: \bar{\boldsymbol{U}}_t^T \boldsymbol{\theta}^* + \rho_t, \tag{11}$$

where $\bar{\boldsymbol{U}} = (1, \boldsymbol{U}) \in \mathbb{R}^{d_x+1}$ incorporates the constant term, $\boldsymbol{\theta}^*$ is the coefficients of the linear projection of f in B scaled by h because of the normalization, i.e., $\boldsymbol{\theta}^* = (\theta_0, h\theta_1, \ldots, h\theta_{d_x})^T$ where $(\theta_0, \ldots, \theta_{d_x})$ is the solution to (6), and $\Delta_t := f(\boldsymbol{X}_t, y) - \bar{\boldsymbol{U}}_t^T \boldsymbol{\theta}^*$ is the approximation error. In other words, we combine the random error ϵ_t and the approximation error Δ_t into ρ_t and transform the problem into a linear regression. It is still not a standard linear regression, as ρ_t is no longer i.i.d. and does not have mean zero. We hope to control Δ_t and thus ρ_t in the subsequent analysis because of Lemma 2.

The new form allows us to utilize the techniques developed for linear regression. More precisely, using the matrix expression, we define the design matrix $A := \frac{1}{\sqrt{n_j}} (\bar{\boldsymbol{U}}_1, \ldots, \bar{\boldsymbol{U}}_{n_j})^T$ and vectorize the observations $\boldsymbol{Z} = \frac{1}{\sqrt{n_j}} (Z_1, \ldots, Z_{n_j})^T$ and the error term $\boldsymbol{\rho} = \frac{1}{\sqrt{n_j}} (\rho_1, \ldots, \rho_{n_j})$. Then (11) can be written as $\boldsymbol{Z} = A\boldsymbol{\theta}^* + \boldsymbol{\rho}$. We also introduce the empirical version of the covariance matrix Ψ defined in Assumption 4, which will be useful in our analysis:

$$\hat{\Psi} = A^T A = \frac{1}{n_j} \sum_{i=1}^{n_j} \bar{\boldsymbol{U}}_i \bar{\boldsymbol{U}}_i^T.$$

We also rearrange the order of the variables so that $J = \{1, \ldots, d_x^*\}$ and $J^c = \{d_x^* + 1, \ldots, d_x\}$ and partition the vectors and matrices into "relevant" and "redundant" blocks:

$$A = \begin{pmatrix} A_{(1)} A_{(2)} \end{pmatrix}, \ \boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{\theta}^*_{(1)} \\ \boldsymbol{\theta}^*_{(2)} \end{pmatrix}, \ \hat{\Psi} = \begin{pmatrix} \hat{\Psi}_{11} & \hat{\Psi}_{12} \\ \hat{\Psi}_{21} & \hat{\Psi}_{22} \end{pmatrix} = \begin{pmatrix} A_{(1)}^T A_{(1)} & A_{(1)}^T A_{(2)} \\ A_{(2)}^T A_{(1)} & A_{(2)}^T A_{(2)} \end{pmatrix}, \tag{12}$$

where the dimensions are clear from the context (e.g., $A_{(1)} \in \mathbb{R}^{n_j \times (d_x^*+1)}$ because of the constant vector $\boldsymbol{e}$).

It is proved in Lemma 1 of Zhao and Yu [2006] that $\boldsymbol{\theta}$ solves (8) if and only if it satisfies the following KKT (Karush-Kuhn-Tucker) conditions:

$$\begin{aligned} (A_{\cdot i})^T (Z - A\boldsymbol{\theta}) &= \lambda \overrightarrow{\text{sign}}(\theta_i) \quad && \text{if} \quad \theta_i \neq 0 \\ \left|(A_{\cdot i})^T (Z - A\boldsymbol{\theta})\right| &\leq \lambda && \text{if} \quad \theta_i = 0 \end{aligned} \tag{13}$$

for all $i = 1, 2, \ldots, d_x$. Here $\overrightarrow{\text{sign}}(\cdot)$ stands for the sign function for each entry of a vector and $A_{\cdot i}$ stands for the i-column of A. Thus, our goal is to show that any $\boldsymbol{\theta}$ satisfying the above equations has the same signs as $\boldsymbol{\theta}^*$, which in turn matches the signs of the partial derivatives of f by Lemma 2. The following parts accomplish this goal.

"Good" Events for Sign Consistency. Suppose $\hat{\boldsymbol{\theta}}$ is the LASSO estimator for (8), or equivalently, a solution to (13). As $\hat{\boldsymbol{\theta}}$ doesn't have a closed form, we then define a set of events Ω_i,

$i = 1, \ldots, 4$, and argue that if $\cap_{i=1}^4 \Omega_i$ occurs, then $\hat{\boldsymbol{\theta}}$ has the same signs as $\boldsymbol{\theta}^*$. The first two events are defined as

$$\Omega_1 \coloneqq \left\{(1-\alpha)\underline{\lambda} \leq \lambda_{\min}(\hat{\Psi}) \leq \lambda_{\max}(\hat{\Psi}) \leq (1+\alpha)\overline{\lambda}\right\}$$
$$\Omega_2 \coloneqq \left\{|(\hat{\Psi}_{21})_{ik}| \leq (1+\delta)\gamma\underline{\lambda}/d_x^*,\ \forall i \in J^c, k \in J\right\},$$

where $\alpha \coloneqq \frac{1-\gamma}{2(1+\gamma)}$ and $\delta \coloneqq \frac{1-\gamma}{4\gamma}$, and $\overline{\lambda}, \underline{\lambda}, \gamma$ are defined in condition two of Assumption 6', a weaker version of Assumption 6 (discussed in Appendix A.2). Compared to Assumption 6', it is clear that Ω_1 and Ω_2 characterize the concentration of the empirical covariance matrix $\hat{\Psi}$ around the mean Ψ. In particular, Ω_1 corresponds to condition one of Assumption 6' and Ω_2 corresponds to condition two. Both events have error margins α and δ to accommodate the random error.

The events Ω_3 and Ω_4 are less straightforward to interpret:

$$\Omega_3 \coloneqq \left\{\left|(\hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho})_i - \lambda(\hat{\Psi}_{11}^{-1} \overrightarrow{\operatorname{sign}}(\boldsymbol{\theta}_{(1)}^*))_i\right| \leq |(\boldsymbol{\theta}_{(1)}^*)_i|,\ \forall i \in J\right\}$$
$$\Omega_4 \coloneqq \left\{\left|\left(\hat{\Psi}_{21}\hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} - A_{(2)}^T \boldsymbol{\rho}\right)_i\right| \leq \frac{1}{2}(1-\gamma)\lambda,\ \forall i \in J^c\right\}.$$

Since LASSO is a shrinkage estimation method, all the estimators $\hat{\boldsymbol{\theta}}$ are biased towards zero. Roughly speaking, Ω_3 guarantees that the estimators for the coefficients of relevant variables are not shrunk too much, while Ω_4 guarantees that the estimators for coefficients of redundant variables are shrunk sufficiently. The degree of the shrinkage is precisely controlled by the penalty term λ. After algebraic manipulations, one can show that $\Omega_3 \cap \Omega_4$ is equivalent to (13). When the joint event $\cap_{i=1}^4 \Omega_i$ occurs, we have

Lemma 3. *On the event $\cap_{i=1}^4 \Omega_i$, the LASSO estimator $\hat{\boldsymbol{\theta}}$ for* (8) *is unique and $\overrightarrow{\operatorname{sign}}(\hat{\boldsymbol{\theta}}) = \overrightarrow{\operatorname{sign}}(\boldsymbol{\theta}^*)$.*

Note that the techniques used in the proof are more or less standard in the LASSO literature. We present the complete proof in Appendix A.3.

Probability Bound for "Good" Events. By Lemma 3, we know the LASSO estimator has the desired property under the "good" events. The last step to prove Proposition 2 is to show $\cap_{i=1}^4 \Omega_i$ occurs with high probability.

Lemma 4. *Under Assumptions 1, 2, 3, 4, 5, and 6, choosing $h \leq b_3$ and $\lambda = b_2 h^2$, we have*

$$\mathbb{P}(\cap_{i=1}^4 \Omega_i) \geq 1 - b_0 \exp(b_1 n_j h^4).$$

The constants in Lemma 4 are the same as Proposition 2, which are presented below

$$
\begin{aligned}
&b_0(d_x) = 2\max\{2(d_x+1), d_x^2/4\},\\
&b_1(d_x,\mu_m,\mu_M,L,\sigma) = 11\mu_m/(10^4(1+d_x/4)) \wedge \mu_m^2/(4608 d_x^2) \wedge 64L^2d_x^2/(2\sigma^2) \wedge 22400\mu_M L^2 d_x^3/\sigma^2,\\
&b_2(d_x,\mu_M) = 64\sqrt{7\mu_M/3}Ld_x,\\
&b_3(d_x,\mu_m,\mu_M,L_\mu,C) = \min\left\{C\mu_m/(768\sqrt{21\mu_m d_x}), \mu_m^2/(3d_xL_\mu)\right\}.
\end{aligned}
$$

Their derivations can be found in the proof, which is provided in Appendix A.3.

Remark 6. *The constants μ_m, μ_M, L_μ, L, and σ appearing in Proposition 2 are defined in Assumptions 1, 3, 4, 5, 6 and the constant C is defined in* (2) *of Lemma 1. To implement the localized LASSO in a bin, the decision maker needs to know μ_M, d_x and L to obtain the penalty λ. To get the misidentification probability p_j for weighted voting, the decision maker in addition needs to know μ_m and σ. The implementation of Algorithm 1 does not need the value of C and L_μ, which appear in the bound for h that is satisfied automatically if n is large enough.*

The proof of Lemma 4 deviates significantly from the LASSO literature, as the error $\boldsymbol{\rho}$ is not i.i.d. due to the approximation error. The bound for $\mathbb{P}(\Omega_1 \cap \Omega_2)$ arises from random matrix concentration inequalities: the empirical covariance matrix $\hat{\Psi}$ can be viewed as the average of independent copies of $\bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T$, whose mean is Ψ. Therefore, we can guarantee that the spectrum (eigenvalues) and entries of the matrix do not deviate too much from the mean. The bound for $\mathbb{P}(\Omega_3 \cap \Omega_4)$ is harder to analyze, as it involves the matrix inverse and multiplications such as $\hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}$. The left-hand sides of the inequalities in Ω_3 and Ω_4 are linear transformations of the error $\boldsymbol{\rho}$, but the coefficients are not tractable. To analyze Ω_3 and Ω_4, we use the bounds for the eigenvalues conditional on Ω_1. In particular, we exploit the following inequalities: for a square matrix A and a vector $\boldsymbol{x}$, we have $\|A\|_2 \le \lambda_{\max}(A)$ and $\|A\boldsymbol{x}\|_2 \le \|A\|_2\|\boldsymbol{x}\|_2$. They help to reduce matrix multiplications to the eigenvalues, which is explicitly bounded in Ω_1. Eventually, we can transform Ω_3 and Ω_4 to a bound for a simple linear combination of sub-Gaussian random variables, for which we can apply standard concentration bounds.

5.2 Analysis of Weighted Voting

Now that we have obtained the probability of making mistakes in selecting relevant variables in a single bin from Proposition 2, we proceed to analyze the effect of weighted voting, i.e., Proposition 3. Note that for a certain variable x_i, the outcome of a bin $\hat{J}_j^{(i)}$ can be treated as a Bernoulli random variable with $\mathbb{P}(\hat{J}_j^{(i)} = 0 | J^{(i)} = 1) < p_j$ and $\mathbb{P}(\hat{J}_j^{(i)} = 1 | J^{(i)} = 0) < p_j$. Therefore, $\hat{J}^{(i)}$ in (10) is a weighted average of Bernoulli random variables. To analyze the probabilities $\mathbb{P}(\hat{J}^{(i)} \ge \xi | J^{(i)} = 0)$ or $\mathbb{P}(\hat{J}^{(i)} < \xi | J^{(i)} = 1)$ for some $\xi > 0$, we can use the concentration inequalities designed for Bernoulli random variables. In particular, using the techniques in proving Chernoff's inequality, we have that for all $\eta > 0$,

$$\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0) = \mathbb{P}(e^{\eta \hat{J}^{(i)}} \geq e^{\eta \xi} | J^{(i)} = 0)$$
$$\leq \exp(-\eta\xi) \prod_{j=1}^{h^{-d_x}} \mathbb{E}[\exp(\eta w_j X_j)] \leq \exp\left\{ \sum_{j=1}^{h^{-d_x}} (e^{\eta w_j} - 1) p_j - \eta \xi \right\}, \tag{14}$$

where X_j is a Bernoulli random variable with $\mathbb{P}(X_j = 1) < p_j$. The last inequality follows form the moment generating function of Bernoulli random variables: $\mathbb{E}[\exp(\eta X_j)] \leq 1 + p_j(\exp(\eta) - 1) \leq \exp(p_j(e^{\eta} - 1))$. The inequality (14) holds for all non-negative η and w_j. Our objective is to find η and w_j that minimize the misidentification error, i.e.,

$$\begin{aligned} \min_{\eta, \boldsymbol{w}} \quad & V(\eta, \boldsymbol{w}) := \exp\left\{ \sum_{j=1}^{h^{-d_x}} (e^{\eta w_j} - 1) p_j - \eta \xi \right\} \\ \text{s.t.} \quad & w_j \geq 0, \ \forall j \in \{1, 2, \ldots, h^{-d_x}\}, \\ & \eta \geq 0, \\ & \sum_{j=1}^{h^{-d_x}} w_j = 1. \end{aligned} \tag{15}$$

Since (15) minimizes a continuous function over a compact set, it has optimal solutions. Although the objective function $V(\eta, \boldsymbol{w})$ may not be convex in $(\eta, \boldsymbol{w})$, we apply the KKT condition as a necessary condition to find all local minima. In the proof of Lemma 5 in Appendix A.4, we prove that the KKT condition admits a unique solution. Thus, it must be the global optimum for problem (15).

Lemma 5. *The optimal solution $\eta^*, \boldsymbol{w}^*$ of the optimization problem* (15) *satisfies:*

1. $\eta^* = \sum_{j=1}^{h^{-d_x}} (\log \xi - \log p_j) \mathbb{I}(p_j < \xi)$*;*
2. *If $p_j < \xi$, then $w_j^* = (\log \xi - \log p_j)/\eta^*$;*
3. *If $p_j \geq \xi$, then $w_j^* = 0$;*
4. *The optimal value $V(\eta^*, \boldsymbol{w}^*) = \exp\left\{ \sum_{j=1}^{h^{-d_x}} (\xi - \xi \log \xi - p_j + \xi \log p_j) \mathbb{I}(p_j < \xi) \right\}$.*

Lemma 5 implies an intuitive structure of the weights. If a bin B_j has a high misidentification error $p_j > \xi$, then the variable selection output by B_j is not counted in the vote ($w_j = 0$). Otherwise, the weight assigned is proportional to $\log(\xi/p_j)$. Clearly, the weights are biased toward the bins with higher confidence (smaller p_j). Moreover, recall that $p_j = b_0 \exp(-b_1 n_j h^4)$. So $\log(\xi/p_j)$ roughly grows in the order of n_j. In other words, the voting power from B_j is almost proportional to the number of observations n_j in each bin. Therefore, each observation contributes equally to the global selector of the covariates.

Lemma 5 provides a weighting mechanism after the covariates have been generated and observed (after calculating p_j). What about the ex ante performance of the mechanism? Note that p_j depends on n_j, the number of observations in a bin. If the distribution of $\boldsymbol{X}$ were known, then p_j might be estimated. However, this is usually too strong an assumption in typical learning problems. Instead, we investigate the worst-case scenario in which $V(\eta, \boldsymbol{w})$ attains the maximum for all possible values of p_j (or equivalently, n_j). Using the form of $V(\eta^*, \boldsymbol{w}^*)$ from Lemma 5, we have

$$\begin{aligned} \max_{\boldsymbol{n}} \quad & V(\boldsymbol{n}) := \exp\left\{ \sum_{j=1}^{h^{-d_x}} (\xi - p_j - \xi \log \xi + \xi \log p_j) \mathbb{I}(p_j < \xi) \right\} \\ \text{s.t.} \quad & p_j = b_0 \exp\left(-b_1 n_j h^4\right) \\ & \sum_{j=1}^{h^{-d_x}} n_j = n \\ & n_j \in N^+, \quad \forall j \in \{1, 2, \ldots, h^{-d_x}\}. \end{aligned} \tag{16}$$

Note that the discontinuity in the objective function introduced by the indicator $\mathbb{I}(p_j < \xi)$ presents a challenge. To address the issue, we treat it as a budget allocation problem. Then, after analyzing the optimal budget allocation rule, we reformulate it as a concave optimization problem. The optimal solution is demonstrated in the following lemma.

Lemma 6. *The optimal solution $\boldsymbol{n}^*$ of the optimization problem (16) satisfies $n_1^* = n_2^* = \ldots = n_{h^{-d_x}}^* = nh^{d_x}$, and the optimal value satisfies $V(\boldsymbol{n}^*) \le \exp\left\{\xi\left(h^{-d_x}(1 + \log b_0 - \log \xi) - b_1 n h^4\right)\right\}$.*

Lemma 6 shows that the worst case occurs when the covariates are equally distributed across the bins. Combining Lemma 5 and Lemma 6 and setting $\xi = 0.5$, we can prove Proposition 3.

6 BV-LASSO for Local Relevance

In this section, we relax Assumption 4 to Assumption 4'. Recall from the analysis in Section 5 that Assumption 4 plays an important role in the successful variable selection (Proposition 2). However, the theoretical guarantee only requires that $|\partial f(\boldsymbol{x}, y) \partial x_i| \ge C$ always holds locally in a bin (see Lemma 1). If this is the case for a large number of bins under Assumption 4', then one may still be able to select the variables by weighing the votes from these bins more, given that there is a mechanism to do so.

To see the intuition, note that the second part of Lemma 1 implies that the hypercube $\mathcal{H}_i$ is contained in the *level set* of variable i, defined as

$$\mathcal{H}_i \subset A_i(C) := \left\{ \boldsymbol{x} : \left| \frac{\partial f(\boldsymbol{x}, y)}{\partial x_i} \right| \ge C, \ \forall\, y \in \mathcal{Y} \right\}, \tag{17}$$

for the same C appearing in Lemma 1. This implies that as $h \to 0$, there are always at least a constant fraction of the h^{-d_x} bins entirely inside $\mathcal{H}_i$, or $A_i(C)$. For those bins, which we refer to as "informative bins", $|\partial f(\boldsymbol{x}, y)\partial x_i| \geq C$ holds locally and the probability guarantee in Proposition 2 holds for the bin. On the other hand, for "uninformative bins" which are partially or entirely outside $A_i(C)$, Assumption 4 fails and we no longer have any theoretical guarantee for the output of localized LASSO.

To formalize the idea, given h and the partition $\mathcal{B}_h$, we define the informative area as the union of informative bins:

$$Q_i(C) \coloneqq \cup \left\{ B_j : B_j \subset A_i(C),\ B_j \in \mathcal{B}_h,\ j = 1, 2, \ldots, h^{-d_x} \right\},$$

while $Q_i^c(C)$ denotes the complimentary area. One would expect that when aggregating the outputs of localized LASSO from the bins, the BV-LASSO algorithm should still work if the area of $Q_i(C)$ does not vanish for $i \in J$. This is indeed the case, as $\mathcal{H}_i$ itself doesn't scale with h.

Proposition 4 (Informative area)**.** *Suppose Assumptions 2, 4', 5 and 6 hold. Then for the constant C, the hypercubes $\{\mathcal{H}_i\}_{i \in J}$ in the second part of Lemma 1 and $h \leq C/(3L)$, we have*

$$\mathbb{P}\left(\boldsymbol{X} \in Q_i(C)\right) \geq \mathbb{P}\left(\boldsymbol{X} \in \mathcal{H}_i\right) \geq \mu_m \left(\frac{C}{3L}\right)^{d_x} =: p_Q,$$

for all $i \in J$. Note that $p_Q \in (0, 1]$ is a constant.

Proposition 4 states that as $h \to 0$, there are always at least a constant fraction of bins for which Proposition 2 holds. Under the stronger Assumption 4, we always have $p_Q = 1$. As we shall see next, as long as p_Q is bounded away from zero, our algorithm can be adjusted to successfully select the relevant variables.

Remark 7. *Although Assumption 4' is very weak and sufficient for Proposition 4, one may be concerned that the hypercube $\mathcal{H}_i$ is small and leads to a small p_Q, which may affect the performance of the algorithm (see Proposition 5 below). In practice, the level set $A_i(C)$ and informative area $Q_i(C)$ can be much larger than $\mathcal{H}_i$ and the value of p_Q in Proposition 4 can be too conservative. Nevertheless, since our algorithm doesn't need to take p_Q as an input, the actual performance may be much better than the theoretical guarantee.*

To give some intuition, consider $f(x_1, x_2, y) = \exp\left(-15(x_1 - 0.5)^2 - 15(x_2 - 0.5)^2\right)$, which is illustrated in the left panel of Figure 1. The partial derivative of x_1 and its contour map are illustrated in the right panel of Figure 1. If we set $C = 0.9$, then the level set $A_1(0.9)$ is the area inside the contour line labelled 0.9 *and* -0.9.

Next consider $Q_1(0.9)$ for given $h = 0.2$ and $h = 0.1$, which is illustrated in Figure 2. The bins completely inside $A_1(0.9)$ are informative bins (heavily shaded bins) and the bins fully outside $A_1(0.9)$ (lightly shaded bins) are uninformative bins. There are some bins (white) intersecting with the boundaries of $A_1(0.9)$, also counted as uninformative. As $h \to 0$, $Q_1(0.9)$ approximates $A_1(0.9)$ and $\mathbb{P}(\boldsymbol{X} \in Q_1(0.9))$ converges to $\mathbb{P}(\boldsymbol{X} \in A_1(0.9))$.

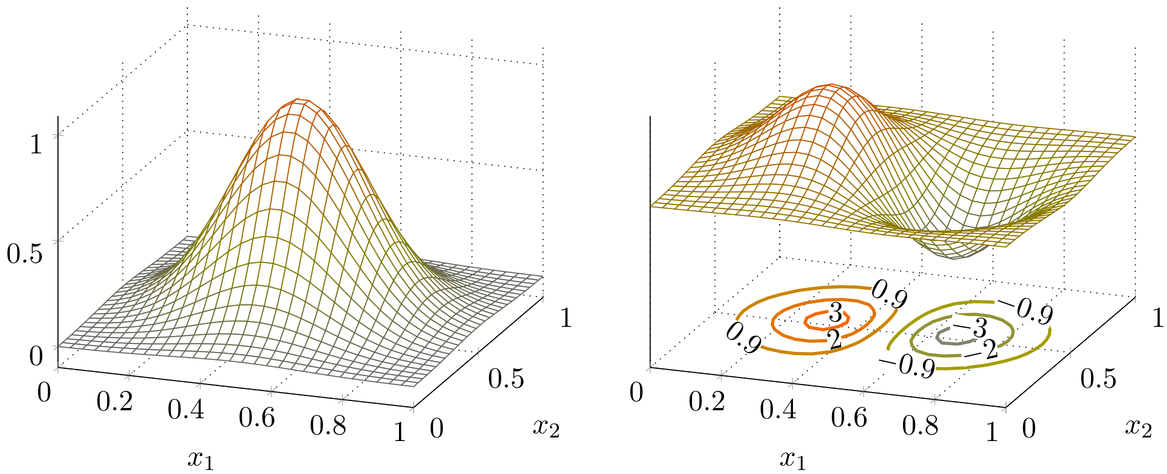

Figure 1: An illustration of the level set.

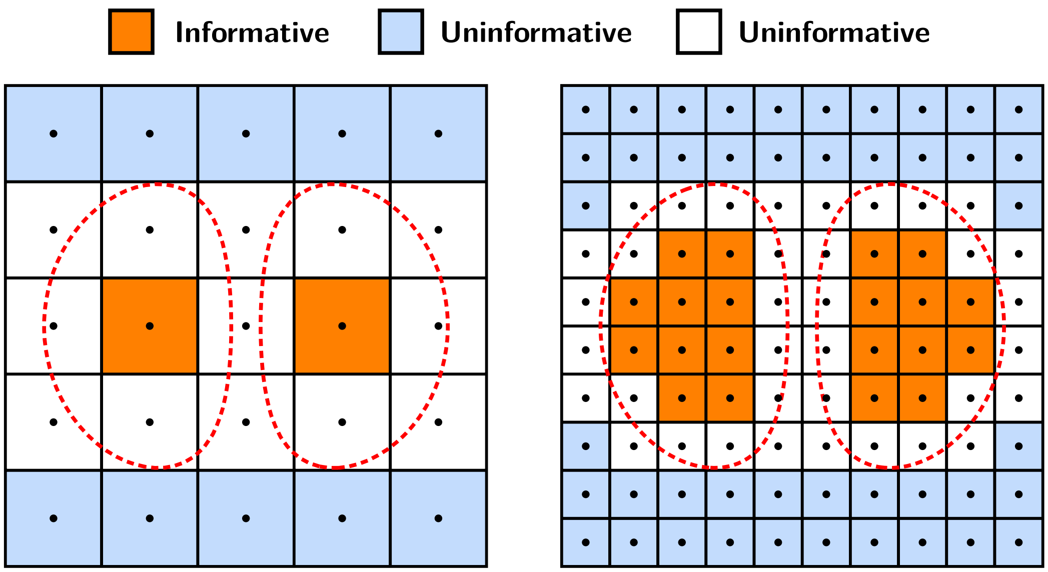

Figure 2: An illustration of the informative area.

Note that Proposition 4 guarantees that roughly at least $p_Q n$ observations fall into informative bins. If the decision maker knows which bins are informative a priori, s/he can assign zero weights to uninformative bins and only allows the informative bins to vote, then the problem is simplified to the problem analyzed in Section 4 with $p_Q n$ observations. The challenge, of course, is that the decision maker does not know which bins are informative. If a majority of bins are uninformative and they vote that the variable is redundant, then it is hard for the decision maker to screen out the noisy votes. In order to bias toward the informative bins in the weighted voting, the key is to tune ξ in Lemma 5 and 6. To see this, note that the threshold ξ balances the false positive probability $\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0)$ and the false negative probability $\mathbb{P}(\hat{J}^{(i)} \leq \xi | J^{(i)} = 1)$. A smaller ξ leads to a higher false positive rate and a lower false negative rate. Because uninformative bins tend to vote "redundant" (or negative) even if $i \in J$, we want to set ξ to be smaller to reduce the false negative rate, which is biased toward informative bins. That is, a small ξ assigns more importance to the bins that vote "relevant" and less importance to the bins that vote "redundant". For a relevant variable, if ξ is sufficiently small, then the "relevant" votes from the informative bins eventually outweigh the "redundant" votes. For a redundant variable, although the importance of "redundant" votes shrinks, there are no bins systematically voting "relevant" and the probability is still guaranteed. The next proposition shows the probability guarantee of the modified voting scheme.

Proposition 5. *Suppose that* $n \geq \log(2b_0)/(b_1 h^{d_x+4})$*,* $h < \min\{C/(3L), b_3/2\}$ *and the weights satisfy*

$$w_j = \begin{cases} \dfrac{\log\xi + \log(1-p_j) - \log p_j - \log(1-\xi)}{\sum_{k:p_k\leq\xi} \log\xi + \log(1-p_k) - \log p_k - \log(1-\xi)}, & \text{if } p_j \leq \xi \\ 0, & \text{if } p_j > \xi \end{cases}.$$

Then, under Assumptions 1, 2, 3, 4', 5 and 6, the misidentification probability of x_i *is bounded by*

$$\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0) \leq \exp\left\{(2\xi\log b_0 - 2\xi\log\xi - (1-\xi)\log(1-\xi))\, h^{-d_x} - \xi b_1 h^4 n\right\}, \tag{18}$$

$$\begin{aligned} \mathbb{P}(\hat{J}^{(i)} \leq \xi | J^{(i)} = 1) \leq{}& \exp\left\{(2(1-\xi)\log b_0 - \xi\log\xi + \xi\log(1-\xi))\, h^{-d_x} - (2p_Q/3 - \xi)\, b_1 h^4 n\right\} \\ &+ \exp\left(-\frac{2}{9}p_Q^2 n\right). \end{aligned} \tag{19}$$

From (18) and (19), we can see how ξ balances the false positive and false negative probabilities. The false positive probability $\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0)$ converges at the rate $O(\exp(-\xi b_1 h^4 n))$ while the false negative probability $\mathbb{P}(\hat{J}^{(i)} \leq \xi | J^{(i)} = 1)$ converges at the rate $O(\exp(-(2p_Q/3 - \xi) b_1 h^4 n))$. If the value of p_Q is known, then setting $\xi = p_Q/3$ leads to a bound of $\exp(-nh^4 p_Q/3)$ for both probabilities $\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0)$ and $\mathbb{P}(\hat{J}^{(i)} \leq \xi | J^{(i)} = 1)$. If p_Q is unknown, we set $\xi = 1/\log(n)$. Then, for a sufficiently large n (or T), ξ is less than $2p_Q/3$.

Corollary 1. *Under the conditions of Proposition 5, we have*

$$\mathbb{P}(\hat{J} = J) \geq \begin{cases} 1 - d_x \exp\left\{2(\log b_0 - \log(\frac{p_Q}{3}))h^{-d_x} - \frac{1}{3}p_Q b_1 h^4 n\right\} - d_x \exp\left(-\frac{2}{9}p_Q^2 n\right), & \text{if } \xi = \frac{p_Q}{3}, \\ 1 - d_x \exp\left\{2(\log b_0 + \log(\log n))h^{-d_x} - \frac{b_1 h^4 n}{\log n}\right\} - d_x \exp\left(-\frac{2}{9}p_Q^2 n\right), & \text{if } \xi = \frac{1}{\log n}. \end{cases}$$

Corollary 1 generalizes the theoretical guarantee of Proposition 3 to local relevance (Assumption 4'). Note that the new bound still guarantees the regret in Theorem 1, since the convergence rate is the same as in Proposition 3 except for the constants.

7 Numerical Experiments

In this section, we conduct numerical experiments to validate the theoretical performances of BV-LASSO. We attempt to address three questions in practice: (1) Can the BV-LASSO algorithm successfully select relevant variables? (2) How does the BV-LASSO and Learning algorithm perform against existing algorithms without considering the sparsity structure? (3) How does BV-LASSO perform when f is a linear function of $\boldsymbol{x}$? We first introduce the setups below.

Reward functions. Supposing $d_x = 3$ and $d_x^* = 1$, we consider two functions. The first function is nonlinear:

$$f_1(\boldsymbol{x}, y) = \exp\left(-10(x_1 - 0.5)^2 - 15(x_1 - y)^2\right), \tag{20}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$ and only x_1 is relevant. Note that its optimal decision $y^*(\boldsymbol{x}) = x_1$, and the optimal value $f_1^*(\boldsymbol{x}) = \exp(-10(x_1 - 0.5)^2)$. The second function is linear with respect to x_1 when y is fixed:

$$f_2(\boldsymbol{x}, y) = 3(1 - 2x_1)y + 3x_1. \tag{21}$$

When $x_1 < 0.5$, its optimal solution $y^*(\boldsymbol{x}) = 1$ and $f_2^*(\boldsymbol{x}) = 3 - 3x_1$; when $x_1 \geq 0.5$, its optimal solution $y^*(\boldsymbol{x}) = 0$ and $f_2^*(\boldsymbol{x}) = 3x_1$. At time t, the covariate $\boldsymbol{X}_t$ is independently sampled from a uniform distribution in $[0, 1]^3$. The noise ϵ_t are generated from a Gaussian distribution $N(0, \sigma^2)$, where we vary the value of σ as a robustness check.

BV-LASSO inputs. To implement the algorithm, we need to specify a set of hyper-parameters: $T, d_x, n, h, \lambda, \xi$. Among them, T and d_x are known to the decision-maker; ξ can be set to 0.5 as the partial derivatives are non-vanishing in most area; n,h are chosen as in Theorem 1. We also set $h = 1/\lfloor n^{1/(2d_x+4)} \rfloor$ for the bin size. To determine the value of λ, the l_1-penalty in localized LASSO, one is required to know L and μ_M as in Proposition 2. To avoid this scenario, we use a heuristic approach by noting that $\lambda = \Theta(h^2)$ in Proposition 2. We set $\lambda = c_\lambda h^2$ for some constant c_λ. We vary c_λ to better understand the sensitivity of the algorithm's performance to the choice.

To choose the weights w_j of the bins, if we follow Proposition 3, then the knowledge of μ_m, μ_M, L, σ are required, which is often unknown in practice. Instead, we simply set w_j to be proportional to n_j (number of observations in bin B_j), $w_j = n_j / \sum_{j=1}^{h^{-d_x}} n_j$, which is still consistent with Propositions 2 and 3 to a large degree. Our numerical results indicate a good performance.

Variable selection. First, we test the performance of BV-LASSO in terms of variable selection. The performance of BV-LASSO is affected by n, σ and λ. As n increases, the space is partitioned more granularly and there are more observations in each bin. Thus, we expect the performance to improve. The sub-Gaussian parameter σ reflects the signal-to-noise ratio. The penalty λ controls the balance between false positive and false negative. We show the results for varying n in Figure 3 while fixing $\sigma = 2$, $c_\lambda = 0.22$ and show the results for varying $\sigma \in \{1, 2, 4\}$ and $c_\lambda \in \{0.1, 0.2, 0.3\}$ in Figure 4.

Figure 3 compares the value of $\hat{J}$ of the three variables according to (10) based on the average of 20 trials, in which only x_1 is relevant. The left (right) panel corresponds to $f_1(x, y)$ ($f_2(x, y)$) and the shaded region corresponds to the 95% confidence interval of 20 trials. The results show $\hat{J}^{(1)}$ is significantly greater than 0.75 and $\hat{J}^{(2)}, \hat{J}^{(3)}$ are significantly less than 0.25. Choosing the threshold as $\xi = 0.5$, the relevant variable can be successfully selected, even if n is not large. The numerical example demonstrates that the BV-LASSO algorithm can successfully select relevant variables.

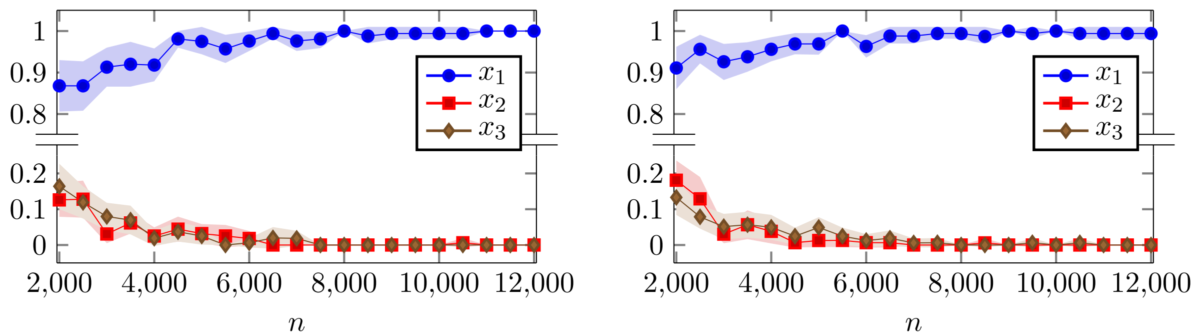

Figure 3: Variable selection of BV-LASSO for f_1 (left) and f_2 (right).

Figure 4 further shows the value of $\hat{J}$ for varying σ and c_λ. As $\hat{J}^{(3)}$ performs similar to $\hat{J}^{(2)}$, we omit $\hat{J}^{(3)}$ and display $\hat{J}^{(1)}$ ($\hat{J}^{(2)}$) in Figure 4. The indicators $\hat{J}^{(1)}$ ($\hat{J}^{(2)}$) for variable x_1 (x_2) are displayed in solid (dashed) curves with filled (hollow) markers. The top row of Figure 4 shows that $\hat{J}^{(1)}$ and $\hat{J}^{(2)}$ are not sensitive to σ as long as it is in a reasonable range. The bottom row of Figure 4 shows the impact of c_λ.

The Regret. Next we compare the regret of our Algorithm 1 with other learning algorithms. Our first benchmark is the Uniform algorithm [Kleinberg, 2005, Lu et al., 2009], which does not learn the sparse structure of the reward function. It incurs regret $O(T^{(d_x+2)/(d_x+3)}\log(T))$ or $\tilde{O}(T^{5/6})$ for functions (20) and (21). Our second benchmark is to first apply the standard LASSO algorithm to select the variables, and then use the Uniform algorithm on the selected variables. It is expected to incur regret $\tilde{O}(T^{3/4})$ for linear function (21) and linear regret for nonlinear function (20) due to model misspecification.

Figure 5 shows the regret of the three algorithms for a range of T values. We note that each point on the curve displays the average regret of 10 independent trials and the shaded region around each curve is the 95% confidence interval. The left (right) panel corresponds to $f_1(x, y)$ ($f_2(x, y)$). In both panels, BV-LASSO and Learning outperforms the other benchmarks. As predicted by the theory, the regret of the Uniform algorithm always grows at rate $\tilde{O}(T^{5/6})$ while the regret of BV-LASSO grows at $\tilde{O}(T^{3/4})$. When the function is nonlinear, the standard LASSO may fail to identify the relevant variable x_1 and incur large regret. When f is linear, the regret of BV-LASSO and Learning almost coincides with that of the standard LASSO, which has been proved to be one of the most effective variable selection methods for the linear setting.

8 Conclusions

In this paper, we study the online learning problem with a high-dimensional covariate. In particular, we address the curse of dimensionality under sparsity as the regret in existing algorithms scales exponentially in the covariate dimension d_x. To our knowledge, we are the first to propose a

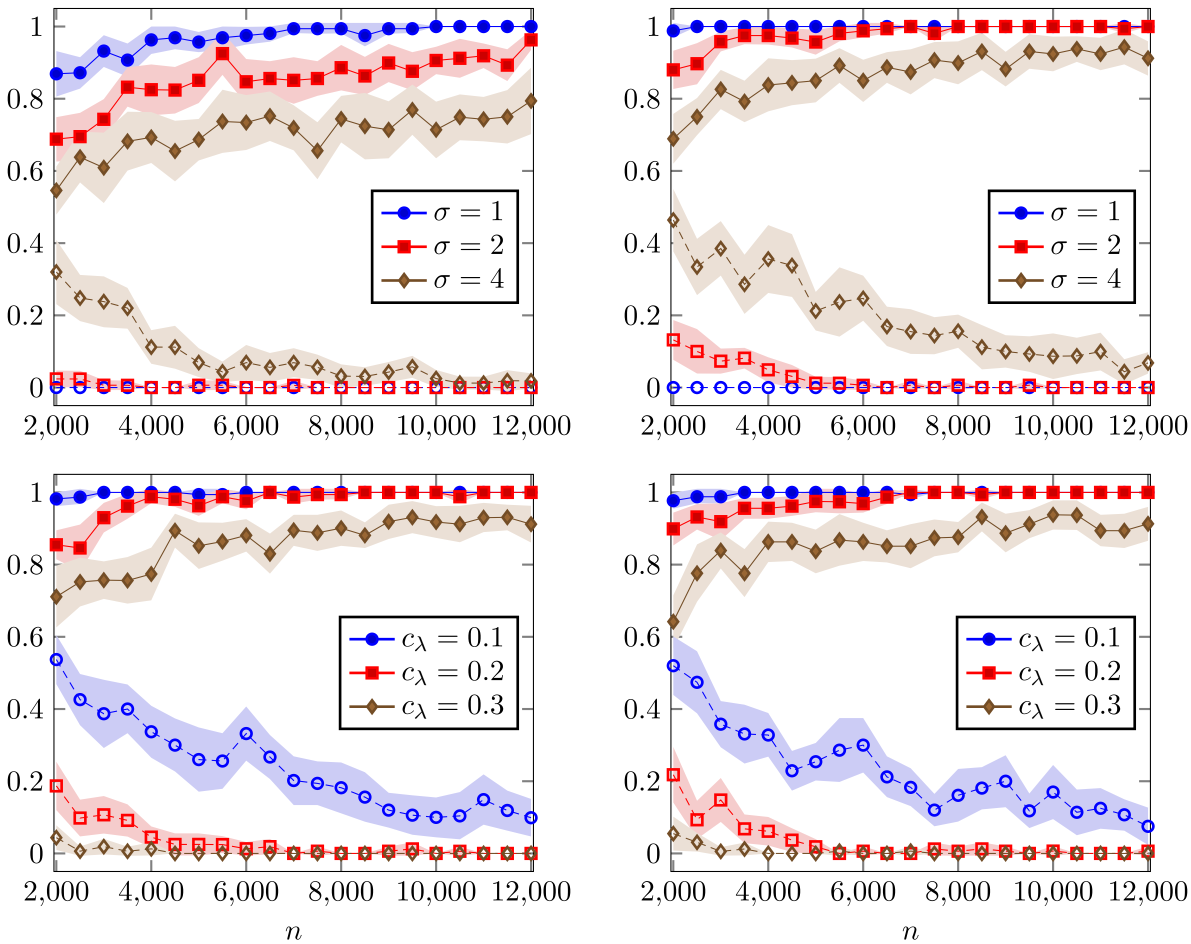

Figure 4: Variable selection of BV-LASSO for varying σ and c_λ. The left and right panels demonstrate the result for f_1 (20) and f_2 (21), respectively.

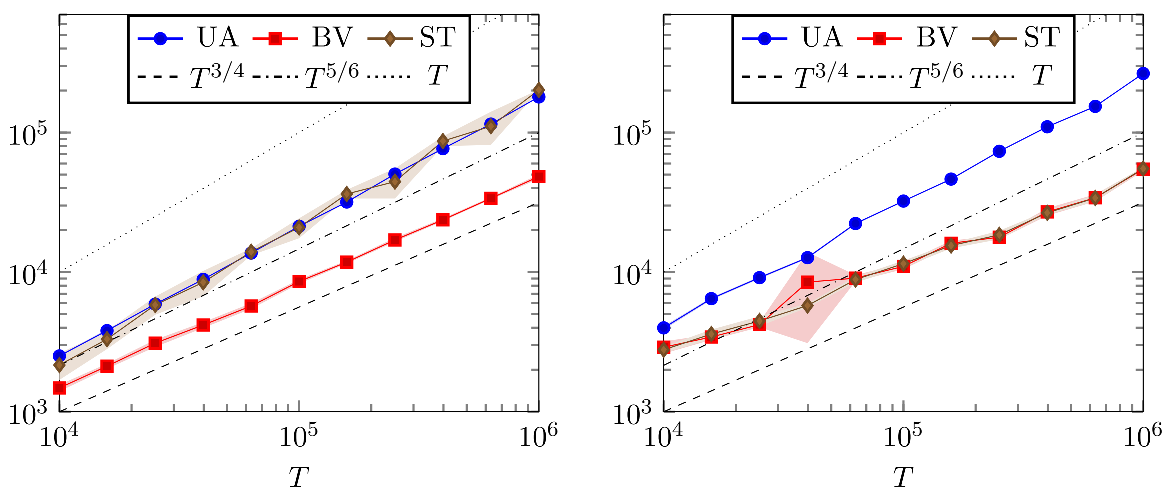

Figure 5: Comparison of regret. The abbreviation UA/BV/ST represents Uniform Algorithm/BV-LASSO/Standard LASSO.

nonparametric variable selection algorithm which takes advantage of the sparsity structure of the covariate. For online learning problems, our algorithm achieves the same order of regret as if the sparsity structure is known in advance. The BV-LASSO algorithm and its two recipes, localized LASSO and weighted voting, may be of independent interest in other nonparametric settings.

We conclude by discussing a few limitations of our algorithm and potential future directions. First, the length of the variable selection phase depends on T. When T is not known, it would be desirable to integrate variable selection with online learning more organically. Second, we assume the sparsity structure remains identical for all decisions. One may consider a setting where different decisions lead to different structures of sparsity. Finally, our algorithm requires the knowledge of several model parameters that are typically unknown and would ideally be optimized for specific applications. An interesting research question is to develop data-driven methods to select these parameters.

References

Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9, 2012.

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

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.

G.-Y. Ban and N. B. Keskin. Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Working paper*, 2020.

R. F. Barber, E. J. Candès, et al. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

R. F. Barber, E. J. Candès, and R. J. Samworth. Robust inference with knockoffs. *Working paper*, 2018.

R. F. Barber, E. J. Candès, et al. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537, 2019.

H. Bastani and M. Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

K. Bertin, G. Lecué, et al. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241, 2008.

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

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

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. X-armed bandits. *Working paper*, 2010.

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

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *Working paper*, 2016.

A. Carpentier and R. Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pages 190–198, 2012.

N. Cesa-Bianchi, P. Gaillard, C. Gentile, and S. Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. *Working paper*, 2017.

N. Chen and G. Gallego. Nonparametric pricing analytics with customer covariates. *Operations Research*, Accepted.

M. C. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. *Management Science*, 2020.

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

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

Y. Deshpande and A. Montanari. Linear bandits in high dimension and recommendation systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1750–1754. IEEE, 2012.

D. Gilton and R. Willett. Sparse linear contextual bandits via relevance vector machines. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 518–522. IEEE, 2017.

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

M. Y. Guan and H. Jiang. Nonparametric stochastic contextual bandits. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Y. Gur, A. Momeni, and S. Wager. Smoothness-adaptive stochastic bandits. *Working paper*, 2019.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

X. He, J. Wang, and S. Lv. Scalable kernel-based variable selection with sparsistency. *Working paper*, 2018.

L. J. Hong, C. Li, and J. Luo. Finite-time regret analysis of kiefer-wolfowitz stochastic approximation algorithm and nonparametric multi-product dynamic pricing with unknown demand. *Naval Research Logistics (NRL)*, 2020.

R. A. Horn and C. R. Johnson. Norms for vectors and matrices. *Matrix analysis*, pages 313–386, 1990.

Y. Hu, N. Kallus, and X. Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. *Working paper*, 2019.

A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

A. Javanmard, H. Nazerzadeh, and S. Shao. Multi-product dynamic pricing in high-dimensions with heterogenous price sensitivity. *Working paper*, 2019.

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

G.-S. Kim and M. C. Paik. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pages 5877–5887, 2019.

R. Kleinberg and A. Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 827–846. SIAM, 2010.

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

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

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *The Journal of Machine Learning Research*, 17(1):8106–8137, 2016.

J. Kwon, V. Perchet, and C. Vernade. Sparse stochastic bandits. In *Conference on Learning Theory*, pages 1269–1270, 2017.

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

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

W. Li, N. Chen, and L. J. Hong. A dimension-free algorithm for contextual continuum-armed bandits. *Working paper*, 2019.

T. Lu, D. Pál, and M. Pál. Showing relevant ads via context multi-armed bandits. In *Proceedings of AISTATS*, 2009.

S. Magureanu, R. Combes, and A. Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. *Working paper*, 2014.

N. Meinshausen, P. Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

H. Miller, P. Hall, et al. Local polynomial regression and variable selection. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 216–233. Institute of Mathematical Statistics, 2010.

M. Nambiar, D. Simchi-Levi, and H. Wang. Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000, 2019.

M.-h. Oh, G. Iyengar, and A. Zeevi. Sparsity-agnostic lasso bandit. *Working paper*, 2020.

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

W. Qian and Y. Yang. Kernel estimation and model combination in a bandit problem with covariates. *The Journal of Machine Learning Research*, 17(1):5181–5217, 2016.

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

H. W. Reeve, J. Mellor, and G. Brown. The k-nearest neighbour ucb algorithm for multi-armed bandits with covariates. *Working paper*, 2018.

Z. Ren and E. Candès. Knockoffs with side information. *Working paper*, 2020.

Z. Ren and Z. Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Working paper*, 2020.

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. *COLT 2010*, page 54, 2010.

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

W. Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

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

J. A. Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

X. Wang, M. Wei, and T. Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208, 2018.

M. Xu, M. Chen, J. Lafferty, et al. Faithful variable screening for high-dimensional convex regression. *The Annals of Statistics*, 44(6):2624–2660, 2016.

L. Yang, S. Lv, and J. Wang. Model-free variable selection in reproducing kernel hilbert space. *The Journal of Machine Learning Research*, 17(1):2885–2908, 2016.

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

G.-B. Ye and X. Xie. Learning sparse gradients for variable selection and dimension reduction. *Machine learning*, 87(3):303–355, 2012.

P. Zhao and B. Yu. Boosted lasso. *Feature Selection for Data Mining*, page 35, 2005.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

A Appendix

A.1 Proofs for Sparsity Assumptions

Proof of Lemma 1. 1. We first prove the first point of Lemma 1. For the simplicity of notation, we denote $\partial f(\boldsymbol{x}, y)\partial x_i$ as $f_i'(\boldsymbol{x}, y)$. By Assumption 2, $f(\boldsymbol{x}, y)$ are continuously differentiable with respect to $\boldsymbol{x} \in \mathcal{X}$, $y \in \mathcal{Y}$. Then $f_i'(\boldsymbol{x}, y)$ is a continuous function for all $i \in \{1, 2, \ldots, d_x\}$. Applying the generalized extreme value theorem [5], we have two constants

$$M_i = \sup_{\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}} f_i'(\boldsymbol{x}, y), \quad m_i = \inf_{\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}} f_i'(\boldsymbol{x}, y),$$

and there exist $\boldsymbol{x}_1 \in \mathcal{X}, y_1 \in \mathcal{Y}$ such that $f_i'(\boldsymbol{x}_1, y_1) = M_i$ and $\boldsymbol{x}_2 \in \mathcal{X}, y_2 \in \mathcal{Y}$ such that $f_i'(\boldsymbol{x}_2, y_2) = m_i$. By Assumption 4, we know $M_i, m_i \neq 0$ for $i \in J$. Then by Theorem 4.22 on page 93 of Rudin et al. [1964], $f_i'(\mathcal{X}, \mathcal{Y})$ is a connected set. If $M_i > 0 > m_i$, then $o \in f_i'(\mathcal{X}, \mathcal{Y})$ and there exist $\boldsymbol{x}_3 \in \mathcal{X}, y_3 \in \mathcal{Y}$ such that $f_i'(\boldsymbol{x}_3, y_3) = 0$, which violates Assumption 4. Thus, for $i \in J$, we have either $M_i > m_i > 0$ or $0 > M_i > m_i$. Let $C = \min_{i \in J}\{|m_i|, |M_i|\}$, we have $|f_i'(\boldsymbol{x}, y)| \geq C$ for all $i \in J, \boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. Thus, we prove the existence of C satisfying (2).

2. Next, we prove the second point in Lemma 1. Since we fix $\boldsymbol{x}_{(i)}$ in $\mathcal{X}$, applying the generalized extreme value theorem, we have the constants

$$M_i = \sup_{y \in \mathcal{Y}} f_i'(\boldsymbol{x}_{(i)}, y), \quad m_i = \inf_{y \in \mathcal{Y}} f_i'(\boldsymbol{x}_{(i)}, y).$$

Let $D = \min_{i \in J}\{|m_i|, |M_i|\}$, like the previous argument, we have $|f_i'(\boldsymbol{x}_{(i)}, y)| \geq D$ for all $i \in J$ and $y \in \mathcal{Y}$.

Furthermore, if f is twice-differentiable with respect to $\boldsymbol{x}$, we will prove that the hypercube $\mathcal{H}_i$ with side length $\overline{h} = D/(2L)$ and centred at $\boldsymbol{x}_{(i)}$ satisfies (3). We omit the argument y when writing $f(\boldsymbol{x}, y)$ as we prove the result for any fixed y. For any $\boldsymbol{x} \in \mathcal{H}_i$, we write $\boldsymbol{x} = \boldsymbol{x}_{(i)} + \boldsymbol{l}$ where $\boldsymbol{l} \in \mathcal{R}^{d_x}$ and $\|\boldsymbol{l}\|_\infty \leq \overline{h}/2$. Define a function $\psi(t) = \nabla f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})$. As assumed in Assumption 5, f is twice-differentiable, thus $\psi(t)$ is continuous differentiable. We have $\psi(t) = \psi(0) + \int_0^1 \psi'(t)\, dt$, i.e.,

$$\nabla f(\boldsymbol{x}_{(i)} + \boldsymbol{l}) = \nabla f(\boldsymbol{x}_{(i)}) + \int_0^1 \nabla^2 f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})\boldsymbol{l}\, dt.$$

Then moving $\nabla f(\boldsymbol{x}_{(i)})$ to the left-hand-side, and taking infinity norm, we have

$$\left\|\nabla f(\boldsymbol{x}_{(i)} + \boldsymbol{l}) - \nabla f(\boldsymbol{x}_{(i)})\right\|_\infty = \left\|\int_0^1 \nabla^2 f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})\boldsymbol{l}\, dt\right\|_\infty. \tag{A-1}$$

[5] See Theorem 4.16 on page 89 of Rudin et al. [1964]

According to the definition of infinity norm, for a matrix A, we have

$$\|A\|_\infty = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}, \text{ and } \|A\boldsymbol{x}\|_\infty \leq \|A\|_\infty \|\boldsymbol{x}\|_\infty.$$

Thus, we have

$$\left\|\nabla^2 f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})\boldsymbol{l}\right\|_\infty \leq \left\|\nabla^2 f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})\right\|_\infty \|\boldsymbol{l}\|_\infty \leq L\overline{h}, \tag{A-2}$$

where the last inequality holds by $\left\|\nabla^2 f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})\right\|_\infty \leq 2L$ (Assumption 5) and $\|\boldsymbol{l}\|_\infty \leq \overline{h}/2$. By (A-1) and (A-2), we have

$$\left\|\nabla f(\boldsymbol{x}_{(i)} + \boldsymbol{l}) - \nabla f(\boldsymbol{x}_{(i)})\right\|_\infty \leq \int_0^1 \left\|\nabla^2 f(\boldsymbol{x}_{(i)} + t\boldsymbol{l})\boldsymbol{l}\right\|_\infty dt \leq \int_0^1 L\overline{h}\, dt = L\overline{h} = D/2, \tag{A-3}$$

where the last equality follows by $\overline{h} = D/(2L)$. We rewrite (A-3) in the form of partial derivatives:

$$\left|\frac{\partial f(\boldsymbol{x})}{\partial x_i} - \frac{\partial f(\boldsymbol{x}_{(i)})}{\partial x_i}\right| \leq D/2, \quad \forall i \in J, \boldsymbol{x} \in \mathcal{H}_i. \tag{A-4}$$

By the previous argument, we have

$$\left|\frac{\partial f(\boldsymbol{x}_{(i)})}{\partial x_i}\right| > D, \quad \forall i \in J. \tag{A-5}$$

Combining (A-4) and (A-5), we have

$$\left|\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right| > D/2, \quad \forall i \in J, \boldsymbol{x} \in \mathcal{H}_i.$$

Finally, let $C = D/2$, we prove the existence of C and $\mathcal{H}_i$ satisfying (3).

□

A.2 Discussion about Assumption 6

In this section, we discuss a weaker version of Assumption 6 and use the weaker version in the proofs of Lemma 3, Lemma 4 and Proposition 2. Recall that $\boldsymbol{U}$ is $\boldsymbol{X}$ being normalized with regard to a bin, introduced in (7).

Assumption 6' (Regular Covariates)**.** *Given any hybercube $B \subset \mathcal{X}$ with side length h such that $\mathbb{P}(\boldsymbol{X} \in B) > 0$ and the normalization (7), we assume that the distribution of $\boldsymbol{X}$ satisfies:*

1. *The conditional covariance matrix $\Psi := \mathbb{E}[\boldsymbol{U}\boldsymbol{U}^T | \boldsymbol{X} \in B]$ satisfies*

$$0 < \underline{\lambda} \leq \lambda_{\min}(\Psi) \leq \lambda_{\max}(\Psi) \leq \overline{\lambda}$$

for some constants $\underline{\lambda}$ and $\overline{\lambda}$, where $\lambda_{\max}$ and $\lambda_{\min}$ represent the maximum and minimum eigenvalues of a matrix.

2. *For any $i \in J$ and $j \in J^c$, there exists a constant $\gamma \in [0,1)$, which may depend on h, such that*

$$(\Psi)_{ij} \leq \gamma \underline{\lambda}/d_x^*.$$

We first give a simple example to show the generality of Assumption 6'. If $\boldsymbol{X}$ follows an independent uniform distribution in $\mathcal{X}$, then $\Psi = \frac{1}{12}\boldsymbol{I}_{d_x}$.[6] It is easy to see that setting $\underline{\lambda} = \overline{\lambda} = \frac{1}{12}$ and $\gamma = 0$ satisfies the assumption.

The first condition of Assumption 6' prevents singular covariate distributions. If the covariates are linearly dependent ($\underline{\lambda} = 0$), then the definition of relevant/redundant variables is ambiguous, as one covariate can be represented by others. In other words, we need sufficient variations in all the dimensions of $\boldsymbol{X}$ in order to estimate the partial derivatives. Similar conditions are imposed in the LASSO literature [Bühlmann and Van De Geer, 2011, Goldenshluger and Zeevi, 2013, Bastani and Bayati, 2020].

The second condition of Assumption 6' states that the pairwise correlation between relevant and redundant variables cannot be too high. It's a sufficient condition for the well-known "Strong Irrepresentable Condition" for LASSO proposed in [Zhao and Yu, 2006]. It prevents any redundant variable to be fully linearly represented by the relevant variables. Note that the condition two is hard to check in practice since d_x^* is unknown. One alternative is to replace d_x^* by d_x and make the assumption stronger: $(\Psi)_{ij} \leq \gamma\underline{\lambda}/d_x$.

Next, we show that if the side length h is small enough, Assumption 6 implies Assumption 6'.

Proposition A1. If Assumption 6 is satisfied and the side length $h < \mu_m^2/(3d_x^* L_\mu)$, then Assumption 6' holds with $\underline{\lambda} = \mu_m/12$, $\overline{\lambda} = \mu_M$ and $\gamma = 3d_x^* L_\mu h/(2\mu_m^2)$.

Proposition A1 states that the first condition of Assumption 6' can be implies by Assumption 6 and the second condition holds automatically when h is sufficiently small, as the requirement of γ diminishes linearly in h.

Proof. Let $\boldsymbol{x}_0$ be a vector in $\mathbb{R}^{d_x}$. To show the first condition of Assumption 6' holds, by the definition of eigenvalues, we only need to provide upper and lower bounds for $\boldsymbol{x}_0^T \Psi \boldsymbol{x}_0/\|\boldsymbol{x}_0\|_2$. Note that

$$\int_{\boldsymbol{x}\in B} \boldsymbol{x}_0^T \boldsymbol{U}(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x})^T \boldsymbol{x}_0 \mu_m \, d\boldsymbol{x} \leq \int_{\boldsymbol{x}\in B} \boldsymbol{x}_0^T \boldsymbol{U}(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x})^T \boldsymbol{x}_0 \mu(\boldsymbol{x}) \, d\boldsymbol{x} \leq \int_{\boldsymbol{x}\in B} \boldsymbol{x}_0^T \boldsymbol{U}(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x})^T \boldsymbol{x}_0 \mu_M \, d\boldsymbol{x}. \tag{A-6}$$

Since $\boldsymbol{U} = (\boldsymbol{X} - C_B)/h$, we have

$$\int_{\boldsymbol{x}\in B} \boldsymbol{U}(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x})^T \, d\boldsymbol{x} = \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{12}\boldsymbol{I}_{d_x} \end{pmatrix}. \tag{A-7}$$

[6] This is the main reason we consider $\boldsymbol{U}$ instead of $\boldsymbol{X}$. The conditional covariance matrix of the normalized covariate is invariant when h changes.

Then, plugging equation (A-7) into (A-6), we have

$$\frac{\mu_m}{12}\|\boldsymbol{x}_0\|_2^2 \le \boldsymbol{x}_0^T \Psi \boldsymbol{x}_0 = \int_{\boldsymbol{x}\in B} \boldsymbol{x}_0^T \boldsymbol{U}(\boldsymbol{x})\boldsymbol{U}(\boldsymbol{x})^T \boldsymbol{x}_0 \mu(\boldsymbol{x})\, d\boldsymbol{x} \le \mu_M \|\boldsymbol{x}_0\|_2^2.$$

Thus, the first condition of Assumption 6' is satisfied by setting $\underline{\lambda} = \frac{\mu_m}{12}, \overline{\lambda} = \mu_M$. To prove the second condition, note that we have

$$\mu_m h^2 = \mu_m \int_{\boldsymbol{x}\in B} d\boldsymbol{x} \le \mathbb{P}(X \in B) = \int_{\boldsymbol{x}\in B} \mu(\boldsymbol{x})\, d\boldsymbol{x} \le \mu_M h^2.$$

Then, for any $i \in J$ and $j \in J^c$,

$$\begin{aligned}
(\Psi)_{ij} &= \mathbb{E}[U_i(\boldsymbol{X})U_j(\boldsymbol{X})|\boldsymbol{X} \in B] \\
&= \frac{1}{\mathbb{P}(X \in B)} \int_{\boldsymbol{x}\in B} U_i(\boldsymbol{x})U_j(\boldsymbol{x})\mu(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \frac{1}{\mathbb{P}(X \in B)} \int_{\boldsymbol{x}\in B} U_i(\boldsymbol{x})U_j(\boldsymbol{x})\left(\mu(\boldsymbol{C_B}) + \mu(\boldsymbol{x}) - \mu(\boldsymbol{C_B})\right) d\boldsymbol{x} \\
&\overset{(a)}{=} \frac{1}{\mathbb{P}(X \in B)} \int_{\boldsymbol{x}\in B} U_i(\boldsymbol{x})U_j(\boldsymbol{x})\left(\mu(\boldsymbol{x}) - \mu(\boldsymbol{C_B})\right) d\boldsymbol{x} \\
&\le \frac{1}{\mathbb{P}(X \in B)} \int_{\boldsymbol{x}\in B} |U_i(\boldsymbol{x})|\,|U_j(\boldsymbol{x})|\,|\mu(\boldsymbol{x}) - \mu(\boldsymbol{C_B})|\, d\boldsymbol{x} \\
&\overset{(b)}{\le} \frac{1}{\mathbb{P}(X \in B)} \int_{\boldsymbol{x}\in B} \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} L_\mu h\, d\boldsymbol{x} \\
&= \frac{L_\mu h^3}{8\mathbb{P}(X \in B)} \\
&\le \frac{L_\mu h}{8\mu_m},
\end{aligned}$$

where (a) holds by (A-7) and (b) follows from $|\mu(\boldsymbol{x}) - \mu(C_B)| \le L_\mu \|\boldsymbol{x} - C_B\|_\infty \le \frac{1}{2}L_\mu h$. Thus, the second condition of Assumption 6' is satisfied by choosing $\gamma = 3d_x^* L_\mu h/(2\mu_m^2)$, and $\gamma < 1$ if $h < 2\mu_m^2/(3d_x^* L_\mu)$. □

A.3 Proofs for Localized LASSO

Proof of Lemma 2: 1. Without loss of generality, we set the geometric centre in bin B as zero. Moreover, we omit the argument y when writing $f(\boldsymbol{x}, y)$ as y is fixed in the proof. In other words, we consider $\boldsymbol{x} \in B = [-\frac{h}{2}, \frac{h}{2}]^{d_x}$ and we have

$$\int_{\boldsymbol{x}\in B} x_i\, d\boldsymbol{x} = 0 \quad \forall i \in \{1, 2, \dots, d_x\}. \tag{A-8}$$

For $i \in J$ and $C > 0$, if $\frac{\partial f(\boldsymbol{x})}{\partial x_i} \geq C$, then

$$
\begin{aligned}
f(x_1, \ldots, x_{d_x}) &= \int_{-h/2}^{x_i} \frac{\partial f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_{d_x})}{\partial z}\, dz + f(x_1, \ldots, x_{i-1}, -\frac{h}{2}, x_{i+1}, \ldots, x_{d_x}) \\
&\geq C \cdot (x_i + \frac{h}{2}) + f(x_1, \ldots, x_{i-1}, -\frac{h}{2}, x_{i+1}, \ldots, x_{d_x}), \qquad \text{(A-9)}
\end{aligned}
$$

where the first equality holds by the differentiability of f (Assumption 2). By the definition of θ_i, we have

$$
\begin{aligned}
\theta_i &= \frac{\int_{\boldsymbol{x} \in B} [f(x_1, \ldots, x_{d_x}) - \theta_0] x_i \, d\boldsymbol{x}}{\int_{\boldsymbol{x} \in B} x_i^2 \, d\boldsymbol{x}} \\
&\overset{(a)}{\geq} \frac{\int_{\boldsymbol{x} \in B} (C x_i^2 + hC x_i/2 + f(x_1, \ldots, x_{i-1}, -h/2, x_{i+1}, \ldots, x_{d_x}) x_i - \theta_0 x_i) \; d\boldsymbol{x}}{\int_{\boldsymbol{x} \in B} x_i^2 \; d\boldsymbol{x}} \\
&\overset{(b)}{=} \frac{\int_{\boldsymbol{x} \in B} C x_i^2 \; d\boldsymbol{x}}{\int_{\boldsymbol{x} \in B} x_i^2 \; d\boldsymbol{x}} \\
&= C.
\end{aligned}
$$

where (a) follows by (A-9) and (b) holds by (A-8). Following the previous argument, we have $\theta_i \leq -C$ if $\frac{\partial f(\boldsymbol{x})}{\partial x_i} \leq -C$.

If $i \notin J$, according to Assumption 3, we have

$$
\begin{aligned}
\theta_i &= \frac{\int_{\boldsymbol{x} \in B} [f(x_1, \ldots, x_{d_x}) - \theta_0] x_i \, d\boldsymbol{x}}{\int_{\boldsymbol{x} \in B} x_i^2 \, d\boldsymbol{x}} \\
&= \frac{\int_{[-h/2, h/2]^{d_x^*}} (g(x_1, \ldots, x_{d_x^*}) - \theta_0) \, dx_1 \ldots dx_{d_x^*} \int_{[-h/2, h/2]^{d_x - d_x^*}} x_i \, dx_{d_x^*+1} \ldots dx_{d_x}}{\int_{\boldsymbol{x} \in B} x_i^2 \, dx_1 \ldots dx_{d_x}} \\
&= 0.
\end{aligned}
$$

In the second equation, we put relevant variables in the first d_x^* dimensions and redundant variables in the remaining $d_x - d_x^*$ dimensions. The last equality follows from $\int_{[-h/2, h/2]^{d_x - d_x^*}} x_i \, dx_{d_x^*+1} \ldots dx_{d_x} = 0$ for $i \notin J$.

2. Let $P(\boldsymbol{x}) = \theta_0 + \sum_{i=1}^{d_x} \theta_i x_i$ and $Q(\boldsymbol{x}) = f(\boldsymbol{x}) - P(\boldsymbol{x})$. We will prove $|Q(\boldsymbol{x})| \leq (2\sqrt{3} + 1) L d_x h^2$ in the following three steps.

First, we claim that there must be a point $\boldsymbol{x_0} \in B$ such that $f(\boldsymbol{x_0}) = P(\boldsymbol{x_0})$. From the definition of θ_0 (6), we know $\int_{\boldsymbol{x} \in B} Q(\boldsymbol{x}) \, d\boldsymbol{x} = \int_{\boldsymbol{x} \in B} f(\boldsymbol{x}) \, d\boldsymbol{x} - \theta_0 = 0$. Also, because $Q(\boldsymbol{x})$ is a continuous function from B to $\mathbb{R}$, there must exist a point $\boldsymbol{x}_0 \in B$ such that $Q(\boldsymbol{x}_0) = 0$.

Second, we approximate $Q(\boldsymbol{x})$ by the Taylor series expansion at point $\boldsymbol{x_0}$,

$$
|Q(\boldsymbol{x}) - Q(\boldsymbol{x_0}) - \nabla Q(\boldsymbol{x_0})^T (\boldsymbol{x} - \boldsymbol{x_0})| \leq \frac{1}{2} \|\nabla^2 Q(\boldsymbol{x_0})\|_\infty \|\boldsymbol{x} - \boldsymbol{x_0}\|_\infty^2.
$$

By $Q(\boldsymbol{x_0}) = 0, \nabla Q(\boldsymbol{x_0}) = \nabla f(\boldsymbol{x_0}) - \boldsymbol{\theta}$, $\nabla^2 Q(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$ and Assumption 5, we have

$$|f(\boldsymbol{x}) - P(\boldsymbol{x}) - (\nabla f(\boldsymbol{x_0}) - \boldsymbol{\theta})^T(\boldsymbol{x} - \boldsymbol{x_0})| \leq L\|\boldsymbol{x} - \boldsymbol{x_0}\|_\infty^2. \tag{A-10}$$

Third, we provide an upper bound for $\|\nabla f(\boldsymbol{x_0}) - \boldsymbol{\theta}\|_\infty$. Recalling the definition of θ_i (6), we have

$$\begin{aligned}
&\left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x_0}} - \theta_i \\
&= \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1} \int_{\boldsymbol{x}\in B} \left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} x_i^2 \, d\boldsymbol{x} - \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1} \int_{\boldsymbol{x}\in B} (f(\boldsymbol{x}) - \theta_0) x_i \, d\boldsymbol{x} \\
&= \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1} \int_{\boldsymbol{x}\in B} \left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} x_i^2 - (f(\boldsymbol{x}) - \theta_0) x_i \, d\boldsymbol{x} \\
&\overset{(a)}{=} \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1} \int_{\boldsymbol{x}\in B} \left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} x_i^2 - f(\boldsymbol{x}) x_i \, d\boldsymbol{x} \\
&\overset{(b)}{=} \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1} \int_{\boldsymbol{x}\in B} \left(\int_{-h/2}^{x_i} \left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x_0}} - \frac{\partial f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_{d_x})}{\partial z} \, dz\right) x_i \, d\boldsymbol{x},
\end{aligned} \tag{A-11}$$

where (a) follows from (A-8) and (b) follows by writing $f(\boldsymbol{x})$ as the integration of x_i's partial derivative and (A-8). Then, by the Cauthy-Schwarz inequality, we have (A-11)

$$\leq \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1/2} \left(\int_{\boldsymbol{x}\in B} \left(\int_{-h/2}^{x_i} \left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x_0}} - \frac{\partial f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_{d_x})}{\partial z} \, dz\right)^2 d\boldsymbol{x}\right)^{1/2}.$$

According to Assumption 5 and following the same argument for (A-3), we have $\|\nabla f(\boldsymbol{x_0}) - \nabla f(\boldsymbol{x})\|_\infty \leq 2Lh$. Then, thus

$$\left| \left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} - \frac{\partial f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_{d_x})}{\partial z} \right| = |(\nabla f(\boldsymbol{x_0}))_i - (\nabla f(\boldsymbol{x}))_i| \leq 2Lh. \tag{A-12}$$

By (A-11), (A-12) and $x_i \leq h/2$, we have

$$\begin{aligned}
\left.\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x}_0} - \theta_i &\leq \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1/2} \left(\int_{\boldsymbol{x}\in B} \left(\int_{-h/2}^{x_i} 2Lh \, dz\right)^2 d\boldsymbol{x}\right)^{1/2} \\
&\leq \left(\int_{\boldsymbol{x}\in B} x_i^2 \, d\boldsymbol{x}\right)^{-1/2} \left(\int_{\boldsymbol{x}\in B} 4L^2h^4 \, d\boldsymbol{x}\right)^{1/2} \\
&= \left(\frac{48L^2h^{d_x+4}}{h^{d_x+2}}\right)^{1/2} = 4\sqrt{3}Lh
\end{aligned}$$

Taking maximum of all $i \in \{1, 2, \ldots, d_x\}$, we have

$$\|\nabla f(\boldsymbol{x_0}) - \boldsymbol{\theta}\|_\infty \leq 4\sqrt{3}Lh.$$

Plugging it into (A-10), we have

$$\begin{aligned}
|f(\boldsymbol{x}) - P(\boldsymbol{x})| &\leq L\|\boldsymbol{x} - \boldsymbol{x_0}\|_\infty^2 + |(\nabla f(\boldsymbol{x_0}) - \boldsymbol{\theta})^T(\boldsymbol{x} - \boldsymbol{x_0})| \\
&\leq L\|\boldsymbol{x} - \boldsymbol{x_0}\|_\infty^2 + \|\nabla f(\boldsymbol{x_0}) - \boldsymbol{\theta}\|_\infty \|\boldsymbol{x} - \boldsymbol{x_0}\|_1 \\
&\leq Lh^2 + 4\sqrt{3}Ld_x h^2 \\
&\leq (4\sqrt{3} + 1)Ld_x h^2.
\end{aligned}$$

Hence, we complete the proof of Lemma 2. □

Proof of Proposition 2: By Lemma 2, we know $\boldsymbol{\theta}^*$ in (6) maintains the sparsity structure of f, i.e.,

$$J = \{i \in \{1, 2, \ldots, d_x\} : \theta_i^* \neq 0\}.$$

As in (8), the variable set selected by LASSO is

$$\hat{J}_j = \{i \in \{1, 2, \ldots, d_x\} : \hat{\theta}_i \neq 0\},$$

where $\hat{\boldsymbol{\theta}}$ is the LASSO estimator. If $\hat{\theta}_i$ has the same sign with θ_i^* for all $i \in \{1, 2, \ldots, d_x\}$, then we have $J = \hat{J}_j$. That's to say, the event $\{J = \hat{J}_j\}$ contains the event $\{\overrightarrow{\mathrm{sign}}(\hat{\boldsymbol{\theta}}) = \overrightarrow{\mathrm{sign}}(\boldsymbol{\theta}^*)\}$. Mathematically,

$$\mathbb{P}\left(J = \hat{J}_j\right) \geq \mathbb{P}\left(\overrightarrow{\mathrm{sign}}(\hat{\boldsymbol{\theta}}) = \overrightarrow{\mathrm{sign}}(\boldsymbol{\theta}^*)\right) \tag{A-13}$$

By Lemma 3, we know the event $\{\overrightarrow{\mathrm{sign}}(\hat{\boldsymbol{\theta}}) = \overrightarrow{\mathrm{sign}}(\boldsymbol{\theta}^*)\}$ contains the event $\{\cap_{i=1}^4 \Omega_i\}$. Thus, we have

$$\mathbb{P}\left(\overrightarrow{\mathrm{sign}}(\hat{\boldsymbol{\theta}}) = \overrightarrow{\mathrm{sign}}(\boldsymbol{\theta}^*)\right) \geq \mathbb{P}\left(\cap_{i=1}^4 \Omega_i\right). \tag{A-14}$$

By Lemma 4, we have

$$\mathbb{P}\left(\cap_{i=1}^4 \Omega_i\right) \geq 1 - b_0 \exp(b_1 n_j h^4). \tag{A-15}$$

Therefore, by (A-13), (A-14) and (A-15), we have

$$\mathbb{P}\left(J = \hat{J}_j\right) \geq 1 - b_0 \exp(b_1 n_j h^4).$$

Hence, we complete the proof of Proposition 2. □

Proof of Lemma 3: Using the notation in Section 5.1, the LASSO estimator (8) can be formulated as

$$\hat{\Theta} = \underset{\boldsymbol{\theta} \in \mathbb{R}_{d_x+1}}{\arg\min} \|\boldsymbol{Z} - A\boldsymbol{\theta}\|_2^2 + 2\lambda\|\boldsymbol{\theta}\|_1. \tag{A-16}$$

Note that $\hat{\Theta}$ can be a set. By (13), we know that $\boldsymbol{\theta} \in \hat{\Theta}$ if and only if it satisfies

$$\begin{cases} A_{(1)}^T (\boldsymbol{Z} - A\boldsymbol{\theta}) = \lambda \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}_{(1)}\right) \\ \left| A_{(2)}^T (\boldsymbol{Z} - A\boldsymbol{\theta}) \right| \preceq \lambda \boldsymbol{e} \end{cases} \tag{A-17}$$

where the notation $|\cdot|$ takes the absolute value of every entry, $\preceq$ conducts entry-wise comparison and $\boldsymbol{e}$ denotes the unit vector in $\mathbb{R}^{d_x - d_x^*}$.

We will first prove that on the event Ω_1, the LASSO estimator $\hat{\Theta}$ is unique, thus denoted as $\hat{\boldsymbol{\theta}}$. Let $\phi(\boldsymbol{\theta}) := \|\boldsymbol{Z} - A\boldsymbol{\theta}\|_2^2 + 2\lambda \|\boldsymbol{\theta}\|_1$ be the objective function in (A-16). Taking the second-order derivative, we have $\phi''(\boldsymbol{\theta}) = 2A^T A = 2\hat{\Psi}$. Under event Ω_1, $\hat{\Psi}$ is positive definite, implying that $\phi(\boldsymbol{\theta})$ is strongly convex with respect to $\boldsymbol{\theta}$. Therefore, the solution to (A-16) exists and is unique.

Next, we will prove on the event $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$, there exists $\boldsymbol{\theta} \in \mathbb{R}^{d_x+1}$ satisfying $\overrightarrow{\text{sign}} \left(\boldsymbol{\theta}\right) = \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}^*\right)$ and $\boldsymbol{\theta} \in \hat{\Theta}$. Thus, by the uniqueness of the LASSO estimator, we have $\overrightarrow{\text{sign}} \left(\hat{\boldsymbol{\theta}}\right) = \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}^*\right)$. The proof mainly follows the line of Proposition 1 in Zhao and Yu [2006]. But the notations and details are somewhat different. So we write down the whole proof.

Let $\boldsymbol{\theta}_{(1)} \in \mathbb{R}^{d_x^*+1}$ and $\boldsymbol{\theta}_{(2)} \in \mathbb{R}^{d_x - d_x^*}$ such that

$$\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}^*_{(1)} + \hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} - \lambda \hat{\Psi}_{11}^{-1} \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}^*_{(1)}\right), \quad \boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}^*_{(2)} = \mathbf{0} \tag{A-18}$$

Then, on the event Ω_3, we have

$$\left|\boldsymbol{\theta}_{(1)} - \boldsymbol{\theta}^*_{(1)}\right| = \left|\hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} - \lambda \hat{\Psi}_{11}^{-1} \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}^*_{(1)}\right)\right| \preceq \left|\boldsymbol{\theta}^*_{(1)}\right|$$

The above inequality implies that $\overrightarrow{\text{sign}} \left(\boldsymbol{\theta}_{(1)}\right) = \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}^*_{(1)}\right)$. Moreover, multiplying both sides of (A-18) by $\hat{\Psi}_{11}$, we have

$$\hat{\Psi}_{11} \left(\boldsymbol{\theta}^*_{(1)} - \boldsymbol{\theta}_{(1)}\right) + A_{(1)}^T \boldsymbol{\rho} = \lambda \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}^*_{(1)}\right) = \lambda \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}_{(1)}\right). \tag{A-19}$$

Similarly, multiplying both sides of (A-18) by $\hat{\Psi}_{21}$ yields

$$\begin{aligned} \left|\hat{\Psi}_{21} \left(\boldsymbol{\theta}^*_{(1)} - \boldsymbol{\theta}_{(1)}\right) + A_{(2)}^T \boldsymbol{\rho}\right| &= \left|\lambda \hat{\Psi}_{21} \hat{\Psi}_{11}^{-1} \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}_{(1)}\right) - \hat{\Psi}_{21} \hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} + A_{(2)}^T \boldsymbol{\rho}\right| \\ &\leq \left|\lambda \hat{\Psi}_{21} \hat{\Psi}_{11}^{-1} \overrightarrow{\text{sign}} \left(\boldsymbol{\theta}_{(1)}\right)\right| + \left|\hat{\Psi}_{21} \hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} - A_{(2)}^T \boldsymbol{\rho}\right| \end{aligned} \tag{A-20}$$

On the event Ω_4, the second term in (A-20), $\left|\hat{\Psi}_{21} \hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} - A_{(2)}^T \boldsymbol{\rho}\right| \preceq \frac{1}{2}(1-\gamma)\lambda \boldsymbol{e}$. On event

$\Omega_1 \cap \Omega_2$, we show an upper bound for the first term in (A-20)

$$\begin{aligned}
\left|\left(\hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_{(1)}\right)\right)_j\right| &= \left|\sum_{k=0}^{d_x^*}(\hat{\Psi}_{21})_{jk}\left(\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}(\boldsymbol{\theta}_{(1)})\right)_k\right| \\
&\overset{(a)}{\leq} \left(\sum_{k=0}^{d_x^*}(\hat{\Psi}_{21})_{jk}^2\right)^{1/2} \left\|\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}(\boldsymbol{\theta}_{(1)})\right\|_2 \\
&\overset{(b)}{\leq} \sqrt{d_x^*}(1+\delta)\gamma\underline{\lambda}/d_x^*\|\hat{\Psi}_{11}^{-1}\|_2\|\overrightarrow{\operatorname{sign}}(\boldsymbol{\theta}_{(1)})\|_2 \\
&\overset{(c)}{\leq} \sqrt{d_x^*}(1+\delta)\gamma\underline{\lambda}/d_x^* \cdot \frac{\sqrt{d_x^*}}{(1-\alpha)\underline{\lambda}} \\
&= \frac{(1+\delta)\gamma}{1-\alpha} = \frac{1}{2}(1+\gamma),
\end{aligned}$$

where (a) follows from the Cauthy-Schwarz inequality, (b) follows from the definition of Ω_2 as well as the matrix inequality $\|AB\|_2 \leq \|A\|_2\|B\|_2$ and (c) is due to $\|A\|_2 \leq \lambda_{\max}(A)$ for square matrices. Combining the two terms in (A-20), we have

$$\left|\hat{\Psi}_{21}\left(\boldsymbol{\theta}_{(1)}^* - \boldsymbol{\theta}_{(1)}\right) + A_{(2)}^T\boldsymbol{\rho}\right| \preceq \frac{1}{2}(1-\gamma)\lambda\boldsymbol{e} + \frac{1}{2}(1+\gamma)\lambda\boldsymbol{e} = \lambda\boldsymbol{e}. \tag{A-21}$$

By (A-19), (A-21) and $\boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}_{(2)}^* = \mathbf{0}$, we have

$$\begin{cases} \hat{\Psi}_{11}\left(\boldsymbol{\theta}_{(1)}^* - \boldsymbol{\theta}_{(1)}\right) + \hat{\Psi}_{12}\left(\boldsymbol{\theta}_{(2)}^* - \boldsymbol{\theta}_{(2)}\right) + A_{(1)}^T\boldsymbol{\rho} = \lambda\overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_{(1)}\right) \\ \left|\hat{\Psi}_{21}\left(\boldsymbol{\theta}_{(1)}^* - \boldsymbol{\theta}_{(1)}\right) + \hat{\Psi}_{22}\left(\boldsymbol{\theta}_{(2)}^* - \boldsymbol{\theta}_{(2)}\right) + A_{(2)}^T\boldsymbol{\rho}\right| \preceq \lambda\boldsymbol{e} \end{cases} \tag{A-22}$$

Notice that (A-17) is equivalent to (A-22) by $\boldsymbol{Z} = A\boldsymbol{\theta}^* + \boldsymbol{\rho}$. Therefore, we have found $\boldsymbol{\theta}$ having the same signs with $\boldsymbol{\theta}^*$ and satisfying (A-17). Further by the uniqueness of the LASSO estimator, we have $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\overrightarrow{\operatorname{sign}}\left(\hat{\boldsymbol{\theta}}\right) = \overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}^*\right)$. Hence, we have proved that on the event $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$, the LASSO estimator has the same signs with the linear approximation. $\square$

Proof of Lemma 4: In this proof, we will show that $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4$ occurs with a high probability. First, we adopt matrix concentration inequalities to give a lower bound for $\mathbb{P}(\Omega_1)$. For any constant $\alpha \in (0,1)$, we have

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\min}(\hat{\Psi}) \leq (1-\alpha)\underline{\lambda}\right) &\leq \mathbb{P}\left(\lambda_{\min}(\hat{\Psi}) \leq (1-\alpha)\lambda_{\min}(\Psi)\right) \\
&\leq (d_x+1)\left(\frac{e^{-\alpha}}{(1-\alpha)^{(1-\alpha)}}\right)^{n\lambda_{\min}(\Psi)/(1+d_x/4)} \\
&\leq (d_x+1)\left(\frac{e^{-\alpha}}{(1-\alpha)^{(1-\alpha)}}\right)^{n\underline{\lambda}/(1+d_x/4)}.
\end{aligned} \tag{A-23}$$

The first and last inequalities follow from $\underline{\lambda} \leq \lambda_{\min}(\hat{\Psi})$. The second inequality follows from Theorem 5.1.1 in Tropp et al. [2015]:

Theorem A1 (Theorem 5.1.1 in Tropp et al. [2015])**.** Consider a finite sequence of i.i.d. random Hermitian matrices $M_t \in \mathbb{R}^{(d_x+1)\times(d_x+1)}$. Assume that

$$0 \le \lambda_{\min}(M_t M_t^T) \quad \textit{and} \quad \lambda_{\max}(M_t M_t^T) \le \lambda_M, \quad \forall\, t \in \{1, 2, \ldots, n\},$$

and

$$\Psi = \mathbb{E}[M_t M_t^T],\ \hat{\Psi} = \frac{1}{n}\sum_{t=1}^{n} M_t M_t^T.$$

Then, we have

$$\mathbb{P}\left(\lambda_{\min}(\hat{\Psi}) \le (1-\alpha)\lambda_{\min}(\Psi)\right) \le (d_x+1)\left(\frac{e^{-\alpha}}{(1-\alpha)^{(1-\alpha)}}\right)^{n\lambda_{\min}(\Psi)/\lambda_M} \qquad \forall \alpha \in [0,1),$$
$$\mathbb{P}\left(\lambda_{\max}(\hat{\Psi}) \ge (1+\alpha)\lambda_{\max}(\Psi)\right) \le (d_x+1)\left(\frac{e^{\alpha}}{(1+\alpha)^{(1+\alpha)}}\right)^{n\lambda_{\max}(\Psi)/\lambda_M} \qquad \forall \alpha \ge 0.$$

Here, to apply Theorem A1, we let $M_t = \bar{\boldsymbol{U}}_t$ and show an upper bound for $\lambda_{\max}(\bar{\boldsymbol{U}}_t \bar{\boldsymbol{U}}_t^T)$. Recalling that $\bar{\boldsymbol{U}}_t$ is the normalized covariates, the absolute of each entry is less than $1/2$ except for the first entry, which is 1. So the ℓ_2-norm $\|\bar{\boldsymbol{U}}_t\|_2^2$ is less than $(1 + d_x/4)$. By the Cauchy-Schwartz inequality, we have

$$\boldsymbol{u}^T \bar{\boldsymbol{U}}_t \bar{\boldsymbol{U}}_t^T \boldsymbol{u} = \left(\boldsymbol{u}^T \bar{\boldsymbol{U}}_t\right)^2 \le \|\boldsymbol{u}\|_2^2 \|\bar{\boldsymbol{U}}_t\|_2^2 \le (1 + d_x/4)\|\boldsymbol{u}\|_2^2$$

for any $\boldsymbol{u} \in \mathcal{R}^{d_x+1}$. Further, considering the characterization of eigenvalues, for a symmetric matrix A, its largest eigenvalue satisfies

$$\lambda_{\max}(A) = \sup_u \frac{u^T A u}{\|u\|_2^2}. \tag{A-24}$$

As a result,

$$\lambda_{\max}(\bar{\boldsymbol{U}}_t \bar{\boldsymbol{U}}_t^T) = \sup_u \frac{u^T \bar{\boldsymbol{U}}_t \bar{\boldsymbol{U}}_t^T u}{\|u\|_2^2} \le (1 + d_x/4).$$

So we set the constant $\lambda_M = 1 + d_x/4$ in the above theorem. In this way, we obtain the second inequality of (A-23). Moreover, we have

$$0 < \frac{e^{-\alpha}}{(1-\alpha)^{(1-\alpha)}} \le e^{-\alpha^2/2} < 1, \quad \text{for } \alpha \in (0,1). \tag{A-25}$$

Similarly, using Theorem A1, we have an probability bound for $\lambda_{\max}(\hat{\Psi})$:

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}(\hat{\Psi}) \ge (1+\alpha)\overline{\lambda}\right) &\le \mathbb{P}\left(\lambda_{\max}(\hat{\Psi}) \ge (1+\alpha)\lambda_{\max}(\Psi)\right) \\
&\le (d_x+1)\left(\frac{e^{\alpha}}{(1+\alpha)^{(1+\alpha)}}\right)^{n\lambda_{\max}(\Psi)/(1+d_x/4)} \\
&\le (d_x+1)\left(\frac{e^{\alpha}}{(1+\alpha)^{(1+\alpha)}}\right)^{n\underline{\lambda}/(1+d_x/4)},
\end{aligned} \tag{A-26}$$

and

$$0 < \frac{e^{\alpha}}{(1+\alpha)^{(1+\alpha)}} < 1, \quad \text{for } \alpha \in (0,1). \tag{A-27}$$

Recall the definition of Ω_1, choosing the constant $\alpha = \frac{1-\gamma}{2(1+\gamma)}$ and by (A-23), (A-26), we have

$$\begin{aligned}
\mathbb{P}(\Omega_1) &= \mathbb{P}\left(\left\{(1-\alpha)\underline{\lambda} \geq \lambda_{\min}(\hat{\Psi})\right\} \cup \left\{\lambda_{\max}(\hat{\Psi}) \leq (1+\alpha)\overline{\lambda}\right\}\right) \\
&\geq 1 - \mathbb{P}\left(\lambda_{\min}(\hat{\Psi}) \leq (1-\alpha)\underline{\lambda}\right) - \mathbb{P}\left(\lambda_{\max}(\hat{\Psi}) \geq (1+\alpha)\overline{\lambda}\right) \\
&\geq 1 - 2(d_x+1)\exp(-c_1 n),
\end{aligned} \tag{A-28}$$

where

$$\begin{aligned}
c_1(\underline{\lambda}, \gamma, d_x) &= \frac{\underline{\lambda}}{(1+d_x/4)} \min\left\{-\log\left(\frac{e^{-\alpha}}{(1-\alpha)^{(1-\alpha)}}\right), -\log\left(\frac{e^{\alpha}}{(1+\alpha)^{(1+\alpha)}}\right)\right\} \\
&= \frac{\underline{\lambda}}{(1+d_x/4)} \min\left\{\alpha + (1-\alpha)\log(1-\alpha), -\alpha + (1+\alpha)\log(1+\alpha)\right\} \\
&= \frac{\underline{\lambda}}{2(1+\gamma)(1+d_x/4)} \min\left\{1-\gamma + (3\gamma+1)\log\left(\frac{3\gamma+1}{2+2\gamma}\right), \gamma - 1 + (3+\gamma)\log\left(\frac{3+\gamma}{2+2\gamma}\right)\right\}.
\end{aligned}$$

By (A-25) and (A-27), we have $c_1 > 0$ as $\gamma \in [0,1)$.

Next, we show the event Ω_2 happens with high probability. Recalling $(\hat{\Psi}_{21})_{ik} = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{U}_j)_i(\boldsymbol{U}_j)_k$, Hoeffding's inequality[7] implies that

$$\mathbb{P}\left(\left|(\hat{\Psi}_{21})_{ik} - (\Psi_{21})_{ik}\right| \geq \delta\gamma\underline{\lambda}/d_x^*\right) \leq 2\exp(-2n\delta^2\gamma^2\underline{\lambda}^2/(d_x^*)^2).$$

According to Assumption 6', $|(\Psi_{21})_{ik}| \leq \gamma\underline{\lambda}/d_x^*$. Thus, we have

$$\mathbb{P}\left(\left|(\hat{\Psi}_{21})_{ik}\right| \geq (1+\delta)\gamma\underline{\lambda}/d_x^*\right) \leq 2\exp(-2n\delta^2\gamma^2\underline{\lambda}^2/(d_x^*)^2).$$

Taking the union bound over $i \in J$ and $k \in J^c$

$$\mathbb{P}(\Omega_2) \geq 1 - 2d_x^*(d_x - d_x^*)\exp(-c_3 n), \tag{A-29}$$

where

$$c_3(\gamma, \underline{\lambda}, d_x^*) = 2\delta^2\gamma^2\underline{\lambda}^2/(d_x^*)^2 = (1-\gamma)^2\underline{\lambda}^2/(8(d_x^*)^2).$$

Next, we show an upper bound for the approximation error of the linear projection. We define the vector

$$\boldsymbol{\Delta} := \frac{1}{\sqrt{n}}\left(\Delta_1, \ldots, \Delta_n\right).$$

[7] See Theorem 2.2.6 on page 18 of Vershynin [2018].

Then by Lemma 2, we have

$$\|\boldsymbol{\Delta}\|_2^2 = \frac{1}{n}\sum_{t=1}^{n}\left(f(\boldsymbol{X}_t) - \bar{\boldsymbol{U}}_t^T\boldsymbol{\theta}^*\right)^2 \le 64L^2 d_x^2 h^4. \tag{A-30}$$

So far, we have provided a lower bound for the probability of the event $\Omega_1 \cap \Omega_2$. It then suffices to bound the probabilities $\mathbb{P}(\Omega_3^c \cap \Omega_1 \cap \Omega_2)$ and $\mathbb{P}(\Omega_4^c \cap \Omega_1 \cap \Omega_2)$. Recall the definition of event Ω_4, the term $\hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}A_{(1)}^T\boldsymbol{\rho} - A_{(2)}^T\boldsymbol{\rho} = \left(\hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}A_{(1)}^T - A_{(2)}^T\right)\left(\boldsymbol{\Delta} + \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}\right)$ is a linear combination of approximation errors $\boldsymbol{\Delta}$ and sub-Gaussian noises $\boldsymbol{\epsilon}$. Denote

$$G \coloneqq \hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}A_{(1)}^T - A_{(2)}^T = (g_{jk})_{d_x^*+1\le j\le d_x; 1\le k\le n}, \tag{A-31}$$

then we have

$$\Omega_4 = \left\{\left|G\left(\boldsymbol{\Delta} + \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}\right)\right| \preceq \frac{1}{2}(1-\gamma)\lambda\boldsymbol{e}\right\}.$$

We want to bound the probability of Ω_4^c

$$\begin{aligned}
\Omega_4^c &= \bigcup_{j=d_x^*+1}^{d_x}\left\{\left(\left|G\left(\boldsymbol{\Delta} + \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}\right)\right|\right)_j \ge \frac{1}{2}(1-\gamma)\lambda\right\} \\
&\subset \bigcup_{j=d_x^*+1}^{d_x}\left\{\left(\left|\frac{1}{\sqrt{n}}G\boldsymbol{\epsilon}\right|\right)_j \ge \frac{1}{2}(1-\gamma)\lambda - (|G\boldsymbol{\Delta}|)_j\right\}.
\end{aligned} \tag{A-32}$$

Note that by (12) and (A-31), we have

$$\begin{aligned}
GG^T &= \left(\hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}A_{(1)}^T - A_{(2)}^T\right)\left(A_{(1)}\hat{\Psi}_{11}^{-1}\hat{\Psi}_{12} - A_{(2)}\right) && (\text{by } \hat{\Psi}_{12} = \hat{\Psi}_{21}^T) \\
&= \hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}\hat{\Psi}_{12} - \hat{\Psi}_{21}\hat{\Psi}_{11}^{-1}A_{(1)}^T A_{(2)} - A_{(2)}^T A_{(1)}\hat{\Psi}_{11}^{-1}\hat{\Psi}_{12} + A_{(2)}^T A_{(2)} && (\text{by } \hat{\Psi}_{11}^{-1} = A_{(1)}^T A_{(1)}) \\
&= -A_{(2)}^T A_{(1)}\hat{\Psi}_{11}A_{(1)}^T A_{(2)} + A_{(2)}^T A_{(2)} && (\text{by } \hat{\Psi}_{12} = A_{(1)}^T A_{(2)}) \\
&= A_{(2)}^T\left(I - A_{(1)}\hat{\Psi}_{11}^{-1}A_{(1)}^T\right)A_{(2)} \\
&= A_{(2)}^T B A_{(2)}
\end{aligned} \tag{A-33}$$

where $B \coloneqq I - A_{(1)}\hat{\Psi}_{11}^{-1}A_{(1)}^T$. Notice that B is symmetric and

$$B^2 = I - 2A_{(1)}\hat{\Psi}_{11}^{-1}A_{(1)}^T + A_{(1)}\hat{\Psi}_{11}^{-1}A_{(1)}^T A_{(1)}\hat{\Psi}_{11}^{-1}A_{(1)}^T = \left(I - A_{(1)}\hat{\Psi}_{11}^{-1}A_{(1)}^T\right) = B.$$

So B is an idempotent matrix whose eigenvalues are either 0 or 1 [Horn and Johnson, 1990]. Since

GG^T is a symmetric matrix, using (A-24), we derive an upper bound for $\lambda_{\max}(GG^T)$,

$$\begin{aligned}
u^T GG^T u &= (A_{(2)}u)^T B(A_{(2)}u) && \text{(by (A-33))} \\
&\le \lambda_{\max}(B)\|A_{(2)}u\|_2^2 && \text{(by (A-24))} \\
&= \lambda_{\max}(B)(u^T\hat{\Psi}_{22}u)^2 && \text{(by } \hat{\Psi}_{22} = A_{(2)}^T A_{(2)}\text{)} \\
&\le \lambda_{\max}(B)\lambda_{\max}(\hat{\Psi}_{22})\|u\|_2^2 && \text{(by (A-24))} \\
&\le \lambda_{\max}(\hat{\Psi}_{22})\|u\|_2^2 && \text{(by } \lambda_{\max}(B) \le 1\text{)} \qquad \text{(A-34)}
\end{aligned}$$

Moreover, on the event Ω_1, the eigenvalue of $\hat{\Psi}_{22} = A_{(2)}^T A_{(2)}$ are smaller than $(1+\alpha)\overline{\lambda}$. Therefore, (A-34) implies the eigenvalues of GG^T are less than $(1+\alpha)\overline{\lambda}$. This implies that

$$\sum_{k=1}^{n} g_{jk}^2 = (GG^T)_{jj} = \boldsymbol{e}_j^T GG^T \boldsymbol{e}_j \le \lambda_{\max}(GG^T)\|\boldsymbol{e}_j\|_2^2 \le (1+\alpha)\overline{\lambda}, \qquad \text{(A-35)}$$

for all $j \in \{d_x^* + 1, d_x^* + 2, \ldots, d_x\}$, where $\boldsymbol{e}_j$ is the j-th standard basis. Thus, we have

$$(|G\boldsymbol{\Delta}|)_j = \left|\sum_{k=1}^{n} g_{jk}\Delta_k\right| \le \left(\sum_{k=1}^{n} g_{jk}^2\right)^{1/2} \|\boldsymbol{\Delta}\|_2 \le \sqrt{(1+\alpha)\overline{\lambda}}\|\boldsymbol{\Delta}\|_2. \qquad \text{(A-36)}$$

By (A-30), we have $\|\boldsymbol{\Delta}\|_2 \le 8Ld_x h^2$, and so

$$\max_{\{j=d_x^*+1,\ldots,d_x\}} (|G\boldsymbol{\Delta}|)_j \le 8\sqrt{(1+\alpha)\overline{\lambda}}Ld_x h^2. \qquad \text{(A-37)}$$

Recalling that we choose

$$\lambda = 32 \cdot \sqrt{\frac{(3+\gamma)\overline{\lambda}}{(1+\gamma)(1-\gamma)^2}}Ld_x h^2 \qquad \text{(A-38)}$$

in Proposition 2, by (A-37), we have

$$\frac{1}{2}(1-\gamma)\lambda - (|G\boldsymbol{\Delta}|)_j \ge \frac{1}{4}(1-\gamma)\lambda.$$

Thus, plugging it into (A-32), we have

$$\begin{aligned}
\Omega_4^c \cap \Omega_1 &\subset \left\{\bigcup_{j=d_x^*+1}^{d_x} \left\{\left(\left|\frac{1}{\sqrt{n}}G\boldsymbol{\epsilon}\right|\right)_j \ge \frac{1}{4}(1-\gamma)\lambda\right\}\right\}\bigcap \Omega_1 \\
&= \left\{\bigcup_{j=d_x^*+1}^{d_x} D_j\right\}\bigcap \Omega_1. \qquad \text{(A-39)}
\end{aligned}$$

where $D_j := \left(\left|\frac{1}{\sqrt{n}}G\boldsymbol{\epsilon}\right|\right)_j > \frac{1}{4}(1-\gamma)\lambda$. Define the realization of normalized covariates as $\mathcal{U}_n := \{\bar{U}_1, \bar{U}_2, \ldots, \bar{U}_n\}$. It provides the full information for the empirical covariance matrix $\hat{\Psi}$ and whether

Ω_1 happens. Note that the covariates and noise are independent, so

$$\left(\frac{1}{\sqrt{n}} G\boldsymbol{\epsilon}\right)_j = \sqrt{\frac{1}{n}} \sum_{k=1}^{n} g_{jk}\epsilon_k,$$

it's a mean-zero $\sqrt{\frac{1}{n}\sum_{k=1}^{n} g_{jk}^2}\sigma$ sub-Gaussian random variable conditional on $\mathcal{U}_n$. So we have

$$\begin{aligned}
\mathbb{P}(\Omega_4^c \cap \Omega_1) &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(\Omega_4^c \cap \Omega_1) \middle| \mathcal{U}_n\right]\right] && \text{(by the tower rule)} \\
&\le \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(\{\cup_j D_j\}\right) \middle| \mathcal{U}_n\right] \cdot \mathbb{I}(\Omega_1)\right] && \text{(by (A-39))} \\
&\le \mathbb{E}\left[\mathbb{E}\left[\sum_{j=d_x^*+1}^{d_x} \mathbb{I}\left(D_j\right) \middle| \mathcal{U}_n\right] \cdot \mathbb{I}(\Omega_1)\right] && \text{(by the union bound)} \\
&\le \sum_{j=d_x^*+1}^{d_x} \mathbb{E}\left[\mathbb{P}\left(\left(\left|\frac{1}{\sqrt{n}} G\boldsymbol{\epsilon}\right|\right)_j > \frac{1}{4}(1-\gamma)\lambda \;\middle|\; \mathcal{U}_n\right) \mathbb{I}(\Omega_1)\right] \\
&\le \sum_{j=d_x^*+1}^{d_x} \mathbb{E}\left[2\exp\left(-\frac{(1-\gamma)^2\lambda^2 n}{32\sum_{k=1}^{n} g_{ik}^2 \sigma^2}\right) \mathbb{I}(\Omega_1)\right] && \text{(sub-Gaussian)} \\
&\le \sum_{j=d_x^*+1}^{d_x} 2\exp\left(-\frac{(1-\gamma)^2\lambda^2 n}{32(1+\alpha)\overline{\lambda}\sigma^2}\right) \mathbb{P}(\Omega_1)
\end{aligned} \tag{A-40}$$

where the last inequality follows from (A-35) on the event Ω_1. Plugging in the value of λ (A-38), we have that (A-40) is upper bounded by $2(d_x - d_x^*)\exp\left(-c_5 n h^4\right)\mathbb{P}(\Omega_1)$, where $c_5 = 64L^2 d_x^2/\sigma^2$.

Similarly, we study event Ω_3. The term $\hat{\Psi}_{11}^{-1} A_{(1)}^T \boldsymbol{\rho} = \hat{\Psi}_{11}^{-1} A_{(1)}^T \left(\boldsymbol{\Delta} + \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}\right)$ is also a linear combination of approximation errors and sub-Gaussian noises. Denote

$$H := \hat{\Psi}_{11}^{-1} A_{(1)}^T = (h_{jk})_{0\le j\le d_x^*; 1\le k\le n}. \tag{A-41}$$

We have

$$\Omega_3^c = \bigcup_{j=0}^{d_x^*} \left\{ \left| \lambda\left(\hat{\Psi}_{11}^{-1} \overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_1^*\right)\right)_j - \left(H\left(\boldsymbol{\Delta} + \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}\right)\right)_j \right| > \left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j \right\} \tag{A-42}$$

$$= \bigcup_{j=0}^{d_x^*} \left\{ \left(\left|\frac{1}{\sqrt{n}} H\boldsymbol{\epsilon}\right|\right)_j > \left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j - \left(|H\boldsymbol{\Delta}|\right)_j - \lambda\left(\left|\hat{\Psi}_{11}^{-1} \overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_1^*\right)\right|\right)_j \right\}. \tag{A-43}$$

Recall that $\left(\boldsymbol{\theta}_{(1)}^*\right)_j$ is the coefficient of j-th relevant variable scaled by h. According to Lemma 2, its absolute value is greater than Ch.[8] Next we analyze the second term of the right-hand side of (A-43).

Note that by (A-41) and (12) we have $HH^T = \hat{\Psi}_{11}^{-1}$. On the event Ω_1, the eigenvalues of $\hat{\Psi}_{11}^{-1}$

[8] Without loss of generality, we assume $|\theta_0| \ge Ch$, since it doesn't matter whether $0 \in J$ or not.

are smaller than $1/\left((1-\alpha)\underline{\lambda}\right)$. Similar to (A-35), we have

$$\sum_{k=1}^{n} h_{jk}^2 \leq \lambda_{\max}(\hat{\Psi}_{11}^{-1}) \leq \frac{1}{(1-\alpha)\underline{\lambda}}. \tag{A-44}$$

Thus, for $j \in \{0, 1, \ldots, d_x^*\}$, we have

$$(|H\mathbf{\Delta}|)_j \leq \left(\sum_{k=1}^{n} h_{jk}^2\right)^{1/2} \|\mathbf{\Delta}\|_2 \leq \sqrt{\frac{1}{(1-\alpha)\underline{\lambda}}}\|\mathbf{\Delta}\|_2.$$

Since (A-30) $\|\mathbf{\Delta}\|_2^2 \leq 64L^2 d_x^2 h^4$, we have

$$\max_{j=1,\ldots,d_x^*} (|H\mathbf{\Delta}|)_j \leq \sqrt{\frac{64}{(1-\alpha)\underline{\lambda}}} L d_x h^2.$$

Moreover, we have

$$\left(\left|\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_{(1)}^*\right)\right|\right)_j \leq \left\|\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_{(1)}^*\right)\right\|_2 \leq \frac{1}{(1-\alpha)\underline{\lambda}}\|\overrightarrow{\operatorname{sign}}(\boldsymbol{\theta}_{(1)}^*)\|_2 \leq \frac{\sqrt{d_x^*}}{(1-\alpha)\underline{\lambda}}.$$

Since we choose λ as in (A-38), we have

$$\begin{aligned}
(|H\mathbf{\Delta}|)_j + \lambda\left(\left|\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_{(1)}^*\right)\right|\right)_j &\leq \left(4\sqrt{\frac{(3+\gamma)\overline{\lambda}d_x^*}{(1+\gamma)(1-\gamma)^2(1-\alpha)\underline{\lambda}}}+1\right) 8Ld_x h^2\sqrt{\frac{1}{(1-\alpha)\underline{\lambda}}} \\
&\leq \left(5\sqrt{\frac{2(3+\gamma)\overline{\lambda}d_x^*}{(1+3\gamma)(1-\gamma)^2\underline{\lambda}}}\right) 8Ld_x h^2\sqrt{\frac{2(1+\gamma)}{(1+3\gamma)\underline{\lambda}}} \\
&= 80Ld_x h^2 \cdot \frac{\sqrt{(3+\gamma)(1+\gamma)\overline{\lambda}d_x^*}}{(1+3\gamma)(1-\gamma)\underline{\lambda}}.
\end{aligned} \tag{A-45}$$

According to Lemma 2 and definition of $\boldsymbol{\theta}^*$, $\left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j \geq Ch$ for $j \in \{0, 1, \ldots, d_x^*\}$. So if h is sufficient small such that

$$h \leq \frac{C(1+3\gamma)(1-\gamma)\underline{\lambda}}{160Ld_x\sqrt{(3+\gamma)(1+\gamma)\overline{\lambda}d_x}},^{9} \tag{A-46}$$

then we have

$$\left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j \geq Ch \geq 160Ld_x h^2 \cdot \frac{\sqrt{(3+\gamma)(1+\gamma)\overline{\lambda}d_x^*}}{(1+3\gamma)(1-\gamma)\underline{\lambda}}. \tag{A-47}$$

Combing (A-45) and (A-47), the right-hand side of (A-43) is at least half of $\left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j$, i.e.,

$$\left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j - (|H\mathbf{\Delta}|)_j - \lambda\left(\left|\hat{\Psi}_{11}^{-1}\overrightarrow{\operatorname{sign}}\left(\boldsymbol{\theta}_{(1)}^*\right)\right|\right)_j \geq \frac{1}{2}\left(\left|\boldsymbol{\theta}_{(1)}^*\right|\right)_j > \frac{1}{2}Ch. \tag{A-48}$$

[9] Since d_x^* is unknown, we replace d_x^* by d_x for a more conservative condition for h.

Notice that all the parameters in right-hand side of (A-43) are known constants. So the validity of (A-48) is assured by choosing a small enough h. As h and λ satisfy (A-43) and (A-46), we have

$$\begin{aligned}\Omega_3^c \cap \Omega_1 \subset & \left\{\bigcup_{j=0}^{d_x^*}\left\{\left(\left|\frac{1}{\sqrt{n}}H\boldsymbol{\epsilon}\right|\right)_j > \frac{1}{2}Ch\right\}\right\}\bigcap \Omega_1. \\ = & \left\{\bigcup_{j=0}^{d_x^*} E_j\right\}\bigcap \Omega_1 \end{aligned} \tag{A-49}$$

where $E_j := \left\{\left(\left|\frac{1}{\sqrt{n}}H\boldsymbol{\epsilon}\right|\right)_j > \frac{1}{2}Ch\right\}$. Recalling the independence of covariates and noise, we have

$$\left(\left|\frac{1}{\sqrt{n}}H\boldsymbol{\epsilon}\right|\right)_j = \sqrt{\frac{1}{n}\sum_{k=1}^{n} h_{ik}\epsilon_k},$$

is a mean-zero $\sqrt{\sum_{k=1}^n h_{ik}^2/n\sigma}$ sub-Gaussian random variable conditional on $\mathcal{U}_n$. Similar to (A-40), we have

$$\begin{aligned}\mathbb{P}(\Omega_3^c \cap \Omega_1) &\leq \sum_{j=0}^{d_x^*}\mathbb{E}\left[\mathbb{P}\left(E_j \middle| \mathcal{U}_n\right)\mathbb{I}(\Omega_1)\right] \\ &\leq \sum_{j=0}^{d_x^*}\mathbb{E}\left[\mathbb{P}\left(\left(\left|\frac{1}{\sqrt{n}}H\boldsymbol{\epsilon}\right|\right)_j > \frac{1}{2}Ch \;\middle|\; \mathcal{U}_n\right)\mathbb{I}(\Omega_1)\right] \\ &\leq \sum_{j=0}^{d_x^*}\mathbb{E}\left[2\exp\left(-\frac{C^2h^2n}{8\sum_{k=1}^n h_{ik}^2\sigma^2}\right)\mathbb{I}(\Omega_1)\right] && \text{(sub-Gaussian)} \\ &\leq \sum_{j=0}^{d_x^*}\mathbb{E}\left[2\exp\left(-\frac{C^2h^2(1-\alpha)\underline{\lambda}n}{8\sigma^2}\right)\mathbb{I}(\Omega_1)\right] && \text{(by (A-44))}\end{aligned} \tag{A-50}$$

where the last inequality follows from (A-44) on the event Ω_1. Plugging the lower bound of Ch (A-47), we have that (A-50) is upper bounded by $2(d_x+1)\exp\left(-c_6nh^4\right)\mathbb{P}(\Omega_1)$, where $c_6 = \dfrac{3200(3+\gamma)\overline{\lambda}L^2d_x^3}{(1-\gamma)\sigma^2}$.

Until now, we have demonstrated the probability lower bounds for event Ω_1, Ω_2, $\Omega_3^c \cap \Omega_1$ and

$\Omega_4^c \cap \Omega_1$. We complete the proof by combining them together,

$$
\begin{aligned}
\mathbb{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4) &= \mathbb{P}(\Omega_1) - \mathbb{P}(\Omega_1 \cap (\Omega_2^c \cup \Omega_3^c \cup \Omega_4^c)) \\
&= \mathbb{P}(\Omega_1) - \mathbb{P}((\Omega_1 \cap \Omega_2^c) \cup (\Omega_1 \cap \Omega_3^c) \cup (\Omega_1 \cap \Omega_4^c)) \\
&\overset{(a)}{\geq} \mathbb{P}(\Omega_1) - \mathbb{P}(\Omega_1 \cap \Omega_2^c) - \mathbb{P}(\Omega_1 \cap \Omega_3^c) - \mathbb{P}(\Omega_1 \cap \Omega_4^c) \\
&\overset{(b)}{\geq} \left[1 - 2(d_x - d_x^*)\exp(-c_5 n h^4) - 2(d_x^* + 1)\exp(-c_6 n h^4)\right] \mathbb{P}(\Omega_1) - \mathbb{P}(\Omega_2^c) \\
&\overset{(c)}{\geq} \left[1 - 2(d_x+1)\exp\left(-(c_5 \wedge c_6) n h^4\right)\right] \cdot \left[1 - 2(d_x+1)\exp(-c_1 n)\right] \\
&\quad - 2 d_x^*(d_x - d_x^*)\exp(-c_3 n) \\
&\geq 1 - 4(d_x+1)\exp\left(-(c_1 \wedge c_5 \wedge c_6) n h^4\right) - 2 d_x^*(d_x - d_x^*)\exp(-c_3 n) \\
&\geq 1 - 2\max\{2(d_x+1), d_x^*(d_x - d_x^*)\}\exp\left(-(c_1 \wedge c_3 \wedge c_5 \wedge c_6) n h^4\right) \\
&= 1 - c_7 \exp(-c_8 n h^4),
\end{aligned}
$$

where $c_7 = 2\max\{2(d_x+1), d_x^*(d_x - d_x^*)\} \leq \max\{2(d_x+1), d_x^2/4\}$, and $c_8 = c_1 \wedge c_3 \wedge c_5 \wedge c_6$. Note that the inequality (a) holds by the union bound, (b) holds by (A-40), (A-44) and (c) holds by (A-28), (A-29).

Finally, we define new constants b_0, b_1, b_2, b_3 to summarize the results,

$$
\begin{aligned}
&b_0(d_x) := 2\max\{2(d_x+1), d_x^2/4\}, \\
&b_1(d_x, \gamma, \underline{\lambda}, \overline{\lambda}, L, \sigma) := c_1 \wedge c_3 \wedge c_5 \wedge c_6, \\
&\qquad\qquad = \left\{c_1 \wedge (1-\gamma)^2 \underline{\lambda}^2/(8d_x^2) \wedge 64L^2 d_x^2/(2\sigma^2) \wedge 3200(3+\gamma)\overline{\lambda} L^2 d_x^3 / \left((1-\gamma)\sigma^2\right)\right\}, \\
&c_1(d_x, \gamma, \underline{\lambda}) := \frac{\underline{\lambda}}{2(1+\gamma)(1+d_x/4)} \min\left\{1 - \gamma + (3\gamma+1)\log\left(\frac{3\gamma+1}{2+2\gamma}\right), \gamma - 1 + (3+\gamma)\log\left(\frac{3+\gamma}{2+2\gamma}\right)\right\}, \\
&b_2(\overline{\lambda}, \gamma, d_x) := 32\sqrt{\frac{(3+\gamma)\overline{\lambda}}{(1+\gamma)(1-\gamma)^2}} L d_x, \\
&b_3(C, \overline{\lambda}, \underline{\lambda}, \gamma, L, d_x) := \frac{C(1+3\gamma)(1-\gamma)\underline{\lambda}}{160 L d_x \sqrt{(3+\gamma)(1+\gamma)\overline{\lambda} d_x}}.
\end{aligned}
$$

Further assuming $h < \mu_m^2/(3d_x^* L_\mu)$, by Proposition A1, we can replace $\underline{\lambda}$, $\overline{\lambda}, \gamma$ by $\mu_m/12$, $\mu_M, 0.5$. Then, given $h \leq b_3$ and $\lambda = b_2 h^2$, we have

$$\mathbb{P}(\cap_{i=1}^4 \Omega_i) \geq 1 - b_0 \exp(b_1 n_j h^4),$$

where the constants

$$
\begin{aligned}
&b_0(d_x) = 2\max\{2(d_x+1), d_x^2/4\},\\
&b_1(d_x,\mu_m,\mu_M,L,\sigma) = 11\mu_m/(10^4(1+d_x/4)) \wedge \mu_m^2/(4608d_x^2) \wedge 64L^2d_x^2/(2\sigma^2) \wedge 22400\mu_M L^2 d_x^3/\sigma^2,\\
&b_2(d_x,\mu_M) = 64\sqrt{7\mu_M/3}Ld_x,\\
&b_3(d_x,\mu_m,\mu_M,L_\mu,C) = \min\left\{C\mu_m/(768\sqrt{21\mu_m d_x}), \mu_m^2/(3d_xL_\mu)\right\}.
\end{aligned}
$$

□

A.4 Proofs for Weighted Voting

Proof of Proposition 3: By Lemma 5 and 6, we obtain an upper bound for the tail probability

$$\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0) \leq \exp\left\{\xi\left(h^{-d_x}(1+\log b_0 - \log\xi) - b_1 nh^4\right)\right\}.$$

By the same argument, if $J^{(i)} = 1$, we have

$$\mathbb{P}(\hat{J}^{(i)} \leq 1-\xi | J^{(i)} = 1) \leq \exp\left\{\xi\left(h^{-d_x}(1+\log b_0 - \log\xi) - b_1 nh^4\right)\right\}.$$

We choose $\xi = 0.5$ to make the two tail probability equivalent. So the variable x_i is classified as relevant if and only if $J^{(i)} \geq 1/2$. And the misidentification probability for x_i has the upper bound

$$\mathbb{P}\left(\left|\hat{J}^{(i)} - J^{(i)}\right| \geq \frac{1}{2}\right) \leq \exp\left\{\frac{1}{2}\left(h^{-d_x}(1+\log b_0 + \log 2) - b_1 nh^4\right)\right\}.$$

Moreover, by the union bound of all variables, we have the probability lower bound for successful variable selection

$$\mathbb{P}(\hat{J} = J) \geq 1 - d_x \exp\left\{\frac{1}{2}\left(h^{-d_x}(1+\log b_0 + \log 2) - b_1 nh^4\right)\right\}.$$

Hence, we complete the proof of Proposition 3. □

Proof of Lemma 5: Since the problem (15) involves minimizing a continuous function over a compact set [10], it has an optimal solution. In the proof of Lemma 5, we will prove the KKT condition admits a unique solution. Thus, it must be the global optimum for problem (15). Considering the optimal η^*, if $\eta^* = 0$, then $V(0, \boldsymbol{w}) = e$ for any $\boldsymbol{w}$. Next we study the local optima with $\eta^* > 0$. Finally, we compare the optimal V in the two cases.

Supposing $\eta^* > 0$, by the first-order optimality condition, we have

$$0 = \frac{\partial V(\eta, \boldsymbol{w})}{\partial \eta} = \left(-\xi + \sum_{j=1}^{h^{-d_x}} p_j w_j e^{\eta w_j}\right) V(\eta, \boldsymbol{w}).$$

[10] It's obvious that $V \to +\infty$ as $\eta \to +\infty$, so the minimum is obtained when η is finite.

Since $V(\eta, \boldsymbol{w}) > 0$, we have

$$\sum_{j=1}^{h^{-d_x}} p_j w_j e^{\eta w_j} = \xi. \tag{A-51}$$

Next, we write down the KKT condition for w_j. Let v_j, u be the Lagrangian multipliers for constraints $w_j \geq 0$ and $\sum_{j=1}^{h^{-d_x}} w_j - 1 = 0$, we have

$$\frac{\partial V(\eta, \boldsymbol{w})}{\partial w_j} - v_j + u = 0, \tag{A-52}$$

$$v_j w_j = 0, \tag{A-53}$$

$$v_j \geq 0, \quad \forall j \in \{1, 2, \ldots, h^{-d_x}\} \tag{A-54}$$

$$\sum_{j=1}^{h^{-d_x}} w_j = 1. \tag{A-55}$$

From (A-53), we know either $v_j = 0$ or $w_j = 0$ for all j. Define a set O including all the subscript j satisfying $w_j > 0$ and define its cardinality as m,

$$O \coloneqq \{j : v_j = 0, w_j > 0\}, \text{ and } m \coloneqq |O|. \tag{A-56}$$

For $j \in O$, plugging v_j into (A-52), we have that

$$-u = \frac{\partial V(\eta, \boldsymbol{w})}{\partial w_j} = \eta e^{\eta w_j} p_j V(\eta, \boldsymbol{w}). \tag{A-57}$$

It is easy to see that

$$e^{\eta w_1} p_1 = e^{\eta w_2} p_2 = \ldots = e^{\eta w_m} p_m = -\frac{u}{\eta V(\eta, \boldsymbol{w})}. \tag{A-58}$$

For $j \notin O$, we have $w_j = 0$ so

$$\sum_{j \in O} w_j = \sum_{j=1}^{h^{-d_x}} w_j = 1. \tag{A-59}$$

Therefore, plugging (A-58), (A-59) into (A-51), we obtain

$$e^{\eta w_j} p_j = \xi, \quad \forall j \in O. \tag{A-60}$$

Because $\eta > 0$ and $w_j > 0$, we have $e^{\eta w_j} > 1$ and thus

$$p_j < \xi, \quad \forall j \in O.$$

Then, taking natural logarithm of both sides of (A-60), we have

$$w_j = (\log \xi - \log p_j) / \eta, \quad \forall j \in O. \tag{A-61}$$

Plugging it into (A-59) and (A-58), we get

$$\eta = \sum_{j \in O} (\log \xi - \log p_j), \quad \forall j \in O, \tag{A-62}$$

and

$$u = -V(\eta, \boldsymbol{w})\eta\xi. \tag{A-63}$$

For $j \notin O$, we have $v_j \geq 0, w_j = 0$. Thus, plugging (A-63) and (A-57) into (A-52), we have

$$v_j = \frac{\partial V(\eta, \boldsymbol{w})}{\partial w_j} + u = \eta V(\eta, \boldsymbol{w}) p_j + u = \eta V(\eta, \boldsymbol{w})(p_j - \xi).$$

As $v_j \geq 0, V(\eta, \boldsymbol{w}) > 0$ and $\eta > 0$, we have

$$p_j \geq \xi, \quad \forall j \notin O.$$

Plugging (A-61) and (A-62) into (15), we get a closed-form solution for $V(\eta^*, \boldsymbol{w}^*)$ if $\eta^* > 0$:

$$V(\eta^*, \boldsymbol{w}^*) = \exp\left(\sum_{j=1}^{h^{-d_x}} (\xi - \xi \log \xi - p_j + \xi \log p_j) \mathbb{I}(p_j < \xi)\right)$$

Define a function

$$H(p) \coloneqq \xi \log p - p, \tag{A-64}$$

and

$$V(\eta^*, \boldsymbol{w}^*) = \exp\left(\sum_{j=1}^{h^{-d_x}} (H(p_j) - H(\xi)) \mathbb{I}(p_j < \xi)\right).$$

Note that $H(\cdot)$ is a concave function, attaining its maximum at ξ. Thus, we have $H(p_j) - H(\xi) \leq 0$ and $V(\eta^*, \boldsymbol{w}^*) \leq e = V(0, \boldsymbol{w})$. So the optimal V is attained when $\eta^* > 0$. Therefore, we have prove the KKT condition admits a unique solution, which must be the global optimum for problem (15).

Finally, we give a summary for the optimal solution $\eta^*, \boldsymbol{w}^*$ of the optimization problem (15):

1. $\eta^* = \sum_{j=1}^{h^{-d_x}} (\log \xi - \log p_j) \mathbb{I}(p_j < \xi)$.
2. If $p_j < \xi$, then $w_i^* = (\log \xi - \log p_j)/\eta^*$.
3. If $p_j \geq \xi$, then $w_i^* = 0$.
4. The optimal value $V(\eta^*, \boldsymbol{w}^*) = \exp\left(\sum_{j=1}^{h^{-d_x}} (\xi - \xi \log \xi - p_j + \xi \log p_j) \mathbb{I}(p_j < \xi)\right)$.

□

Proof of Lemma 6: Recalling the definition of $H(\cdot)$ in (A-64), the objective function of (16) can be rewritten as

$$V(\boldsymbol{n}) = \exp\left(\sum_{j=1}^{h^{-d_x}} \left(H(p_j) - H(\xi)\right) \mathbb{I}(p_j < \xi)\right). \tag{A-65}$$

Note that $H(\cdot)$ is a negative and concave function, attaining its maximum at ξ. Moreover, $H(p_j)$ increases with p_j when $p_j < \xi$. Since p_j is a monotone decreasing function of n_j, there exists a threshold

$$\underline{n} := \max\{n : b_0 \exp(-b_1 n h^4) \geq \xi\}, \tag{A-66}$$

such that $H\left(p_j(n_j)\right)$ (denoted as $H(n_j)$ for simplicity) decreases with n_j when $n_j > \underline{n}$. In particular, we have n budgets and h^{-d_x} bins. We divide all the bins into two groups: active bins $A := \{j : \mathbb{I}(p_j < \xi)\}$ and non-active bins $A^c := \{j : \mathbb{I}(p_j \geq \xi)\}$. For active bins, $H(n_j)$ decreases as more budgets allocated to the bin. For non-active bins, they only consume budgets but have no contribution to the objective function (A-65). To maximize $V(\boldsymbol{n})$, the non-active bins should consume as much budgets as possible. So their optimal budgets should equal to the threshold that $n_j^* = \underline{n}$. Thus, if $n \leq \underline{n} h^{-d_x}$ (equivalent to $n \leq \log(2b_0)/(b_1 h^{d_x+4})$), then all the bins are non-active bins and $n_j^* = n h^{d_x}$, thus we have $V(\boldsymbol{n}) = 1$. If $n > \underline{n} h^{-d_x}$, then there must exist active bins. We assume the cardinality for active bins is $m := |A|$ and their indices are from 1 to m. Then, we can fully separate the budgets for active and non-active bins, and reformulate (16) as

$$\begin{aligned}
\max_{\boldsymbol{n}} \quad & \log\left(V(\boldsymbol{n})\right) = \sum_{j=1}^{m} H(p_j) - H(\xi) \\
\text{s.t.} \quad & p_j = b_0 \exp\left(-b_1 n_j h^4\right) \\
& n_j \geq \underline{n} \\
& \sum_{j=1}^{m} n_j = n - \underline{n}(h^{-d_x} - m) \\
& n_j \in N^+, \quad \forall j \in \{1, 2, \dots, m\}.
\end{aligned} \tag{A-67}$$

Relaxing n_j to $\mathbb{R}^+$, it's a concave and continuous optimization problem. By the KKT condition, let v_j, u be the Lagrangian multipliers for $\underline{n} - n_j \leq 0$ and $\sum_{j=1}^{m} n_j = n - \underline{n}(h^{-d_x} - m)$, we have

$$\frac{\partial \log\left(V(\boldsymbol{n})\right)}{\partial n_j} + v_j + u = 0 \tag{A-68}$$

$$v_j(\underline{n} - n_j) = 0 \tag{A-69}$$

$$v_j \geq 0$$

$$\sum_{j=1}^{m} n_j + \underline{n}(h^{-d_x} - m) = n. \tag{A-70}$$

From (A-69), either $v_j = 0$ or $n_j = \underline{n}$ for all j. Define a set

$$O \coloneqq \{j : v_j = 0,\ n_j > \underline{n}\},$$

for $j \in O$. Plugging $v_j = 0$ into (A-68), we have

$$u = -\frac{\partial \log\left(V(\boldsymbol{n})\right)}{\partial n_j} = b_1 h^4(\xi - p_j). \tag{A-71}$$

By the definition of $\underline{n}$ in (A-66), $n_j > \underline{n}$ implies $p_j < \xi$. Then $u > 0$. For $k \in O^c$, we have $v_k \geq 0$, $n_k = \underline{n}$ and

$$u = -\frac{\partial \log\left(V(\boldsymbol{n})\right)}{\partial n_k} - v_k = b_1 h^4(\xi - p_k) - v_k. \tag{A-72}$$

In fact, (A-71) and (A-72) cannot hold simultaneously. Recalling that for $j \in O$ and $k \in O^c$, $n_j > \underline{n} = n_k$. Thus we have $p_j < p_k$ because p_j is decreasing in n_j. By (A-71) and (A-72), we have

$$u = b_1 h^4(\xi - p_j) > b_1 h^4(\xi - p_k) \geq b_1 h^4(\xi - p_k) - v_k = u.$$

That's to say, either O or O^c is empty. Since we have supposed $n > \underline{n} h^{-d_x}$, there's at least in one bin $n_j > \underline{n}$. So O^c is empty and (A-71) is satisfied for all $j \in 1, 2, \ldots, m$, further implying $p_1 = p_2 = \ldots = p_m$. As p_j is a strictly decreasing function of n_j, by (A-70), we get the optimal solution

$$n_1^* = n_2^* = \ldots = n_m^* = \left(n - \underline{n}(h^{-d_x} - m)\right)/m. \tag{A-73}$$

Moreover, the optimal value in (A-67) is

$$\begin{aligned}
\log(V(\boldsymbol{n})) &= m(\xi \log p_j - p_j - \xi \log \xi + \xi) \\
&\leq m\xi(\log p_j - \log \xi + 1) \\
&= \xi\left(-b_1 h^4 n + b_1 \underline{n} h^{4-d_x} + \left(\log b_0 - \log \xi + 1 - b_1 h^4 \underline{n}\right) m\right),
\end{aligned} \tag{A-74}$$

where the last equality follows by (A-73) and $p_j = b_0 \exp(-b_1 n_j h^4)$. By the definition of $\underline{n}$ in (A-66), we have

$$\log b_0 - \log \xi + 1 - b_1 h^4 \underline{n} > 0,$$

which implies that the term in (A-74) will increase as m. Therefore, when $m = h^{-d_x}$, the term in (A-74) attains its maximum, which also gives an upper bound for the optimal $V(\boldsymbol{n}^*)$ in (16),

$$V(\boldsymbol{n}^*) \leq \exp\left\{\xi\left(h^{-d_x}(1 + \log b_0 - \log \xi) - b_1 n h^4\right)\right\}.$$

$\square$

A.5 Proofs for the Regret Bound

Proof of Proposition 1. Supposing the dimension of decision space is d_y, we prove the stronger version stated in Remark 3. Recall that the total regret in T periods can be upper bounded by

$$R(T) \le 2n \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + \mathbb{P}(\hat{J}=J)R_2(T-n) + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)|\mathbb{P}(\hat{J}\neq J)(T-n),$$

where $n \le T^{(d_x^*+d_y+1)/(d_x^*+d_y+2)}$ and $\mathbb{P}(\hat{J}\neq J) \le n^{-1/(d_x^*+d_y+1)}$. Further relaxing the right-hand side, we have

$$\begin{aligned}
R(T) &\le 2n \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + \mathbb{P}(\hat{J}=J)R_2(T-n) + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| n^{-1/(d_x^*+d_y+1)}(T-n)\\
&= 2n \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + R_2(T) + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| n^{-1/(d_x^*+d_y+1)}T\\
&= O(T^{1-1/(d_x^*+d_y+2)}) + R_2(T)\\
&= O\left(T^{1-1/(d_x^*+d_y+2)}\log(T)\right).
\end{aligned}$$

The first equality follows from $\mathbb{P}(\hat{J}=J)\le 1$ and $T-n\le T$, the second equality follows from $\mathbb{P}(\hat{J}\neq J)\le n^{-1/(d_x^*+d_y+1)} \le \left(T^{(d_x^*+d_y+1)/(d_x^*+d_y+2)}\right)^{-1/(d_x^*+d_y+1)} = O(T^{-1/(d_x^*+d_y+2)})$ and the last equality is supported by (4). $\square$

Proof of Theorem 1. Recall that the total regret in T periods can be upper bounded by

$$R(T) \le 2n \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + \mathbb{P}(\hat{J}=J)R_2(T-n) + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)|\mathbb{P}(\hat{J}\neq J)(T-n).$$

where $n=(\log(T))^{(2+4/d_x)}$ and $h=(\log(T))^{-1/d_x}$. It can be checked that $n\ge \log(2b_0)/(b_1h^{d_x+4})$ and $h\le b_3$. Then by Proposition 3, we have

$$\mathbb{P}(\hat{J}\neq J)\le d_x\exp\left\{\frac{1}{2}\left(h^{-d_x}(1+\log b_0+\log 2)-b_1nh^4\right)\right\}.$$

Further relaxing the right-hand side, we have

$$\begin{aligned}
R(T) &\le 2(\log T)^{(2+4/d_x)} \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + R_2(T-n)\\
&\quad + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| d_x\exp\left\{\frac{1}{2}\left(h^{-d_x}(1+\log b_0+\log 2)-b_1nh^4\right)\right\}T\\
&\le 2(\log T)^6 \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + R_2(T-n) + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| d_x\exp(-\log T)T\\
&= 2(\log T)^6 \max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| + R_2(T-n) + 2\max_{\boldsymbol{x}\in\mathcal{X},y\in\mathcal{Y}} |f(\boldsymbol{x},y)| d_x\\
&= O(R_2(T)).
\end{aligned}$$

The second inequality follows from $d_x\ge 1$ and $T\ge \exp\{(3+\log b_0+\log 2)/b_1\}$. $\square$

A.6 Proofs for Local Relevance

Proof of Proposition 4: Recall that in the proof of Lemma 1, the hypercube $\mathcal{H}_i$ with side length $\overline{h} = C/L$ and centred at $\boldsymbol{x}_{(i)}$ satisfies (3). We will show that $Q_i(C)$ covers at least $(1/3)^{d_x}$ proportion of $\mathcal{H}_i$, if choosing $h \le \overline{h}/3$. Note that $\mathcal{H}_i$ is covered by $Q_i(C)$ and bins intersected with the boundary of $\mathcal{H}_i$. We consider the worst case that as more areas covered by the intersected bins as possible. When all the boundaries of the intersected bins exactly coincide with the boundary of $\mathcal{H}_i$, the intersected bins take up the most proportion of $\mathcal{H}_i$. In this case, $2/3$ proportion of each side length is covered by the intersected bins, and $(1/3)^{d_x}$ proportion of $\mathcal{H}_i$ is covered by the bins in $Q_i(C)$. Then, by Assumption 6, the probability density has a lower bound μ_m and

$$\mathbb{P}(\boldsymbol{X} \in Q_i(C)) \ge \mathbb{P}(\boldsymbol{X} \in \mathcal{H}_i) \ge \mu_m (\overline{h}/3)^{d_x} = \mu_m \left(\frac{C}{3L}\right)^{d_x}.$$

Hence, we complete the proof of Proposition 4. $\square$

Proof of Proposition 5: Step one: In the first step, we consider the case that the true indicator $J^{(i)} = 0$, that is, x_i is redundant. We show an upper bound for the misidentification probability:

$$\begin{aligned}
\mathbb{P}(\hat{J}^{(i)} \ge \xi | J^{(i)} = 0) &= \mathbb{P}(e^{\eta \hat{J}^{(i)}} \ge e^{\eta\xi} | J^{(i)} = 0) \\
&\le e^{-\eta\xi} \mathbb{E}\left[e^{\eta \hat{J}^{(i)}} | J^{(i)} = 0\right] && \text{(by Markov's inequality)} \\
&= e^{-\eta\xi} \prod_{j=1}^{h^{-d_x}} \mathbb{E}\left[e^{\eta w_j \hat{J}_j^{(i)}} | J^{(i)} = 0\right] && \text{(by the definition of } \hat{J}^{(i)} \text{ (10))} \\
&= e^{-\eta\xi} \prod_{j=1}^{h^{-d_x}} (1 + (e^{\eta w_j} - 1)p_j) && \text{(by } \hat{J}_j^{(i)} \sim \textit{ Bernoulli}(p_j)) \\
&= \exp\left\{-\eta\xi + \sum_{j=1}^{h^{-d_x}} \log(1 + (e^{\eta w_j} - 1)p_j)\right\} && \text{(A-75)}
\end{aligned}$$

Note that we use tighter inequalities than (14), which turns out to be too loose for Proposition 5.

Since (A-75) holds for arbitrary non-negative η and w_j, we need to find η, w_j to minimize the probability:

$$\begin{aligned}
\min_{\eta, \boldsymbol{w}} \quad & V(\eta, \boldsymbol{w}) = -\eta\xi + \sum_{j=1}^{h^{-d_x}} \log(1 + (e^{\eta w_j} - 1)p_j) \\
\text{s.t.} \quad & w_j \ge 0, \ \forall j \in \{1, 2, \dots, h^{-d_x}\}, \\
& \eta \ge 0, \\
& \sum_{j=1}^{h^{-d_x}} w_j = 1.
\end{aligned} \tag{A-76}$$

Similar to the proof of Lemma 5, we apply the KKT optimality condition. If $\eta^* = 0$, then $V(0, \boldsymbol{w}) = 0$ for any $\boldsymbol{w}$, which is clearly not optimal. So $\eta^* > 0$ and the first-order condition holds.

That implies the optimal η solving

$$0 = \frac{\partial V(\eta, \boldsymbol{w})}{\partial \eta} = -\xi + \sum_{j=1}^{h^{-d_x}} \frac{w_j p_j e^{\eta w_j}}{1 + (e^{\eta w_j} - 1)p_j} \tag{A-77}$$

Next, we write down the KKT condition for w_j. Let v_j, u be the Lagrangian multipliers for constraints $w_j \geq 0$ and $\sum_{j=1}^{h^{-d_x}} w_j - 1 = 0$, we have

$$\frac{\partial V(\eta, \boldsymbol{w})}{\partial w_j} - v_j + u = 0, \tag{A-78}$$

$$v_j w_j = 0, \tag{A-79}$$

$$v_j \geq 0, \quad \forall j \in \{1, 2, \ldots, h^{-d_x}\}$$

$$\sum_{j=1}^{h^{-d_x}} w_j = 1. \tag{A-80}$$

From (A-79), we know that either $v_j = 0$ or $w_j = 0$ for any j. We define a set O for all j satisfying $w_j > 0$ and define its cardinality as m,

$$O \coloneqq \{j : v_j = 0, w_j > 0\}, \text{ and } m \coloneqq |O|. \tag{A-81}$$

For $j \in O$, plugging $v_j = 0$ into (A-78), we have

$$-u = \frac{\partial V(\eta, \boldsymbol{w})}{\partial w_j} = \frac{\eta p_j e^{\eta w_j}}{1 + (e^{\eta w_j} - 1)p_j}. \tag{A-82}$$

After a simple transform, we have

$$\frac{p_1 e^{\eta w_1}}{1 + (e^{\eta w_1} - 1)p_1} = \frac{p_2 e^{\eta w_2}}{1 + (e^{\eta w_2} - 1)p_2} = \ldots = \frac{p_m e^{\eta w_m}}{1 + (e^{\eta w_m} - 1)p_m} = -u/\eta. \tag{A-83}$$

Plugging it into (A-77), we have

$$\xi = -\frac{u}{\eta} \sum_{j \in O} w_j = -\frac{u}{\eta}. \tag{A-84}$$

This is because for $j \notin O$, we have $w_j = 0$, so

$$\sum_{j \in O} w_j = \sum_{j=1}^{h^{-d_x}} w_j = 1. \tag{A-85}$$

By (A-83) and (A-84), we have

$$\eta w_1 = \eta w_2 = \ldots = \eta w_m = \log \xi + \log(1 - p_j) - \log p_j - \log(1 - \xi). \tag{A-86}$$

Since $\eta w_1 > 0$, we have $p_j < \xi$. Therefore, plugging (A-86), into (A-85), we obtain

$$\eta = \sum_{j \in O} (\log \xi + \log(1 - p_j) - \log p_j - \log(1 - \xi)), \tag{A-87}$$

and

$$w_j = \frac{1}{\eta} \left(\log \xi + \log(1 - p_j) - \log p_j - \log(1 - \xi)\right), \quad \forall j \in O. \tag{A-88}$$

For $j \notin O$, we have $v_j \geq 0, w_j = 0$. Then, plugging (A-87), (A-84), (A-82) into (A-78), we have

$$v_j = \eta(p_j - \xi),$$

and

$$p_j \geq \xi, \quad \forall j \notin O.$$

Plugging (A-87) and (A-88) into (A-76), we have

$$V(\eta^*, \boldsymbol{w}^*) = \sum_{j=1}^{h^{-d_x}} \left(\xi \log p_j + (1 - \xi) \log(1 - p_j) - \xi \log \xi - (1 - \xi) \log(1 - \xi)\right) \mathbb{I}\left(p_j < \xi\right). \tag{A-89}$$

Therefore, we have prove the KKT condition admits a unique solution, which must be the global optimum for problem (A-76).

We give a summary for the optimal solution $\eta^*, \boldsymbol{w}^*$ of the optimization problem (A-76):

1. $\eta^* = \sum_{j=1}^{h^{-d_x}} \left(\log \xi + \log(1 - p_j) - \log p_j - \log(1 - \xi)\right) \mathbb{I}\left(p_j < \xi\right)$.

2. If $p_j < \xi$, then $w_j^* = (\log \xi + \log(1 - p_j) - \log p_j - \log(1 - \xi))/\eta^*$.

3. If $p_j \geq \xi$, then $w_j^* = 0$.

4. The optimal value $V(\eta^*, \boldsymbol{w}^*)$ shows in (A-89).

The optimal V of (A-89) depends on p_j, which is the probability bound derived in Proposition 2. The condition of Proposition 2 holds since $h \leq b_3/2 \leq b_3$ and $\lambda = b_2 h^2$. Plugging p_j into (A-89) and because of $\log(1 - p_j) < 0$, we have

$$\begin{aligned}
V(\eta^*, \boldsymbol{w}^*) &\leq \sum_{j=1}^{h^{-d_x}} \left(\xi \log p_j - \xi \log \xi - (1 - \xi) \log(1 - \xi)\right) \mathbb{I}\left(p_j < \xi\right) \\
&\leq \sum_{j=1}^{h^{-d_x}} \left(-\xi b_1 n_j h^4 + \xi \log b_0 - \xi \log \xi - (1 - \xi) \log(1 - \xi)\right) \mathbb{I}\left(p_j < \xi\right) \\
&\leq \left(\xi \log b_0 - \xi \log \xi - (1 - \xi) \log(1 - \xi)\right) h^{-d_x} - \xi b_1 h^4 \cdot \left(\sum_{j=1}^{h^{-d_x}} n_j \mathbb{I}\left(p_j < \xi\right)\right).
\end{aligned} \tag{A-90}$$

The first inequality follows from $\log(1-p_j) < 0$; the second follows from $p_j = b_0 \exp(-b_1 n_j h^4)$. Since p_j is a monotone decreasing function of n_j, there exists a threshold

$$\underline{n} \coloneqq \max\{n : b_0 \exp(-b_1 n h^4) \geq \xi\}, \tag{A-91}$$

such that $p_j < \xi$ for $n_j > \underline{n}$. By (A-91), we have

$$b_0 \exp(-b_1 \underline{n} h^4) > \xi \Longrightarrow b_1 \underline{n} h^4 \leq \log b_0 - \log \xi \tag{A-92}$$

So we get a lower bound for the last term in (A-90),

$$\sum_{j=1}^{h^{-d_x}} n_j \mathbb{I}\left(p_j < \xi\right) \geq n - h^{-d_x} \underline{n} \geq n - \frac{h^{-d_x}(\log b_0 - \log \xi)}{b_1 h^4}. \tag{A-93}$$

Therefore, plugging (A-93) into (A-90), we have

$$\begin{aligned} V(\eta^*, \boldsymbol{w}^*) &\leq \left(\xi \log b_0 - \xi \log \xi - (1-\xi)\log(1-\xi)\right) h^{-d_x} - \xi b_1 h^4 \left(n - h^{-d_x} \underline{n}\right) \\ &\leq \left(\xi \log b_0 - \xi \log \xi - (1-\xi)\log(1-\xi)\right) h^{-d_x} + \xi(\log b_0 - \log \xi) h^{-d_x} - \xi b_1 h^4 n \\ &= \left(2\xi \log b_0 - 2\xi \log \xi - (1-\xi)\log(1-\xi)\right) h^{-d_x} - \xi b_1 h^4 n. \end{aligned}$$

The second inequality holds by (A-92). Recalling the tail probability in (A-75), we have

$$\mathbb{P}(\hat{J}^{(i)} \geq \xi | J^{(i)} = 0) \leq \exp\left\{\left(2\xi \log b_0 - 2\xi \log \xi - (1-\xi)\log(1-\xi)\right) h^{-d_x} - \xi b_1 h^4 n\right\}. \tag{A-94}$$

So far, we show a tail probability upper bound for the variable x_i satisfying $J^{(i)} = 0$.

Step two: Next we consider the case when $J^{(i)} = 1$. We start with the following bound

$$\begin{aligned} \mathbb{P}(\hat{J}^{(i)} \leq \xi | J^{(i)} = 1) &= \mathbb{P}(1 - \hat{J}^{(i)} \geq 1 - \xi | J^{(i)} = 1) \\ &\leq e^{-\eta(1-\xi)} \mathbb{E}\left[\exp\left(\eta\left(1 - \hat{J}^{(i)}\right)\right) | J^{(i)} = 1\right] \\ &= e^{-\eta(1-\xi)} \prod_{j=1}^{h^{-d_x}} \mathbb{E}\left[\exp\left(\eta w_j \left(1 - \hat{J}_j^{(i)}\right)\right) | J^{(i)} = 1\right] \\ &= \exp\left\{-\eta(1-\xi) + \sum_{j=1}^{h^{-d_x}} \log\left(1 + (e^{\eta w_j} - 1) p_j\right)\right\}. \end{aligned} \tag{A-95}$$

Notice that (A-95) is the same as (A-75) except that ξ is replaced by $1-\xi$. Thus, replacing ξ by $1-\xi$ in (A-87) and (A-88), we get the optimal solution for (A-95),

$$\eta^* = \sum_{j=1}^{h^{-d_x}} \left(\log(1-\xi) + \log(1-p_j) - \log p_j - \log \xi\right) \mathbb{I}\left(p_j < 1 - \xi\right), \tag{A-96}$$

and

$$w_j^* = \left(\log(1-\xi) + \log(1-p_j) - \log p_j - \log \xi\right)/\eta^* \qquad \text{for } p_j < 1-\xi, \tag{A-97}$$

$$w_j^* = 0 \qquad \text{for } p_j \geq 1-\xi. \tag{A-98}$$

Plugging them into (A-95), we have

$$V(\eta^*, \boldsymbol{w}^*) = \exp\left\{-\eta^*(1-\xi) + \sum_{j=1}^{h^{-d_x}} \log\left(1 + (e^{\eta^* w_j^*} - 1)p_j\right) \mathbb{I}\left(p_j < 1-\xi\right)\right\} \tag{A-99}$$

Note that on the event $p_j < 1-\xi$,

$$\begin{aligned}
\log\left(1 + (e^{\eta^* w_j^*} - 1)p_j\right) &= \log\left(1 + \left(\exp\left(\log(1-\xi) + \log(1-p_j) - \log p_j - \log(\xi)\right) - 1\right) p_j\right) \\
&= \log\left(1 + \left(\frac{(1-\xi)(1-p_j)}{p_j \xi} - 1\right) p_j\right) \\
&= \log(1-p_j) - \log \xi.
\end{aligned}$$

Moreover, for the uninformative bins $j \in Q^c$ (with a slight abuse, we omit C and i in $Q_i(C)$ and use j to represent B_j), we simply use an upper bound $p_j \leq 1$ and $\log\left(1 + (e^{\eta^* w_j^*} - 1)p_j\right) \leq \eta^* w_j^*$. Plugging them to (A-99), we have

$$\log V(\eta^*, \boldsymbol{w}^*) \leq -\eta^*(1-\xi) + \sum_{j=1}^{h^{-d_x}} \left(\eta^* w_j^* \mathbb{I}(j \in Q^c) + \left(\log(1-p_j) - \log \xi\right) \mathbb{I}(j \in Q)\right) \mathbb{I}\left(p_j < 1-\xi\right).$$

Like the previous argument, we define

$$O \coloneqq \left\{j : p_j < 1-\xi\right\}.$$

Plugging in the form of η^* and w_j^* from (A-96) and (A-97), we have

$$\begin{aligned}
&\log V(\eta^*, \boldsymbol{w}^*) \\
&\quad \leq -\eta^*(1-\xi) + \sum_{j=1}^{h^{-d_x}} \left(\eta^* w_j^* \mathbb{I}(j \in Q^c) + \left(\log(1-p_j) - \log \xi\right) \mathbb{I}(j \in Q)\right) \mathbb{I}\left(j \in O\right) \\
&\quad = \sum_{j=1}^{h^{-d_x}} -(1-\xi)\left(\log(1-\xi) + \log(1-p_j) - \log p_j - \log \xi\right) \mathbb{I}(j \in O) \\
&\qquad + \left(\log(1-\xi) + \log(1-p_j) - \log p_j - \log \xi\right) \mathbb{I}(j \in O \cap Q^c) + \left(\log(1-p_j) - \log \xi\right) \mathbb{I}(j \in O \cap Q).
\end{aligned} \tag{A-100}$$

Recombining the terms in (A-100) by $\log(1-\xi)$, $\log\xi$, $\log(1-p_j)$ and $\log p_j$, we have (A-100)

$$\begin{aligned}
=\sum_{j=1}^{h^{-d_x}} &(-(1-\xi)\mathbb{I}(j\in O)+\mathbb{I}(j\in O\cap Q^c))\log(1-\xi)\\
&+((1-\xi)\mathbb{I}(j\in O)-\mathbb{I}(j\in O\cap Q^c)-\mathbb{I}(j\in O\cap Q))\log\xi\\
&+(-(1-\xi)\mathbb{I}(j\in O)+\mathbb{I}(j\in O\cap Q^c)+\mathbb{I}(j\in O\cap Q))\log(1-p_j)\\
&+((1-\xi)\mathbb{I}(j\in O)-\mathbb{I}(j\in O\cap Q^c))\log p_j.
\end{aligned}$$

Further by $\mathbb{I}(j\in O)=\mathbb{I}(j\in O\cap Q)+\mathbb{I}(j\in O\cap Q^c)$, (A-100) is simplified to

$$\begin{aligned}
=\sum_{j=1}^{h^{-d_x}} &(-(1-\xi)\mathbb{I}(j\in O)+\mathbb{I}(j\in O\cap Q^c))\log(1-\xi)-\xi\mathbb{I}(j\in O\cap Q^c)\log\xi\\
&+\xi\mathbb{I}(j\in O)\log(1-p_j)+((1-\xi)\mathbb{I}(j\in O\cap Q)-\xi\mathbb{I}(j\in O\cap Q^c))\log p_j.
\end{aligned}\tag{A-101}$$

Note that $\log\xi\le 0$, $\log(1-\xi)\le 0$, $\log(1-p_j)\le 0$ and $\mathbb{I}(j\in O\cap Q^c)\le\mathbb{I}(j\in O)$, we have (A-101)

$$\begin{aligned}
&\le\sum_{j=1}^{h^{-d_x}}(-(1-\xi)\log(1-\xi)\mathbb{I}(j\in O))-\xi\mathbb{I}(j\in O\cap Q^c)\log\xi\\
&\quad+((1-\xi)\mathbb{I}(j\in O\cap Q)-\xi\mathbb{I}(j\in O\cap Q^c))\log p_j\\
&\le\sum_{j=1}^{h^{-d_x}}\mathbb{I}(j\in O)\left(-\xi\log\xi-(1-\xi)\log(1-\xi)\right)+((1-\xi)\mathbb{I}(j\in O\cap Q)-\xi\mathbb{I}(j\in O\cap Q^c))\log p_j\\
&\le h^{-d_x}\left(-\xi\log\xi-(1-\xi)\log(1-\xi)\right)+\sum_{j=1}^{h^{-d_x}}\mathbb{I}(j\in O\cap Q)(1-\xi)\log p_j-\mathbb{I}(j\in O\cap Q^c)\xi\log p_j.
\end{aligned}\tag{A-102}$$

The last inequality follows from $\mathbb{I}(j\in O)\le 1$.

Next, we give an upper bound for (A-102). Recalling that Q $(Q_i(C))$ is defined as the union of bins in $A_i(C)$, then the probability bound in Proposition 2 holds as long as $h\le b_3/2$ and $\lambda=b_2h^2$. Plugging $p_j=b_0\exp\left(-b_1n_jh^4\right)$ into the last two terms, we have

$$\begin{aligned}
&\sum_{j=1}^{h^{-d_x}}\mathbb{I}(j\in O\cap Q)(1-\xi)\log p_j-\mathbb{I}(j\in O\cap Q^c)\xi\log p_j\\
&=\sum_{j=1}^{h^{-d_x}}(\mathbb{I}(j\in O\cap Q)(1-\xi)-\mathbb{I}(j\in O\cap Q^c)\xi)\log b_0-(\mathbb{I}(j\in O\cap Q)(1-\xi)n_j-\mathbb{I}(j\in O\cap Q^c)\xi n_j)\,b_1h^4\\
&\le(1-\xi)h^{-d_x}\log b_0-(1-\xi)b_1h^4\cdot\sum_{j=1}^{h^{-d_x}}\mathbb{I}(j\in O\cap Q)n_j+\xi b_1h^4\cdot\sum_{j=1}^{h^{-d_x}}\mathbb{I}(j\in Q^c)n_j.
\end{aligned}\tag{A-103}$$

The last inequality follows from $\mathbb{I}(j \in O \cap Q) \leq 1$ and $\log b_0 \geq 0$.

Notice that $\sum_{j=1}^{h^{-d_x}} \mathbb{I}(j \in Q^c) n_j = \sum_{k=1}^{n} \mathbb{I}(\boldsymbol{X}_k \in Q^c)$ is the number of covariates falling in Q^c, which is a binomial distribution. According to Proposition 4, the mean probability $\mathbb{P}(\boldsymbol{X} \in Q^c) < 1 - p_Q$. Then, applying the Hoeffding's inequality for binomial random variable, we have

$$\mathbb{P}\left(\sum_{k=1}^{n} \mathbb{I}(\boldsymbol{X}_k \in Q^c) - n(1-p_Q) \geq \frac{1}{3} p_Q n\right) \leq e^{-\frac{2}{9} p_Q^2 n}. \tag{A-104}$$

Thus, with probability no less than $1 - e^{-\frac{2}{9} p_Q^2 n}$, we have

$$\sum_{j=1}^{h^{-d_x}} \mathbb{I}(j \in Q^c) n_j \leq (1 - \frac{2}{3} p_Q) n. \tag{A-105}$$

Similar to (A-91) and (A-93) we define the threshold

$$\underline{n} \coloneqq \max\{n : b_0 \exp(-b_1 n h^4) \geq 1 - \xi\},$$

and we have

$$\sum_{j=1}^{h^{-d_x}} \mathbb{I}(j \in O) n_j \geq n - h^{-d_x} \underline{n} \geq n - \frac{h^{-d_x}}{b_1 h^4} (\log b_0 - \log(1-\xi)). \tag{A-106}$$

By (A-105) and (A-106), we have

$$\begin{aligned}
\sum_{j=1}^{h^{-d_x}} \mathbb{I}(j \in O \cap Q) n_j &\geq \sum_{j=1}^{h^{-d_x}} \mathbb{I}(j \in O) n_j - \sum_{j=1}^{h^{-d_x}} \mathbb{I}(j \in Q^c) n_j \\
&\geq n - h^{-d_x} \underline{n} - (1 - \frac{2}{3} p_Q) n \\
&= \frac{2}{3} p_Q n - h^{-d_x} \underline{n} \\
&\geq \frac{2}{3} p_Q n - \frac{h^{-d_x}}{b_1 h^4} (\log b_0 - \log(1-\xi)).
\end{aligned} \tag{A-107}$$

Plugging (A-105) and (A-107) into (A-103) also (A-100), we have

$$\begin{aligned}
&\log V(\eta^*, \boldsymbol{w}^*) \\
&\quad \leq h^{-d_x} \left((1-\xi) \log b_0 - \xi \log \xi - (1-\xi) \log(1-\xi)\right) + (1-\xi) \underline{n} b_1 h^4 h^{-d_x} - \left(\frac{2}{3} p_Q - \xi\right) b_1 h^4 n \\
&\quad \leq h^{-d_x} \left(2(1-\xi) \log b_0 - \xi \log \xi + \xi \log(1-\xi)\right) - \left(\frac{2}{3} p_Q - \xi\right) b_1 h^4 n.
\end{aligned}$$

Plugging it and (A-104) into (A-95), we have

$$
\begin{aligned}
&\mathbb{P}(\hat{J}^{(i)} \leq \xi | J^{(i)} = 1) \\
&\leq \exp\left\{h^{-d_x}\left(2(1-\xi)\log b_0 - \xi\log\xi + \xi\log(1-\xi)\right) - \left(\frac{2}{3}p_Q - \xi\right)b_1 h^4 n\right\} + \exp\left(-\frac{2}{9}p_Q^2 n\right).
\end{aligned}
$$

Hence, we complete the proof of Proposition 5. □