

# Local Inference in Additive Models with Decorrelated Local Linear Estimator\*

Zijian Guo   Cun-Hui Zhang

July 31, 2019

## Abstract

Additive models, as a natural generalization of linear regression, have played an important role in studying nonlinear relationships. Despite of a rich literature and many recent advances on the topic, the statistical inference problem in additive models is still relatively poorly understood. Motivated by the inference for the exposure effect and other applications, we tackle in this paper the statistical inference problem for  $f_1'(x_0)$  in additive models, where  $f_1$  denotes the univariate function of interest and  $f_1'(x_0)$  denotes its first order derivative evaluated at a specific point  $x_0$ . The main challenge for this local inference problem is the understanding and control of the additional uncertainty due to the need of estimating other components in the additive model as nuisance functions. To address this, we propose a decorrelated local linear estimator, which is particularly useful in reducing the effect of the nuisance function estimation error on the estimation accuracy of  $f_1'(x_0)$ . We establish the asymptotic limiting distribution for the proposed estimator and then construct confidence interval and hypothesis testing procedures for  $f_1'(x_0)$ . The variance level of the proposed estimator is of the same order as that of the local least squares in nonparametric regression, or equivalently the additive model with one component, while the bias of the proposed estimator is jointly determined by the statistical accuracies in estimating the nuisance functions and the relationship between the variable of interest and the nuisance variables. The method is developed for general additive models and is demonstrated in the high-dimensional sparse setting.

*Keywords:* High dimension; Decorrelation; Double Estimation Accuracy; Derivative; Exposure Effect; Extreme value location.

---

\*The research of Z Guo was supported in part by the NSF DMS 1811857; The research of C Zhang was supported in part by the NSF Grants DMS-1513378, DMS-1721495 and IIS-1741390.

# 1 Introduction

Additive models play an important role in modern data analysis [4, 20, 41], as a generalization of two popular statistical models, namely the multiple linear model and univariate nonparametric model. A main advantage of the additive model is its relaxation of the stringent linearity assumption imposed in the multiple linear model but at the same time dramatically mitigates the curse of dimensionality in the more complex multiple nonparametric regression. In the low-dimensional setting, additive models have been carefully investigated from both the methodological and theoretical perspectives [4, 20, 21, 25, 27, 41]. Recently, there has been a growing interest in additive models in high dimensions, which generalizes high-dimensional linear regression. Much progress has been made to understand the prediction performance of various proposals, including [23, 26, 30, 31, 37, 38, 42, 43]. However, the statistical inference problem in the high-dimensional additive model is far less understood from both methodological and theoretical perspectives.

Throughout the paper, we consider a general form of the additive model,

$$y_i = f_1(X_{i1}) + f_2(X_{i2}) + \epsilon_i, \text{ for } 1 \leq i \leq n, \quad (1)$$

where  $X_{i1} \in \mathbb{R}$ ,  $X_{i2} \in \mathbb{R}^p$ ,  $f_1 : \mathbb{R} \rightarrow \mathbb{R}$  and  $f_2 : \mathbb{R}^p \rightarrow \mathbb{R}$  are unknown functions and  $\epsilon_i$  is the regression error with mean zero and variance  $\sigma_1^2$ . The observed data  $(y_i, X_{i1}, X_{i2}^\top)$ ,  $1 \leq i \leq n$ , are assumed to be i.i.d. Here, the variable  $X_{i1}$  is singled out to represent a generic variable of interest and  $X_{i2}$  denotes the set of all other variables collected for the data analysis. Typically, in scientific studies, the variable of interest is determined by the scientific goal, for example, a given treatment, exposure to a certain dose level or an economic or climate factor. We treat  $X_{i2}$  as the collection of all other observed variables that are possibly associated with the outcome variable. As a remark,  $X_{i2}$  can be univariate, multivariate, or high dimensional, and in the case that  $X_{i2}$  is high dimensional, additional additive conditions may also be imposed on the generic  $f_2$ . We adopt the additive model in the general form (1) to include both low- and high-dimensional  $X_{i2}$ . In terms of terminology, we shall refer to  $X_{i1}$  and  $f_1$  as the variable and function of interest, respectively, and to  $X_{i2}$  and  $f_2$  as nuisance variables and function.

The current paper is focused on the statistical inference problem for  $f_1'(x_0)$ , the derivative of the function of interest at a local point  $x_0$ . In the following, we shall provide some motivating examples for our study.

**Effect to exposure.** In observational studies, a major concern is the existence of unmeasured confounders, which are associated with both the variable of interest (exposure variable) and the outcome. To address this, a commonly used method is to condition on certain measured confounders so that the exposure variable is plausibly exogenous as in ran-

domized trials. Since it is challenging to know which exact measured confounders should be conditioned on to achieve this goal, a large number of measured confounders are conditioned but only some of them are associated with the outcome [3]. Such a general idea has been carefully investigated in the framework of linear regression, under the specific assumption that the exposure has a linear effect on the outcome. However, nonlinear effects to exposure have been commonly observed in scientific studies, including return to schooling [7], climate on crop yields [33] and the climate change on the economic outcomes [12,13]. As a relaxation of the linear effect of the exposure variable, the additive model (1) captures the non-linear effects by treating the function of interest  $f_1$  as an unknown smooth function. In such a general model, the effect of exposure  $X_{i1}$  at a pre-specified level  $x_0$  can be captured by the rate of change  $(f_1(x_0 + a) - f_1(x_0)) / a$  for a small  $a$ . With  $a$  approaching zero, the effect of exposure can be captured by the first order derivative  $f_1'(x_0)$ . Instead of assuming a constant effect, the exposure effect  $f_1'(\cdot)$  depends on the value of the exposure variable  $x_0$ . Such a definition of treatment effect has been introduced in [2]. More generally, we allow a non-linear relationship between the outcome and the nuisance variables.

**Location of extreme values.** Another important motivation for studying the first order derivative is to locate the extreme value of the function of interest  $f_1$ . The location of maximum values have found many applications in different industries, including identifying the extrema of profile expressions in genetic studies [29,34] and searching for the range of burden distribution indices of blast furnace to optimize the iron extraction from large quantities of iron-bearing materials [8,9]. Under the model formulation in (1), we can check whether  $x_0$  is a local extreme values of  $f_1$  via the hypothesis testing problem  $H_0 : f_1'(x_0) = 0$ . The testing procedure developed in this paper would be useful in locating the extreme value of  $f_1$  in the presence of nuisance covariates  $X_{i2}$ .

Despite the usefulness of making statistical inference for  $f_1'(x_0)$ , there is a lack of methods and theoretical justification for the problem under the additive model (1), especially when the nuisance variables  $X_{i2}$  are of high-dimension. The following section will discuss the challenges of this inference problem in additive models and also provide an overview of the proposed method from both methodological and theoretical perspectives.

## 1.1 Results and Contributions

Inference for the function derivative has been carefully studied in the classical nonparametric regression [15–17, 28, 45]. However, inference for the derivative of one component  $f_1$  in the additive model is a more challenging problem due to the fact that we have to estimate the unknown nuisance function  $f_2$  without a direct observation of the function of interest  $f_1$ . To illustrate this, we use  $\hat{f}_2$  to denote a reasonably good estimator of  $f_2$  and

then calculate the residual  $R_i = y_i - \hat{f}_2(X_{i2})$  for  $i = 1, 2, \dots, n$ . A natural idea is to use this residual as a proxy outcome for  $f_1(X_{i1})$  and apply the classical method developed in nonparametric regression to the pair of new observations  $(X_{i1}, R_i)$ . However, such simple plug-in methods are problematic as it directly inherits the error  $\hat{f}_2 - f_2$  of estimating  $f_2$ .

As a remedy, we propose a decorrelated local linear (DLL) estimator to reduce the bias inherited from estimating the nuisance function  $f_2$ . In nonparametric regression where  $f_2 = 0$ , [15] has shown that the local linear estimator of the derivative can be expressed as a ratio of two weighted sums, a weighted sum of the outcome over a weighted sum of the covariate of interest with the same weights. See (2) for the exact form. To account for the error of estimating the nuisance function  $f_2$ , the DLL estimator uses certain weights which are correlated with  $X_{i1} \in \mathbb{R}$  but not  $X_{i2} \in \mathbb{R}^p$ , at least approximately. These weights are referred to as “decorrelated weights” to reflect the fact that they enjoy the “decorrelation property”, that is, they are (nearly) uncorrelated with the difference between  $\hat{f}_2$  and  $f_2$ . As a result, if we treat  $(X_{i1}, R_i)$  as the observed data, the decorrelated weights would be particularly useful in reducing the bias inherited from estimating the nuisance function  $f_2$ .

To provide theoretical justifications for the proposed method, we establish the rate of convergence for estimating  $f'_1(x_0)$  by decomposing the estimation error into three errors, stochastic error, approximation error and nuisance error. The stochastic error is shown to have an asymptotic normal limiting distribution after rescaling and approximation error and nuisance error represent the random error of approximating the nonparametric by a linear function at a local neighborhood and estimating the nuisance function, respectively. While the stochastic error and approximation error are of the same order of magnitude as that in the classical nonparametric regression setting [16], the nuisance error captures the additional difficulty induced by the presence of the nuisance function  $f_2$ . The nuisance error is determined by two factors, 1) statistical accuracy of estimating the conditional expectation of certain functions of  $X_{i1}$  given other nuisance variables  $X_{i2}$ ; and 2) statistical accuracy of estimating the nuisance function  $f_2$ . In the ideal case where both the conditional expectation and the nuisance function are estimated with sufficient accuracy, the stochastic error dominates the nuisance error and we can establish the asymptotic limiting distribution for the proposed DLL estimator. Based on this asymptotic limiting distribution, we construct confidence interval for  $f'_1(x_0)$  and test for the hypothesis  $H_0 : f'_1(x_0) = 0$ .

We observe two interesting phenomenons in our theoretical study. First, accurate estimation of the conditional distribution of the variable of interest  $X_{i1}$  given the nuisance variables  $X_{i2}$  is crucial for statistical inference of the derivative of the single component  $f'_1(x_0)$  in our approach. In the most extreme case where the conditional distribution is known, the proposed DLL estimator is asymptotically normal as long as we start with any consistent estimator  $\hat{f}_2$  of  $f_2$ . Thus, the required property for the initial estimator  $\hat{f}_2$  is

significantly weakened due to the prior knowledge of this conditional distribution. More generally, the more accurately we can estimate the conditional distribution of  $X_{i1}$  given  $X_{i2}$ , the less stringent the accuracy we require in the initial estimation of  $f_2$ , and vice versa, and the nuisance error of the DLL estimator converges to zero at a faster rate than either the rate in estimating the conditional distribution or the rate in estimating  $f_2$ . We shall refer to this synergy of estimation accuracy as *double estimation accuracy* as it is closely related to the double robustness [1, 32] in causal inference, cf. Section 4.

The second interesting phenomenon is about the sample size requirement for valid inference in terms of model complexity parameters such as sparsity and smoothness level. In the high-dimension setting, constructing the confidence interval/set typically requires much stronger conditions than the corresponding estimation problem, due to the fact that the confidence interval requires not only a good estimator as the center but also consistent quantification of the uncertainty for this center. These additional conditions will be referred to as *uncertainty-quantification* conditions as they are sufficient conditions only imposed for conducting statistical inference beyond a point estimator. In the high-dimensional sparse linear regression, the *uncertainty-quantification* condition for a single regression coefficient  $\beta_i$  has been imposed in [22, 39, 44] and this condition has shown to be necessary for adaptive inference for a single regression coefficient in [5]. As a comparison, we consider a special case of the general additive model (1), the high-dimensional sparse additive model. Surprisingly, we observe that the *uncertainty-quantification* condition can be weakened even if we are considering the more complex additive model. In contrast to the high-dimensional linear regression, the nonlinearity structure imposed by the additive model increases both the magnitudes of the stochastic error and the nuisance error. The striking phenomenon of a weaker *uncertainty-quantification* condition in the additive model happens because of a careful decorrelation step through parametric modeling of the relationship between the variable of interest and the nuisance variables. More specifically, the proposed decorrelation step leads to smaller increase in the nuisance error than the increase in the stochastic error when the model becomes more complex.

In summary, the contribution of the current paper is two-folded:

1. We develop statistical inference for the derivative of a component of interest in the general additive models by introducing a new decorrelation step to reduce the error inherited from estimating the nuisance function. This decorrelation idea is of independent interest in extending the classical nonparametric regression techniques to other inference problems in the additive model.
2. We carry out a rigorous theoretical investigation of the proposed estimator and establish the rate of convergence for estimating the derivative of the component function of interest. We have identified a set of sufficient conditions for establishing the asymp-

otic limiting distribution. The theoretical analysis has revealed the phenomenon of *double estimation accuracy* as a synergetic effect of the accuracies in the estimation of the conditional distribution given the nuisance variable and the estimation of the nuisance function.

## 1.2 Literature Review and Comparison

Inference for function derivative has been actively studied in the nonparametric modeling, including local linear estimator [15], regression spline [45], kernel methods [17], empirical likelihood based methods [28] and others cited therein. However, as we have discussed, the presence of unknown nuisance function in the additive model poses great challenges to statistical inference for the function derivative at a local point. There are also significant recent progresses in studying the high-dimensional additive models [23, 26, 37, 38], but the main focus there is the prediction accuracy instead of statistical inference.

There is a recent line of active research on statistical inference in high-dimensional linear regression. Debiased estimators were developed in [44], [22] and [39] to study the inference problem for a single regression coefficient  $\beta_i$ . While the linear model can be viewed as a special case of the additive model, where the regression coefficient  $\beta_1$  represents the function derivative, that is,  $\beta_1 = f'_1(x_0)$ , the inference problem in the additive model is much more challenging. Specifically, novel methodology is required to address the nonlinearity, and both the rate of convergence and the sufficient conditions for confidence interval construction are quite different from those established in the high-dimensional linear regression. Beyond the high-dimensional linear regression, [10] and [46] studied the inference procedure to the partial linear model. However, the focus is still on the inference problem for the linear component, instead of the nonparametric component addressed here.

Two of the most relevant works to the current paper are [24] and [18]. Specifically, [24] considered the high-dimensional sparse additive model and constructed confidence bands for one component of the additive model. The method proposed in [24] is to approximate the nonparametric function by a set of basis functions and then apply the debiasing method for the corresponding linear model of the basis functions. Regarding [18], a two-step procedure was developed, where in the first step, a pre-smoothing estimator was obtained for each component by applying the group-Lasso penalization together with debiasing technique developed for high-dimensional regression; in the second step, the pre-smoothing estimator is taken as the proxy outcome and standard univariate nonparametric technique was then applied. These two related works in high-dimensional sparse additive models either focused on different problems or proposed different methods for the related statistical inference problem. Additionally, although the outcome model considered in the current paper and

these two papers [18, 24] are closely related, there is a significant difference in terms of modeling the relationship between the variable of interest and the nuisance variables. The current paper is imposing a parametric relationship or known general relationship while the relationship is modeled in nonparametric frameworks in [18, 24]. A careful utilization of the parametric model assumption leads to a significant relaxation of the sample size condition required for confidence interval construction, which is much weaker than those imposed in [18, 24]; See Section 5.3 for details.

### 1.3 Paper Organization and Notations

In Section 2, we introduce the decorrelated local linear estimator; In Section 3, we establish the theoretical guarantee of the proposed estimator; In Section 4, we present results in the case where additional information is available about the conditional distribution of the variable of interest given nuisance variables; In Section 5, we consider the high-dimensional sparse additive model as a special case; In Section 6, we provide conclusion and discussion; In Section 7, we provide the technical analysis to illustrate the effect of decorrelation. Additional proofs are presented as supplementary materials.

**Notations.** For a sequence of random variables  $X_n$  indexed by  $n$ , we use  $X_n \xrightarrow{p} X$  and  $X_n \xrightarrow{d} X$  to represent that  $X_n$  converges to  $X$  in probability and in distribution, respectively. For a sequence of random variables  $X_n$  and numbers  $a_n$ , we define  $X_n = o_p(a_n)$  if  $X_n/a_n$  converges to zero in probability and  $X_n = O_p(a_n)$  if for every  $c > 0$ , there exists a finite constant  $C$  such that  $\mathbf{P}(|X_n/a_n| \geq C) \leq c$ . We use  $c$  and  $C$  to denote generic positive constants that may vary from place to place. For two positive sequences  $a_n$  and  $b_n$ ,  $a_n \lesssim b_n$  means  $a_n \leq Cb_n$  for all  $n$  and  $a_n \gtrsim b_n$  if  $b_n \lesssim a_n$  and  $a_n \asymp b_n$  if  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ , and  $a_n \ll b_n$  if  $\limsup_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$  and  $a_n \gg b_n$  if  $b_n \ll a_n$ .

## 2 Inference in Additive Models

We review the local polynomial method in Section 2.1 and then we propose the Decorrelated Local Linear (DLL) estimator for  $f'_1(x_0)$  in Section 2.2. In Section 2.3, we construct a confidence interval for  $f'_1(x_0)$  using the DLL estimator as the center and also a solution to the hypothesis testing problem related to identifying the extreme value of  $f_1$ .

## 2.1 Local Polynomial: A Review

In classical (univariate) nonparametric regression, the local polynomial estimator has been developed for analyzing the data  $\{(y_i, X_{i1})\}_{1 \leq i \leq n}$  generated in the following model,

$$y_i = f_1(X_{i1}) + \epsilon_i,$$

where  $f_1 : \mathbb{R} \rightarrow \mathbb{R}$  is an unknown function belonging to a certain class of smooth functions. The main idea can be seen by taking a Taylor expansion of  $f_1(x)$  near  $x_0$ ,

$$f_1(x) = \sum_{l=0}^L \beta_l \psi_l(x) + r_L(x) \quad \text{for } x_0 - h \leq x \leq x_0 + h,$$

where  $\beta_l = f_1^{(l)}(x_0)/l!$ ,  $\psi_l(x) = (x - x_0)^l$ , and  $r_L(x)$  is the remainder term. We consider the above expansion with  $L = 1$  and assume that  $f_1''(x)$  is continuous at  $x_0$ . Define the kernel function  $k(x) = \mathbf{1}(|x| \leq 1)$ , which satisfies the following properties:  $\int k(x)dx = 2$ ,  $\int xk(x)dx = 0$  and  $\int x^2k(x)dx = \frac{1}{3}$ . Given a bandwidth  $h > 0$ , we define a rescaled kernel function

$$K_h(x) = \frac{1}{h}k\left(\frac{x - x_0}{h}\right) = \begin{cases} 1/h & \text{if } |x - x_0| \leq h \\ 0 & \text{otherwise} \end{cases}$$

The local linear estimator of  $f_1'(x_0)$  [11, 14, 15, 35] is of the form

$$\frac{\sum_{i=1}^n W_i y_i K_h(X_{i1})}{\sum_{i=1}^n W_i (X_{i1} - x_0) K_h(X_{i1})} \quad (2)$$

where  $W_i = (X_{i1} - x_0) - \frac{\sum_{i=1}^n (X_{i1} - x_0) K_h(X_{i1})}{\sum_{i=1}^n K_h(X_{i1})}$ . The main intuition here is that instead of using the whole data  $\{y_i, X_{i1}\}_{1 \leq i \leq n}$ , we select a subset of the data whose corresponding  $X_{i1}$  values are within a small neighborhood of  $x_0$ . For this selected subset of data, the relationship between  $y_i$  and  $X_{i1}$  can be viewed as an approximate linear regression due to the Taylor expansion. As a result, the form of estimator in (2) can be achieved by applying the standard linear regression argument, where  $W_i$  is computed as centered  $X_{i1} - x_0$  by the weighted average with weights  $\{K_h(X_{i1})\}_{1 \leq i \leq n}$ .

The bandwidth in the kernel function  $K_h(x)$  is useful in deciding the effective sample size, which measures the number of the selected data points with  $X_{i1}$  falling into the interval  $[x_0 - h, x_0 + h]$ . For the case that the marginal density function  $\pi$  for  $X_{i1}$  is continuous near  $x_0$  and has a positive marginal density  $\pi(x_0)$ , the expected number of observations falling into  $[x_0 - h, x_0 + h]$  is

$$\mathbf{E}|\{1 \leq i \leq n : x_0 - h \leq X_{i1} \leq x_0 + h\}| = n \cdot \int_{x_0-h}^{x_0+h} \pi(x)dx \approx 2nh \cdot \pi(x_0), \quad (3)$$

where  $|\mathcal{A}|$  of a set  $\mathcal{A}$  denotes the set cardinality. That is to say, although we have a total of  $n$  observations, only part of the data, with the size  $2nh \cdot \pi(x_0)$ , is effective in estimating the first order derivative due to the non-linearity of the function.



## 2.2 Decorrelated Local Linear Estimator

Although the local polynomial estimator has been proven to enjoy both methodological and theoretical advantages in nonparametric regression, it is challenging to extend the local linear estimator to the additive model in the presence of the nuisance function  $f_2(X_{i2})$ . In the following, we propose the DLL estimator to address the additional challenges through a novel method of decorrelating the weights in (2).

The DLL estimator of  $f'_1(x_0)$  is constructed in two steps. The first step is to obtain a certain good initial estimator of the nuisance function  $\hat{f}_2$ . To highlight the main idea, we assume in the current section that we have some reasonably good initial estimator  $\hat{f}_2$  of  $f_2$ , and we will specify the exact requirements for such a good estimator in Section 3. These requirements are compatible with a large class of initial estimators  $\hat{f}_2$ , which have been proposed in the literature in both low- and high-dimensional additive models. In Section 5, we focus on the high-dimensional sparse additive model and show that certain existing estimators in the literature are sufficient for our use in the high-dimensional setting.

The focus of the following discussion is on the second step, that is the construction of an accurate estimator of  $f'_1(x_0)$  by utilizing the initial estimator  $\hat{f}_2$  from the first step. We compute the residual of outcome variable after adjusting for the estimator  $\hat{f}_2$ ,

$$R_i = y_i - \hat{f}_2(X_{i2}) = f_1(X_{i1}) - (\hat{f}_2(X_{i2}) - f_2(X_{i2})) + \epsilon_i. \quad (4)$$

In contrast to the univariate regression, the above residual form has an additional term  $\hat{f}_2(X_{i2}) - f_2(X_{i2})$ , which is the error of the data-dependent estimator for  $f_2$ . This additional error term may bias a direct application of the local linear estimator to  $(X_{i1}, R_i)$  in the sense that the additional error would be carried over in the estimation of the first order derivative  $f'_1(x_0)$  and this carried-over error may blow up the final estimation error of the local linear estimator proposed in (2). This motivates us to develop new methods to take this additional term into consideration. To motivate our propose estimator, we introduce a generic estimator of  $f'_1(x_0)$  in the following form,

$$\frac{\frac{1}{n} \sum_{i=1}^n D_{i1} R_i K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})} \quad (5)$$

where the weights  $\{D_{i1}\}_{1 \leq i \leq n}$  are to be specified. As a comparison to the local linear estimator (2), we replace the outcome  $y_i$  with the residual variable  $R_i$  and the weights  $W_i$  with the generic weights  $D_{i1}$ .

The next main step is to construct the weights  $D_{i1}$  such that the proposed estimator enjoys similar properties as the local linear estimator defined in (2) while at the same time reduces the error due to estimating the nuisance function  $f_2$  as much as possible. More

explicitly, we decompose the estimation error of the estimator defined in (5) as follows,

$$\begin{aligned} & \frac{\frac{1}{n} \sum_{i=1}^n D_{i1} R_i K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})} - f'_1(x_0) \\ = & \frac{\frac{1}{n} \sum_{i=1}^n D_{i1} [f_1(x_0) + r_1(X_{i1}) + \epsilon_i] K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})} + \frac{\frac{1}{n} \sum_{i=1}^n D_{i1} (\hat{f}_2(X_{i2}) - f_2(X_{i2})) K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})}. \end{aligned}$$

Regarding the above decomposition, the first term on the righthand side is the same as the error in the standard univariate local linear estimator while the second term is due to the estimation error of all other nuisance functions expressed as  $f_2$ . In the construction of  $D_{i1}$ , we need to achieve the following three goals simultaneously,

- (i) **Stochastic error:** Construct  $D_{i1}$  such that compared with the classical local linear estimator, the stochastic error  $\frac{\frac{1}{n} \sum_{i=1}^n D_{i1} \epsilon_i K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})}$  is not inflated.
- (ii) **Approximation error:** Construct  $D_{i1}$  such that the numerator of the approximation error,  $\frac{1}{n} \sum_{i=1}^n D_{i1} [f_1(x_0) + r_1(X_{i1})] K_h(X_{i1})$ , is of a small order of magnitude than that of the stochastic error.
- (iii) **Nuisance function error:** Construct  $D_{i1}$  such that the numerator of the nuisance error,  $\frac{1}{n} \sum_{i=1}^n D_{i1} (\hat{f}_2(X_{i2}) - f_2(X_{i2})) K_h(X_{i1})$ , is also of a small order of magnitude than that of the stochastic error.

Note that goal (ii) is to make sure that the linear approximation is accurate near the neighborhood of  $x_0$  and goal (iii) is to make sure that the estimation error from  $\hat{f}_2$  vanishes at a sufficiently fast speed. Goals (i) and (ii) are satisfied for the standard local linear estimator defined in (2) while goal (iii) is more challenging to achieve simultaneously.

Since goal (ii) is relatively easy to achieve as long as  $f_1$  is smooth and  $D_{i1}$  is empirically centered, we first consider goals (i) and (iii). To this end, we focus on the generic form

$$D_{i1} = (X_{i1} - x_0) - e(X_{i2})$$

where  $e(X_{i2})$  is a function of  $X_{i2}$ . Regarding goal (iii), we construct  $D_{i1}$  such that

$$\mathbf{E}(D_{i1} \Delta(X_{i2}) K_h(X_{i1}) | X_{i2}) = 0, \quad \text{for any function } \Delta : \mathbb{R}^p \rightarrow \mathbb{R}. \quad (6)$$

If  $\Delta$  is taken as  $\hat{f}_2 - f_2$  and  $(X_{i1}, X_{i2}^\top, y_i)$  is not used to construct  $\hat{f}_2$ , (6) implies

$$\mathbf{E} D_{i1} (\hat{f}_2(X_{i2}) - f_2(X_{i2})) K_h(X_{i1}) = 0. \quad (7)$$

We would refer this to as the *decorrelation property* of the weights  $D_{i1}$ . With this property, the empirical sum  $\frac{1}{n} \sum_{i=1}^n D_{i1} (\hat{f}_2(X_{i2}) - f_2(X_{i2})) K_h(X_{i1})$  would vanish at a fast rate due to the standard concentration result. A sufficient condition to guarantee (6) is

$$\mathbf{E}(D_{i1} K_h(X_{i1}) | X_{i2}) = \mathbf{E}([X_{i1} - x_0 - e(X_{i2})] K_h(X_{i1}) | X_{i2}) = 0.$$

Through solving the above equation, we obtain the closed form of the function  $e(X_{i2})$  as

$$e(X_{i2}) = \frac{\mathbf{E}([X_{i1} - x_0]K_h(X_{i1})|X_{i2})}{\mathbf{E}(K_h(X_{i1})|X_{i2})}. \quad (8)$$

Then we identify the following expression of the variable  $D_{i1}$ ,

$$D_{i1} = (X_{i1} - x_0) - \frac{\mathbf{E}([X_{i1} - x_0]K_h(X_{i1})|X_{i2})}{\mathbf{E}(K_h(X_{i1})|X_{i2})}. \quad (9)$$

When the conditional distribution of  $X_{i1}$  given  $X_{i2}$  is known or can be well estimated, we can compute the  $D_{i1}$  in (9) with these available information. More detailed discussion and theoretical justification regarding this setting will be provided in Section 4.

A more challenging setting is that we need to estimate the unknown conditional distribution of  $X_{i1}$  given  $X_{i2}$  using the given data. To study this, we borrow the strength of approximating this conditional distribution by utilizing certain model assumption for the relationship between  $X_{i1}$  and  $X_{i2}$ . Specifically, we expand the variable of interest  $X_{i1}$  as a sum of its population linear projection to the other covariates  $X_{i2}$  and an error term,

$$X_{i1} = X_{i2}^\top \gamma + \delta_i,$$

where  $\gamma = [\mathbf{E}(X_{i2}X_{i2}^\top)]^{-1}\mathbf{E}(X_{i2}X_{i1})$  and  $\sigma_2^2 = \text{Var}(\delta_i)$ . We assume that the error  $\delta_i$  is independent of  $X_{i2}$  and the normalized error  $\delta_i/\sigma_2$  has the density function  $\phi(t)$ . We then derive the following explicit expression for  $e(X_{i2})$  in (8),

$$e(X_{i2}) = \sigma_2 \frac{\int_{\mu_i - L_i}^{\mu_i + L_i} (t - \mu_i) \phi(t) dt}{\int_{\mu_i - L_i}^{\mu_i + L_i} \phi(t) dt} := l(X_{i2}, \gamma, \sigma_2) \quad (10)$$

where  $\mu_i = (x_0 - X_{i2}^\top \gamma)/\sigma_2$  and  $L_i = h/\sigma_2$ .

By further assuming the error  $\delta_i$  to follow a Gaussian distribution, we simplify  $l(X_{i2}, \gamma, \sigma_2)$  defined in (10) as

$$\begin{aligned} l(X_{i2}, \gamma, \sigma_2) &= \sigma_2 \frac{\int_{-L_i}^{L_i} t(1 - t\mu_i - t^2/2 + \mu_i^2 t^2/2) dt + O_p(h^5(\log n)^{3/2})}{\int_{-L_i}^{L_i} (1 - t\mu_i) dt + O_p(h^3 \log n)} \\ &= \frac{h^2}{3\sigma_2^2} (X_{i2}^\top \gamma - x_0) + O_p(h^4(\log n)^{3/2}) \end{aligned} \quad (11)$$

In this expression, the Gaussian assumption of  $\delta_i$  is used here to provide a simplified expression for the function  $l(X_{i2}, \gamma, \sigma_2)$ . Since the dominating part in this expression is linear in  $\gamma$  and also  $X_{i2}^\top \gamma - x_0$ , we refer to the above expression as the *linear approximation*. This is the main place that we make use of the Gaussian error assumption. As a remark, the decorrelation method can be applied using the expression in (10) even without this Gaussian error assumption; There is much room to relax the Gaussian error assumption as we essentially only require a good approximation by a linear function in  $\gamma$ , as in (11).

With some reasonably good estimator  $(\hat{\gamma}, \hat{\sigma}_2)$  for the parameters  $(\gamma, \sigma_2)$ , we estimate  $l(X_{i2}, \gamma, \sigma_2)$  by  $l(X_{i2}, \hat{\gamma}, \hat{\sigma}_2)$  and then estimate  $D_{i1}$  by  $\tilde{D}_{i1} = (X_{i1} - x_0) - l(X_{i2}, \hat{\gamma}, \hat{\sigma}_2)$ . For the case of Gaussian error or the linear approximation in (11) holds, we can estimate  $l(X_{i2}, \gamma, \sigma_2)$  by  $\frac{h^2}{3\hat{\sigma}_2^2} (X_{i2}^\top \hat{\gamma} - x_0)$  and then estimate  $D_{i1}$  by

$$\tilde{D}_{i1} = (X_{i1} - x_0) - \frac{h^2}{3\hat{\sigma}_2^2} (X_{i2}^\top \hat{\gamma} - x_0). \quad (12)$$

To achieve goal (ii) for controlling the approximation error, we propose an additional “centering” step in construction of the decorrelated weights  $\hat{D}_{i1}$ ,

$$\hat{D}_{i1} = \tilde{D}_{i1} - \frac{\frac{1}{n} \sum_{i=1}^n \tilde{D}_{i1} K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})}, \quad (13)$$

so that the weights  $\{\hat{D}_{i1}\}_{1 \leq i \leq n}$  are empirically centered with respect to the kernel  $K_h(\cdot)$ .

Two remarks about the decorrelation step are in order. First, from a higher perspective, we work on the inference problem in a semi-parametric framework. Specifically, the outcome model is assumed to be in the general additive form while the relationship model between  $X_{i1}$  and  $X_{i2}$  is treated with a parametric model to decouple the relationship between the covariates. The corresponding parametric modeling assumption of the error  $\delta_i$  is mainly used to provide an approximation of the function  $e(X_{i2})$  defined in (8) by a simple form, for example, the linear approximation in (11) in the case of Gaussian error  $\delta_i$ .

Second, when  $X_{i2}$  is univariate or of low dimension, classical nonparametric density estimator can be used to estimate the density of  $\phi(\cdot)$  and hence  $e(X_{i2})$  or  $l(X_{i2}, \gamma, \sigma_2)$  in (10). However, if  $X_{i2}$  is of high dimension, it is in general a challenging problem to estimate the conditional expectation  $\mathbf{E}([X_{i1} - x_0]K_h(X_{i1})|X_{i2})$  and  $\mathbf{E}(K_h(X_{i1})|X_{i2})$  without additional modeling assumption between  $X_{i1}$  and  $X_{i2}$ . Since we are interested in a general theory for additive models, including both low- and high-dimensional  $X_{i2}$ , we introduce this additional parametric model to provide a unified treatment. More interestingly, our analysis reveals that a careful decorrelation procedure making use of the parametric assumption on the conditional distribution of  $X_{i1}$  given  $X_{i2}$  would significantly weaken the sample size requirement. See Section 5.3 for details.

### 2.3 Point and Interval Estimators

By combining the generic estimator defined in (5) and estimator  $\hat{D}_{i1}$  defined in (13), we propose our final estimator for  $f'_1(x_0)$  as

$$\widehat{f'_1(x_0)} = \frac{1}{n\hat{S}_n} \sum_{i=1}^n \hat{D}_{i1} R_i K_h(X_{i1}) \quad \text{where} \quad \hat{S}_n = \frac{1}{n} \sum_{i=1}^n \hat{D}_{i1} (X_{i1} - x_0) K_h(X_{i1}). \quad (14)$$

As mentioned earlier, we refer to our estimator as Decorrelated Local Linear (DLL) estimator. In Section 3, we will specify a set of required conditions for the initial estimators  $\widehat{f}_2, \widehat{\gamma}, \widehat{\sigma}_2$  and provide a careful theoretical analysis of this estimator.

The construction of confidence interval directly follows from the asymptotic limiting distribution for the estimator  $\widehat{f'_1(x_0)}$  in (14) together with a consistent estimator of the variance level. Denote by  $\widehat{\sigma}_1^2$  a consistent estimator of the variance of  $\epsilon_i$  in the additive model (1). We estimate the variance of the proposed DLL estimator  $\widehat{f'_1(x_0)}$  in (14) by

$$\widehat{V} = \frac{\widehat{\sigma}_1^2}{n^2 \widehat{S}_n^2} \sum_{i=1}^n \widehat{D}_{i1}^2 K_h^2(X_{i1}).$$

This leads to the following  $1 - \alpha$  confidence interval for  $f'_1(x_0)$ ,

$$\text{CI}_{x_0} = \left( \widehat{f'_1(x_0)} - z_{\alpha/2} \widehat{V}, \widehat{f'_1(x_0)} + z_{\alpha/2} \widehat{V} \right) \quad (15)$$

where  $z_{\alpha/2}$  is the upper  $\alpha/2$  quantile of the standard normal distribution. We can also conduct the hypothesis testing for  $H_0 : f'_1(x_0) = 0$  with the following testing procedure

$$\psi_{x_0} = \mathbf{1} \left( |\widehat{f'_1(x_0)}| \geq z_{\alpha/2} \widehat{V} \right). \quad (16)$$

Theoretical justifications for the confidence interval in (15) and hypothesis testing procedure in (16) are provided in the next section.

### 3 Theory for Additive Models

In this section, we present the theoretical justification for the statistical inference based on the proposed  $\widehat{f'(x_0)}$ . In Section 3.1, we describe some regularity conditions and the concept of *double estimation accuracy* as briefly mentioned earlier. After this, we present several important intermediate results for studying the asymptotic limiting distribution of the proposed estimator, including the estimation accuracy for the weights in Section 3.2 and the bias-variance tradeoff with the proposed estimator  $\widehat{f'(x_0)}$  in Section 3.3. Most interestingly, in Section 3.4, we carefully characterize how the decorrelated weights  $\widehat{D}_{i1}$  help reduce the estimation error inherited from estimating the nuisance function  $f_2$ . These technical results can be of independent interest for studying related inference problems in additive models. In Section 3.5, we present detailed properties of the DLL estimator  $\widehat{f'(x_0)}$  and justify the validity of the related confidence interval and hypothesis testing procedures.

#### 3.1 Conditions and Estimation Accuracy

We first introduce the conditions (A1) – (A2) for the statistical modeling,

(A1) The additive model (1) holds with  $f_1''(x)$  being continuous at  $x_0$  and  $\mathbf{E}|\epsilon_i|^{2+\tau} \leq C$  for some constants  $\tau > 0$  and  $C > 0$ .

(A2) The bandwidth satisfies  $nh\pi(x_0) \rightarrow \infty$  and  $hC_u \rightarrow 0$ , where  $\pi(x_0)$  denotes the marginal probability density of  $X_{i1}$  at  $x_0$  and

$$C_u = \frac{1}{\sigma_2} \left( x_0 + h + \|\gamma\|_2 \sqrt{\log n} \right). \quad (17)$$

Conditions (A1) requires the additive model structure along with a mild moment condition on the error term. In addition, Conditions (A1) imposes the smoothness condition on the function  $f_1$  such that the approximation error of  $f_1$  by a local linear estimator is negligible in comparison to the stochastic error when the bandwidth  $h$  is of the order  $(n\pi(x_0))^{1/5}$ . Here, we are not imposing specific smoothness and other conditions on the nuisance function  $f_2$  as the description of these specific conditions is deferred to Section 5 where such conditions are used to provide error bounds for suitable estimators  $\hat{f}_2$ . As pointed out in (3), the expected number of observations in the local neighborhood  $[x_0 - h, x_0 + h]$  of  $x_0$  is  $nh\pi(x_0)$ . Condition (A2) requires that there are (asymptotically) infinitely many observations in the local neighborhood of  $x_0$  with bandwidth  $h$ . The condition  $hC_u \rightarrow 0$  is mild since  $h$  is usually set as  $n^{-c}$  for some positive constant  $c > 0$  and  $C_u$  is of the order  $\sqrt{\log n}$ . Both Conditions (A1) and (A2) are regularity conditions needed for analyzing the local linear estimator in univariate nonparametric regression. Our results are established under more general conditions for  $\pi(x_0)$  than the standard nonparametric regression, where the marginal density  $\pi(x_0)$  is allowed to vanish to zero at certain rates. We shall also remark that our theoretical results are explicit in terms of keeping the dependence on marginal density  $\pi(x_0)$  but do not sharpen this dependence on  $\pi(x_0)$ .

To facilitate the discussion, we introduce accuracy measures for estimating the weights  $\{D_{i1}\}_{1 \leq i \leq n}$  and the nuisance function  $f_2$  as follows. We define  $\bar{\mu}_D$  as the weighted sample average of  $D_{i1}$ ,  $\bar{\mu}_D = \frac{\frac{1}{n} \sum_{i=1}^n D_{i1} K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})}$ . We use  $\text{Err}(\hat{D})$  to denote the accuracy measure of estimating  $D_{i1}$ , defined as

$$\text{Err}(\hat{D}) = \sqrt{\frac{1}{n} \sum_{i=1}^n \left( \hat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right)^2 K_h(X_{i1})}. \quad (18)$$

Specifically,  $\text{Err}(\hat{D})$  measures the average accuracy of  $\hat{D}_{i1}$  with the corresponding kernel weights  $K_h(X_{i1})$ . In addition to the estimation accuracy of  $\hat{D}_{i1}$ , we define  $\text{Err}(\hat{f}_2)$  as the estimation error for the nuisance function as follows

$$\mathbf{P} \left( \sqrt{\mathbf{E}_{X_{0,2}} (\hat{f}_2(X_{0,2}) - f_2(X_{0,2}))^2} > \text{Err}(\hat{f}_2) \right) \leq \gamma(n),$$

where  $\gamma(n) \rightarrow 0$  and  $X_{0,2}$  is an independent copy to the i.i.d. data  $\{X_{i,2}\}_{1 \leq i \leq n}$  used to produce  $\hat{f}_2$ . Note that  $\mathbf{E}_{X_{0,2}}(\hat{f}_2(X_{0,2}) - f_2(X_{0,2}))^2$  denotes the expectation taken with respect to the independent copy  $X_{0,2}$ ; The outside probability is with respect to the randomness of the estimator  $\hat{f}_2$ .

We will show in Section 3.5 that the estimation accuracies as measured by  $\text{Err}(\hat{f}_2)$  and  $\text{Err}(\hat{D})$  jointly determine the theoretical performance of the proposed estimator.

### 3.2 Estimation Accuracy for $D_{i1}$

In this section, we provide a careful study of estimating the weights  $\{D_{i1}\}_{1 \leq i \leq n}$ . Since the goal of the current paper is to estimate one component  $f_1$  instead of the summation  $f_1 + f_2$  as in the literature, we need to decouple the variable of interest  $X_{i1}$  with all other nuisance variables  $X_{i2}$ . We impose the following model assumption for the relationship between  $X_{i1}$  and  $X_{i2}$ .

(A3) In the decomposition

$$X_{i1} = X_{i2}^\top \gamma + \delta_i \quad \text{with} \quad \gamma = [\mathbf{E}(X_{i2}X_{i2}^\top)]^{-1} \mathbf{E}(X_{i2}X_{i1}) \quad (19)$$

we assume that  $\|\gamma\|_0 \leq k$  and the error  $\delta_i$  follows a centered Gaussian distribution with variance  $\sigma_2^2$  and independent of  $X_{i2}$ . Additionally, we assume  $X_{i2}$  is a Sub-gaussian random vector.

We shall provide some remarks here on this model assumption. First, the expression in (19) is valid as long as  $\mathbf{E}(X_{i2}X_{i2}^\top)$  is invertible and  $\mathbf{E}(X_{i2}X_{i1})$  exists, where both are mild conditions. The key assumptions in Condition (A3) are the parametric modeling of the error  $\delta_i$  and the sparsity of  $\gamma$ . The sparsity condition  $\|\gamma\|_0 \leq k$  is only imposed in the case where the dimension  $p$  is larger than the sample size  $n$ . In the case that  $X_{i2}$  is of low dimension, this assumption automatically holds with  $k = p$ .

The distributional part of Condition (A3) will be automatically satisfied if  $(X_{i1}, X_{i2}^\top)^\top$  follows a multivariate Gaussian distribution. The essential part of Condition (A3) is to introduce a specific parametric modeling assumption for the error  $\delta_i$ . The Gaussianity of the error  $\delta_i$  is imposed as an example of the parametric modeling but can be easily replaced with other specified parametric assumptions on  $\delta_i$ . Even the parametric modeling assumption can be further weakened as long as we can provide a good approximation to  $l(X_{i2}, \gamma, \sigma_2)$  as in (11).

We now state the specific accuracy requirements for estimating  $\gamma$  and  $\sigma_2$ , which can be achieved by applying the existing results for low- and high-dimensional linear regression.

(A4) With probability larger than  $1 - n^{-c}$ , the initial estimators  $(\hat{\gamma}, \hat{\sigma}_2)$  satisfy

$$\frac{1}{n} \sum_{i=1}^n [X_{i2}^\top (\hat{\gamma} - \gamma)]^2 K_h(X_{i1}) \lesssim \sqrt{\frac{k \log p}{n}}, \quad |\hat{\sigma}_2 - \sigma_2| \lesssim \frac{1}{\sqrt{n}} + \frac{k \log p}{n} \quad (20)$$

Under the modeling condition (A3), condition (A4) can usually be guaranteed under regularity conditions. In particular, (A4) holds in low-dimensional settings if  $\hat{\gamma}$  is the least square estimator and  $\hat{\sigma}_2$  is the variance estimator based on the residual of linear models; and in high-dimensional settings (A4) holds with the scaled Lasso estimator  $\hat{\gamma}, \hat{\sigma}_2$ . As a side remark, the weighted prediction error  $\frac{1}{n} \sum_{i=1}^n [X_{i2}^\top (\hat{\gamma} - \gamma)]^2 K_h(X_{i1})$  is typically not controlled for the penalized estimator unless we assume independence between the initial estimator  $\hat{\gamma}$  and the covariates  $X_{i2}$ . In Section 3.4, a data swap technique is used to verify that the constructed initial estimators satisfy condition (A4).

The following lemma shows that the estimation accuracy for the estimation of  $D_{i1}$  is mainly determined by those for  $\gamma$  and  $\sigma_2$ , as indicated in the condition (A4).

**Lemma 1** *Suppose that conditions (A3) and (A4) hold and  $h \|\gamma\|_2 \sqrt{\log n} \rightarrow 0$ . Then the estimation error  $\text{Err}(\hat{D})$  defined in (18) satisfies*

$$\mathbf{P} \left( \text{Err}(\hat{D}) \lesssim h^2 \sqrt{k \log p / n} + h^4 (\sqrt{\log n})^3 \right) \geq 1 - n^{-c}.$$

for some positive constants  $C, c > 0$ .

There are two terms in the estimation accuracy of  $\hat{D}$ , where  $h^2 \sqrt{k \log p / n}$  comes out of estimating the regression vector  $\gamma$  by  $\hat{\gamma}$  and the other term  $h^4 (\sqrt{\log n})^3$  results from approximating  $l(X_{i2}, \gamma, \sigma_2)$  by the linear component  $\frac{h^2}{3\sigma_2^2} (X_{i2}^\top \gamma - x_0)$ , as in (12). Though this lemma only considers the case where the linear approximation in (11) holds, we can also establish a similar result for  $\text{Err}(\hat{D})$  under a more general parametric modeling assumption on  $X_{i1} \mid X_{i2}$ , where  $l(X_{i2}, \hat{\gamma}, \hat{\sigma}_2)$  is used to estimate  $l(X_{i2}, \gamma, \sigma_2)$

### 3.3 Error decomposition

Since  $\sum_{i=1}^n \hat{D}_{i1} K_h(X_{i1}) = 0$ , the estimation error of  $\widehat{f'(x_0)}$  can be decomposed as follows,

$$\widehat{f'(x_0)} - f'(x_0) = \frac{1}{n \hat{S}_n} \sum_{i=1}^n \hat{D}_{i1} (\epsilon_i + r(X_{i1}) + \Delta(X_{i2})) K_h(X_{i1})$$

where  $r(X_{i1}) = f_1(X_{i1}) - f_1(x_0) - f'_1(x_0)(X_{i1} - x_0)$  and  $\Delta(X_{i2}) = f_2(X_{i2}) - \hat{f}_2(X_{i2})$  and  $\hat{S}_n = n^{-1} \sum_{i=1}^n \hat{D}_{i1} (X_{i1} - x_0) K_h(X_{i1})$  is as in (14). More explicitly, we decompose



$\widehat{f'(x_0)} - f'_1(x_0)$  as

$$\underbrace{\frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} \epsilon_i K_h(X_{i1})}_{\text{Stochastic Error}} + \underbrace{\frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} r(X_{i1}) K_h(X_{i1})}_{\text{Approximation Error}} + \underbrace{\frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} \Delta(X_{i2}) K_h(X_{i1})}_{\text{Nuisance Error}} \quad (21)$$

As this decomposition indicates, there exist three sources of estimation errors for the proposed estimator, denoted as “Stochastic Error”, “Approximation Error” and “Nuisance Error”. Here, “Stochastic Error” represents a random component with mean zero and, after rescaling, following an asymptotic normal limiting distribution as long as the Lindeberg condition can be verified; “Approximation Error” represents the impact of approximating the non-linear function  $f_1$  by a linear function at a local neighborhood of  $x_0$ ; “Nuisance Error” represents the error due to estimating the nuisance function by  $\widehat{f}_2$ . The first two components in the decomposition (21), the stochastic and approximation errors, also appear in classical nonparametric regression.

The following lemma establishes the limiting distribution for the stochastic error and establishes the rate of convergence for the approximation error. The rate of convergence for the nuisance error is deferred to the next subsection.

**Lemma 2** *Suppose that conditions (A1)–(A4) hold, then the approximation error satisfies*

$$\frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} \left[ r(X_{i1}) - \frac{(X_{i1} - x_0)^2}{2} f''(x_0) \right] K_h(X_{i1}) = o_p \left( \frac{\text{Err}(\widehat{D})}{\sqrt{\pi(x_0)}} + h \right) \quad (22)$$

and

$$\frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) = O_p \left( \frac{\text{Err}(\widehat{D})}{\sqrt{\pi(x_0)}} + c_n h \right) \quad (23)$$

with  $c_n = hC_u + (nh\pi(x_0))^{-1/4} \rightarrow 0$ . Additionally, if  $\text{Err}(\widehat{D}) \ll h\sqrt{\pi(x_0)}$  and  $\widehat{\gamma}$  satisfies

$$\mathbb{P} \left( \max_{1 \leq i \leq n} |X_{i2}^I(\widehat{\gamma} - \gamma)| \gtrsim \sqrt{k \log p \log n/n} \right) \rightarrow 0, \quad (24)$$

then we have

$$\frac{1}{\sqrt{V}} \frac{\sum_{i=1}^n \widehat{D}_{i1} \epsilon_i K_h(X_{i1})}{n\widehat{S}_n} \xrightarrow{d} N(0, 1) \quad (25)$$

where  $V = \frac{\sigma_1^2}{n^2 \widehat{S}_n^2} \sum_{i=1}^n \widehat{D}_{i1} K_h^2(X_{i1}) \xrightarrow{P} \frac{3}{2nh^3 \cdot \pi(x_0)} \sigma_1^2$ .

A combination of (22) and (23) establishes the order of magnitude of the approximation error  $\frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} r(X_{i1}) K_h(X_{i1})$  as  $O_p \left( \frac{\text{Err}(\widehat{D})}{\sqrt{\pi(x_0)}} \right) + o_p(h)$ . One sufficient condition for confidence intervals construction is that the stochastic error dominates the approximation error,

which is reduced to the following conditions,

$$\frac{h}{\sqrt{V}} = O_p(1) \quad \text{and} \quad \frac{\text{Err}(\widehat{D})}{\sqrt{\pi(x_0)}\sqrt{V}} = o_p(1) \quad (26)$$

By choosing the bandwidth as  $h \asymp (n\pi(x_0))^{-\frac{1}{5}}$ , we have  $h/\sqrt{V} = O_p(1)$ . Combined with Lemma 1, we can show that (26) holds with high probability and hence the approximation error is negligible in comparison to the stochastic error.

### 3.4 Analysis of Nuisance Error

In this section, we control the error of estimating the nuisance function  $f_2$ . This is the exact place where the decorrelation weights play a crucial role. The main step in controlling the nuisance error is to provide a sharp bound for the following quantity,

$$\frac{1}{n} \sum_{i=1}^n \widehat{D}_{i1} \Delta(X_{i2}) K_h(X_{i1}) \quad \text{where } \Delta(X_{i2}) = f_2(X_{i2}) - \widehat{f}_2(X_{i2}).$$

As discussed in Section 2.2, especially (7), we construct  $D_{i1}$  satisfying the *decorrelation* property. We show that the same goal will be achieved if the weights  $D_{i1}$  are estimated by  $\widehat{D}_{i1}$ . This *decorrelation* property is essential in reducing the estimation error related to the nuisance function.

To rigorously control  $\frac{1}{n} \sum_{i=1}^n \widehat{D}_{i1} \Delta(X_{i2}) K_h(X_{i1})$ , we introduce a specific version of the initial estimators  $(\widehat{f}_2, \widehat{\gamma})$  for technical reasons. Recall the places to use the initial estimators  $\widehat{f}_2$  and  $\widehat{\gamma}$ , where  $\widehat{f}_2$  is used in calculating the residual  $R_i = y_i - \widehat{f}_2(X_{i2})$  defined in (12) and  $\widehat{\gamma}$  is used in construction of weights  $\widetilde{D}_{i1} = (X_{i1} - x_0) - \frac{h^2}{3\sigma_2^2} (X_{i2}^\top \widehat{\gamma} - x_0)$  in (12). As noted in the expression  $D_{i1} \Delta(X_{i2}) K_h(X_{i1}) = D_{i1} (f_2(X_{i1}) - \widehat{f}_2(X_{i2})) K_h(X_{i1})$ , if  $\widehat{f}_2$  is estimated based on  $X_{i2}$ , then the decorrelation property cannot be directly applied due to the complex dependence structure between  $\widehat{f}_2$  and  $X_{i2}$ . To avoid this technical difficulty, we construct initial estimators  $(\widehat{f}_2, \widehat{\gamma})$  such that they are independent of the corresponding  $X_{i2}$ .

In the case where historical data is available, we can simply estimating  $(\widehat{f}_2, \widehat{\gamma})$  using the historical data and apply these constructed estimators to the current data. This ensures that the independence assumption is satisfied for the technical analysis. If no historical data is available, we can actually use the “data swapping” idea detailed in the following to create the independence required for the proof but does not lead to loss of efficiency.

We split the data into two random disjoint subsets with approximately equal sample size,  $\mathcal{I}_a$  and  $\mathcal{I}_b$  with  $\mathcal{I}_a \cap \mathcal{I}_b$  empty and  $\mathcal{I}_a \cup \mathcal{I}_b = \{1, 2, \dots, n\}$ . As illustrated in Figure 1, we use data in  $\mathcal{I}_a$  to produce the initial estimator  $\widehat{f}_2^a, \widehat{\gamma}^a$  and use data in  $\mathcal{I}_b$  to produce the initial estimator  $\widehat{f}_2^b, \widehat{\gamma}^b$ . After obtaining these two initial estimators, we *swap* the data and the initial estimators as illustrated by the bolded arrow in Figure 1.

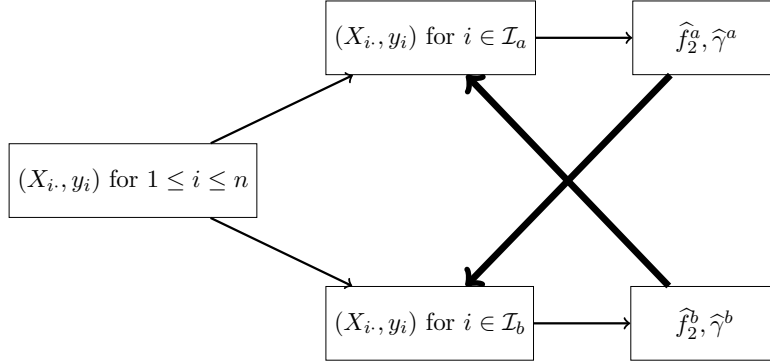


Figure 1: Illustration of Data-Swapping Estimators.

Specifically, this data swapping idea is characterized in the following definitions of  $\hat{f}_2(X_{i2})$  and  $\tilde{D}_{i1}$ ,

$$\hat{f}_2(X_{i2}) = \begin{cases} \hat{f}_2^b(X_{i2}) & \text{for } i \in \mathcal{I}_a \\ \hat{f}_2^a(X_{i2}) & \text{for } i \in \mathcal{I}_b \end{cases} \quad \tilde{D}_{i1} = \begin{cases} (X_{i1} - x_0) - \frac{h^2}{3\hat{\sigma}_2^2} (X_{i2}^\top \hat{\gamma}^b - x_0) & \text{for } i \in \mathcal{I}_a \\ (X_{i1} - x_0) - \frac{h^2}{3\hat{\sigma}_2^2} (X_{i2}^\top \hat{\gamma}^a - x_0) & \text{for } i \in \mathcal{I}_b \end{cases} \quad (27)$$

Note that  $\hat{\sigma}_2$  can be constructed based on the whole data as the corresponding dependence won't cause troubles for the technical analysis. The name “swap” is coming from the fact that the initial estimators applied in the decorrelation step to the data with indexes in  $\mathcal{I}_a$  is constructed based on the other part of the data with indexes in  $\mathcal{I}_b$ . After applying the data swapping technique, for the  $i$ -th observation, the corresponding estimator  $(\hat{f}_2, \hat{\gamma})$  is independent of the corresponding observation  $X_{i2}$  although  $(\hat{f}_2, \hat{\gamma})$  depends on the other half of data excluding  $X_{i2}$ . We can slightly modify the definition of  $\text{Err}(\hat{f}_2)$  as follows

$$\mathbf{P} \left( \max \left\{ \sqrt{\mathbf{E}_{X_{0,2}} (\hat{f}_2^a(X_{0,2}) - f_2(X_{0,2}))^2}, \sqrt{\mathbf{E}_{X_{0,2}} (\hat{f}_2^b(X_{0,2}) - f_2(X_{0,2}))^2} \right\} > \text{Err}(\hat{f}_2) \right) \leq \gamma(n), \quad (28)$$

where  $\gamma(n) \rightarrow 0$ . In particular, the following theorem characterizes exactly how much the estimation error can be reduced after applying the decorrelation step.

**Theorem 1** *Suppose that conditions (A2) – (A4) hold. For  $\Delta(X_{i2}) = f_2(X_{i2}) - \hat{f}_2(X_{i2})$  where  $\hat{f}_2$  is defined in (27), then with probability larger than  $1 - \gamma(n) - \frac{1}{t} - \frac{1}{n^c}$ ,*

$$\left| \frac{1}{n} \sum_{i=1}^n D_{i1} \Delta(X_{i2}) K_h(X_{i1}) \right| \leq Ct \sqrt{h/n} \cdot \text{Err}(\hat{f}_2), \quad (29)$$

where  $\text{Err}(\hat{f}_2)$  is defined in (28). In addition, we can further establish that, with probability larger than  $1 - \gamma(n) - (nh\pi(x_0))^{-\frac{1}{4}} - \frac{1}{t} - \frac{1}{n^c}$ ,

$$\frac{1}{n\hat{S}_n} \sum_{i=1}^n \hat{D}_{i1} \Delta(X_{i2}) K_h(X_{i1}) \leq t \left( \sqrt{\frac{1}{nh^3\pi^2(x_0)}} + \frac{\text{Err}(\hat{D})}{h^2\pi(x_0)} \right) \text{Err}(\hat{f}_2). \quad (30)$$

In the above theorem, the error reduction by decorrelation is achieved in the error bound (29), where *decorrelation* property in (7) is used to guarantee a fast convergence rate for the sum  $\frac{1}{n} \sum_{i=1}^n D_{i1} \Delta(X_{i2}) K_h(X_{i1})$ . Then (30) follows from the concentration result in (29), together with the estimation accuracy of  $D_{i1}$  in Lemma 1 and the order of magnitude of  $\widehat{S}_n$ . The above theorem characterizes the effect of decorrelation and is of independent interest to study other inference problems in the additive modeling.

As a remark, we believe that the independence structure required in the proof of Theorem 1 is only a technical condition and a more refined analysis is likely to remove this technical condition. To focus on the main point, we are not pursuing further here and using data swapping idea to guarantee the independence structure and retain the statistical efficiency.

### 3.5 Properties of Proposed Estimators

Finally, we establish the asymptotic limiting distribution for the proposed estimator  $\widehat{f'(x_0)}$  in the following theorem by applying the results obtained in the previous subsections.

**Theorem 2** *Suppose that conditions (A1)-(A4) hold,  $nh^5\pi(x_0) \leq c$  for some positive constant  $c > 0$  and the final estimator in (14) is constructed using  $\widehat{f}_2(X_{i2})$  and  $\widehat{D}_{i1}$  defined in (27). If  $\widehat{\gamma}$  satisfies (24) and  $\text{Err}(\widehat{D})$  defined in (18) and  $\text{Err}(\widehat{f}_2)$  defined in (28) satisfy*

$$\frac{\text{Err}(\widehat{D})}{h^2} \frac{\text{Err}(\widehat{f}_2)}{\pi(x_0)} = o_p \left( \frac{1}{\sqrt{nh^3 \cdot \pi(x_0)}} \right) \text{ and } \max \left\{ \text{Err}(\widehat{f}_2), \text{Err}(\widehat{D})/\sqrt{V} \right\} \ll \sqrt{\pi(x_0)}, \quad (31)$$

*then the following asymptotic limiting distribution holds,*

$$\frac{1}{\sqrt{V}} \left( \widehat{f'(x_0)} - f'(x_0) \right) \xrightarrow{d} N(0, 1) \text{ with } V = \frac{\sigma_1^2}{n^2 \widehat{S}_n^2} \sum_{i=1}^n \widehat{D}_{i1}^2 K_h^2(X_{i1}), \quad (32)$$

*where  $\widehat{S}_n$  is defined as (14).*

The above theorem provides the theoretical guarantee of the proposed estimator through establishing the limiting distribution as in (32). We have a few remarks regarding this limiting distribution result. First, the asymptotic variance depends on the value  $x_0$  implicitly since all three terms  $\widehat{D}_{i1}$ ,  $K_h(X_{i1})$  and  $\widehat{S}_n$  depend on the value of  $x_0$ . Second, beyond the conditions that we have already discussed, the additional condition (31) is the *double estimation accuracy* condition, which is imposed on the estimation accuracy of the weights  $D$  and the nuisance function  $f_2$ . The condition (31), involved with *double estimation accuracy*, comes from the fact that, to construct valid confidence intervals, the nuisance function

error established in Theorem 1 has to be smaller than the standard deviation level  $\sqrt{V}$ . In Section 5, we show that (31) is reduced to a condition for the sample size and model complexity.

We will present several additional important propositions and corollaries that we can obtain from Theorem 2. The following proposition establishes the estimation rate of the proposed estimator  $\widehat{f'(x_0)}$ .

**Theorem 3** *Suppose that conditions (A1)-(A4) hold, then with probability larger than  $1 - \gamma(n) - (nh\pi(x_0))^{-\frac{1}{4}} - \frac{1}{t} - \frac{1}{n^c}$ , the proposed estimator in (14) satisfies*

$$\left| \widehat{f'(x_0)} - f'(x_0) \right| \lesssim \frac{t}{\sqrt{nh^3 \cdot \pi(x_0)}} + c_* h + \frac{\text{Err}(\widehat{D})}{\sqrt{\pi(x_0)}} + t \left( \sqrt{\frac{1}{nh^3 \pi^2(x_0)}} + \frac{\text{Err}(\widehat{D})}{h^2 \pi(x_0)} \right) \text{Err}(\widehat{f}_2)$$

where  $c_* = o(1)$ .

The above proposition establishes the rate of convergence for the proposed estimator  $\widehat{f'(x_0)}$ . In contrast to Theorem 2, the condition of establishing the rate of convergence is much weaker by removing the condition (31) and the condition on the bandwidth  $nh^5\pi(x_0) \leq c$ . The main intuition is that the estimation accuracy in Theorem 3 is a summation of the stochastic error, the approximation error and the nuisance error while the limiting distribution can be established only when the stochastic error dominates both the approximation error and the nuisance error.

With the estimation accuracy  $\text{Err}(\widehat{D})$  obtained in Lemma 1, Theorem 3 can be simplified as a condition for the estimation error of  $\widehat{f}_2$ . The following corollary of Theorem 2 establishes such a result and presents a more explicit condition on  $\text{Err}(\widehat{f}_2)$ .

**Corollary 1** *Assume conditions (A1)-(A4) hold,  $h \asymp (n\pi(x_0))^{-\frac{1}{5}}$ , and  $\pi(x_0) \gg n^{-\frac{2}{7}}$ . Suppose that  $\widehat{\gamma}$  satisfies (24) and  $\text{Err}(\widehat{f}_2)$  defined in (28) satisfies the following condition,*

$$\text{Err}(\widehat{f}_2) = o \left( \sqrt{\pi(x_0)} \min \left\{ \sqrt{\frac{1}{h^3 k \log p}}, 1 \right\} \right) \quad (33)$$

then the limiting distribution in (32) holds.

As a consequence of Corollary 1, we can establish both the coverage and the precision properties of the constructed confidence interval  $\text{CI}_{x_0}$  defined in (15), where  $\mathbf{L}(\text{CI}_{x_0})$  denotes the length of the proposed confidence interval.

**Corollary 2** *Suppose that the same conditions as in Corollary 1 hold and  $\widehat{\sigma}_1^2$  is a consistent estimator of  $\sigma_1^2$ , then the constructed confidence interval  $\text{CI}_{x_0}$  defined in (15) satisfies the following properties,*

$$\liminf_{n \rightarrow \infty} \mathbb{P}(f'(x_0) \in \text{CI}_{x_0}) \geq 1 - \alpha$$

and

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \mathbf{L}(\text{CI}_{x_0}) \geq (2 + \delta_0) z_{\alpha/2} \sqrt{\frac{3}{2nh^3 \cdot \pi(x_0)}} \sigma_1 \right) = 0$$

for any positive constant  $\delta_0 > 0$ .

The above corollary justifies the validity of the proposed confidence interval defined in (15) and also controls the length of the proposed confidence interval. Similarly, we can establish the validity of the proposed testing procedure  $\psi_{x_0}$  in (16).

**Corollary 3** *Under the same condition as in Corollary 2, then for any  $f_1$  and  $x_0$  such that  $f'_1(x_0) = 0$ , the proposed testing procedure  $\psi_{x_0}$  defined in (16) controls the type I error,*

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\psi_{x_0} = 1) \leq \alpha.$$

## 4 Inference with Additional Information of $X_{i1} \mid X_{i2}$

In this section, we present the results on inference for  $f'_1(x_0)$  with additional information of the conditional distribution  $X_{i1} \mid X_{i2}$ . The relationship between  $X_{i1}$  and  $X_{i2}$  can be known in certain machine learning applications, including compressed sensing, as the covariates are generated by the users. In addition, a more interesting intermediate regime appears in the semi-supervised setting [6, 40], where we have the supervised data  $(X_{i1}, X_{i2}^\top, y_i)^\top\}_{1 \leq i \leq n}$  and also have access to a large number of unsupervised observations  $\{(X_{i1}, X_{i2}^\top)^\top\}_{n+1 \leq i \leq n+N}$ . Here, the additional sample size  $N$  can be much larger than the sample size  $n$ , even the dimension  $p$ . In such a scenario, we can utilize this large set of unlabelled data  $\{(X_{i1}, X_{i2}^\top)^\top\}_{n+1 \leq i \leq n+N}$  to provide an accurate estimation of the conditional distribution of  $X_{i1} \mid X_{i2}$  and also the conditional expectations  $\mathbf{E}([X_{i1} - x_0]K_h(X_{i1}) \mid X_{i2})$  and  $\mathbf{E}(K_h(X_{i1}) \mid X_{i2})$  used in (8). Procedurewise, we modify  $\tilde{D}_{i1}$  defined in (12) as  $\tilde{D}_{i1} = (X_{i1} - x_0) - e(X_{i2})$ , with  $e(X_{i2})$  defined in (8) and the other parts of the proposed estimator keep unchanged as in (14).

The following theorem establishes the rate of convergence and also the limiting distribution for the case of known conditional distribution  $X_{i1} \mid X_{i2}$ .

**Theorem 4** *Assume the conditions (A1)-(A2) hold and the conditional distribution  $X_{i1} \mid X_{i2}$  is known, then with probability larger than  $1 - (nh\pi(x_0))^{-\frac{1}{4}} - \frac{1}{t} - \frac{1}{n^c}$ ,*

$$\left| \widehat{f'(x_0)} - f'(x_0) \right| \lesssim \frac{t}{\sqrt{nh^3 \cdot \pi(x_0)}} \left( 1 + \sqrt{\text{Err}^2(\hat{f}_2)/\pi(x_0)} \right) + c_* h.$$

where  $c_* = o(1)$ . In addition, if  $nh^5\pi(x_0) \leq c$  for some positive constant  $c > 0$  and

$\text{Err}(\widehat{f}_2) \ll \sqrt{\pi(x_0)}$ , then

$$\frac{1}{\sqrt{V}} \left( \widehat{f'_1(x_0)} - f'_1(x_0) \right) \xrightarrow{d} N(0, 1) \quad \text{where } V = \frac{\sigma_1^2}{n^2 \widehat{S}_n^2} \sum_{i=1}^n \widehat{D}_{i1}^2 K_h^2(X_{i1}).$$

The knowledge of the conditional distribution  $X_{i1} \mid X_{i2}$  has a significant effect on both point estimation and statistical inference for  $f'_1(x_0)$ . In particular, in contrast to Theorem 3, the rate of convergence of estimating  $f'_1(x_0)$  is much faster as the terms involved with the estimation error  $\text{Err}(\widehat{D})$  disappear; in contrast to Theorem 2, the limiting distribution holds under much weaker accuracy requirement on the initial estimator  $\widehat{f}_2$ ; as observed in Section 5, these weaker conditions on initial estimators will lead to weaker sample size conditions.

We shall also highlight the *double estimation accuracy* phenomenon here using the above theorem. The interesting observation here is the statistical inference results are almost the same if either the nuisance function  $f_2$  is known or  $D_{i1}$  is known a priori, where the later is true if the conditional distribution  $X_{i1} \mid X_{i2}$  is known a priori. That is, even if we do not know the nuisance function  $f_2$  but known the relationship between  $X_{i1}$  and  $X_{i2}$  accurately enough, we can achieve the same statistical accuracy by utilizing the information of the relationship between  $X_{i1}$  and  $X_{i2}$ , as if the nuisance function  $f_2$  is known.

## 5 Inference in High-dimensional Sparse Additive Model

We consider high-dimensional sparse additive model to demonstrate the inference results developed for the general additive model (1). We assume that the nuisance function itself is of an additive structure and with slight abuse of notation, we rewrite the model (1) as

$$y_i = f_1(X_{i1}) + \sum_{j=2}^p f_j(X_{ij}) + \epsilon_i, \text{ for } 1 \leq i \leq n. \quad (34)$$

Here, the nuisance function is an additive form of  $p-1$  univariate nonparametric functions, where each  $f_j$  is a univariate function of  $X_{ij}$ . To apply the inference method developed in previous subsections, we need to construct initial estimators  $\widehat{f}_2$  satisfying (33) and  $(\widehat{\gamma}, \widehat{\sigma}_2)$  satisfying Condition (A4) and (24). We particularly take the doubly penalized estimation approach developed in [38] and apply the prediction accuracy result in [38], together with results established in [19], to establish the nuisance function estimation error  $\text{Err}(\sum_{j=2}^p \widehat{f}_j)$ . Additionally, the proposed DLL estimator is also compatible with the estimators proposed in [23, 26, 30, 31, 37].

We detail the exact technical assumptions for the sparse additive model in Section 5.1 and present the initial estimator construction in Section 5.2. In Section 5.3, we present the inference results for  $f'_1(x_0)$  in high-dimensional sparse additive model and also discuss *uncertainty-quantification* conditions.

## 5.1 Statistical Modeling

The estimation methods in additive models are mainly developed for variables of compact supports, say  $[0, 1]$ . To apply the existing methods directly, we need to define a transformation  $G_j$  from  $X_{ij} \in \mathbb{R}$  to the compact set  $[0, 1]$ . Specifically,  $Z_{ij} = G_j(X_{ij}) \in [0, 1]$  denotes the transformed variable, where  $G_j : (-\infty, +\infty) \rightarrow [0, 1]$  is to be specified later. To directly apply the theoretical results in [38], we impose the following model assumption for the high-dimensional additive model in (34).

(E1) The additive model (34) can be expressed as

$$y_i = \sum_{j=1}^p g_j(Z_{ij}) + \epsilon_i, \text{ for } 1 \leq i \leq n, \quad (35)$$

where  $\epsilon_i$  is sub-Gaussian random variable,  $g_j$  belongs to a Sobolev space  $\mathcal{W}_r^{m_j}$  on  $[0, 1]$  with the corresponding norm  $\|g_j\|_{F,j} = (\int_0^1 |g_j^{(m_j)}|^r)^{\frac{1}{r}}$  and the marginal density  $q_j$  of  $Z_{ij} = G_j(X_{ij})$  is uniformly bounded away from zero for  $1 \leq j \leq p$ . We use  $\mathcal{S}$  to denote the support set  $\mathcal{S} = \{j : \int f_j^2(t)dt > 0\}$  and define  $M_F = \sum_{j=1}^p \|g_j\|_{F,j}$ .

The condition (E1) imposes three types of modeling conditions, sub-gaussian tail for the error, the model complexity condition, including both smoothness and sparsity conditions, and the lower bound on the marginal density of the transformed variables. We shall supply detailed discussions on these conditions and also give examples of transformation  $G_{ij}$  satisfying the above condition. Since the smoothness condition is imposed on the functions of the transformed variable  $Z_{ij}$ , we can view this as assuming  $f_j$  to be a composite function of  $f_j = g_j \circ G_j$ , where  $g_j$  satisfies certain smoothness conditions and  $G_j$  is the pre-specified transformation. Here, the cardinality of the signal set,  $|\mathcal{S}|$ , and a measure of total smoothness,  $M_F$ , are allowed to depend on  $(n, p)$ , where  $|\mathcal{S}|$  and  $M_F$  capture the sparsity and the smoothness of the additive model, respectively. For the case that  $\|g_j\|_{F,j} \leq C$ , then the smoothness parameter can be upper bounded by the sparsity level, that is,  $M_F \leq Cs$ .

In addition, (E1) assumes that the marginal density of the transformed variable  $Z_{ij}$  is lower bounded by a small positive constant. We will present examples of transformations such that this lower bound for marginal density condition holds. We use  $F_j(\cdot)$  to denote the cumulative density function of  $X_{ij}$  and  $q_j$  to denote the marginal density of  $Z_{ij}$  over the support  $[0, 1]$ .



*Example 1: Copula Model.* The additive model (35) on the transformed random variables can be viewed from the perspective of copula model. The transformation  $G_j$  is set as the corresponding marginal CDF  $F_j$  of  $X_{ij}$  and then  $Z_{ij} = G_j(X_{ij})$  follows uniform distribution on  $[0, 1]$ . Hence, the lower bound condition on the marginal density holds with  $\min_{t \in [0, 1]} q_j(t) = 1$  for all  $1 \leq j \leq p$ . For the case that the marginal CDF  $F_j$  is known, the model (35) is exactly reduced to an additive copula model. For the case that  $F_j$  is unknown, we can estimate the CDF  $F_j$  by the empirical CDF as  $\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_{ij} \leq x)$  and define  $\hat{Z}_{ij} = \hat{F}_j(X_{ij})$ . In this case, the transformed variable  $Z_{ij}$  is not directly observed but its empirical estimator  $\hat{Z}_{ij}$  is observed and can be shown by standard concentration results that the observed  $\hat{Z}_{ij}$  is closed to  $Z_{ij}$ . If  $F_j$  belongs to certain parametric families, we can also estimate the transformed variable  $Z_{ij}$  through estimating the corresponding parameters of  $F_j$ .

*Example 2: Heavier Tail Transformation.* Instead of using the quantile transformation in copula model, we also show that a heavier tail transformation  $G_j$  is sufficient for our use. Specifically, we consider the case that  $X_{ij}$  is of mean zero and variance  $\Sigma_{jj}$  and has a marginal sub-gaussian tail. We define the following transformation  $G_j(x) = \Phi(T(x))$ , where  $\Phi$  is the CDF of the standard normal distribution and  $T(x)$  satisfies the following condition with a positive constant  $C > 0$ ,

$$T^2(x) \geq C \frac{x^2}{\Sigma_{jj}} \quad \text{for large value of } |x|. \quad (36)$$

Two specific examples of  $T(x)$  include 1)  $T(x) = x \cdot \log \log n$ ; and 2)  $T(x) = \text{sign}(x) \cdot |x|^c$  for any  $c > 1$ . Since the property of the transformation  $G_j(\cdot)$  only matters for large values of  $|x|$ , we further generalize the above heavier tail transformation and define  $G_j(x) = (1 - c_0)G_0(x) + c_0\Phi(T(x))$ , where  $0 < c_0 \leq 1$ ,  $G_0 : (-\infty, \infty) \rightarrow [0, 1]$ ,  $G'_0(x) = 0$  for  $|x| \geq C$  and  $T(x)$  satisfies the condition (36). Hence, we have the flexibility of adding a fraction of function  $G_0$  as long as  $G_0$  has the range  $[0, 1]$  and has a vanishing derivative out of a bounded support. See more discussion in Section B in the supplementary materials.

In addition to the condition (E1), the other condition needed for controlling  $\text{Err}(\sum_{j=2}^p \hat{f}_j)$  is on the theoretical restricted eigenvalue or compatibility condition, which intuitively guarantees the “invertibility” of the additive modeling. Here we introduce one version of the theoretical restricted eigenvalue, which was used in [23, 37]. For the centered functions  $\mathbf{E}f_j(X_{ij}) = 0$  for  $j = 1, 2, \dots, p$ ,

$$\text{if } \sum_{j \in \mathcal{S}^c} \sqrt{\mathbf{E}f_j^2(X_{ij})} \leq \xi_0^* \sum_{j \in \mathcal{S}} \sqrt{\mathbf{E}f_j^2(X_{ij})}, \text{ then } c_0 \sum_{j \in \mathcal{S}} \mathbf{E}f_j^2(X_{ij}) \leq \mathbf{E}(\sum_{j=1}^p f_j(X_{ij}))^2. \quad (37)$$

This implies the theoretical compatibility condition stated as Assumption 5 of [38]. It has been shown that [19] that the condition (37) will hold for a large class of distributions as

long as the underlying correlation structure between  $\{X_{ij}\}_{1 \leq j \leq p}$  is generated by a pairwise Gaussian. We restate the Corollary 4 in [19] in the following form.

**Corollary 4** *Suppose  $(X_{i1}, X_{i2}, \dots, X_{ip})$  follows a hidden Gaussian distribution with  $X_{ij} = T_j(Q_{ij})$  for a pairwise Gaussian vector  $(Q_{i1}, \dots, Q_{ip})$  with  $\text{Corr}(Q_{i1}, \dots, Q_{ip}) = \Sigma^Q$  and some deterministic functions  $T_j$  with  $0 < \text{Var}(T_j(Q_{ij})) < \infty$ . Then, the condition (37) holds with  $\kappa_0 = \lambda_{\min}(\Sigma^Q)$ .*

As implied by the above corollary, a special case is that  $(X_{i1}, X_{i2}, \dots, X_{ip})$  follows a joint Gaussian distribution, then any transformed variables  $(Z_{i1}, Z_{i2}, \dots, Z_{ip})$ , including those in Examples 1 and 2, will satisfy the theoretical restricted eigenvalue condition (37).

## 5.2 Initial Estimator Construction

Define the empirical  $L_2$  norm as  $\|g_j\|_n = \{\frac{1}{n} \sum_{j=1}^n g_j^2(Z_{ij})\}^{\frac{1}{2}}$ . The double penalized estimator in [38] is stated as follows,

$$\{\hat{g}_j\}_{1 \leq j \leq p} = \arg \min \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^p g_j(Z_{ij}))^2 + \sum_{j=1}^p (\rho_{nj} \|g_j\|_{F,j} + \lambda_{nj} \|g_j\|_n), \quad (38)$$

where  $\rho_{nj}$  and  $\lambda_{nj}$  are tuning parameters. As shown in [38], with proper chosen tuning parameters, the above proposed estimator attains the optimal rate of convergence in the prediction problem. The algorithm (38) is implemented with respect to the transformed variables  $\{Z_{ij}\}_{1 \leq j \leq p}$  and we define the estimators of  $f_j$  as the composite function  $\hat{f}_j = \hat{g}_j \circ G_j$ . Additionally, we implement scaled Lasso estimator [36] to decouple the relation between  $X_{i1}$  and  $X_{i,-1}$ ,

$$(\hat{\gamma}, \hat{\sigma}_2) = \arg \min_{\gamma \in \mathbb{R}^{p-1}, \sigma_2 \in \mathbb{R}^+} \frac{1}{2n\sigma_2} \sum_{i=1}^n \left( X_{i,1} - X_{i,-1}^\top \gamma \right)^2 + \frac{\sigma_2}{2} + \sqrt{\frac{2A \log p}{n}} \|\gamma\|_1,$$

for some pre-specified constant  $A > 1$ . We then construct the final estimator  $\widehat{f'_1(x_0)}$  as in (14) with the initial estimator  $\sum_{j=2}^p \hat{g}_j \circ G_j(X_{ij})$  for the nuisance function and the estimator of the regression vector  $\hat{\gamma}$ . The following two lemmas characterize the accuracy of the initial estimators,  $\hat{\gamma}$  and  $\{\hat{g}_j\}_{1 \leq j \leq p}$ .

**Lemma 3** *Suppose that Condition (A3) holds and  $X_{i \cdot} \in \mathbb{R}^p$  is a sub-gaussian random vector with covariance matrix  $\Sigma$  satisfying  $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$  for some positive constants  $c_0, C_0 > 0$ . For the initial estimators  $\hat{\gamma}$  defined in the data-swapping way as in (27), then with probability larger than  $1 - n^{-c}$ , the initial estimators  $(\hat{\gamma}, \hat{\sigma}_2)$  satisfy (24) and (20) for  $h = n^{-\delta_0}$  with  $0 < \delta_0 < \frac{1}{2}$ .*

The following lemma is established by combining Proposition 4 and Theorem 2 in [38] and Corollaries 4 and 5 in [19].

**Lemma 4** *Suppose that Condition (E1) holds for  $r = 2$  and  $m = \min_{1 \leq j \leq p} m_j \geq 1$  and  $X_i \in \mathbb{R}^p$  is a pairwise Gaussian random vector with covariance matrix  $\Sigma$  satisfying  $\lambda_{\min}(\Sigma) \geq c_0$  for some positive constants  $c_0 > 0$ . Then with probability larger than  $1 - \frac{1}{p}$ , we have*

$$\mathbf{E}(\sum_{j=2}^p \hat{f}_j - \sum_{j=2}^p f_j)^2 \lesssim \sum_{j \in \mathcal{S}} \left( n^{-\frac{m_j}{2m_j+1}} + \sqrt{\log p/n} \right) \left( [1 + \|g_j\|_{F,j}] n^{-\frac{m_j}{2m_j+1}} + \sqrt{\log p/n} \right) \quad (39)$$

under the condition

$$\left\{ w_n^*(0)^{-\frac{1}{2m}} \gamma_n^*(0) + w_n^*(0)^{-\frac{1}{2m-1}} \sqrt{\log p/n} \right\} (1 + M_F + |\mathcal{S}|) = o(1) \quad (40)$$

where  $w_n^*(0) = \max \left\{ n^{-\frac{m}{2m+1}}, \sqrt{\log p/n} \right\}$  and  $\gamma_n^*(0) = \min \left\{ n^{-\frac{m}{2m+1}}, n^{-1/2} (\log p/n)^{-1/4m} \right\}$ .

Lemma 4 guarantees that, for bounded norm  $\|g_j\|_{F,j}$ , the accuracy of estimating  $f_2$  satisfies  $\text{Err}(\sum_{j=2}^p \hat{f}_j) \lesssim \sum_{j \in \mathcal{S}} (n^{-\frac{m_j}{2m_j+1}} + \sqrt{\log p/n})^2$ .

### 5.3 Inference for $f'(x_0)$ and Uncertainty-Quantification Conditions

Finally, we can combine Lemmas 3 and 4 with Corollary 1 to establish the limiting distribution for  $f'_1(x_0)$  in the high-dimensional sparse additive model.

**Theorem 5** *Suppose that conditions (A1)-(A3) hold,  $h \asymp (n\pi(x_0))^{-\frac{1}{5}}$ , and  $\pi(x_0) \gg n^{-\frac{2}{7}}$ , (E1) holds with  $r = 2$  and  $m \geq 1$ , the model complexity condition (40) holds,  $X_i \in \mathbb{R}^p$  is a pairwise Gaussian random vector with covariance matrix  $\Sigma$  satisfying  $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$  for some positive constants  $c_0, C_0 > 0$ . Then the limiting distribution in (32) holds under the additional condition*

$$\sqrt{\sum_{j \in \mathcal{S}} \left( n^{-\frac{m_j}{2m_j+1}} + \sqrt{\log p/n} \right)^2} = o \left( \sqrt{\pi(x_0)} \min \left\{ \sqrt{\frac{1}{h^3 k \log p}}, 1 \right\} \right). \quad (41)$$

We provide some discussion on the *uncertainty-quantification* condition (41), which is the extra condition imposed for establishing the distributional results. The condition (41) follows from the combination of (33) and (39) and this is the condition to ensure the nuisance estimator error, due to estimating the high-dimensional nuisance function  $f_2$ , is negligible, in comparison to the stochastic error. To highlight interesting observations implied by condition (41), we focus on one of the most interesting regimes,  $\pi(x_0)$  is at constant level,

$h \log p = o(1)$ ,  $\max_{1 \leq j \leq p} \|g_j\|_{F,j} \leq C$  and  $m_1 = 2$  and  $m_2 = \dots = m_p := m_0$  and drop all log terms. Then we simplify (41) as

$$k \cdot |\mathcal{S}| \ll n^{\frac{2m_0}{2m_0+1}} h^{-3} \text{ and } \max\{k, |\mathcal{S}| \cdot n^{\frac{1}{2m_0+1}}\} \ll n \text{ up to a polynomial order of } \log p$$

Here, the second condition is a standard one to guarantee that we have enough data in comparison to the significant variables. In the case that  $f_1$  has a continuous second order derivative near  $x_0$ , the optimal rate of choosing the bandwidth is  $h \asymp n^{-1/5}$ , then we have

$$k \cdot |\mathcal{S}| \ll n^{\frac{2m_0}{2m_0+1} + \frac{3}{5}} \text{ and } \max\{k, |\mathcal{S}| \cdot n^{\frac{1}{2m_0+1}}\} \ll n \text{ up to a polynomial order of } \log p \quad (42)$$

For  $m \geq 1$ , the power of  $n$  in  $n^{\frac{2m_0}{2m_0+1} + \frac{3}{5}}$  is always larger than  $\frac{19}{15} > 1$ .

A few remarks are in order for this sample size condition (42). First, this is only a sufficient *uncertainty-quantification* condition that we can conduct adaptive inference and establish the asymptotic normal limiting distribution. In the high-dimensional sparse linear regression where all  $\{f_j\}_{1 \leq j \leq p}$  are assumed to be linear, a similar form of the *uncertainty-quantification* condition for sample size and model complexity can be established as

$$k \cdot |\mathcal{S}| \ll n / \log p. \quad (43)$$

Through comparing (43) and (42), we have observed a striking phenomenon that the *uncertainty-quantification* condition required for the sparse additive model is weaker than that for the sparse regression model. The main reason of this phenomenon is due to the fact that the inflation of nuisance error in the additive models is not as large as that for the stochastic error, where the stochastic error increases from  $1/\sqrt{n}$  to  $1/\sqrt{nh^3 \cdot \pi(x_0)}$  and the nuisance error increases from  $\sqrt{k \cdot |\mathcal{S}|} \log p/n$  to

$$\frac{\sqrt{k \cdot |\mathcal{S}|} \log^2 p}{n^{\frac{2m_0+0.5}{2m_0+1}} \sqrt{\pi(x_0)}} + \frac{\sqrt{k \cdot |\mathcal{S}|} \log^{\frac{5}{2}} p}{n \sqrt{\pi(x_0)}} + \frac{\sqrt{k} \log^2 p}{n^{0.9} \sqrt{\pi(x_0)}}.$$

For this nuisance error, the first term would be the dominating term in most settings. We shall remark that part of this striking relaxation of the *uncertainty-quantification* condition for sample size and model complexity is due to condition (A3), the parametric model relationship between the variable of interest and the nuisance variables. In contrast, for the high-dimensional sparse linear regression, even though the parametric model condition (A3) is imposed, there is not such a phenomenon of significantly relaxing the corresponding *uncertainty-quantification* condition in (43).

Second, we consider two special cases and highlight some interesting conclusions obtained by applying (41). If the Sobolev smoothness level is  $m_0 = 2$ , we can apply (41) and obtain the *uncertainty-quantification* condition as

$$k \cdot |\mathcal{S}| \ll n^{7/5} \text{ and } \max\{k, |\mathcal{S}| \cdot n^{\frac{1}{2m_0+1}}\} \ll n \text{ up to a polynomial order of } \log p \quad (44)$$

which is much weaker than the sufficient *uncertainty-quantification* condition used for linear model in (43). Another interesting case is to consider the semi-parametric outcome model with assuming  $f_2$  to be of additional linear structure, that is,

$$y_i = f_1(X_{i1}) + X_{i2}^\top \eta + \epsilon_i, \text{ for } 1 \leq i \leq n.$$

The linear components can be viewed as belonging to the Sobolev space with  $m_0 = \infty$  and hence the *uncertainty-quantification* (41) for sample size is reduced to be

$$k \cdot |\mathcal{S}| \ll n^{\frac{8}{5}} \text{ and } \max\{k, |\mathcal{S}| \cdot n^{\frac{1}{2m_0+1}}\} \ll n \text{ up to a polynomial order of } \log p$$

Third, we shall compare the obtained results for high-dimensional sparse additive model with those obtained in [18]. The most significant difference is that [18] considers the relation between the variable of interest and all other nuisance variables from nonparametric perspectives by imposing the assumption that any basis function of variable of interest can be well approximated by the basis functions defined on a sparse set of  $k$  nuisance variables. However, the current paper is considering a completely different parametric assumption between the variable of interest and nuisance variables. After carefully developing the decorrelated linear estimator, we establish a much weaker *uncertainty-quantification* condition. If we set  $m = 2$  and  $h \asymp n^{-\frac{1}{5}}$  and use the current paper notation, then the sample size condition obtained in [18] is reduced to

$$|\mathcal{S}| \ll n^{\frac{3}{10}} \text{ and } k \ll n^{\frac{4}{15}} \text{ up to a polynomial order of } \log p. \quad (45)$$

To compare the condition (45) with (44), we take  $|\mathcal{S}| = \frac{1}{\log n} n^{\frac{3}{10}}$  in both conditions (44) and (45) and then (44) is further simplified as  $k \ll n$ , which is significantly weaker than the requirement for (45). This is to say, if we utilize the parametric assumption between  $X_{i1}$  and  $X_{i2}$ , then the method allows for a much larger number of nuisance variables to be associated with the variable of interest.

## 6 Conclusion and Discussion

In conclusion, we study the local inference problem in the general additive model, including both confidence interval construction for  $f'_1(x_0)$  and hypothesis testing related to  $f'_1(x_0)$ . We have developed general method and theory and demonstrate it in the high-dimensional sparse additive model. The key challenge posed by the inference problem is the uncertainty of estimating the nuisance function. To address this challenge, we develop a novel decorrelated local linear estimator to conduct statistical inference for  $f'_1(x_0)$  in presence of other unknown nuisance functions. Such a decorrelation step is particularly useful in diminishing

the effect of estimating the nuisance function and can be of independent interest in solving other inference problems in additive models.

An important perspective of the current paper is to impose a parametric modeling assumption between the variable of interest and the nuisance variables. This is definitely facilitating the statistical inference result by avoiding studying the conditional density between the variable of interest and all other nuisance variables through additional nonparametric techniques. Interestingly, a careful utilization of this parametric modeling assumption significantly reduces the *uncertainty-quantification* condition for sample size and model complexity, which is even weaker than the corresponding assumption for the high-dimensional linear model. To the most extreme case where the distribution  $X_{i1} \mid X_{i2}$  is known a priori, we can achieve the statistical inference accuracy as if we know the nuisance function  $f_2$ . It would be interesting to relax this parametric modeling assumption and work out the corresponding *uncertainty-quantification* condition for more general relationship between  $X_{i1}$  and  $X_{i2}$ . It is conjectured that such a weak *uncertainty-quantification* as in (42) would not generally hold without the parametric modeling condition (A3). This is left for future research.

The local inference problem for  $f_1'(x_0)$  considered in this paper is motivated from studying the treatment effect using the general additive modeling. There are many other interesting related inference problems, including inference for  $f_1(x_0)$  and also the significance test  $H_0 : f_1 = 0$ , which are left for future research.

## 7 Analysis of Nuisance Error: Proof of Theorem 1

The proof is divided into two parts, proof of (29) by applying the decorrelation property of the constructed weights  $D_{i1}$  and proof of (30) by approximating  $D_{i1}$  by  $\hat{D}_{i1}$ .

Proof of (29) The following technical proof relies on independence created by data swapping. Recall  $\mathcal{I}_a$  and  $\mathcal{I}_b$  are two disjoint subset with approximately equal sample size, with  $\mathcal{I}_a \cap \mathcal{I}_b$  empty and  $\mathcal{I}_a \cup \mathcal{I}_b = \{1, 2, \dots, n\}$ ;  $\hat{f}_2^a$  and  $\hat{f}_2^b$  denote the initial estimator  $\hat{f}_2$  based on the data  $(X_{i\cdot}, y_i)_{i \in \mathcal{I}_a}$  and  $(X_{i\cdot}, y_i)_{i \in \mathcal{I}_b}$ , respectively. We define  $\Delta^a(X_{i2}) = \hat{f}_2^a(X_{i2}) - f_2(X_{i2})$  and  $\Delta^b(X_{i2}) = \hat{f}_2^b(X_{i2}) - f_2(X_{i2})$ . We write  $\mathbf{E}_{\mathcal{I}_a}, \text{Var}_{\mathcal{I}_a}$  and  $\mathbf{P}_{\mathcal{I}_a}$  as the expectation, variance and probability taken with respect to the sample  $(X_{i\cdot}, y_i)_{i \in \mathcal{I}_a}$ , respectively. Similarly, we can define  $\mathbf{E}_{\mathcal{I}_b}, \text{Var}_{\mathcal{I}_b}$  and  $\mathbf{P}_{\mathcal{I}_b}$  with respect to  $(X_{i\cdot}, y_i)_{i \in \mathcal{I}_b}$ . We define  $\mathcal{A}_{3,i} = \{\|X_{i2}^\top \gamma\|_2 \lesssim \|\gamma\|_2 \sqrt{\log n}\}$  for  $1 \leq i \leq n$  and have the following decomposition

$$\frac{1}{n} \sum_{i=1}^n D_{i1} \Delta(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} = \frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} + \frac{1}{n} \sum_{i \in \mathcal{I}_b} D_{i1} \Delta^a(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}. \quad (46)$$

In the following, we control the first term  $\frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}$  and the second term can be controlled by symmetry. Note that there are two sources of randomness in

$\frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}$ , one from the initial estimator  $\Delta^b$  and the other from the data  $\{X_{i\cdot}\}_{i \in \mathcal{I}_a}$ . Since the randomness of  $\Delta^b$  is induced from the data  $(X_{i\cdot}, y_i)_{i \in \mathcal{I}_b}$ , the independence between  $\Delta^b$  and  $\{X_{i\cdot}\}_{i \in \mathcal{I}_a}$  can be used here.

Since  $D_{i1}$  is constructed such that (6) holds, the summation  $\frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}$  satisfies  $\mathbf{E}_{\mathcal{I}_a} \left( \frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right) = 0$ . We control the variance as  $\text{Var}_{\mathcal{I}_a} \left( \frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right) = \frac{|\mathcal{I}_a|}{n^2} \mathbf{E}_{\mathcal{I}_a} \left( D_{i1}^2 (\Delta^b(X_{i2}))^2 K_h^2(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right)$ . By (115) in the supplement, we have  $\left| \frac{(\Delta^b(X_{i2}))^2 \mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) | X_{i2}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}}{(\Delta^b(X_{i2}))^2 \frac{2}{3} h q(x_0 | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}} - 1 \right| \xrightarrow{P} 0$  and hence

$$\mathbf{E}_{\mathcal{I}_a} \left[ (\Delta^b(X_{i2}))^2 D_{i1}^2 K_h^2(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right] \lesssim h \mathbf{E}_{\mathcal{I}_a} \left( \hat{f}_2^b(X_{i2}) - f_2(X_{i2}) \right)^2.$$

Since  $\mathbf{P}_{\mathcal{I}_a} \left( \left| \frac{1}{n} \sum_{i=1}^n D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \right| \neq \left| \frac{1}{n} \sum_{i=1}^n D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right| \right) \leq n^{-c}$ ,

$$\mathbf{P}_{\mathcal{I}_a} \left( \left| \frac{1}{n} \sum_{i \in \mathcal{I}_a} D_{i1} \Delta^b(X_{i2}) K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right| \leq t \sqrt{h/n} \cdot \text{Err}(\hat{f}_2) \right) \geq 1 - \frac{1}{t} - \frac{1}{n^c},$$

where  $\text{Err}(\hat{f}_2)$  is defined in (28). By symmetry and (46), we establish (29).

Proof of (30) We decompose the expression  $\frac{1}{n} \sum_{i=1}^n \hat{D}_{i1} \Delta(X_{i2}) K_h(X_{i1})$  as

$$\frac{1}{n} \sum_{i=1}^n \left( \hat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right) \Delta(X_{i2}) K_h(X_{i1}) + \frac{1}{n} \sum_{i=1}^n D_{i1} \Delta(X_{i2}) K_h(X_{i1}) - \bar{\mu}_D \cdot \frac{1}{n} \sum_{i=1}^n \Delta(X_{i2}) K_h(X_{i1}) \quad (47)$$

By Cauchy-Schwarz inequality, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \left( \hat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right) \Delta(X_{i2}) K_h(X_{i1}) \right| \leq \text{Err}(\hat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \Delta(X_{i2})^2 K_h(X_{i1})} \quad (48)$$

and

$$\frac{1}{n} \sum_{i=1}^n |\Delta(X_{i2})| K_h(X_{i1}) \leq \sqrt{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \Delta(X_{i2})^2 K_h(X_{i1})} \quad (49)$$

Hence, it is sufficient to control  $\sqrt{\frac{1}{n} \sum_{i=1}^n \Delta(X_{i2})^2 K_h(X_{i1})}$ . Similar to (46), we have

$$\frac{1}{n} \sum_{i=1}^n \Delta(X_{i2})^2 K_h(X_{i1}) = \frac{1}{n} \sum_{i \in \mathcal{I}_a} \left| \Delta^b(X_{i2}) \right|^2 K_h(X_{i1}) + \frac{1}{n} \sum_{i \in \mathcal{I}_b} \left| \Delta^a(X_{i2}) \right|^2 K_h(X_{i1})$$

and it is sufficient to control  $\frac{1}{n} \sum_{i \in \mathcal{I}_a} \left| \Delta^b(X_{i2}) \right|^2 K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}$ . Note that

$$\mathbf{E}_{\mathcal{I}_a} \left( \frac{1}{n} \sum_{i \in \mathcal{I}_a} \left| \Delta^b(X_{i2}) \right|^2 K_h^2(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right) \leq \mathbf{E}_{\mathcal{I}_a} \left[ \left| \Delta^b(X_{i2}) \right|^2 \cdot \mathbf{E}[K_h(X_{i1}) | X_{i2}] \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right]$$

By (57) in the supplement, we have  $\mathbf{E}[K_h(X_{i1}) | X_{i2}] \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \lesssim q(x_0 | X_{i2})$ . Since  $q(x_0 | X_{i2})$  is upper bounded by a constant, we have  $\mathbf{E}_{\mathcal{I}_a} \left( \frac{1}{n} \sum_{i \in \mathcal{I}_a} \left| \Delta^b(X_{i2}) \right|^2 K_h^2(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \right) \lesssim$

$\text{Err}^2(\hat{f}_2^b)$  and hence  $\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\Delta(X_{i2})^2K_h(X_{i1})\right|\leq t^2\text{Err}^2(\hat{f}_2)\right)\geq 1-\frac{1}{t^2}-\frac{1}{n^c}-\gamma(n)$ . Combined with (48), we show  $\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\left(\hat{D}_{i1}-(D_{i1}-\bar{\mu}_D)\right)\Delta(X_{i2})K_h(X_{i1})\right|\lesssim t\text{Err}(\hat{D})\cdot\text{Err}(\hat{f}_2)\right)\geq 1-\frac{1}{t^2}-\frac{1}{n^c}-\gamma(n)$ . By combining (58) and (63) in the supplement with (49), we establish  $\mathbf{P}\left(\left|\bar{\mu}_D\cdot\frac{1}{n}\sum_{i=1}^n\Delta(X_{i2})K_h(X_{i1})\right|\lesssim t\text{Err}(\hat{f}_2)\sqrt{h/n}\right)\geq 1-\frac{1}{t^2}-\frac{1}{n^c}-\gamma(n)$ . By the decomposition (47), we have  $\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\hat{D}_{i1}\Delta(X_{i2})K_h(X_{i1})\right|\lesssim t\left(\sqrt{h/n}+\text{Err}(\hat{D})\right)\text{Err}(\hat{f}_2)\right)\geq 1-\frac{1}{t^2}-\frac{1}{n^c}-\gamma(n)$ . Together with Lemma 8 (specifically, (67)) in the supplement, we establish (30).

## References

- [1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- [2] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.
- [3] Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.
- [4] Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.
- [5] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.
- [6] T Tony Cai and Zijian Guo. Semi-supervised inference for explained variance in high-dimensional linear regression and its applications. *arXiv preprint arXiv:1806.06179*, 2018.
- [7] David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, 2001.
- [8] GM Changchien. Optimization of blast furnace burden distribution. In *Proceedings of the 1990 Taipei Symposium in Statistics*, pages 63–78, 1990.
- [9] Hung Chen, Mong-Na Lo Huang, and Wen-Jang Huang. Estimation of the location of the maximum of a regression function using extreme order statistics. *Journal of multivariate analysis*, 57(2):191–214, 1996.



- [10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.
- [11] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.
- [12] Melissa Dell, Benjamin F Jones, and Benjamin A Olken. What do we learn from the weather? the new climate-economy literature. *Journal of Economic Literature*, 52(3):740–98, 2014.
- [13] Olivier Deschênes and Michael Greenstone. The economic impacts of climate change: evidence from agricultural output and random fluctuations in weather: reply. *American Economic Review*, 102(7):3761–73, 2012.
- [14] Jianqing Fan. Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004, 1992.
- [15] Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993.
- [16] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press, 1996.
- [17] Theo Gasser and Hans-Georg Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, pages 171–185, 1984.
- [18] Karl Gregory, Enno Mammen, and Martin Wahl. Optimal estimation of sparse high-dimensional additive models. *arXiv preprint arXiv:1603.07632*, 2016.
- [19] Zijian Guo and Cun-Hui Zhang. Extreme nonlinear correlation for multiple random variables and stochastic processes with applications to additive models. *arXiv preprint arXiv:1904.12897*, 2019.
- [20] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.
- [21] Joel L Horowitz and Enno Mammen. Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, 32(6):2412–2443, 2004.
- [22] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

- [23] Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.
- [24] Junwei Lu, Mladen Kolar, and Han Liu. Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *arXiv preprint arXiv:1503.02978*, 2015.
- [25] Enno Mammen, Oliver Linton, and J Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27(5):1443–1490, 1999.
- [26] Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.
- [27] Jean D Opsomer. Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73(2):166–179, 2000.
- [28] Gengsheng Qin and Min Tsao. Empirical likelihood based inference for the derivative of the nonparametric regression function. *Bernoulli*, 11(4):715–735, 2005.
- [29] MK Raghuraman, Elizabeth A Winzeler, David Collingwood, Sonia Hunt, Lisa Wodicka, Andrew Conway, David J Lockhart, Ronald W Davis, Bonita J Brewer, and Walton L Fangman. Replication dynamics of the yeast genome. *science*, 294(5540):115–121, 2001.
- [30] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012.
- [31] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.
- [32] James M Robins and Andrea Rotnitzky. Comment on the bickel and kwon article, “inference for semiparametric models: Some questions and an answer”. *Statistica Sinica*, 11(4):920–936, 2001.
- [33] Wolfram Schlenker and Michael J Roberts. Estimating the impact of climate change on crop yields: The importance of nonlinear temperature effects. Technical report, National Bureau of Economic Research, 2008.
- [34] Peter X-K Song, Xin Gao, Rui Liu, and Wen Le. Nonparametric inference for local extrema with application to oligonucleotide microarray data in yeast genome. *Biometrics*, 62(2):545–554, 2006.

- [35] Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.
- [36] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101(2):269–284, 2012.
- [37] Taiji Suzuki and Masashi Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, pages 1381–1405, 2013.
- [38] Zhiqiang Tan and Cun-Hui Zhang. Doubly penalized estimation in additive regression with high-dimensional data. *arXiv preprint arXiv:1704.07229*, 2017.
- [39] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.
- [40] Larry Wasserman and John D Lafferty. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pages 801–808, 2008.
- [41] Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.
- [42] Yun Yang, Surya T Tokdar, et al. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.
- [43] Ming Yuan, Ding-Xuan Zhou, et al. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.
- [44] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.
- [45] Shanggang Zhou and Douglas A Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108, 2000.
- [46] Ying Zhu, Zhuqing Yu, and Guang Cheng. High dimensional inference in partially linear models. *Available at SSRN 3015397*, 2017.

## A Proof

In this section, we provide all remaining proofs. We recall notations and introduce several useful lemmas in Section A.1; we present the proof of Lemma 1 in Section A.2; we present the proof of Lemma 2 in Section A.3; we present the proofs of Theorems 2, 3 and 4 in Section A.4; we present the proof of Lemma 3 in Section A.5.

### A.1 Preliminary Analysis

We introduce the notation of conditional distribution of  $X_{i1}$  given  $X_{i2}$  as  $q(X_{i1} \mid X_{i2})$ . Then the marginal density of  $X_{i1}$  is expressed as

$$\pi(x_0) = \mathbf{E}q(X_{i1} = x_0 \mid X_{i2}) = \mathbf{E}\phi\left(\frac{x_0 - X_{i2}^\top \gamma}{\sigma_2}\right), \quad (50)$$

where the last equality follows from the assumption that  $X_{i1} - X_{i2}^\top \gamma$  follows a Gaussian distribution. We define the following event,

$$\begin{aligned} \mathcal{A}_1 &= \left\{ \|\hat{\gamma} - \gamma_2\| \lesssim \sqrt{\frac{k \log p}{n}} \right\} \\ \mathcal{A}_2 &= \left\{ |\hat{\sigma}_2^2 - \sigma_2^2| \lesssim \frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right\} \\ \mathcal{A}_{3,i} &= \left\{ \|X_{i2}^\top \gamma\|_2 \lesssim \|\gamma\|_2 \sqrt{\log n} \right\} \end{aligned} \quad (51)$$

and  $\mathcal{A}_3 = \cap_{i=1}^n \mathcal{A}_{3,i}$ . Define  $\mathcal{A} = \cap_{i=1}^3 \mathcal{A}_i$  and under the conditions (A1) and (A3), we can apply the maximal inequality and establish the following high probability result,

$$\mathbf{P}(\mathcal{A}) \geq 1 - n^{-c} \text{ for some constant } c > 1. \quad (52)$$

We introduce the following lemmas to facilitate the proof.

**Lemma 5** *If  $0 \leq b - a \leq 1$ , then*

$$\begin{aligned} \frac{1}{b-a} \int_a^b \phi(z) dz &\geq e^{-\frac{3}{2}} \min \left\{ 1, \frac{1}{(b-a) \cdot \min\{|a|, |b|\}} \right\} \phi(\min\{|a|, |b|\}) \\ &\geq e^{-2} \min \left\{ 1, \frac{1}{(b-a) \cdot \min\{|a|, |b|\}} \right\} \max_{z \in [a, b]} \phi(z) \end{aligned} \quad (53)$$

**Lemma 6** *Suppose that the bandwidth  $h$  satisfies  $hC_u \leq 1$  with  $C_u$  defined in (17), then*

$$\frac{\mathbf{E}K_h(X_{i1})}{2\pi(x_0)} \rightarrow 1. \quad (54)$$

$$\frac{\mathbf{E}D_{i1}^2 K_h^2(X_{i1})}{\frac{2}{3}h\pi(x_0)} \rightarrow 1 \quad (55)$$

$$\frac{\mathbf{E}D_{i1}(X_{i1} - x_0)K_h(X_{i1})}{\frac{2}{3}h^2\pi(x_0)} \rightarrow 1 \quad (56)$$

In particular, we have

$$\left| \frac{\mathbf{E}[K_h(X_{i1}) | X_{i2}]}{2q(x_0 | X_{i2})} - 1 \right| \cdot \mathbf{1}_{A_{3,i}} \lesssim h^2 (1 + C_u^2) \exp\left(C_u \cdot \frac{h}{\sigma_2}\right) \quad (57)$$

**Lemma 7** Suppose that the bandwidth  $h$  satisfies  $hC_u \leq 1$  with  $C_u$  defined in (17). For a sufficiently large  $n$ , with probability  $1 - \frac{1}{t}$ ,

$$c\pi(x_0) \left[ 1 - \frac{t}{\sqrt{nh\pi(x_0)}} \right] \leq \left| \frac{1}{n} \sum_{i=1}^n K_h(X_{i1}) \right| \lesssim C\pi(x_0) \left[ 1 + \frac{t}{\sqrt{nh\pi(x_0)}} \right]. \quad (58)$$

$$\left| \frac{1}{n} \sum_{i=1}^n D_{i1} K_h(X_{i1}) \right| \lesssim Ct \sqrt{\frac{h}{n} \pi(x_0)} \quad (59)$$

$$\left| \frac{1}{n} \sum_{i=1}^n (X_{i1} - x_0) K_h(X_{i1}) \right| \lesssim C_u h^2 \pi(x_0) + t \sqrt{\frac{h}{n} \pi(x_0)} \quad (60)$$

$$h^2 \pi(x_0) \left( 1 - t \sqrt{\frac{1}{4nh\pi(x_0)}} \right) \lesssim \frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1}) \lesssim h^2 \pi(x_0) \left( 1 + t \sqrt{\frac{1}{4nh\pi(x_0)}} \right) \quad (61)$$

$$\left| \frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right| \lesssim C_u h^4 \pi(x_0) + t \sqrt{\frac{h^5}{4n} \pi(x_0)} \quad (62)$$

Combining (58) and (59), we have

$$|\bar{\mu}_D| \leq t \sqrt{\frac{2h}{3n\pi(x_0)}} \quad (63)$$

In addition, if  $X_{i,-1}$  is Sub-gaussian random vector, then with probability larger than  $1 - \frac{1}{t}$ ,

$$\frac{1}{n} \sum_{i=1}^n [X_{i,-1}^\top \gamma]^2 K_h(X_{i1}) \lesssim \left( 1 + \frac{t}{\sqrt{nh}} \right) \|\gamma\|_2^2 \quad (64)$$

## A.2 Proof of Lemma 1

We start with deriving the explicit formula for  $\mathbf{E}([X_{i1} - x_0]K_h(X_{i1})|X_{i2})$  and  $\mathbf{E}(K_h(X_{i1})|X_{i2})$ ,

$$\mathbf{E}([X_{i1} - x_0]K_h(X_{i1})|X_{i2}) = \int_{\left|\frac{X_{i1}-x_0}{h}\right| \leq 1} \left[ \frac{X_{i1} - x_0}{h} \cdot q(X_{i1} | X_{i2}) \right] dX_{i1}.$$

We transform the variable  $X_{i1}$  to the standardized variable  $t = \frac{X_{i1} - X_{i2}^\top \gamma}{\sigma_2}$ , then we have

$$\int_{\left|\frac{X_{i1}-x_0}{h}\right| \leq 1} \left[ \frac{X_{i1} - x_0}{h} \cdot q(X_{i1} | X_{i2}) \right] dX_{i1} = \frac{\sigma_2^2}{h} \int_{\mu_i - L_i}^{\mu_i + L_i} (t - \mu_i) \phi(t) dt,$$

where

$$\mu_i = \frac{x_0 - X_{i2}^\top \gamma}{\sigma_2} \quad \text{and} \quad L_i = \frac{h}{\sigma_2}.$$

Similarly, we have

$$\mathbf{E}(K_h(X_{i1})|X_{i2}) = \int_{\left|\frac{X_{i1}-x_0}{h}\right| \leq 1} \left[ \frac{1}{h} \cdot q(X_{i1} | X_{i2}) \right] dX_{i1} = \frac{\sigma_2}{h} \int_{\mu_i - L_i}^{\mu_i + L_i} \phi(t) dt.$$

Hence, we establish (10). In the following, we shall approximate  $\int_{\mu_i - L_i}^{\mu_i + L_i} (t - \mu_i) \phi(t) dt$  and  $\int_{\mu_i - L_i}^{\mu_i + L_i} \phi(t) dt$ . By change of variable to  $s = t - \mu_i$ , then we have

$$\int_{\mu_i - L_i}^{\mu_i + L_i} (t - \mu_i) \phi(t) dt = \int_{-L_i}^{L_i} s \phi(\mu_i + s) ds = \phi(\mu_i) \int_{-L_i}^{L_i} s \exp(-\mu_i s - \frac{s^2}{2}) ds$$

where the last equality follows from the fact that  $\phi$  is Gaussian. We apply a Taylor expansion and establish

$$\int_{-L_i}^{L_i} s \exp(-\mu_i s - \frac{s^2}{2}) ds = \int_{-L_i}^{L_i} s \left( 1 - \mu_i s - \frac{s^2}{2} + \frac{1}{2} \left( \mu_i s + \frac{s^2}{2} \right)^2 + C \left( \mu_i s + \frac{s^2}{2} \right)^3 \right) ds$$

for some positive constant  $C > 0$ . Since  $\int_{-L_i}^{L_i} s^q = 0$  for an odd  $q$  and  $|\mu_i| \lesssim \|\gamma\|_2 \sqrt{\log n}$  on the event  $\mathcal{A}_{3,i}$ , we have

$$\int_{-L_i}^{L_i} s \exp(-\mu_i s - \frac{s^2}{2}) ds = -\frac{2}{3} \mu_i L_i^3 + O_p \left( h^5 \left( \sqrt{\log n} \right)^3 \right)$$

Similarly, for  $\int_{\mu_i - L_i}^{\mu_i + L_i} \phi(t) dt$ , we have

$$\int_{\mu_i - L_i}^{\mu_i + L_i} \phi(t) dt = \int_{-L_i}^{L_i} \phi(\mu_i + s) ds = \phi(\mu_i) \int_{-L_i}^{L_i} \exp(-\mu_i s - \frac{s^2}{2}) ds$$

Note that

$$\int_{-L_i}^{L_i} \exp(-\mu_i s - \frac{s^2}{2}) ds = \int_{-L_i}^{L_i} \left( 1 - \mu_i s - \frac{s^2}{2} + C \left( \mu_i s + \frac{s^2}{2} \right)^2 \right) ds = 2L_i + O_p(h^3 \log n)$$

Hence

$$\sigma_2 \frac{\int_{\mu_i-L_i}^{\mu_i+L_i} (t-\mu_i) \phi(t) dt}{\int_{\mu_i-L_i}^{\mu_i+L_i} \phi(t) dt} = \sigma_2 \frac{-\frac{2}{3}\mu_i L_i^3 + O_p(h^5(\sqrt{\log n})^3)}{2L_i + O_p(h^3 \log n)} = -\frac{\sigma_2}{3}\mu_i L_i^2 + O_p(h^4(\sqrt{\log n})^3)$$

Then we have

$$\left| l(X_{i2}, \hat{\gamma}, \hat{\sigma}_2) - \frac{h^2}{3\hat{\sigma}_2^2} (X_{i2}^\top \hat{\gamma} - x_0) \right| \lesssim \frac{h^2}{\sigma_2^4} \left[ \left( 1 + \left| \frac{\sigma_2^2}{\hat{\sigma}_2^2} - 1 \right| \right) |X_i^\top (\hat{\gamma} - \gamma)| + \left| \frac{\sigma_2^2}{\hat{\sigma}_2^2} - 1 \right| |X_i^\top \gamma| \right] + O_p(h^4(\sqrt{\log n})^3)$$

Under the assumption (A4), we have

$$\begin{aligned} \sqrt{\frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1})^2 K_h(X_{i1})} &\lesssim \frac{h^2}{\sigma_2^4} \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i^\top (\hat{\gamma} - \gamma))^2 K_h(X_{i1})} \\ &+ \frac{h^2}{\sigma_2^4} \left( \frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i^\top \gamma)^2 K_h(X_{i1}) + O_p(h^4(\sqrt{\log n})^3)} \end{aligned} \quad (65)$$

Note that

$$\left( \hat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right)^2 \lesssim \left( \tilde{D}_{i1} - D_{i1} \right)^2 + \left( \frac{\frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1}) K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} \right)^2$$

By Cauchy-Schwarz inequality

$$\left( \frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1}) K_h(X_{i1}) \right)^2 \leq \left( \frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1})^2 K_h(X_{i1}) \right) \cdot \left( \frac{1}{n} \sum_{i=1}^n K_h(X_{i1}) \right),$$

we have

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{\frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1}) K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} \right)^2 K_h(X_{i1}) \leq \frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1})^2 K_h(X_{i1})$$

and hence

$$\begin{aligned} \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{D}_{i1} - (D_{i1} - \bar{\mu}_D))^2 K_h(X_{i1})} &\lesssim \sqrt{\frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1})^2 K_h(X_{i1})} \\ &+ \sqrt{\frac{1}{n} \sum_{i=1}^n \left( \frac{\frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1}) K_h(X_{i1})}{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} \right)^2 K_h(X_{i1})} \lesssim \sqrt{\frac{1}{n} \sum_{i=1}^n (\tilde{D}_{i1} - D_{i1})^2 K_h(X_{i1})} \end{aligned}$$

Combined with (64) and (65), we establish the rate of convergence for  $\text{Err}(\hat{D})$ .

### A.3 Proof of Lemma 2

We will separate the proof of Lemma 2 into two parts, analysis of stochastic error in Section A.3.1 and analysis of approximation error in Section A.3.2.

### A.3.1 Analysis of stochastic error

We introduce the following lemma to facilitate the proof and present the corresponding proof in Section A.3.3.

**Lemma 8** *Under the condition that  $\text{Err}(\hat{D}) \ll h\sqrt{\pi(x_0)}$ ,  $hC_u \leq 1$  and  $nh\pi(x_0) \rightarrow \infty$  where  $C_u$  is defined in (17), then*

$$\frac{\frac{1}{n} \sum_{i=1}^n \hat{D}_{i1}^2 K_h^2(X_{i1})}{\frac{2}{3} h \pi(x_0)} \xrightarrow{p} 1 \quad (66)$$

$$\frac{\hat{S}_n}{\frac{2}{3} h^2 \pi(x_0)} \xrightarrow{p} 1 \quad (67)$$

In addition, with probability larger than  $1 - (nh\pi(x_0))^{-\frac{1}{4}}$ ,

$$\left| \frac{\hat{S}_n}{\frac{2}{3} h^2 \pi(x_0)} - 1 \right| \lesssim \frac{\text{Err}(\hat{D})}{h \pi(x_0)} + (nh\pi(x_0))^{-\frac{1}{4}} + \frac{(nh\pi(x_0))^{\frac{1}{4}}}{n} \quad (68)$$

In the following, we establish the asymptotic limiting distribution by first conditioning on  $X$ . Set

$$W_i = \frac{1}{\sqrt{V}} \frac{\hat{D}_{i1} \epsilon_i K_h(X_{i1})}{n \hat{S}_n}$$

and then we have  $\mathbf{E}_{\epsilon|X} \sum_{i=1}^n W_i^2 = 1$ . It is sufficient to check the Lindeberg's condition

$$\begin{aligned} \sum_{i=1}^n \mathbf{E}_{\epsilon|X} [W_i^2 \cdot \mathbf{1}\{|W_i| > \delta_0\}] &\leq \sum_{i=1}^n \frac{\hat{D}_{i1}^2 K_h^2(X_{i1})}{n^2 \hat{S}_n^2 V} \mathbf{E}_{\epsilon|X} \epsilon_i^2 \mathbf{1}\left\{ \max_{1 \leq i \leq n} \left| \frac{1}{\sqrt{V}} \frac{\hat{D}_{i1} \epsilon_i K_h(X_{i1})}{n \hat{S}_n} \right| > \delta_0 \right\} \\ &= \mathbf{E}_{\epsilon|X} \frac{\epsilon_i^2}{\sigma_1^2} \mathbf{1}\left\{ \max_{1 \leq i \leq n} \left| \frac{1}{\sqrt{V}} \frac{\hat{D}_{i1} \epsilon_i K_h(X_{i1})}{n \hat{S}_n} \right| > \delta_0 \right\} \lesssim \left( \mathbf{P} \left\{ \max_{1 \leq i \leq n} \left| \frac{1}{\sqrt{V}} \frac{\hat{D}_{i1} \epsilon_i K_h(X_{i1})}{n \hat{S}_n} \right| > \delta_0 \right\} \right)^{\frac{\tau}{2+\tau}} \end{aligned}$$

where the last inequality follows from the bounded  $2 + \tau$  moments for  $\epsilon_i$ . To bound the last term in the above inequality, we use the bounded  $2 + \tau$  moments for  $\epsilon_i$ ,

$$\begin{aligned} \mathbf{P} \left\{ \max_{1 \leq i \leq n} \left| \frac{1}{\sqrt{V}} \frac{\hat{D}_{i1} \epsilon_i K_h(X_{i1})}{n \hat{S}_n} \right| > \delta_0 \right\} &\lesssim \left( \frac{\sqrt{V} n \hat{S}_n}{h + \max_{1 \leq i \leq n} |\hat{D}_{i1} - D_{i1}|} \right)^{-(2+\tau)} \\ &\lesssim \left( \frac{\sqrt{\sum_{i=1}^n \hat{D}_{i1}^2 K_h^2(X_{i1})}}{h + \max_{1 \leq i \leq n} |\hat{D}_{i1} - D_{i1}|} \right)^{-(2+\tau)} \end{aligned}$$

Note that  $\max_{1 \leq i \leq n} |\hat{D}_{i1} - D_{i1}| \leq |\bar{\mu}_D| + \max_{1 \leq i \leq n} |\tilde{D}_{i1} - D_{i1}|$  and

$$\max_{1 \leq i \leq n} |\tilde{D}_{i1} - D_{i1}| \lesssim h^2 \max_{1 \leq i \leq n} |X_{i2}^T(\hat{\gamma} - \gamma)| + h^4 (\sqrt{\log n})^3 \lesssim h^2 \sqrt{\frac{k \log p \log n}{n}} + h^4 (\sqrt{\log n})^3.$$



Then we have  $\max_{1 \leq i \leq n} |\hat{D}_{i1} - D_{i1}| \leq h$ . Define

$$\Omega_{\eta_0} = \left\{ \left| \frac{\frac{1}{n} \sum_{i=1}^n \hat{D}_{i1}^2 K_h^2(X_{i1})}{\frac{2}{3} h \pi(x_0)} - 1 \right| \leq \eta_0 \text{ and } \max_{1 \leq i \leq n} |\hat{D}_{i1} - D_{i1}| \leq h \right\}$$

with  $0 < \eta_0 < 1/10$ . By (66) and Lemma 1, then we have

$$\mathbf{P}(\Omega_{\eta_0}) \rightarrow 1. \quad (69)$$

On the event  $\Omega_{\eta_0}$ , we can check that Lindeberg's condition is satisfied and hence we have

$$\sum_{i=1}^n \frac{1}{\sqrt{V}} \frac{\hat{D}_{i1} \epsilon_i K_h(X_{i1})}{n \hat{S}_n} \mid X \xrightarrow{d} N(0, 1).$$

For any bound function  $b$ , then we have

$$\mathbf{E} \left[ b \left( \sum_{i=1}^n W_i \right) \right] = \mathbf{E}_X \mathbf{E} \left[ b \left( \sum_{i=1}^n W_i \right) \mid X \right] \cdot \mathbf{1}_{\Omega_{\eta_0}} + \|b\|_{\infty} \mathbf{P}(\Omega_{\eta_0}^c).$$

Note that

$$\mathbf{E}_X \mathbf{E} \left[ b \left( \sum_{i=1}^n W_i \right) \mid X \right] \cdot \mathbf{1}_{\Omega_{\eta_0}} \rightarrow \mathbf{E}[b(Z)] \mathbf{P}(\Omega_{\eta_0}),$$

where  $Z$  follows standard normal distribution. By (69), we establish (25). Combining (66) and (67), we establish the asymptotic limit of  $V$ .

### A.3.2 Analysis of approximation error

The control of the approximation error in (22) follows from a combination of (67) in Lemma 8 and the control of  $\frac{1}{n} \sum_{i=1}^n \hat{D}_{i1} r(X_{i1}) K_h(X_{i1})$ . Note that

$$\begin{aligned} r(X_{i1}) &= f_1(X_{i1}) - f(x_0) - (X_{i1} - x_0) f'(x_0) \\ &= \frac{(X_{i1} - x_0)^2}{2} f''(x_0) + \frac{(X_{i1} - x_0)^2}{2} [f''(x_0 + c(X_{i1} - x_0)) - f''(x_0)] \end{aligned} \quad (70)$$

for some  $c \in (0, 1)$ . Hence, we have

$$\frac{2}{h^2} \left| r(X_{i1}) \mathbf{1} \left( \left| \frac{X_{i1} - x_0}{h} \right| \leq 1 \right) - \frac{(X_{i1} - x_0)^2}{2} f''(x_0) \mathbf{1} \left( \left| \frac{X_{i1} - x_0}{h} \right| \leq 1 \right) \right| \leq |f''(x_1) - f''(x_0)|, \quad (71)$$

for some  $x_1$  satisfying  $x_0 - h \leq x_1 \leq x_0 + h$ . Since  $h = h(n) \rightarrow 0$  and  $f''(x)$  is continuous at  $x_0$ , then we have

$$\frac{2}{h^2} \left| r(X_{i1}) \mathbf{1} \left( \left| \frac{X_{i1} - x_0}{h} \right| \leq 1 \right) - \frac{(X_{i1} - x_0)^2}{2} f''(x_0) \mathbf{1} \left( \left| \frac{X_{i1} - x_0}{h} \right| \leq 1 \right) \right| \rightarrow 0. \quad (72)$$

Hence we have

$$\frac{2}{h^2} \left| \frac{1}{n} \sum_{i=1}^n \hat{D}_{i1} r(X_{i1}) K_h(X_{i1}) - \frac{1}{n} \sum_{i=1}^n \hat{D}_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \right| = o \left( \frac{1}{n} \sum_{i=1}^n |\hat{D}_{i1}| K_h(X_{i1}) \right). \quad (73)$$

With the above calculation, the problem of controlling the approximation error is reduced to the control of the two terms  $\frac{1}{n} \sum_{i=1}^n \hat{D}_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1})$  and  $\frac{1}{n} \sum_{i=1}^n |\hat{D}_{i1}| K_h(X_{i1})$ , which are established in the following lemma. The proof of the following lemma is present in Section A.3.4.

**Lemma 9** *Suppose that  $hC_u \rightarrow 0$  and  $nh\pi(x_0) \rightarrow \infty$ , then with probability larger than  $1 - (nh\pi(x_0))^{-\frac{1}{4}}$ ,*

$$\left| \frac{1}{nh^2\pi(x_0)} \sum_{i=1}^n \hat{D}_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \right| \lesssim \frac{\text{Err}(\hat{D})}{\sqrt{\pi(x_0)}} + h \left( hC_u + \frac{1}{(nh\pi(x_0))^{\frac{1}{4}}} \right) \quad (74)$$

$$\frac{1}{n\pi(x_0)} \sum_{i=1}^n |\hat{D}_{i1}| K_h(X_{i1}) \lesssim \frac{\text{Err}(\hat{D})}{\sqrt{\pi(x_0)}} + h \quad (75)$$

Combination of (67) and (74) leads to (23). Combining (67), (73) and (75), we establish (22).

### A.3.3 Proof of Lemma 8

To establish (66), we decompose the error between  $\frac{1}{n} \sum_{i=1}^n \hat{D}_{i1}^2 K_h^2(X_{i1})$  and its corresponding estimand,

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n \hat{D}_{i1}^2 K_h^2(X_{i1}) - \frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D)^2 K_h^2(X_{i1}) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n \left[ 2(D_{i1} - \bar{\mu}_D) \cdot (\hat{D}_{i1} - (D_{i1} - \bar{\mu}_D)) + (\hat{D}_{i1} - (D_{i1} - \bar{\mu}_D))^2 \right] K_h^2(X_{i1}) \right| \quad (76) \\ &\leq \frac{1}{h} \left( \text{Err}^2(\hat{D}) + \text{Err}(\hat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D)^2 K_h^2(X_{i1})} \right) \end{aligned}$$

where the inequality follows from triangle inequality and Cauchy-Schwarz inequality. To establish (67), we approximate  $\hat{S}_n$  by its corresponding estimand,

$$\begin{aligned} & \left| \hat{S}_n - \frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D) (X_{i1} - x_0) K_h(X_{i1}) \right| \\ &\leq \text{Err}(\hat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (X_{i1} - x_0)^2 K_h(X_{i1})} \leq h \cdot \text{Err}(\hat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})}. \quad (77) \end{aligned}$$

We apply Law of Large Numbers and Lemma 6,

$$\frac{\frac{1}{n} \sum_{i=1}^n D_{i1}^2 K_h^2(X_{i1})}{\frac{2}{3} h \pi(x_0)} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})}{\frac{2}{3} h^2 \pi(x_0)} \xrightarrow{p} 1. \quad (78)$$

We bound the difference between the sum of centered variables and that of uncentered variables,

$$\left| \frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D)^2 K_h^2(X_{i1}) - \frac{1}{n} \sum_{i=1}^n D_{i1}^2 K_h^2(X_{i1}) \right| \leq 2 |\bar{\mu}_D| \cdot \left| \frac{1}{n} \sum_{i=1}^n D_{i1} K_h^2(X_{i1}) \right| + 2 \bar{\mu}_D^2 \left| \frac{1}{n} \sum_{i=1}^n K_h^2(X_{i1}) \right|$$

It follows from Lemma 7 that, with probability larger than  $1 - \frac{1}{t}$ ,

$$\left| \frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D)^2 K_h^2(X_{i1}) - \frac{1}{n} \sum_{i=1}^n D_{i1}^2 K_h^2(X_{i1}) \right| \lesssim 2t \sqrt{\frac{2}{3nh\pi(x_0)}} \sqrt{\frac{h}{n} \pi(x_0)} + \frac{2}{3n} \lesssim \frac{t}{n} \quad (79)$$

Similarly, we have

$$\left| \frac{1}{n} \sum_{i=1}^n \bar{\mu}_D (X_{i1} - x_0) K_h(X_{i1}) \right| = |\bar{\mu}_D| \cdot \left| \frac{1}{n} \sum_{i=1}^n (X_{i1} - x_0) K_h(X_{i1}) \right|$$

It follows from Lemma 7 that, with probability larger than  $1 - \frac{1}{t}$ ,

$$\left| \frac{1}{n} \sum_{i=1}^n \bar{\mu}_D (X_{i1} - x_0) K_h(X_{i1}) \right| \lesssim 2 \sqrt{\frac{2h}{3n\pi(x_0)}} \cdot \left( C_u h^2 \pi(x_0) + t \sqrt{\frac{h}{n} \pi(x_0)} \right) \lesssim C_u \cdot \sqrt{\frac{h^5}{n} \cdot \pi(x_0)} + t \frac{h}{n}. \quad (80)$$

By taking  $t = \sqrt{nh\pi(x_0)}$ , we combine (79), (80) and (78) and establish

$$\frac{\frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D)^2 K_h^2(X_{i1})}{\frac{2}{3} h \pi(x_0)} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D) (X_{i1} - x_0) K_h(X_{i1})}{\frac{2}{3} h^2 \pi(x_0)} \xrightarrow{p} 1. \quad (81)$$

Also note the following fact

$$\left| \frac{\frac{1}{h} \left( \text{Err}^2(\hat{D}) + \text{Err}(\hat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (D_{i1} - \bar{\mu}_D)^2 K_h^2(X_{i1})} \right)}{\frac{2}{3} h \pi(x_0)} \right| \lesssim \frac{\text{Err}^2(\hat{D})}{h^2 \pi(x_0)} + \frac{\text{Err}(\hat{D})}{h \sqrt{\pi(x_0)}}$$

and

$$\left| \frac{h \cdot \text{Err}(\hat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})}}{\frac{2}{3} h^2 \pi(x_0)} \right| \lesssim \frac{\text{Err}(\hat{D})}{h \sqrt{\pi(x_0)}} \quad (82)$$

Combined with (76) and (77), we establish (66) and (67). In addition, together with (80) and (82), we apply (61) with  $t = (4nh\pi(x_0))^{\frac{1}{4}}$  and establish (68).

### A.3.4 Proof of Lemma 9

By the expression  $\widehat{D}_{i1} = (D_{i1} - \bar{\mu}_D) + \widehat{D}_{i1} - (D_{i1} - \bar{\mu}_D)$ , we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \widehat{D}_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) &= \frac{1}{n} \sum_{i=1}^n \left[ \widehat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right] \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \\ &+ \frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) - \bar{\mu}_D \frac{1}{n} \sum_{i=1}^n \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \end{aligned} \quad (83)$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n \left[ \widehat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right] \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \right| \\ &\lesssim |f''(x_0)| \text{Err}(\widehat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(X_{i1} - x_0)^4}{2} K_h(X_{i1})} \leq |f''(x_0)| h^2 \text{Err}(\widehat{D}) \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} \end{aligned} \quad (84)$$

where the last inequality follows from the fact that  $\frac{(X_{i1} - x_0)^4}{2} K_h(X_{i1}) \leq h^4 K_h(X_{i1})$ . In addition, we have

$$\left| \frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \right| = |f''(x_0)| \cdot \left| \frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right| \quad (85)$$

and

$$\begin{aligned} \left| \bar{\mu}_D \frac{1}{n} \sum_{i=1}^n \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \right| &= |\bar{\mu}_D| \cdot |f''(x_0)| \cdot \left| \frac{1}{n} \sum_{i=1}^n \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right| \\ &\leq h^2 |\bar{\mu}_D| \cdot |f''(x_0)| \cdot \frac{1}{n} \sum_{i=1}^n K_h(X_{i1}) \end{aligned} \quad (86)$$

where the last inequality follows from the fact that  $\frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \leq h^2 K_h(X_{i1})$ . Together with Lemma 7, we have

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n \widehat{D}_{i1} \frac{(X_{i1} - x_0)^2}{2} f''(x_0) K_h(X_{i1}) \right| \lesssim h^2 \text{Err}(\widehat{D}) \sqrt{\pi(x_0)} + C_u h^4 \pi(x_0) \\ &+ t \sqrt{\frac{h^5}{4n} \pi(x_0)} + h^2 \sqrt{\frac{2h}{3n\pi(x_0)}} \pi(x_0) \lesssim \left[ \frac{\text{Err}(\widehat{D})}{\sqrt{\pi(x_0)}} + h \left( h C_u + \frac{1+t}{\sqrt{nh\pi(x_0)}} \right) \right] h^2 \pi(x_0). \end{aligned}$$

Taking  $t = (nh\pi(x_0))^{\frac{1}{4}}$ , then we establish (74). The proof of (75) follows from the following inequality,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \left| \widehat{D}_{i1} \right| K_h(X_{i1}) &\leq \frac{1}{n} \sum_{i=1}^n \left| \widehat{D}_{i1} - (D_{i1} - \bar{\mu}_D) \right| K_h(X_{i1}) + \frac{1}{n} \sum_{i=1}^n (|D_{i1}| + |\bar{\mu}_D|) K_h(X_{i1}) \\
&\leq \text{Err}(\widehat{D}) \sqrt{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} + \frac{2}{n} \sum_{i=1}^n |D_{i1}| K_h(X_{i1}) \\
&\leq \text{Err}(\widehat{D}) \sqrt{\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})} + 2h \left( \frac{1}{n} \sum_{i=1}^n K_h(X_{i1}) \right)
\end{aligned} \tag{87}$$

where the last inequality follows from the fact that  $|D_{i1}| K_h(X_{i1}) \leq h K_h(X_{i1})$ . Together with Lemma 7 with  $t = (nh\pi(x_0))^{\frac{1}{4}}$ , we establish (75).

#### A.4 Proof of Theorems 2, 3 and 4

By(22), (23), (26) and the conditions  $\text{Err}(\widehat{D}) \ll \sqrt{V} \sqrt{\pi(x_0)}$  and  $nh^5\pi(x_0) \leq c$ , we have

$$\frac{1}{\sqrt{V}} \frac{1}{n\widehat{S}_n} \sum_{i=1}^n \widehat{D}_{i1} r(X_{i1}) K_h(X_{i1}) = o_p(1) \tag{88}$$

It follows from (30) that

$$\frac{1}{\sqrt{V}} \frac{\frac{1}{n} \sum_{i=1}^n \widehat{D}_{i1} \Delta(X_{i2}) K_h(X_{i1})}{\widehat{S}_n} = O_p \left( \sqrt{\frac{\text{Err}^2(\widehat{f}_2)}{\pi(x_0)}} + \frac{\text{Err}(\widehat{D})}{h^2} \frac{\text{Err}(\widehat{f}_2)}{\pi(x_0)} \cdot \sqrt{nh^3 \cdot \pi(x_0)} \right). \tag{89}$$

Combining (88) and (89), we establish the limiting distribution (32). The proof of Theorem 3 follows from a combination of Lemma 2 and Theorem 1. Theorem 4 follows from Theorems 2 and 3 by taking  $\text{Err}(\widehat{D}) = 0$ .

#### A.5 Proof of Lemma 3

By [36], we can show that event  $\mathcal{A}$  happens with probability larger than  $1 - n^{-c}$  for some positive constant  $c$ . Hence, the results  $|\widehat{\sigma}_2 - \sigma_2| \lesssim \frac{1}{\sqrt{n}} + \frac{k \log p}{n}$  in (20) and

$$\mathbb{P} \left( \max_{1 \leq i \leq n} \left| X_{i,-1}^\top (\widehat{\gamma} - \gamma) \right| \lesssim \sqrt{k \log p \log n/n} \right) \rightarrow 0$$

follow from the definition of event  $\mathcal{A}$ . In the following, we shall control  $\frac{1}{n} \sum_{i=1}^n \left[ X_{i,-1}^\top (\hat{\gamma} - \gamma) \right]^2 K_h(X_{i1})$ . By the definition of data-swapping, we have

$$\sum_{i=1}^n \left( X_{i,-1}^\top (\hat{\gamma} - \gamma) \right)^2 K_h(X_{i1}) = \sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) + \sum_{i \in \mathcal{I}_b} \left( X_{i,-1}^\top (\hat{\gamma}^a - \gamma) \right)^2 K_h(X_{i1}). \quad (90)$$

By symmetry, it is sufficient to control  $\sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}}$ . Note that

$$\begin{aligned} \mathbf{E}_{\mathcal{I}_a} \sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} &\leq \mathbf{E}_{\mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} \\ &= \mathbf{E}_{X_{i,-1}} \left( \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 \mathbf{E}[K_h(X_{i1}) \mid X_{i,-1}] 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} \right) \end{aligned}$$

By (57), we have  $\mathbf{E}[K_h(X_{i1}) \mid X_{i,-1}] 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} \lesssim q(x_0 \mid X_{i,-1}) \lesssim C$ , then we further upper bound the above equation by

$$\mathbf{E}_{X_{i,-1}} \left( \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 \mathbf{E}[K_h(X_{i1}) \mid X_{i,-1}] 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} \right) \lesssim \frac{k \log p}{n}.$$

Note that  $\sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}}$  can be viewed as a summation of independent sub-exponential random variables with exponential norm  $\frac{1}{h} \frac{k \log p}{n}$ . By applying the Bernstein inequality, we establish that with probability larger than  $1 - n^{-1}$ , then

$$\left| \frac{1}{n_a} \sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} - \mathbf{E}_{\mathcal{I}_a} \sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} \right| \lesssim \sqrt{\frac{\log n}{n_a}} \frac{1}{h} \frac{k \log p}{n}$$

So if  $h = n^{-\delta_0}$  for  $0 < \delta_0 < \frac{1}{2}$ , we have  $\frac{1}{n_a} \sum_{i \in \mathcal{I}_a} \left( X_{i,-1}^\top (\hat{\gamma}^b - \gamma) \right)^2 K_h(X_{i1}) 1_{\mathcal{A}_2 \cap \mathcal{A}_{3,i}} \lesssim \frac{k \log p}{n}$  and hence we establish that with probability larger than  $1 - n^{-c}$ ,

$$\frac{1}{n} \sum_{i=1}^n \left( X_{i,-1}^\top (\hat{\gamma} - \gamma) \right)^2 K_h(X_{i1}) \lesssim \frac{k \log p}{n}.$$

## B Additional Theory on Variable Transformation

In the following, we will present a more general proposition of guaranteeing the lower bound for the marginal density function of the transformed variable, which includes the Example 2 in the main paper as a special case.

**Proposition 1** *Suppose that  $F$  is CDF for the random variable  $X$ ,  $C > 0$  and  $0 < c_0 < 1$  are some given constants,  $H(x)$  is an increasing differentiable function in  $x$  with  $H(-\infty) =$*

0 and  $H(\infty) = 1$  and  $G_0$  is an increasing differentiable function with  $G(-\infty) = 0$  and  $G(\infty) = 1$  and  $G'_0(x) = 0$  for  $|x| \geq C$ . For the transformation  $G$  defined as

$$G(X) = (1 - c_0) \cdot G_0(x) + c_0 H(X) \in [0, 1], \quad (91)$$

the marginal density  $q$  of the random variable  $G(X) \in [0, 1]$  satisfying

$$\min_{t \in [0, 1]} q(t) \geq \min \left\{ \min_{|x| \geq C} \frac{F'(x)}{2c_0 H'(x)}, \min_{|x| \leq C} \frac{F'(x)}{(1 - c_0)G'_0(x) + c_0 H'(x)} \right\} \quad (92)$$

The above results reveal that a key factor to determine the lower bound of the marginal density  $q$  defined in (91) is the ratio  $\frac{F'(x)}{2c_0 H'(x)}$  for large  $|x|$ . Since the minimum value over the bounded support,  $\min_{|x| \leq C} \frac{F'(x)}{(1 - c_0)G'_0(x) + c_0 H'(x)}$ , is relatively easy to be lower bounded, the above proposition provides the insight that we need to pay attention to the tail part of the derivative  $H'(x)$ .

## Proof of Proposition 1

We first note that  $G(X)$  is an increasing and differentiable function. To derive the density function of the transformed variable  $G(X)$ , we start with the CDF for  $G(X)$ .

$$\mathbf{P}(G(X) \leq t) = F(G^{-1}(t)) \quad (93)$$

and we take derivative and establish

$$q(t) = \frac{F'(G^{-1}(t))}{G'(G^{-1}(t))} \quad (94)$$

For  $|x| = |G^{-1}(t)| \geq C$ , we have

$$G'(G^{-1}(t)) = c_0 H'(G^{-1}(t)). \quad (95)$$

We establish (92) by combining (94) and (95).

## C Proof of Additional Lemmas

We present the proofs of additional lemmas in this section.

## C.1 Proof of Lemma 5

We first assume that  $|b| \geq |a|$  and hence

$$\int_a^b \phi(x) dx \geq (b-a)\phi(\max\{|a|, |b|\}) = (b-a)\phi(b).$$

In addition, we have

$$\frac{\phi(b)}{\phi(a)} = \exp\left(-\frac{(b-a)^2 + 2a(b-a)}{2}\right) \geq \exp\left(-\frac{1}{2} - |a(b-a)|\right). \quad (96)$$

We will separate the proof into two cases,

(a) We first consider  $|a(b-a)| \leq 1$ , then we have  $\frac{\phi(b)}{\phi(a)} \geq e^{-\frac{3}{2}}$  and hence

$$\int_a^b \phi(x) dx \geq e^{-\frac{3}{2}}(b-a)\phi(\min\{|a|, |b|\}). \quad (97)$$

(b) We then consider  $|a(b-a)| \geq 1$  and have  $b \geq a + \frac{1}{|a|}$ . Then we have

$$\int_a^b \phi(x) dx \geq \int_a^{a+\frac{1}{|a|}} \phi(x) dx \geq e^{-\frac{3}{2}} \frac{1}{|a|} \phi(\min\{|a|, |b|\}) \quad (98)$$

Combining (97) and (98), we establish

$$\frac{1}{b-a} \int_a^b \phi(z) dz \geq e^{-\frac{3}{2}} \min\left\{1, \frac{1}{(b-a)|a|}\right\} \phi(|a|) \quad \text{for } |b| \geq |a|.$$

Similarly, we establish

$$\frac{1}{b-a} \int_a^b \phi(z) dz \geq e^{-\frac{3}{2}} \min\left\{1, \frac{1}{(b-a)|b|}\right\} \phi(|b|) \quad \text{for } |b| \leq |a|.$$

Moreover, when  $ab > 0$ , we have  $\max_{z \in [a,b]} \phi(z) = \phi(\min\{|a|, |b|\})$ ; otherwise, since  $0 \leq a-b \leq 1$ , we have  $\max_{z \in [a,b]} \phi(z) \geq \frac{1}{\sqrt{e}} \phi(\min\{|a|, |b|\})$ . Then we establish the lemma.

## C.2 Proof of Lemma 6

### C.2.1 Proof of (54) and (57)

We focus on the analysis of  $\mathbf{E}[K_h(X_{i1}) | X_{i2}]$  in the following. We first characterize  $\mathbf{E}[K_h(X_{i1}) | X_{i2}]$  by its exact expression,

$$\mathbf{E}[K_h(X_{i1}) | X_{i2}] = \int_{\left|\frac{X_{i1}-x_0}{h}\right| \leq 1} \frac{1}{h} q(X_{i1} | X_{i2}) dX_{i1}$$



By setting  $z = \frac{X_{i1} - x_0}{h}$ , we can simplify  $\mathbf{E}[K_h(X_{i1}) | X_{i2}]$  as

$$\int_{|z| \leq 1} q(x_0 + hz | X_{i2}) dz = \int_{|z| \leq 1} \left[ q(x_0 | X_{i2}) + hz q'(x_0 | X_{i2}) + \frac{h^2 z^2}{2} q''(x_0 + c(z)hz | X_{i2}) \right] dz \quad (99)$$

for some  $c(z) \in (0, 1)$ . As a remark, we shall use  $c(z)$  as a generic function of  $z$  throughout the proof and the specific function  $c(z)$  can vary from place to place. Hence, we have

$$|\mathbf{E}[K_h(X_{i1}) | X_{i2}] - 2q(x_0 | X_{i2})| \leq \frac{2}{3} h^2 \max_{|c| \leq 1} q''(x_0 + ch | X_{i2}) \quad (100)$$

where

$$q''(x | X_{i2}) = \left( \frac{(x - X_{i2}^\top \gamma)^2}{\sigma_2^2} - 1 \right) \phi \left( \frac{x - X_{i2}^\top \gamma}{\sigma_2} \right).$$

On the event  $A_{3,i}$ , we have

$$\max_{|c| \leq 1} \left| \frac{x_0 + ch - X_{i2}^\top \gamma}{\sigma_2} \right| \leq C_u, \quad (101)$$

where  $C_u$  is defined in (17). A simple fact to facilitate the proof is

$$\frac{\phi(b)}{\phi(a)} = \exp \left( -\frac{(b-a)^2 + 2a(b-a)}{2} \right) \leq \exp(|a(b-a)|). \quad (102)$$

By applying (102), we have

$$\left| \frac{\max_{|c| \leq 1} q''(x_0 + ch | X_{i2}) \cdot \mathbf{1}_{A_{3,i}}}{q(x_0 | X_{i2})} \right| \lesssim (1 + C_u^2) \exp \left( C_u \cdot \frac{h}{\sigma_2} \right) \quad (103)$$

Together with (100), we establish (57). Then we have

$$\left| \frac{\mathbf{E}(K_h(X_{i1}) | X_{i2}) \mathbf{1}_{A_{3,i}}}{2q(x_0 | X_{i2})} - 1 \right| \leq \mathbf{1}_{A_{3,i}^c} + h^2 (1 + C_u^2) \exp \left( C_u \cdot \frac{h}{\sigma_2} \right).$$

and hence

$$\left| \frac{\mathbf{E}(K_h(X_{i1}) \mathbf{1}_{A_{3,i}})}{2\pi(x_0)} - 1 \right| \lesssim \frac{1}{n^c \pi(x_0)} + h^2 (1 + C_u^2) \exp \left( C_u \cdot \frac{h}{\sigma_2} \right). \quad (104)$$

In addition, we have

$$\left| \frac{\mathbf{E}(K_h(X_{i1}) \mathbf{1}_{A_{3,i}^c})}{2\pi(x_0)} \right| \leq \frac{\mathbf{E} \mathbf{1}_{A_{3,i}^c}}{2h\pi(x_0)} \leq \frac{1}{n^c h\pi(x_0)} \rightarrow 0$$

Together with (104), we have (54).

In addition to the previous analysis, we can also provide the following bound for the conditional expectation  $\mathbf{E}[K_h(X_{i1}) \mid X_{i2}] \mathbf{1}_{\mathcal{A}_{3,i}}$ . By Lemma 5, we have

$$\begin{aligned}
\mathbf{E}[K_h(X_{i1}) \mid X_{i2}] \mathbf{1}_{\mathcal{A}_{3,i}} &= \int_{\left|\frac{X_{i1}-x_0}{h}\right| \leq 1} \frac{1}{h} q(X_{i1} \mid X_{i2}) dX_{i1} \mathbf{1}_{\mathcal{A}_{3,i}} \\
&\geq C \min \left\{ 1, \frac{\sigma_2}{h \cdot \min\{|\mu_i - L_i|, |\mu_i + L_i|\}} \right\} \max_{w: |w - \mu_i| \leq L_i} \phi(w) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \\
&\geq C \min \left\{ 1, \frac{1}{h C_u} \right\} \max_{w: |w - \mu_i| \leq L_i} \phi(w) \\
&= C \min \left\{ 1, \frac{1}{h C_u} \right\} \max_{|c| \leq 1} q(x_0 + ch \mid X_{i2})
\end{aligned} \tag{105}$$

where  $\mu_i = (x_0 - X_{i2}^\top \gamma) / \sigma_2$ ,  $L_i = h / \sigma_2$  and  $C_u$  is defined as (17).

### C.2.2 Proof of (55)

We start with the following iterated expectation,

$$\begin{aligned}
\mathbf{E}(D_{i1}^2 K_h^2(X_{i1})) &= \mathbf{E}_{X_{i2}} \mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) \\
&= \mathbf{E}_{X_{i2}} [\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}] + \mathbf{E}_{X_{i2}} [\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}^c}]
\end{aligned}$$

We first analyze  $\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}$ , by noting that

$$\begin{aligned}
\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) &= \frac{1}{h} \mathbf{E}(D_{i1}^2 K_h(X_{i1}) \mid X_{i2}) \\
&= \frac{1}{h} \left( \mathbf{E}[(X_{i1} - x_0)^2 K_h(X_{i1}) \mid X_{i2}] - \frac{\mathbf{E}^2[(X_{i1} - x_0) K_h(X_{i1}) \mid X_{i2}]}{\mathbf{E}[K_h(X_{i1}) \mid X_{i2}]} \right),
\end{aligned} \tag{106}$$

where the last equality follows from the definition of  $D_{i1}$ . In the following, we provide upper bounds for  $\mathbf{E}[(X_{i1} - x_0) K_h(X_{i1}) \mid X_{i2}]$  and  $\mathbf{E}[(X_{i1} - x_0)^2 K_h(X_{i1}) \mid X_{i2}]$ .

Analysis of  $\mathbf{E}[(X_{i1} - x_0)^2 K_h(X_{i1}) \mid X_{i2}]$ . Similar to (99), we write down the following explicit expression,

$$\mathbf{E}[(X_{i1} - x_0)^2 K_h(X_{i1}) \mid X_{i2}] = \int_{\left|\frac{X_{i1}-x_0}{h}\right| \leq 1} [X_{i1} - x_0]^2 \frac{1}{h} q(X_{i1} \mid X_{i2}) dX_{i1}$$

By setting  $z = \frac{X_{i1}-x_0}{h}$ , we further have

$$\begin{aligned}
\mathbf{E}[(X_{i1} - x_0)^2 K_h(X_{i1}) \mid X_{i2}] &= \int_{|z| \leq 1} h^2 z^2 q(x_0 + hz \mid X_{i2}) dz \\
&= \int_{|z| \leq 1} h^2 z^2 \left[ q(x_0 \mid X_{i2}) + hz q'(x_0 \mid X_{i2}) + \frac{h^2 z^2}{2} q''(x_0 + c(z)hz \mid X_{i2}) \right] dz
\end{aligned} \tag{107}$$

Hence, we have where

$$\left| \mathbf{E} [(X_{i1} - x_0)^2 K_h(X_{i1}) | X_{i2}] - \frac{2}{3} h^2 q(x_0 | X_{i2}) \right| \leq \frac{4}{5} h^4 \max_{|c| \leq 1} q''(x_0 + ch | X_{i2}) \quad (108)$$

Analysis of  $\mathbf{E} [(X_{i1} - x_0) K_h(X_{i1}) | X_{i2}]$ . Similar to (99), we write down the following explicit expression,

$$\mathbf{E} [(X_{i1} - x_0) K_h(X_{i1}) | X_{i2}] = \int_{\left| \frac{X_{i1} - x_0}{h} \right| \leq 1} [X_{i1} - x_0] \frac{1}{h} q(X_{i1} | X_{i2}) dX_{i1}$$

Then we have

$$\begin{aligned} \mathbf{E} [(X_{i1} - x_0) K_h(X_{i1}) | X_{i2}] &= \int_{|z| \leq 1} h z q(x_0 + h z | X_{i2}) dz \\ &= \int_{|z| \leq 1} h z \left[ q(x_0 | X_{i2}) + h z q'(x_0 | X_{i2}) + \frac{h^2 z^2}{2} q''(x_0 | X_{i2}) + \frac{h^3 z^3}{6} q'''(x_0 + c(z) h z | X_{i2}) \right] dz \end{aligned}$$

Hence, we have

$$\left| \mathbf{E} [(X_{i1} - x_0) K_h(X_{i1}) | X_{i2}] - \frac{2}{3} h^2 q'(x_0 | X_{i2}) \right| \leq \frac{1}{15} h^4 \max_{|c| \leq 1} q'''(x_0 + ch | X_{i2}) \quad (109)$$

where

$$q'(x_0 | X_{i2}) = -\frac{x_0 - X_{i2}^\top \gamma}{\sigma_2} q(x_0 | X_{i2}) \quad (110)$$

and

$$q'''(x | X_{i2}) = \frac{x - X_{i2}^\top \gamma}{\sigma_2} \left( 3 - \frac{(x - X_{i2}^\top \gamma)^2}{\sigma_2^2} \right) \phi \left( \frac{x - X_{i2}^\top \gamma}{\sigma_2} \right). \quad (111)$$

With a similar argument as (103), we can establish

$$\left| \frac{q'(x_0 | X_{i2}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}}{q(x_0 | X_{i2})} \right| \leq C_u.$$

and

$$\left| \frac{\max_{|c| \leq 1} q'''(x_0 + ch | X_{i2}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}}}{q(x_0 | X_{i2})} \right| \lesssim C_u (1 + C_u^2) \exp \left( C_u \cdot \frac{h}{\sigma_2} \right) \quad (112)$$

Hence, we have

$$\mathbf{E} [(X_{i1} - x_0) K_h(X_{i1}) | X_{i2}] \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \leq \frac{2}{3} h^2 C_u q(x_0 | X_{i2}) \left[ 1 + C h^2 (1 + C_u^2) \exp \left( C_u \cdot \frac{h}{\sigma_2} \right) \right] \quad (113)$$

for some constant  $C$ . Then we can further provide upper bounds for the expression in (106),

$$\begin{aligned} &\left| \mathbf{E} (D_{i1}^2 K_h(X_{i1}) | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} - \frac{2}{3} h^2 q(x_0 | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} \right| \\ &\leq \left| \mathbf{E} [(X_{i1} - x_0)^2 K_h(X_{i1}) | X_{i2}] - \frac{2}{3} h^2 q(x_0 | X_{i2}) \right| \mathbf{1}_{\mathcal{A}_{3,i}} + \frac{\mathbf{E}^2 [(X_{i1} - x_0) K_h(X_{i1}) | X_{i2}]}{\mathbf{E} [K_h(X_{i1}) | X_{i2}]} \mathbf{1}_{\mathcal{A}_{3,i}} \end{aligned}$$

By applying (105), (108) and (109), then the previous inequality can be further upper bounded by

$$\frac{4}{5}h^4 \max_{|c| \leq 1} q''(x_0 + ch \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} + \frac{\left(\frac{2}{3}h^2 q'(x_0 \mid X_{i2}) + \frac{1}{15}h^4 \max_{|c| \leq 1} q'''(x_0 + ch \mid X_{i2})\right)^2}{\min\left\{1, \frac{1}{hC_u}\right\} \max_{|c| \leq 1} q(x_0 + ch \mid X_{i2})}$$

Hence, if  $\frac{1}{\sigma_2}hC_u \leq 1$ , then we have

$$\left| \frac{\mathbf{E}(D_{i1}^2 K_h(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}}{\frac{2}{3}h^2 q(x_0 \mid X_{i2})} - 1 \right| \leq \mathbf{1}_{\mathcal{A}_{3,i}^c} + [h^2(1 + C_u^2) + h^6 C_u^2(1 + C_u^4)] \exp\left(C_u \cdot \frac{h}{\sigma_2}\right). \quad (114)$$

and

$$\left| \frac{\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}}{\frac{2}{3}h q(x_0 \mid X_{i2})} - 1 \right| \leq \mathbf{1}_{\mathcal{A}_{3,i}^c} + [h^2(1 + C_u^2) + h^6 C_u^2(1 + C_u^4)] \exp\left(C_u \cdot \frac{h}{\sigma_2}\right) \quad (115)$$

Hence, we further have

$$\begin{aligned} \left| \mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mathbf{1}_{\mathcal{A}_{3,i}}) - \frac{2}{3}h\pi(x_0) \right| &\leq \int \left| \frac{\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}}{\frac{2}{3}h q(x_0 \mid X_{i2})} - 1 \right| \frac{2}{3}h q(x_0 \mid X_{i2}) p(X_{i2}) dX_{i2} \\ &\lesssim \int \mathbf{1}_{\mathcal{A}_{3,i}^c} \frac{2}{3}h q(x_0 \mid X_{i2}) p(X_{i2}) dX_{i2} + [h^2(1 + C_u^2) + h^6 C_u^2(1 + C_u^4)] \exp\left(C_u \cdot \frac{h}{\sigma_2}\right) h\pi(x_0) \\ &\lesssim h\mathbf{P}(\mathcal{A}_{3,i}^c) + [h^2(1 + C_u^2) + h^6 C_u^2(1 + C_u^4)] \exp\left(C_u \cdot \frac{h}{\sigma_2}\right) h\pi(x_0) \end{aligned}$$

where the last inequality follows from  $q(x_0 \mid X_{i2}) \leq 1$ . Hence, together with (52), we establish

$$\left| \frac{\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mathbf{1}_{\mathcal{A}_{3,i}})}{\frac{2}{3}h\pi(x_0)} - 1 \right| \lesssim \frac{1}{n^c \pi(x_0)} + [h^2(1 + C_u^2) + h^6 C_u^2(1 + C_u^4)] \exp\left(C_u \cdot \frac{h}{\sigma_2}\right) \quad (116)$$

Since  $|D_{i1}|K_h(X_{i1}) \leq 1$ , we have

$$\left| \frac{\mathbf{E}(D_{i1}^2 K_h^2(X_{i1}) \mathbf{1}_{\mathcal{A}_{3,i}^c})}{\frac{2}{3}h\pi(x_0)} \right| \leq \frac{\mathbf{E}\mathbf{1}_{\mathcal{A}_{3,i}^c}}{\frac{2}{3}h\pi(x_0)} \leq \frac{1}{n^c h\pi(x_0)} \rightarrow 0 \quad (117)$$

Combining (116) and (117), we establish (55).

### C.2.3 Proof of (56)

The proof of (56) is similar to that of (55). We first have the following decomposition,

$$\mathbf{E}D_{i1}(X_{i1} - x_0)K_h(X_{i1}) = \mathbf{E}D_{i1}(X_{i1} - x_0)K_h(X_{i1})\mathbf{1}_{\mathcal{A}_{3,i}} + \mathbf{E}D_{i1}(X_{i1} - x_0)K_h(X_{i1})\mathbf{1}_{\mathcal{A}_{3,i}^c}.$$

Based on the following relation,

$$\begin{aligned} & \mathbf{E} [D_{i1}(X_{i1} - x_0)K_h(X_{i1}) \mid X_{i2}] \mathbf{1}_{\mathcal{A}_{3,i}} \\ &= \left( \mathbf{E} [(X_{i1} - x_0)^2 K_h(X_{i1}) \mid X_{i2}] - \frac{\mathbf{E}^2 [(X_{i1} - x_0)K_h(X_{i1}) \mid X_{i2}]}{\mathbf{E} [K_h(X_{i1}) \mid X_{i2}]} \right) \mathbf{1}_{\mathcal{A}_{3,i}} = \mathbf{E} (D_{i1}^2 K_h(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}}. \end{aligned}$$

By applying (106) and (114), we have

$$\frac{\mathbf{E} [D_{i1}(X_{i1} - x_0)K_h(X_{i1})\mathbf{1}_{\mathcal{A}_{3,i}}]}{\frac{2}{3}h^2\pi(x_0)} \rightarrow 1 \quad (118)$$

Since  $D_{i1}(X_{i1} - x_0)K_h(X_{i1}) \leq h$ , then

$$\left| \frac{\mathbf{E} D_{i1}(X_{i1} - x_0)K_h(X_{i1})\mathbf{1}_{\mathcal{A}_3^c}}{\frac{2}{3}h^2\pi(x_0)} \right| \leq \frac{\mathbf{E} \mathbf{1}_{\mathcal{A}_{3,i}^c}}{\frac{2}{3}h\pi(x_0)} \leq \frac{1}{n^c h \pi(x_0)} \rightarrow 0,$$

Together with (118), we establish (56).

### C.3 Proof of Lemma 7

Proof of (58)

The term  $\frac{1}{n} \sum_{i=1}^n K_h(X_{i1})$  satisfies

$$\mathbf{E} \left( \frac{1}{n} \sum_{i=1}^n K_h(X_{i1}) \right) = \mathbf{E} (K_h(X_{i1})), \quad \text{Var} \left( \frac{1}{n} \sum_{i=1}^n K_h(X_{i1}) \right) \leq \frac{1}{nh} \mathbf{E} (K_h(X_{i1})) \quad (119)$$

Together with (54), there exists  $0 < c < 1/2$  such that

$$(2 - c)\pi(x_0) \leq \mathbf{E} (K_h(X_{i1})) \leq (2 + c)\pi(x_0). \quad (120)$$

Then we establish (58).

Proof of (59)

The term  $\frac{1}{n} \sum_{i=1}^n D_{i1}K_h(X_{i1})$  satisfies

$$\mathbf{E} \left( \frac{1}{n} \sum_{i=1}^n D_{i1}K_h(X_{i1}) \right) = 0, \quad \text{Var} \left( \frac{1}{n} \sum_{i=1}^n D_{i1}K_h(X_{i1}) \right) = \frac{1}{n} \mathbf{E} (D_{i1}^2 K_h^2(X_{i1})). \quad (121)$$

By (55), we establish (59).

Proof of (60)

Note that

$$\mathbf{E} \left( \frac{1}{n} \sum_{i=1}^n (X_{i1} - x_0)K_h(X_{i1}) \right) = \mathbf{E} (X_{i1} - x_0)K_h(X_{i1}) \quad (122)$$

and

$$\text{Var} \left( \frac{1}{n} \sum_{i=1}^n (X_{i1} - x_0) K_h(X_{i1}) \right) \leq \frac{1}{nh} \mathbf{E} (X_{i1} - x_0)^2 K_h(X_{i1}) \quad (123)$$

By (109), we have

$$\begin{aligned} |\mathbf{E}(X_{i1} - x_0) K_h(X_{i1})| &\leq \frac{2}{3} h^2 \mathbf{E} q'(x_0 | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} + \frac{1}{15} h^4 \cdot \max_{|c| \leq 1} q'''(x_0 + ch | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} + \mathbf{P}(A_{3,i}^c) \\ &\lesssim \frac{2}{3} C_u h^2 \pi(x_0) + h^4 C_u (1 + C_u^2) \pi(x_0) + \frac{1}{n^c} \lesssim C_u h^2 \pi(x_0) \end{aligned} \quad (124)$$

where the second inequality follows from (110), (111) and (101). By (108), we have

$$\begin{aligned} \mathbf{E}(X_{i1} - x_0)^2 K_h(X_{i1}) &\leq \frac{2}{3} h^2 \mathbf{E} q(x_0 | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} + \frac{4}{5} h^4 \cdot \mathbf{E} \max_{|z| \leq 1} q''(x_0 + cz | X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} + \mathbf{P}(A_{3,i}^c) \\ &\leq \frac{2}{3} h^2 \pi(x_0) + h^4 (1 + C_u^2) \pi(x_0) + \frac{1}{n^c} \lesssim h^2 \pi(x_0) \end{aligned} \quad (125)$$

where the second inequality follows from (111) and (101). Hence, we establish (60).

Proof of (61)

The term  $\frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1})$  satisfies

$$\mathbf{E} \left( \frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1}) \right) = \mathbf{E} [D_{i1} (X_{i1} - x_0) K_h(X_{i1})] \quad (126)$$

and

$$\text{Var} \left( \frac{1}{n} \sum_{i=1}^n D_{i1} (X_{i1} - x_0) K_h(X_{i1}) \right) \leq \frac{1}{nh} \mathbf{E} (D_{i1}^2 (X_{i1} - x_0)^2 K_h(X_{i1})) \leq \frac{h^3}{4n} \mathbf{E} (K_h(X_{i1})) \quad (127)$$

Combined with (56), we establish that (61).

Proof of (62)

The term  $\frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1})$  satisfies

$$\mathbf{E} \left( \frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right) = \mathbf{E} \left[ D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right] \quad (128)$$

and

$$\text{Var} \left( \frac{1}{n} \sum_{i=1}^n D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right) \leq \frac{1}{nh} \mathbf{E} \left( D_{i1}^2 \frac{(X_{i1} - x_0)^4}{4} K_h(X_{i1}) \right) \leq \frac{h^5}{4n} \mathbf{E} (K_h(X_{i1})) \quad (129)$$

In the following, we shall prove that

$$\left| \mathbf{E} D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \right| \leq C C_u h^4 \pi(x_0) \quad (130)$$

Combined with (128), (129) and (120), we establish (62). Now let's complete the proof of (130). Note that

$$\mathbf{E} D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) = \mathbf{E} D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} + \mathbf{E} D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}^c} \quad (131)$$

Note that

$$\mathbf{E} D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mathbf{1}_{\mathcal{A}_{3,i}} = \mathbf{E}_{X_{i2}} \left[ \mathbf{E} \left( D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mid X_{i2} \right) \mathbf{1}_{\mathcal{A}_{3,i}} \right]$$

and

$$\begin{aligned} & \mathbf{E} \left( D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mid X_{i2} \right) \\ &= \mathbf{E} \left( \frac{(X_{i1} - x_0)^3}{2} K_h(X_{i1}) \mid X_{i2} \right) - l(X_{i2}) \mathbf{E} \left( \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mid X_{i2} \right) \\ &= \mathbf{E} \left( \frac{(X_{i1} - x_0)^3}{2} K_h(X_{i1}) \mid X_{i2} \right) - \frac{\mathbf{E}((X_{i1} - x_0) K_h(X_{i1}) \mid X_{i2}) \mathbf{E} \left( \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mid X_{i2} \right)}{\mathbf{E}(K_h(X_{i1}) \mid X_{i2})}, \end{aligned}$$

Then it is sufficient to control the terms  $\mathbf{E}[(X_{i1} - x_0)^3 K_h(X_{i1}) \mid X_{i2}] \mathbf{1}_{\mathcal{A}_{3,i}}$  and

$$\frac{\mathbf{E}((X_{i1} - x_0) K_h(X_{i1}) \mid X_{i2}) \mathbf{E} \left( \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mid X_{i2} \right)}{\mathbf{E}(K_h(X_{i1}) \mid X_{i2})} \mathbf{1}_{\mathcal{A}_{3,i}}.$$

The second term can be upper bounded by  $\frac{h^2}{2} \mathbf{E}((X_{i1} - x_0) K_h(X_{i1}) \mid X_{i2})$  since  $\frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \leq \frac{h^2}{2}$ . It follows from (113) and the condition  $hC_u \rightarrow 0$  that

$$\frac{h^2}{2} \mathbf{E}((X_{i1} - x_0) K_h(X_{i1}) \mid X_{i2}) \mathbf{1}_{\mathcal{A}_{3,i}} \lesssim \frac{1}{3} h^4 C_u q(x_0 \mid X_{i2}).$$

We control the first term  $\mathbf{E}[(X_{i1} - x_0)^3 K_h(X_{i1}) \mid X_{i2}]$  in the following.

$$\begin{aligned} \mathbf{E}[(X_{i1} - x_0)^3 K_h(X_{i1}) \mid X_{i2}] &= \int_{|z| \leq 1} h^3 z^3 q(x_0 + hz \mid X_{i2}) dz \\ &= \int_{|z| \leq 1} h^3 z^3 \left[ q(x_0 \mid X_{i2}) + hz q'(x_0 \mid X_{i2}) + \frac{h^2 z^2}{2} q''(x_0 \mid X_{i2}) + \frac{h^3 z^3}{6} q'''(x_0 + c(z)hz \mid X_{i2}) \right] dz \end{aligned}$$

and then we have

$$\left| \mathbf{E}[(X_{i1} - x_0)^3 K_h(X_{i1}) \mid X_{i2}] - \frac{2}{5} h^4 q'(x_0 \mid X_{i2}) \right| \leq \frac{1}{21} h^6 \max_{|c| \leq 1} q'''(x_0 + ch \mid X_{i2}). \quad (132)$$

We then have

$$\mathbf{E}[(X_{i1} - x_0)^3 K_h(X_{i1}) \mid X_{i2}] = \frac{2}{5} h^4 q(x_0 \mid X_{i2}) \cdot \frac{-x_0 + X_{i2}^\top \gamma}{\sigma_2} + O(h^6) \cdot \max_{|c| \leq 1} q'''(x_0 + ch \mid X_{i2})$$

By (111), we have

$$\left| \mathbf{E} \left( D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \mid X_{i2} \right) \mathbf{1}_{\mathcal{A}_{3,i}} \right| \leq \left( C_u \frac{2}{5} h^4 + O(h^6) C_u (3 + C_u^2) \right) q(x_0 \mid X_{i2})$$

where  $C_u$  is defined in (17). Together with

$$\left| \mathbf{E} D_{i1} \frac{(X_{i1} - x_0)^2}{2} K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}^c} \right| \leq h^2 \mathbf{P}(\mathcal{A}_{3,i}^c) = h^2 \cdot n^{-c},$$

we have (130).

Proof of (64) We control the mean and variance of  $\frac{1}{n} \sum_{i=1}^n \left( X_{i,-1}^\top \gamma \right)^2 K_h(X_{i1})$  as follows,

$$\begin{aligned} \mathbf{E} \frac{1}{n} \sum_{i=1}^n \left( X_{i,-1}^\top \gamma \right)^2 K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} &= \mathbf{E} \left( X_{i,-1}^\top \gamma \right)^2 K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \\ &= \mathbf{E}_{X_{i2}} \left( X_{i,-1}^\top \gamma \right)^2 \mathbf{E} (K_h(X_{i1}) \mid X_{i2}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \lesssim \mathbf{E}_{X_{i2}} \left( X_{i,-1}^\top \gamma \right)^2 \end{aligned}$$

and

$$\begin{aligned} \text{Var} \frac{1}{n} \sum_{i=1}^n \left( X_{i,-1}^\top \gamma \right)^2 K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} &\leq \frac{1}{nh} \mathbf{E} \left( X_{i,-1}^\top \gamma \right)^4 K_h(X_{i1}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \\ &= \frac{1}{nh} \mathbf{E}_{X_{i2}} \left( X_{i,-1}^\top \gamma \right)^4 \mathbf{E} (K_h(X_{i1}) \mid X_{i2}) \cdot \mathbf{1}_{\mathcal{A}_{3,i}} \lesssim \frac{1}{nh} \mathbf{E}_{X_{i2}} \left( X_{i,-1}^\top \gamma \right)^4 \end{aligned}$$

Since  $X_{i,-1}$  is Sub-gaussian random variable, we have  $\mathbf{E}_{X_{i2}} \left( X_{i,-1}^\top \gamma \right)^2 \lesssim \|\gamma\|_2^2$  and  $\mathbf{E}_{X_{i2}} \left( X_{i,-1}^\top \gamma \right)^2 \lesssim \|\gamma\|_2^4$ . Hence, with probability larger than  $1 - \frac{1}{t}$ , we have

$$\frac{1}{n} \sum_{i=1}^n \left( X_{i,-1}^\top \gamma \right)^2 K_h(X_{i1}) \lesssim \left( 1 + \frac{t}{\sqrt{nh}} \right) \|\gamma\|_2^2.$$